LECTURE 8

Last time:

- Source coding
- Huffman code, Elias Code.

Lecture outline

- Discrete Memoryless Channels
- Channel capacity
- Binary symmetric channels and Erasure channels
- Joint AEP

Reading: Reading: Scts. 8.1-8.6.

Discrete Memoryless Channel

Definition: Discrete Channel

- We can assume a discrete input alphabet $\mathcal{X}$ and a discrete output alphabet $\mathcal{Y}$.
- We can describe a channel by a set of transition probabilities

$P_{\underline{Y}^n|\underline{X}^n}(\underline{y}^n|\underline{x}^n)$, for all n

Definition Discrete Memoryless Channel(DMC)

Let us restrict ourselves the channels with

$P_{\underline{Y}^n|\underline{X}^n}(\underline{y}^n|\underline{x}^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$

- The distribution of Y_i depends only on the current input.
- We assume the transition probability $P_{Y|X}$ is time-invariant.

Channel Capacity

- The capacity of a DMC channel is defined as

$$C = \max_{P_X(x)} I(X;Y)$$

- The operational meaning of the capacity is the maximum rate of information that can be transferred over the channel reliably.
- Our goal now is to find P_X to maximize $I(X;Y)$, and later use this distribution to achieve the maximum communication rate.
- We can also consider the capacity for n uses of the channel,

$$C^{(n)} = \frac{1}{n} \max_{P_{\underline{X}^n}(\underline{x}^n)} I(\underline{X}^n;\underline{Y}^n)$$

Channel capacity

- Use the memoryless assumption,

$$\begin{aligned} I(\underline{X}^n;\underline{Y}^n) &= H(\underline{Y}^n) - H(\underline{Y}^n|\underline{X}^n) \\ &= \sum_{i=1}^{n} H(Y_i|\underline{Y}^{i-1}) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i;Y_i) \end{aligned}$$

- The inequality can be met with equality if we take the X_is to be independent, because the Y_is then are also independent

- If $I(X;Y)$ is maximized by a distribution $P_X(x)$, then the taking X_i's to be i.i.d. with P_X maximizes each term on the RHS.

- For a memoryless channel, we can focus on maximizing the mutual information of one channel use. This does not mean we can communicate reliably with just one channel use.

Binary Symmetric Channel (BSC)

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x=0,1} P_X(x) H(Y|X=x) \\ &= H(Y) - H(\epsilon) \\ &\leq 1 - H(\epsilon) \end{aligned}$$

where $H(\epsilon) = -(\epsilon \log(\epsilon) + (1-\epsilon)\log(1-\epsilon))$

- The optimal input distribution is P_X being equiprobable on 0 and 1.
- The resulting channel capacity is $1 - H(\epsilon)$.
- Intuitively, we can think of a correction data with entropy $H(\epsilon)$.

Binary Erasure Channel (BEC)

E indicator variable that is 1 if there is an error and is 0 otherwise

$$
\begin{aligned}
C &= \max_{P_X(x)} I(X;Y) \\
&= \max_{P_X(x)} (H(Y) - H(Y|X)) \\
&= \max_{P_X(x)} (H(Y,E) - H(Y|X)) \\
&= \max_{P_X(x)} (H(E) + H(Y|E) - H(Y|X))
\end{aligned}
$$

$H(E) = \mathbf{H}(\epsilon)$

$$
\begin{aligned}
& H(Y|E) \\
= \ & P(E=0)H(Y|E=0) + P(E=1)H(Y|E=1) \\
= \ & (1-\epsilon)H(X)
\end{aligned}
$$

$H(Y|X) = \mathbf{H}(\epsilon)$

Thus $C = \max_{P_X(x)}(H(Y|E)) = 1 - \epsilon$

Symmetric channels

Let us consider the transition matrix T the $|\mathcal{X}|\times|\mathcal{Y}|$ matrix whose elements are $P_{Y|X}(y|x)$

Definition A DMC is symmetric iff all the rows are permutations of each other, and the columns are permutations of each other.

Denote a row of T as $r = [r_1, \ldots r_{|\mathcal{Y}|}]$, and the corresponding entropy $H(r) = -\sum_i r_i \log r_i$.

Optimal Input Distribution for Symmetric Channels

$$\begin{aligned} I(X;Y) &= h(Y) - H(Y \mid X) \\ &= H(Y) - E_X[H(Y \mid X = x)] \\ &= H(Y) - H(r) \\ &\leq \log|\mathcal{Y}| - H(r) \end{aligned}$$

The equality holds only if Y is uniformly distributed.

- Let X be uniformly distributed,

$$\begin{aligned} P_Y(y) &= \sum_{x\in\mathcal{X}} P_X(x) P_{Y|X}(y|x) \\ &= \frac{1}{|\mathcal{X}|}\sum_x P_{Y|X}(y|x) \\ &= \frac{c}{|\mathcal{X}|} \end{aligned}$$

Therefore the uniform input distribution is optimal.

An Alternative Approach

- Consider an arbitrary input distribution $P^{(1)} = [p_1, p_2, \ldots, p_n]$, where $n = |\mathcal{X}|$. Let the corresponding mutual information be $I^{(1)}$.

- Now since the channel is symmetric, any permutation of $P^{(1)}$ gives the same mutual information. Denote all the permutations as $P^{(2)}, \ldots, P^{(n!)}$.

- Define a new input distribution $P^* = \frac{1}{n!}\sum_i P^{(i)}$. P^* is the uniform distribution.

- By the concavity of I as a function of the input distribution:

$$\begin{aligned} I(P^*) &= I\left(\frac{1}{n!}\sum_i P^{(i)}\right) \\ &\geq \frac{1}{n!}\sum_i I(P^{(i)}) = I(P^{(1)}) \end{aligned}$$

Finding the Optimal Input for Asymmetric Channel

- Let the input distribution of X be $[P, Q, Q]$, easily check that the distribution of Y is also $[P, Q, Q]$. Goal: find P and Q to maximize the mutual information.

- Compute

$$
\begin{aligned}
H(X) &= -P \log P - 2Q \log Q \\
H(X|Y) &= P(Y = 1) \times 0 \\
&\quad +2P(Y = 2)H(X|Y = 2) \\
&= 2QH(\epsilon)
\end{aligned}
$$

- Maximize $I(X;Y) = H(X) - H(X|Y)$ subject to the constraint $P + 2Q = 1$, define

$$J = -P \log P - 2Q \log Q - 2QH(\epsilon) + \lambda(P + 2Q)$$

we have

$$
\begin{aligned}
\frac{\partial J}{\partial P} &= -1 - \log P + \lambda = 0 \\
\frac{\partial J}{\partial Q} &= -2 - 2 \log Q - 2H(\epsilon) + 2\lambda = 0
\end{aligned}
$$

Solve for : $\log P = \log Q + H(\epsilon)$.

Let $\alpha = e^{H(\epsilon)}$, we have

$$P = \frac{\alpha}{\alpha+2}, Q = \frac{1}{\alpha+2}$$
$$C = \log \frac{\alpha+2}{\alpha}$$

Check:

- If $\epsilon = 0$, $\alpha = 1$, input is uniform on $\mathcal{X}$, capacity is $\log 3$.
- If $\epsilon = \frac{1}{2}$, $H(\epsilon) = \log 2 (nats)$, and $\alpha = 2$. Capacity $\log 2$.
- The larger $H(\epsilon)$ is, the higher P is. We rely more on distinguishing the two groups to convey information.

Why the Mutual Information is Important

- Reliable communication requires disjoint partitioning in the received signal space, corresponding to the different possible transmitted signals.
- For each (typical) input sequence $\underline{X}^n$, there are approximately $2^{nH(Y|X)}$ possible Y sequences.
- There are in total $2^{nH(Y)}$ (typical) Y sequences. Divide this set into sets of size $2^{nH(Y|X)}$, each corresponding to one input X sequences.
- The number of disjoint sets is no more than $2^{nH(Y)-nH(Y|X)} = 2^{nI(X;Y)}$. Therefore we can send at most $2^{nI(X;Y)}$ different sequences that are distinguishable.

Joint AEP

Definition The set of the joint typical sequences $\{(\underline{x}^n, \underline{y}^n)\}$ is

$$
\begin{aligned}
A_\epsilon^{(n)} \;=\; & \Big\{(\underline{x}^n, \underline{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
& \left| -\frac{1}{n} \log P_{\underline{X}^n}(\underline{x}^n) - H(X) \right| < \epsilon \\
& \left| -\frac{1}{n} \log P_{\underline{Y}^n}(\underline{y}^n) - H(Y) \right| < \epsilon \\
& \left| -\frac{1}{n} \log P_{\underline{X}^n, \underline{Y}^n}(\underline{x}^n, \underline{y}^n) - H(X, Y) \right| < \epsilon \Big\}
\end{aligned}
$$

Theorem

Consider sequences $(\underline{X}^n, \underline{Y}^n)$ drawn i.i.d. from $P_{X,Y}$, then

- $P((\underline{X}^n, \underline{Y}^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
- $|A_\epsilon^{(n)}| \leq 2^{nH(X,Y)+\epsilon}$

Simple extension of the AEP for single random variable.

Joint AEP

Question:
If I randomly pick a typical sequence $\underline{\tilde{X}}^n \in A_\epsilon^{(n)}(X)$, and a $\underline{\tilde{Y}}^n \in A_\epsilon^{(n)}(Y)$, are they joint typical?

$$A_\epsilon^{(n)}(X) \quad \times A_\epsilon^{(n)}(Y) \stackrel{?}{=} \quad A_\epsilon^{(n)}(X,Y)$$

$$2^{nH(X)} \quad \times 2^{nH(Y)} \geq \quad 2^{nH(X,Y)}$$

Theorem If $(\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n)$ are independent with the same marginal distributions, i.e., $(\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \sim P_{\underline{X}^n}(\underline{x}^n) P_{\underline{Y}^n}(\underline{y}^n)$,

$$P((\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \in A_\epsilon^{(n)}(X,Y)) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

and for n large enough,

$$P((\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \in A_\epsilon^{(n)}(X,Y)) \geq (1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)}$$

Proof of the Theorem

$$
\begin{aligned}
& P((\underline{\tilde{X}}^n, \underline{\tilde{Y}}^n) \in A_\epsilon^{(n)}(X,Y)) \\
= & \sum_{A_\epsilon^{(n)}(X,Y)} P_{\underline{X}^n}(\underline{x}^n) P_{\underline{Y}^n}(\underline{y}^n) \\
\leq & \; 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\
= & \; 2^{-n(I(X;Y)-3\epsilon)}
\end{aligned}
$$

- Fix a typical $\underline{y}^n$, out of the $2^{nH(X)}$ typical X sequences, there are approximately $2^{nH(X|Y)}$ sequences that are jointly typical with $\underline{y}^n$.
- Pass $\underline{x}^n$ through the channel to obtain a joint typical pair $(\underline{x}^n, \underline{y}^n)$. Assume now that only $\underline{y}^n$ is observed. If we pick randomly another $\underline{\tilde{x}}^n$ and ask if this is the originally transmitted sequence, checking joint typically gives a probability of confusion $2^{-nI(X;Y)}$.
- Next time, decoding with joint AEP.