Noisy-Channel Theorem: Forward Direction

With the tools from the previous section, we are ready to prove the forward direction of Shannon's noisy-channel coding theorem, which states that any rate strictly below the channel capacity is achievable:

Theorem: Shannon's noisy-channel coding theorem (forward direction)

For a discrete memoryless channel with capacity C, any rate R < C is achievable. Concretely, for any $\varepsilon > 0$ and any rate R < C, for large enough n there exists a $(2^{n \cdot R}, n)$ code with maximal error $\lambda^{(n)} < \varepsilon$.

Proof

Given a channel $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ with capacity $C = \max P_X I(X;Y)$, let R < C and $\epsilon > 0$. We will first show that for big enough n, a randomly constructed code with rate R has a low error probability. We will then argue that a low error probability on average of codes implies the existence of some specific code with low error probability.

Fix an input distribution P_X that maximizes I(X;Y). For any n, construct a $(2^{n\cdot R},n)$ -code $\mathcal C$ by choosing a codebook at random according to P_X . That is, for every message $w\in[2^{n\cdot R}]$, sample n times from the distribution P_X , creating a codeword $\mathcal C(w)=(\mathcal C_1(w),\mathcal C_2(w),\dots,\mathcal C_n(w))$ by concatenating the n independent samples $\mathcal C_i(w)\sim P_X$.

Since the channel is memoryless, if w is sent over the channel using \mathcal{C} , the output distribution Y^n is given by:

$$P_{Y^n\mid X^n}(y^n|\mathcal{C}(w)) = \prod_{i=1}^n P_{Y\mid X}(y_i\mid \mathcal{C}_i(w)).$$

What is the probability that the decoded message \hat{w} is incorrect, i.e., not equal to w? This depends on the decoding method used by the receiver. The optimal decoding procedure is **maximum-likelihood decoding**, where the input message that is most likely with respect to $P_{X|Y}$ is selected as the decoding \hat{w} . However, it

created: 2018-12-12

Information Theory | Noisy-Channel Theorem: Forward Direction

is hard to analyze the error probability for this decoding method. Instead, we will assume that the receiver applies **jointly typical decoding**, which has a slightly higher probability of decoding to the wrong message, but still small enough for our analysis. Jointly typical decoding works as follows: upon receiving an output y^n , the receiver looks for a *unique* message \hat{w} such that the pair $(\mathcal{C}(\hat{w}), y^n)$ is jointly typical. If there exists no such message, or if it is not unique, the receiver declares a failure by decoding to $\hat{w}=0$ (which is always wrong because $w\in[2^{n\cdot R}]=\{1,2,\ldots,2^{n\cdot R}\}$).

With this decoding procedure in mind, we analyze the average error probability $P[\mathtt{error}]$, where the average is taken over both the randomly constructed code $\mathcal C$ and the uniformly randomly selected message w. Defining $\lambda_w(\mathcal C) := P[\hat w \neq w \mid \mathcal C(w) \text{ was sent over the channel}]$ to be the probability that a message w (encoded using $\mathcal C$) is decoded incorrectly, we get:

$$egin{aligned} P[\mathtt{error}] &= \sum_{\mathcal{C}} P[\mathcal{C}] \cdot \left(\sum_{w=1}^{2^{n \cdot R}} rac{1}{2^{n \cdot R}} \cdot \lambda_w(\mathcal{C})
ight) \ &= rac{1}{2^{n \cdot R}} \sum_{w=1}^{2^{n \cdot R}} \sum_{\mathcal{C}} P[\mathcal{C}] \cdot \lambda_w(\mathcal{C}). \end{aligned}$$

Since we average over all randomly constructed codes $\mathcal C$, and the codewords for all messages are sampled independently, the value $\sum_{\mathcal C} P[\mathcal C] \cdot \lambda_w(\mathcal C)$ does not depend on the particular message w. Hence if we set, for example, $w_0=1$, then for all $w\in[2^{n\cdot R}]$,

$$\sum_{\mathcal{C}}P[\mathcal{C}]\lambda_w(\mathcal{C})=\sum_{\mathcal{C}}P[\mathcal{C}]\lambda_{w_0}(\mathcal{C}).$$

This simplifies the calculation of $P[\mathtt{error}]$ significantly:

$$egin{aligned} P[exttt{error}] &= rac{1}{2^{n \cdot R}} \sum_{w=1}^{2^{n \cdot R}} \sum_{\mathcal{C}} P[\mathcal{C}] \cdot \lambda_{w_0}(\mathcal{C}) \ &= \sum_{\mathcal{C}} P[\mathcal{C}] \cdot \lambda_{w_0}(\mathcal{C}). \end{aligned}$$

That is, the average probability of error is the probability (over the selection of the code \mathcal{C} , and over the randomness in the channel) that the message w_0 is decoded incorrectly. There are two possible reasons for an error in the decoding:

1. The output of the channel is not jointly typical with $C(w_0)$. By the first item of the joint AEP, this probability approaches zero as n goes to infinity. Hence, for

created: 2018-12-12

Information Theory | Noisy-Channel Theorem: Forward Direction big enough n, the probability of an error for this reason is smaller than ϵ .

2. There is some $w' \neq w_0$ such that the output of the channel is (also) jointly typical with $\mathcal{C}(w')$. Since \mathcal{C} is a random code (and so $\mathcal{C}(w')$ is independent from the channel output y^n), by the third item of the joint AEP the probability that this occurs is at most

$$\sum_{w'
eq w_0} 2^{-n(I(X;Y)-3\epsilon)} = (2^{n \cdot R} - 1)2^{-n(I(X;Y)-3\epsilon)}.$$

We can thus bound the average probability of error, using the union bound and the bounds in the above analysis, by

$$egin{aligned} P[exttt{error}] & \leq \epsilon + (2^{n \cdot R} - 1) 2^{-n(I(X;Y) - 3\epsilon)} \ & \leq \epsilon + 2^{n \cdot R} 2^{-n(I(X;Y) - 3\epsilon)} \ & = \epsilon + 2^{-n(I(X;Y) - R - 3\epsilon)} \end{aligned}$$

As long as R < I(X;Y), one can choose n large enough so that $P[\mathtt{error}] \leq 2\epsilon$.

This analysis upper bounds the (expected) average error probability for a random code \mathcal{C} . However, if this expected probability is low, there must be some specific code \mathcal{C}^* that also has low average error probability.

Finally, in \mathcal{C}^* , we aim to bound the *maximal* error probability, i.e., the probability of error for the worst message. We can do so by noting that at least half of the messages w has error probability $\lambda_w(\mathcal{C}^*) \leq 4\epsilon$: if not, then the total error probability of these messages would already exceed $2^{n \cdot R} \cdot 2\epsilon$, contradicting the upper bound of 2ϵ to the average error probability. Thus, we can construct a better code by discarding the worst half of the codewords, and using the remaining $2^{n \cdot R - 1}$ codewords to construct a new code, with rate

$$\frac{\log(2^{n\cdot R-1})}{n} = \frac{n\cdot R-1}{n} = R - \frac{1}{n}$$

and maximal probability of error $\lambda^{(n)} \leq 4\epsilon$.

In the above proof, we implicitly assumed that $2^{n\cdot R}$ is an integer. You can try to redo the proof for the case when it is not: construct $\mathcal C$ as a $(\lceil 2^{n\cdot R} \rceil, n)$ code, and verify that the average probability of error $P[\mathtt{error}]$ is still sufficiently small. Also compute a lower bound on the rate of the final code.

created: 2018-12-12