Noisy Channels and Capacity

Uri Shaham

1 Noisy channels

1.1 Motivation

Compression asks: how many bits are needed to describe data? Communication asks: how many bits can be sent reliably through noise?

The surprising answer is that noise does not make reliable communication impossible. It imposes a maximum reliable rate, called channel capacity.

1.2 Discrete memoryless channels

Definition 1.1 (Discrete memoryless channel)**.** A discrete memoryless channel consists of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and transition probabilities

$$P_{Y|X}(y \mid x).$$

Memoryless means that over n uses,

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

Example 1.2 (Binary symmetric channel)**.** The binary symmetric channel BSC(p) has input $X \in \{0, 1\}$ and output

$$Y = X \oplus N,$$

where $N \sim$ Bernoulli(p). With probability p, the bit flips.

Example 1.3 (Binary erasure channel)**.** The binary erasure channel BEC(α) has output alphabet $\{0, 1, ?\}$. With probability $1 - \alpha$, the output equals the input. With probability α, the output is an erasure symbol ?. The receiver knows which symbols were erased.

1.3 Channel codes

An (n, M) channel code has:

- a message W uniformly distributed over $\{1, \ldots, M\}$;
- an encoder mapping each message w to a codeword $x^n(w) \in \mathcal{X}^n$;
- a decoder mapping received sequences y^n to estimates $\hat{W}$.

The rate is

$$R = \frac{1}{n} \log M$$

bits per channel use. The probability of error is

$$P_e = \mathbb{P}[\hat{W} \neq W].$$

1.4 Mutual information through a channel

For a chosen input distribution P_X (determined by the encoder), the joint distribution is

$$P_{XY}(x, y) = P_X(x) P_{Y|X}(y \mid x).$$

The mutual information $I(X;Y)$ measures how much information one channel use conveys about its input.

Definition 1.4 (Channel capacity)**.** The capacity of a discrete memoryless channel is

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

Capacity is measured in bits per channel use.

1.5 Capacity of the BSC and BEC

Proposition 1.5 (Capacity of the BSC)**.** *The capacity of BSC(p) is*

$$C = 1 - H_2(p).$$

Proof sketch. For the BSC, once X is known, the randomness in Y comes from the noise, which is Bernoulli(p).

$$H(Y \mid X) = H_2(p).$$

Thus

$$I(X;Y) = H(Y) - H_2(p) \leq 1 - H_2(p),$$

as the entropy of a binary random variable is at most one bit. As we know, this entropy is maximized (and equals 1) when Y is uniform. This happens when $X \sim \text{Bernoulli}(1/2)$. □

Proposition 1.6 (Capacity of the BEC)**.** *The capacity of BEC(α) is*

$$C = 1 - \alpha.$$

Proof sketch. The channel output either reveals the input perfectly or erases it. With uniform input,

$$I(X;Y) = H(X) - H(X \mid Y) = 1 - \alpha.$$

No input distribution can convey more than one bit when not erased, and erasures occur a fraction α of the time. □

2 The noisy-channel coding theorem

Theorem 2.1 (Shannon noisy-channel coding theorem, informal)**.** *For a discrete memoryless channel of capacity C:*

(i) If $R < C$, then there exists a sequence of length-n codes of rate approaching R whose error probability tends to zero.

(ii) If $R > C$, then no sequence of length-n codes of rate R can have error probability tending to zero.

This theorem is existential: it says good codes exist, but it does not by itself give a practical encoder or decoder. To prove the theorem, we will use the following corollary of AEP.

Lemma 2.2 (Joint Typicality Lemma, informal)**.** *Let $(X, Y) \sim P_{XY}$, and let P_X, P_Y be the marginals.*

Define the jointly typical set by

$$\mathcal{J}_\epsilon^{(n)} = \mathcal{T}_\epsilon^{(n)}(X) \cap \mathcal{T}_\epsilon^{(n)}(Y) \cap \mathcal{T}_\epsilon^{(n)}((X, Y)).$$

Then:

1. *If $(X^n, Y^n) \sim P_{XY}^n$, then*
$$\Pr\big((X^n, Y^n) \in \mathcal{J}_\epsilon^{(n)}\big) \to 1.$$
2. *If $\widetilde{X}^n \sim P_X^n$ and $\widetilde{Y}^n \sim P_Y^n$ are independent, then*
$$\Pr\big((\widetilde{X}^n, \widetilde{Y}^n) \in \mathcal{J}_\epsilon^{(n)}\big) \lesssim 2^{-nI(X;Y)}.$$

Proof. For the first part, apply the AEP from Lecture 2 to the three IID sources

$$X^n, \qquad Y^n, \qquad (X^n, Y^n).$$

Each is typical with probability tending to 1. Therefore all three typicality conditions hold simultaneously with probability tending to 1.

For the second part, suppose $\widetilde{X}^n$ and $\widetilde{Y}^n$ are independent. If (x^n, y^n) is jointly typical, then

$$P_X^n(x^n) \approx 2^{-nH(X)}$$

and

$$P_Y^n(y^n) \approx 2^{-nH(Y)}.$$

Hence, by independence,

$$P_X^n(x^n) P_Y^n(y^n) \approx 2^{-n(H(X)+H(Y))}.$$

Also, by the typical-set size bound applied to the pair source (X, Y), the number of jointly typical pairs is approximately

$$2^{nH(X,Y)}.$$

Therefore,

$$\Pr\big((\widetilde{X}^n, \widetilde{Y}^n) \text{ is jointly typical}\big) \approx 2^{nH(X,Y)} 2^{-n(H(X)+H(Y))}.$$

Thus

$$\Pr\big((\widetilde{X}^n, \widetilde{Y}^n) \text{ is jointly typical}\big) \approx 2^{-n(H(X)+H(Y)-H(X,Y))}.$$

Since

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

we get

$$\Pr\big((\widetilde{X}^n, \widetilde{Y}^n) \text{ is jointly typical}\big) \approx 2^{-nI(X;Y)}.$$

□

2.1 Proof of Achievability, Part (i)

Proof. Let the channel be a discrete memoryless channel with transition law $P_{Y|X}$, and let its capacity be

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

Fix a rate $R < C$. Then there exists an input distribution P_X such that

$$R < I(X;Y).$$

We will show that for this P_X, there exists a sequence of codes of rate approaching R whose probability of error goes to zero.

Generate a random codebook with

$$M = 2^{nR}$$

codewords. For each message $w \in \{1, \dots, M\}$, independently draw

$$X^n(w) \sim P_X^n.$$

That is, each symbol of each codeword is generated independently according to P_X.

To send message w, the encoder transmits the codeword $X^n(w)$. The channel output is Y^n, generated according to

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

The decoder looks for a unique message $\hat{w}$ such that

$$\big(X^n(\hat{w}), Y^n\big)$$

is jointly typical with respect to $P_{XY} = P_X P_{Y|X}$, where the pair (x^n, y^n) is jointly typical if the empirical frequencies of pairs (x_i, y_i) are close to their theoretical probabilities under P_{XY}. If there is no such unique message, the decoder declares an error.

By symmetry of the random code construction, assume without loss of generality that message $W = 1$ was sent. An error can occur in two ways:

$$\mathcal{E}_1 = \big\{\big(X^n(1), Y^n\big) \text{ is not jointly typical}\big\},$$

or

$$\mathcal{E}_2 = \left\{\exists m \neq 1 : \left(X^n(m), Y^n\right) \text{ is jointly typical}\right\}.$$

By the law of large numbers, since $X^n(1)$ and Y^n are generated according to P_{XY}^n,

$$\Pr(\mathcal{E}_1) \to 0 \qquad \text{as } n \to \infty.$$

Now consider $\mathcal{E}_2$. For $m \neq 1$, the wrong codeword $X^n(m)$ is independent of Y^n. By the joint typicality lemma, for every $\delta > 0$ and sufficiently large n,

$$\Pr\left[\left(X^n(m), Y^n\right) \text{ is jointly typical}\right] \leq 2^{-n(I(X;Y)-\delta)}.$$

Using the union bound over all $m \neq 1$,

$$\Pr(\mathcal{E}_2) \leq \sum_{m=2}^{M} \Pr\left[\left(X^n(m), Y^n\right) \text{ is jointly typical}\right].$$

Therefore,

$$\Pr(\mathcal{E}_2) \leq (M-1)2^{-n(I(X;Y)-\delta)}.$$

Since $M = 2^{nR}$, we get

$$\Pr(\mathcal{E}_2) \leq 2^{nR}2^{-n(I(X;Y)-\delta)} = 2^{-n(I(X;Y)-R-\delta)}.$$

Choose $\delta > 0$ small enough so that

$$R < I(X;Y) - \delta.$$

Then

$$\Pr(\mathcal{E}_2) \to 0 \qquad \text{as } n \to \infty.$$

Combining the two error events,

$$P_e^{(n)} \leq \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) \to 0.$$

Thus the average probability of error, averaged over the random choice of codebook, goes to zero. Hence, for sufficiently large n, there must exist at least one deterministic codebook whose probability of error also goes to zero.

Therefore, for every $R < C$, there exists a sequence of length-n codes of rate R whose probability of error tends to zero. This proves part (i). □

2.2 Converse intuition: Fano's inequality

The converse says rates above capacity are impossible. The key is that if the receiver can decode W from Y^n with small error, then Y^n must contain almost all the information in W.

The proof uses the following theorem.

Theorem 2.3 (Fano's Inequality)**.** *Let W be uniformly distributed on $\{1, \ldots, M\}$, and let $\widehat{W}$ be an estimate of W. Define the probability of error*

$$P_e = \Pr(\widehat{W} \neq W).$$

Then

$$H(W \mid \widehat{W}) \leq H_2(P_e) + P_e \log(M-1),$$

where

$$H_2(p) = -p \log p - (1-p) \log(1-p)$$

is the binary entropy function.

Proof. Define the error indicator

$$E = \begin{cases} 1, & \widehat{W} \neq W, \\ 0, & \widehat{W} = W. \end{cases}$$

Then

$$\Pr(E = 1) = P_e.$$

We will bound $H(W \mid \widehat{W})$. Since E is determined by the pair $(W, \widehat{W})$, we have

$$H(E \mid W, \widehat{W}) = 0.$$

Therefore,

$$H(W, E \mid \widehat{W}) = H(W \mid \widehat{W}).$$

On the other hand, by the chain rule,

$$H(W, E \mid \widehat{W}) = H(E \mid \widehat{W}) + H(W \mid E, \widehat{W}).$$

Thus

$$H(W \mid \widehat{W}) = H(E \mid \widehat{W}) + H(W \mid E, \widehat{W}).$$

Now,

$$H(E \mid \widehat{W}) \leq H(E) = H_2(P_e).$$

Next, condition on whether an error occurred. If $E = 0$, then

$$W = \widehat{W},$$

so

$$H(W \mid E = 0, \widehat{W}) = 0.$$

If $E = 1$, then $W \neq \widehat{W}$, so once $\widehat{W}$ is known, W can take at most $M - 1$ possible values. Hence

$$H(W \mid E = 1, \widehat{W}) \leq \log(M-1).$$

Therefore,

$$H(W \mid E, \widehat{W}) \leq (1 - P_e) \cdot 0 + P_e \log(M-1).$$

So

$$H(W \mid E, \widehat{W}) \leq P_e \log(M-1).$$

Combining the two bounds gives

$$H(W \mid \widehat{W}) \leq H_2(P_e) + P_e \log(M-1).$$

This proves Fano's inequality. □

We can now prove the converse direction.

Proof of Part (ii) - small errors imply $R \leq C$. For a channel code,

$$W \to X^n \to Y^n \to \hat{W}.$$

By data processing,

$$I(W; \hat{W}) \leq I(X^n; Y^n).$$

For a memoryless channel,

$$I(X^n; Y^n) \leq nC.$$

If P_e is small, Fano's inequality implies

$$I(W; \hat{W}) = H(W) - H(W \mid \hat{W}) \approx \log M = nR.$$

Thus $nR \lesssim nC$, so $R \leq C$.

2.3 Repetition coding as a first error-correcting code

For BSC(p), a length-3 repetition code sends

$$0 \mapsto 000, \qquad 1 \mapsto 111.$$

The decoder uses majority vote. The rate is $R = 1/3$.
The decoding error probability is the probability that at least two of the three bits flip:

$$P_e = \binom{3}{2} p^2(1-p) + p^3 = 3p^2 - 2p^3.$$

For small p, this is about $3p^2$, which is much smaller than p. Reliability improves, but rate decreases. This is the key tradeoff:

$$\text{more redundancy} \quad \Rightarrow \quad \text{more reliability but lower rate.}$$

□