

Channel Coding Notes

Adam Kirsch

CS 222 - Fall 2006

This document gives a brief overview of the mathematical formulation of communication channels in information theory. In particular, we give and motivate the mathematical model of communication channels, state and interpret Shannon's channel coding theorem (the fundamental result in this area), and prove a weaker result that captures the basic idea of the main theorem.

We start by giving a mathematical model for a communication channel. Intuitively, a channel is a black box that takes an input symbol in some alphabet \mathcal{X} and generates an output symbol in some (possibly different) alphabet \mathcal{Y} . This process most likely allows for some uncertainty between the input and output symbols, which we model as randomness inside the box. Formally, we have the following definition.

Definition 1. A *channel*¹ with input alphabet \mathcal{X} and output alphabet \mathcal{Y} is a conditional probability distribution $p(y|x)$ defined for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

In class, we saw two concrete examples of channels: the binary symmetric channel and the binary erasure channel. The binary symmetric channel takes a bit as input and flips it with some probability q , so we have $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $p(0|0) = p(1|1) = 1 - q$ and $p(1|0) = p(0|1) = q$. The binary erasure channel erases each bit with probability q , where *erasing* means that we do not learn anything about the transmitted bit. Thus, in this case we have $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, ?\}$, where $?$ is a special erasure symbol. Then $p(0|0) = p(1|1) = 1 - q$ and $p(?|0) = p(?|1) = q$.

Note that we can connect a channel to any random source X that takes values in \mathcal{X} in a natural way. We simply imagine sending X over the channel, and let Y denote the value that comes out. The conditional probability that $Y = y$ given $X = x$ is then just $p(y|x)$. If we wish to send n symbols across the channel, then we just connect the channel to some random source X_n that takes values in \mathcal{X}^n , send X_n across the channel symbol-by-symbol, and let $Y_n \in \mathcal{Y}^n$ denote the string that comes out. (Here, we assume that the randomness used by the channel for each input symbol is independent, so the i th symbol of Y_n is conditionally independent of X_n given the i th symbol of X_n .)

If we are given some communication channel, then intuitively our objective should be to transmit as much information across the channel per input symbol as possible. In information-theoretic terms, we should choose X so that Y tells us as many bits of information about X as possible. Now, recall that $I(X; Y)$ is the mathematical formulation of the amount of information that Y tells us about X . Thus, our objective should be to choose X to maximize $I(X; Y)$. This observation leads to the next definition.

¹For readers who are already familiar with this material, note that we only consider discrete memoryless channels in this document.

Definition 2. The *information capacity* of a channel is

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

where the maximum is taken over all possible probability distributions for X (that is, all probability distributions defined on \mathcal{X}).

Definition 2 is certainly a nice theoretical definition, but it doesn't really capture any concrete, functional notion of what it means to send information over a channel. In practice, we don't really care about how many bits of information we can send across the channel. Rather, we care about whether we can send an encoding of one of (hopefully) many possible messages across the channel symbol-by-symbol and then expect to be able to determine what message was sent from the received sequence. To formalize this objective, we first introduce the notion of a (channel) code.

Definition 3. An (M, n) code for the channel $p(y|x)$ is defined by:

- An *encoding function* $\text{Enc} : \{1, \dots, M\} \rightarrow \mathcal{X}^n$. In this context, the elements of $\{1, \dots, M\}$ are called the *messages*, the set $\{\text{Enc}(1), \text{Enc}(2), \dots, \text{Enc}(M)\}$ is called the *codebook*, and the strings in the codebook are called *codewords*.
- A *decoding function* $\text{Dec} : \mathcal{Y}^n \rightarrow \{1, \dots, M\}$. In this context, the strings in \mathcal{Y}^n are called *possible received sequences*, and for $y \in \mathcal{Y}^n$, the message $\text{Dec}(y)$ is called the *decoded message for y* .

Definition 3 gives us a formal notion of what it means to encode one of several possible messages, send it across the channel symbol-by-symbol, and then decode the received sequence to (hopefully) obtain the original message. Now intuitively, if we send a random message from an (M, n) code across the channel and then successfully decode it, then we have sent $\log M$ bits of information across the channel. Since each message is encoded by an n -symbol string, we have received about $(\log M)/n$ bits of information across the channel for every symbol transmitted. Thus, we have the natural definition:

Definition 4. The *rate* of an (M, n) code is

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

Of course, the intuition above relies on us successfully decoding the transmitted message. But an error may occur in this process, as the transmitted and decoded messages might not be equal. If the code is well-designed, then these errors can be attributed to the randomness (such as errors) introduced by the channel during transmission. In any case, we need formal definitions and expressions for various probabilities related to errors.

Definition 5. Fix an (M, n) code and some $i \in \{1, \dots, M\}$. Let $Z_i \in \mathcal{Y}^n$ denote the string that is received when $\text{Enc}(i)$ is sent across the channel. Next, let

$$\lambda_i = \Pr(\text{Dec}(Z_i) \neq i)$$

denote the probability that sending $\text{Enc}(i)$ across the channel and decoding the received sequence results in an error. Also, let

$$\bar{\lambda} = \frac{1}{M} \sum_{i=1}^M \lambda_i$$

denote the average probability of an error. (Note that $\bar{\lambda}$ is exactly the probability of an error when i is chosen uniformly at random from $\{1, \dots, M\}$.) Finally, let

$$\lambda^* = \max_{i \in \{1, \dots, M\}} \lambda_i$$

denote the maximal probability of error for a message in the code.

If the error probabilities for an (M, n) code are small, then the intuition behind the definition of the rate $R = (\log M)/n$ of the code is valid, and therefore about R bits of information are transmitted across the channel for every symbol sent. It is then natural to ask how big R (and therefore M) can be before the intuition no longer holds (implying that the error probabilities are no longer small). But we know from before that, in information-theoretic terms, the maximal number of bits of information that can be transmitted across the channel per symbol sent is exactly C , the information-theoretic capacity of the channel. Thus, it is natural to conjecture that the maximal value of R for which our intuition works is exactly C . Indeed, this is exactly the case; this result is called the channel coding theorem.

Definition 6. A rate R is *achievable* if there is a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that $\lambda^* \rightarrow 0$ as $n \rightarrow \infty$. The *operational capacity* of a channel is the supremum of all achievable rates.

Theorem 1 (Shannon's Channel Coding Theorem). *The information capacity and the operational capacity are equal.*

As an aside, we show below that Theorem 1 still holds if we replace λ^* with $\bar{\lambda}$ in Definition 6. Typically, the proof of Theorem 1 is given in two parts:

Theorem 2. *The operational capacity is at most the information capacity.*

Theorem 3. *The operational capacity is at least the information capacity.*

Theorem 2 can be proven directly using basic information theory (using material covered in this course), but we omit the full proof here for reasons of space. However, to show that the intuition above really is amenable to a formal argument, we give a brief sketch.

Proof Sketch for Theorem 2. Suppose that R is an achievable rate, so there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes with $\lambda^* \rightarrow 0$ as $n \rightarrow \infty$. We need to show that $R \leq C$. To this end, fix some n , and let W be a uniformly chosen random element of $\{1, \dots, \lceil 2^{nR} \rceil\}$ (that is, W is a random message of the code). Then

$$nR \leq \log_2 \lceil 2^{nR} \rceil = H(W) = I(W; Z) + H(W | Z). \quad (1)$$

Now we hand-wave a bit. Recall that $I(W; Z)$ is the information that the output of the channel tells us about the input. Since there are n input symbols, we should have $I(W; Z) \leq nC$ by definition of C . Furthermore, since $\Pr(\text{Dec}(Z) \neq W) = \bar{\lambda} \rightarrow 0$ as $n \rightarrow \infty$, we have that W is essentially determined by Z , and therefore we should have $H(W | Z)/n \rightarrow 0$ as $n \rightarrow \infty$. Substituting into (1) now gives

$$R \leq C + \frac{H(W | Z)}{n} \rightarrow C \quad \text{as } n \rightarrow \infty,$$

which completes the proof. □

The proof of Theorem 2 goes through even if we relax the definition of an achievable rate to allow for rates R where there are sequences of $(\lceil 2^{nR} \rceil, n)$ codes with $\bar{\lambda} \rightarrow 0$ as $n \rightarrow \infty$. (This is because we only need $\bar{\lambda} \rightarrow 0$, as opposed to the stronger condition $\lambda^* \rightarrow 0$.) In other words, the channel coding theorem still holds if we replace the notion of maximal error with average error in Definition 6.

The proof of Theorem 3 is outside the scope of this class. However, we give a proof for the binary symmetric channel that is in the same spirit as the proof of the general result. In fact, Theorem 3 can be proven by an appropriate generalization of this argument.

Recall that the binary symmetric channel takes a bit as input and flips it with some probability q , so we have $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $p(0|0) = p(1|1) = 1 - q$ and $p(1|0) = p(0|1) = q$.

Theorem 4. *For the binary symmetric channel, if we replace λ^* with $\bar{\lambda}$ in Definition 6, then the operational capacity is at least the information capacity.*

Proof. Fix any $R < C$ and let $M = \lceil 2^{nR} \rceil$. We wish to show that there is a sequence of (M, n) codes with $\bar{\lambda} \rightarrow 0$ as $n \rightarrow \infty$. We do this by choosing a codebook of strings of length n randomly, defining the functions Enc and Dec appropriately, and then showing that for this code $\mathbf{E}[\bar{\lambda}] \leq f(n)$, for some function f with $\lim_{n \rightarrow \infty} f(n) = 0$. This shows the existence of a sequence of (M, n) codes such that for sufficiently large n , we have $\bar{\lambda} \leq f(n) \rightarrow 0$, which completes the proof.

We now proceed formally. Let X_1, X_2, \dots, X_M be independent random strings in $\{0, 1\}^n$; these are our codewords. Define $\text{Enc} : \{1, \dots, M\} \rightarrow \{0, 1\}^n$ by $\text{Enc}(i) = X_i$. Now we need to figure out how to choose $\text{Dec} : \{0, 1\}^n \rightarrow \{1, \dots, M\}$. Recall that when a string x is sent across the channel, we get out some string Z_x where each bit of x is flipped independently with probability q . Let F denote the number of bits that are flipped by the channel. By the Chernoff bound given in Lemma 1 (in the appendix), we have that for any $\delta \in (0, 1)$,

$$\Pr(|F - nq| > \delta nq) < 2e^{-\delta^2 nq/3}.$$

This observation suggests the following choice of Dec. We fix some small $\delta \in (0, 1)$, to be specified later. Then, for each $y \in \{0, 1\}^n$, we choose $\text{Dec}(y)$ to be some value of i such that X_i differs from y in at least $(1 - \delta)nq$ bits, but no more than $(1 + \delta)nq$ bits. If there is no such value of i , then we set $\text{Dec}(y)$ arbitrarily.

Now recall that $\bar{\lambda}$ is the conditional probability, given the codewords, that the transmission of a randomly chosen message results in an error. Thus, if J is chosen randomly from $\{1, \dots, M\}$, then $\mathbf{E}[\bar{\lambda}]$ is simply the probability that transmission of X_J yields a string Z_J with $\text{Dec}(Z_J) \neq J$. By our previous observations, we have that $\text{Dec}(Z_J) \neq J$ only if $|F - nq| > \delta nq$ or there is some $i \neq J$ for which X_i differs from Z_{X_J} in at least $(1 - \delta)nq$ bits, but no more than $(1 + \delta)nq$ bits.

We have already bounded the probability that $|F - nq| > \delta nq$, so we focus on the second event. The set S of strings that differ from Z_{X_J} by at least $(1 - \delta)nq$ bits, but no more than $(1 + \delta)nq$ bits, has cardinality at most

$$\sum_{i=\lceil (1-\delta)nq \rceil}^{\lfloor (1+\delta)nq \rfloor} \binom{n}{i} \triangleq s.$$

Now, for any $i \neq J$, the string X_i is random and independent of Z_J , and therefore $\Pr(X_i \in S \mid i \neq J) \leq s2^{-n}$. A union bound now gives

$$\Pr(\exists i \neq J : X_i \in S) \leq \lceil 2^{nR} \rceil s2^{-n} \leq 2^{n(R-1)+\log s+1},$$

(where the final 1 is just to eliminate the ceiling). Applying another union bound now gives

$$\begin{aligned} \mathbf{E}[\bar{\lambda}] &\leq \mathbf{Pr}(|F - nq| > \delta nq) + \mathbf{Pr}(\exists i \neq J : X_i \in S) \\ &\leq 2e^{-\delta^2 nq/3} + 2^{n(R-1)+\log s+1} \end{aligned} \quad (2)$$

Now, choose an integer $m \in [\lfloor (1 - \delta)nq \rfloor, \lceil (1 + \delta)nq \rceil]$ to maximize $\binom{n}{m}$. Define r so that $m = n(q + r)$, and note that $r \leq \delta$. Then, using the fact that

$$\log(t!) = \sum_{i=1}^t \log i = \int_1^t \log x \, dx \pm O(\log t) = t \log t - \frac{t}{\ln 2} \pm O(\log t) \quad \text{as } t \rightarrow \infty,$$

we have

$$\begin{aligned} \log s &\leq \log \left(O(n) \binom{n}{m} \right) \\ &= \log(O(n)) + \log(n!) - \log(m!) - \log((n - m)!) \\ &= (n \log n - n/\ln 2) - (m \log m - m/\ln 2) - ((n - m) \log(n - m) - (n - m)/\ln 2) + O(\log n) \\ &= n \log n - m \log m - (n - m) \log(n - m) + O(\log n) \\ &= n \log n - \left(m \log n + m \log \frac{m}{n} \right) - \left((n - m) \log n + (n - m) \log \frac{n - m}{n} \right) + O(\log n) \\ &= -m \log \frac{m}{n} - (n - m) \log \frac{n - m}{n} + O(\log n) \\ &= -n(q + r) \log(q + r) - n(1 - (q + r)) \log(1 - (q + r)) + O(\log n) \\ &= nH(q + r) + O(\log n). \end{aligned}$$

Substituting $\log s \leq nH(q + r) + O(\log n)$ into (2) now gives

$$\mathbf{E}[\bar{\lambda}] \leq 2e^{-\delta^2 nq/3} + 2^{n(R-1-H(q+r))+O(\log n)} = 2e^{-\delta^2 nq/3} + 2^{n[(R-C)+(H(q+r)-H(q))+O(\log n)]},$$

where we have used the fact that $C = 1 - H(q)$ (proven in class). Recalling that $r \leq \delta$ and $R < C$, the continuity of $H(\cdot)$ tells us that for sufficiently small $\delta > 0$, we have $|H(q+r) - H(q)| \leq (C - R)/2$. It follows that we can choose $\delta > 0$ so that

$$\mathbf{E}[\bar{\lambda}] \leq 2e^{-\delta^2 nq/3} + 2^{n(R-C)/2+O(\log n)} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

completing the proof. □

A Chernoff Bound

The following Chernoff bound is often useful for arguing that a sum of independent random variables is tightly concentrated around its expectation.

Lemma 1. *Let X_1, \dots, X_n be independent binary random variables, let $X = \sum_{i=1}^n X_i$, and let $\mu = \mathbf{E}[X]$. Then for any $\delta \in (0, 1)$*

$$\mathbf{Pr}(|X - \mu| > \delta\mu) < 2e^{-\delta^2 \mu/3}.$$