Final Project - Shannon's Noisy Channel

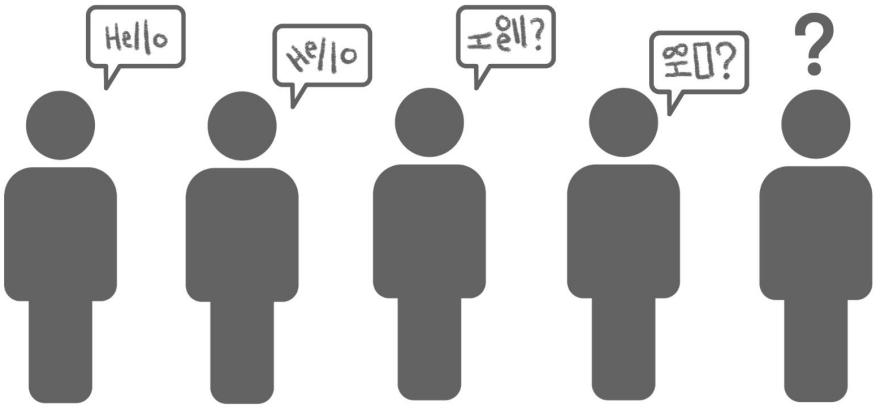

Figure 1: The classic game *telephone* provides a nice example of a noisy channel.[10]

1 Background

In 1948, Claude Shannon authored "A Mathematical Theory of Communication"[1], and later co-authored the book "The Mathematical Theory of Communication" with Warren Weaver in 1998[2]. These texts' ideas are foundational to the field of information theory. Most significantly, they introduce the noisy channel theorem (Part of which is covered in Theorem 8) and the notion of channel capacity (Definition 12). I am proving part of the noisy channel theorem, a result which shows that we can communicate efficiently despite noise.

Communication through noise and information theory arises in a variety of contexts. The application which is related most directly, and on which the field was initially motivated with, is communication over distances: if I am sending you a message, how can I ensure that you can understand it it. There are other applications and generalizations as well. For example, a 2008 paper showed that it was mathematically possible for the brain to receive signals correctly through neuronal noise[3]. It has ties to statistics, notions of lossless compression of information[4], statistical mechanics[5], musical analysis[6] and more.

I am most interested in this topic is its relation to statistics. For example, we may view observations of an individual as observations through a noisy channel—while this channel is much more complex than the one we will deal with here, the question still stands: when we make a conclusion based off of these observations, how do we know that we are fitting something significant instead of noise? Furthermore, how confident can we be in sparse data of groups of individuals? In a similar vein, if our data generating process evolves over time so that later observations may not be *iid* from earlier observations, and filtering through noise becomes even more important.

The following set of definitions and theorems are largely adapted from Blitzstein & Hwang's "Introduction to Probability"[7] section on information theory as well as from class-work and common knowledge, however I will denote where information explicitly comes from other sources. Shannon and Weaver's book mentioned earlier[2] is a good reference

for information theory. For modern sources, Henry McKean's "Probability: The classical limit theorems"[8] has a chapter on noisy channels and David MacKay's "Information Theory, Inference, and Learning Algorithms"[9] has numerous chapters introducing information theory, touching on concepts here, before going into specific applications.

2 Theory

Definition 1: (Entropy) Let X be a discrete random variable with pmf $\mathbf{p}$. We define the *entropy* of X by the formula

$$H[X] = -\sum_k \mathbf{p}(k) \log\big(\mathbf{p}(k)\big).$$

$\triangledown$

For notational purposes, we will often refer to the entropy of a Bernoulli random variable with probability p as $H(p) = -p\log(p) - (1-p)\log(1-p)$. The main following properties from the above formula is that (a) the entropy equals zero exactly when the probability of a single event is one, and (b) the entropy is maximized when the pmf is uniform. Entropy can be thought of as a measure of uncertainty present within a random variable. When we have a degenerate random variable, its entropy is zero. Furthermore, we have maximal entropy when we are the most uncertain about the result of an experiment.

In Shannon's original paper, entropy was defined in terms of the base-2 logarithm[1], however it can be equivalently defined using the natural logarithm, as seen on Midterm 1, question 3—the two definitions only vary by a factor of $\log_2(e)$, so the choice of logarithm base is arbitrary; we will use the base-2 logarithm. Entropy also has an analogous definition for continuous random variables, however we will not be interacting with this definition.

Definition 2: (Vocabulary) A *message* $x \in \{0,1\}^k$ is a binary string. A *code* is a function $c : \{0,1\}^k \to \{0,1\}^n$ that brings messages, to what we call *signals* or *encoded messages*. We call the quantity $\frac{k}{n}$ the *transmission rate* of the code. Generally we call any $w \in \{0,1\}^m$ a *string*; messages and signals are both strings with possibly different m. Finally introduce the indicator function $I : A \times A \to \{0,1\}$ for nonempty set A by the relation $I_{j,k} = 1$ if $j = k$ and zero otherwise. $\triangledown$

The choice to use the binary set $\{0,1\}^m$ is arbitrary, this can switched for other "alphabets"—say, the set of letters {a,b,. . . , y, z}—however from here on, we will strictly use the binary set. The transmission rate of a code can be thought of as how much information about a message x is carried by any entry in a signal $c(x)$.

For example, if I gave a code c the message "011", and it returns the signal "011011011", we would say that this code is the repetition code with transmission rate of one third; further, on average we are getting one third of a character of the message per character in the signal. Generally we the rate of a code is less than or equal to one, since otherwise we have ineffective communication.

Definition 3: (Distance) Define $d : \{0,1\}^m \times \{0,1\}^m \to \mathbb{R}$ to be the *distance* between two strings, that is, for $w, v \in \{0,1\}^m$, we say that

$$d(w,v) = \sum_{k=1}^{m} |w_k - v_k|.$$

▽

This definition can similarly be extended to other alphabets of messages (as discussed earlier), blurring the definition of strings, by defining the distance between strings v and w of length m as the sum $d(w,v) = \sum_{k=1}^{m} I_{w_k,v_k}$. The distance is a metric of how many entries are equal. For example, the vectors $(0,0,1)$ and $(1,0,1)$ have a distance of 1 since only one entry is different between them. Distance is useful for measuring how much a string changes after going through some process. We now introduce such a process.

Definition 4: (Noisy Channel, Received Signal) A *noisy channel* is a medium through which, passing signals independently get entries flipped randomly according to some noise parameter p. Mathematically, we define a noisy channel to be a function $Q : \{0,1\}^n \times [0,1] \to \{0,1\}^n$ taking an input of a string w and noise parameter p satisfying $0 < p < 0.5$. At each entry k of $Q(w;p)$ we have $P_{Noise}\Big(\big[Q(w;p)\big]_k = w_k\Big) = 1 - p$. We call the string $Q(w;p)$ the *received signal*. The subscript $Noise$ denotes that we are taking the probability with respect to the noise in Q; we will similarly define $\mathbb{E}_{Noise}$.

▽

We defined that p can be in the unit interval but specify that $0 < p < 0.5$ so discussions where $0.5 \leq p$ or $p = 0$ make sense. Specifying that $0 < p < 0.5$ instead of $0 < p < 1$ comes down to a matter of simplicity: the case where $p = 0.5$ results in a string which is independent of the original as we will show in the follow example. Moreover, it is unnecessary to consider p with $p > 0.5$ since that is similar to having p less than 0.5, requiring only that we flip every single bit of the received message to get an equivalent $p < 0.5$. The next two examples show first the claim mentioned above, and then how the expected distance between a string w and $Q(w;p)$ behaves.

Example 5: ($Q(w;0.5)$ is independent of w) Let w be a string, the string $Q(w;0.5)$ resulting from passing w through a noisy channel with noise parameter $p = 0.5$ is independent of w.

Proof.
For ease call $w' = Q(w;0.5)$. Define random variables W_k and W'_k by $P_{Noise}(W_k = w_k) = 1$, $P_{Noise}(W'_k = w_k) = 0.5$, and hence $P_{Noise}(W'_k = |w_k - 1|) = 0.5$. So W_k is a degenerate random variable always returning the value of w_k, and W'_k is a Bernoulli random variable we may use to model w'_k. With these, for $b, b' \in \{0,1\}$, $P_{Noise}(W_k = b \text{ and } W'_k = b') = 0.5 I_{W_k,b}$. Though observe that these are the products of probabilities of each event:

$$P_{Noise}(W_k = b \text{ and } W'_k = b') = P_{Noise}(W_k = b) P_{Noise}(W'_k = b').$$

Thus each entry of w_k and w'_k are independent; Further, since Q applies noise independently to each entry of w, this holds for all entries and we have that the two strings are independent.
□

To call w and $Q(w;0.5)$ independent is a slight abuse of notation, however this can be corrected by saying that W and $Q(w;0.5)$ are independent where W is a degenerate string, always returning w. Since w is independent of $Q(w;0.5)$ in this sense, we are not able to gain any insight about w by investigating $Q(w;0.5)$. We now move to the next example.

Example 6: (Expected Distance in a noisy channel) Let $w \in \{0,1\}^m$ be a string sent across a noisy channel with noise parameter p. Then $\mathbb{E}\left[d(w, Q(w;p))\right] = mp$.

Proof.
Consider the indicator variable $\mathcal{I}_k = 1 - I_{[Q(w;p)]_k, w_k}$. Observe that we may also write that $\mathcal{I}_k = |[Q(w;p)]_k - w_k|$. By this characterization, $\mathcal{I}_k$ is a binomial random variable with probability p and size m. The desired result follows by the expectation of the binomial.

□

This example will help us in our proof of theorem 8, giving us expected behavior of distance in noisy channels, thus allowing us to invoke the law of large numbers.

Definition 7: (Decoding & Decoder Failure) Say we have a received signal Y—that is, for some message x, code c, and noise parameter p, $Y = Q(c(x);p)$. Define another probability p' with $p < p' < 0.5$. In the event where there is a unique message z satisfying $d(c(z), Y) < np'$, we say that Y decodes to z. If this $z = x$, then we have successfully decoded Y. If there is not a unique message satisfying the inequality or if $z \neq x$, declare *decoder failure*. There are two events which can trigger a decoder failure:

(DF1) we are unable to recover the message x from the received signal; in other words, $d(c(x), Y) > np'$, or

(DF2) For some message $z \neq x$, we decode Y to z. In other words, $d(c(z), Y) < np'$.

▽

Two things to note about this definition. First, this notion of failure has inherent ambiguity: we have defined failure with respect to some unknown p' which is *close* to p. While we will leverage this p' in the proof of Theorem 8 (Shannon's noisy channel), it is worth noticing this ambiguity. With the above definition in combination with Definition 2 and Definition 4, we can now understand the following schematic of a noisy channel:

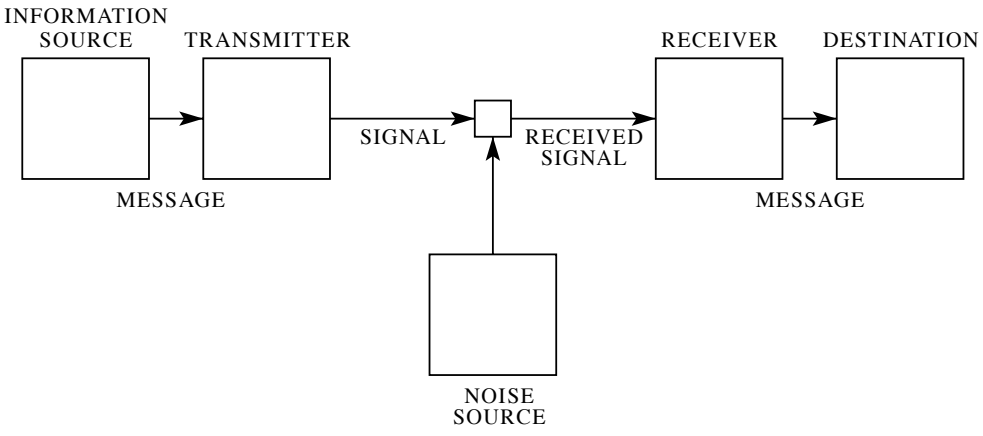

Figure 2: Schematic Diagram for a noisy channel.[1]

When dealing with the probability of decoder failure, we will often use the identity that $P(DF1 \text{ or } DF2) \leq P(DF1) + P(DF2)$. Citing the principle of inclusion/exclusion we

calculate that the $P(DF1 \text{ or } DF2) = P(DF1) + P(DF2) - P(DF1 \text{ and } DF2)$, and neglect the third term. Now we move to the main result discussed here.

Theorem 8: (Shannon's Noisy Channel) Consider a noisy channel with noise parameter $0 < p < 0.5$. Then for any positive ε, there is a code with transmission rate at least $1 - H(p) - \varepsilon$ that can be decoded with probability of decoder failure less than ε.

$\triangledown$

To illustrate why this is significant, consider the repetition code from earlier: For any message $x \in \{0,1\}^k$, $c_m(x) = \bigcup_{i=1}^{m} x$—that is, c repeats x m times. If we want to get the probability that we make some error when decoding c to be smaller and smaller, the only way for this to happen is to make m larger and larger, meaning that for the repetition code, the rate $\frac{1}{m}$ tends towards zero as we demand our probability of error go towards zero. This theorem states that we can communicate, provided strings are long enough, with reasonable efficiency—although it is difficult to show such an example explicitly. In order to show this theorem, we will use something called the probabilistic method: to prove that there is an item with a property, we uniformly randomly draw from the set of such items and show that the probability that our draw has the property is nonzero: the following simple example illustrates this:

Example 9: (Probabilistic Method) Let P_3 be the (unordered) elements of length three in the power set of $\{1, \ldots, n\}$ for $3 < n$. Show that the there is an element in P_3 which contains 1 and 2 but not n.

Proof.
Call S the event that the first two values of an element are 1 and 2 and the last is not n, this is the required condition. It is simple enough to nominate the element $\{1,2,3\}$, but we will instead use the probabilistic method. Take a uniformly random X in P_3. Assume that the first two entries are 1 and 2, call this event S'. Since the first two values are fixed, there are only $n-3$ possible values for the last entry that are not n. Since n is at least 4, there must be at least one possible value for X. Then citing Cardano's rule we have that

$$P(X \text{ satisfies } S \mid X \text{ satisfies } S') = \frac{n-3}{n-2}.$$

Similarly, counting the number of X which satisfy S' is equivalent to counting the number of elements which go in the third entry (fixing the first two). This is $n-2$ and so

$$P(X \text{ satisfies } S') = \frac{n-2}{\binom{n}{3}} = \frac{n-2}{\binom{n}{3}}.$$

It follows that by using the rule for conditional probabilities, noting that $S' \subseteq S$,

$$P(X \text{ satisfies } S')P(X \text{ satisfies } S \mid X \text{ satisfies } S') = P(X \text{ satisfies } S).$$

$$\frac{n-2}{\binom{n}{3}}\frac{n-3}{n-2} = \frac{n-3}{\binom{n}{3}}$$

Since this is a nonzero probability, there must be events where it occurs. Furthermore, since n choose 3 is the size of P_3, we see that there are $n-3$ such elements in P_3.

$\square$

The following lemmas will prove useful in our proof of Shannon's noisy channel:

Lemma 10: (Good Score Principle) For a discrete random variable X and positive $0 \leq c$, if $\mathbb{E}[X] < c$ then $P(X < c) \neq 0$, or equivalently, if $P(X < c) = 0$ then $c \leq \mathbb{E}[X]$.

Proof.
The second implication follows from the definition of expected value: if $P(X < c) = 0$, then

$$\mathbb{E}[X] = \sum_k k\mathbf{p}_X(k) = \sum_{c \leq k} k\mathbf{p}_X(k) \text{ and so } c = c\sum_{c \leq k} \mathbf{p}_X(k) \leq \sum_{c \leq k} k\mathbf{p}_X(k).$$

□

Lemma 11: (Upper bounding the binomial coefficient) For integers j, n with $j < n$ and generic real number $0 < r < 1$, we have the relation

$$\binom{n}{j} < r^{-j}(1-r)^{-(n-j)}.$$

Proof.
Let $0 < r < 1$, the binomial theorem tells us that

$$1 = 1^n = (1 - r + r)^n = \sum_{j=0}^{n} \binom{n}{j} r^j (1-r)^{n-j}.$$

The summand is positive for each value of j so picking out any term in the sum, we have $\binom{n}{j} r^j (1-r)^{n-j} < 1$, and the result follows.

□

Now we may move onto the proof of Theorem 8

Proof. (Theorem 8)
To begin, let's establish our setting: We have some noise parameter p, a message $x \in \{0,1\}^k$, a code c, and a received signal $Y = Q(c(x); p) \in \{0,1\}^n$. Define another probability $0 < p' < 0.5$ so that $H(p') - H(p) < \frac{\varepsilon}{2}$. Also, without loss of generality, choose p' so that np' is an integer—that we may do this follows from $\mathbb{R}$ satisfying the Archimedean property and our later stipulations on the size of n. We also assume ε is small enough so that the rate of our code is positive.

Our claim is that there is some code c with a sufficiently large transmission rate that can be decoded with probability of error less than ε. Let C be a *uniformly random* code in the set of possible codes with transmission rate $\lceil 1 - H(p) - \varepsilon \rceil$, this means equivalently that $k = n\lceil (1 - H(p) - \varepsilon) \rceil$. Notice that for our purposes $C(x)$ is a random sring with each entry being a Bernoulli random variable with probability equal to 0.5 independent of the others.

Now recall that we define decryption error as the events (DF1) $d(C(x), Y) > np'$ or (DF2) for another string $z \neq x$, $d(C(z), Y) < np'$. Since each of these events relies both on our random code C and the noisy process Q, the probabilities $P_{Noise}(DF1)$ or $P_{Noise}(DF2)$ are still random variables since these do not account for the codes. Therefore, we will consider the expectations $\mathbb{E}_{Code}[P_{Noise}(DF1)]$ and $\mathbb{E}_{Code}[P_{Noise}(DF2)]$.

That we may minimize the expected probability of DF1, or have $\mathbb{E}_{Code}[P(DF1)] < \varepsilon/2$ follows directly from example 6 and the law of large numbers: the expected value of this

distance (across the channel noise) is np, so if we choose a large enough n, the LLN tells us that the distance between $d(C(x), Y) = d(C(x), Q(C(x), p)$ and np will become as small as we wish—choose n to be large enough such that the probability that this occurs is less than $\varepsilon/2$. Observe also that even though we did not address the randomness of C, the associated randomness vanished when considering the distance, this is because Example 6 considers general w, and so setting $w = C(x)$, we have the same result regardless of C. Therefore, $\mathbb{E}_{Code}[P_{Noise}(DF1)] < \varepsilon/2$.

The proof for (DF2) is a bit more involved. To show $\mathbb{E}_{Code}\left[P_{Noise}(DF2)\right] < \varepsilon/2$, or explicitly, $\mathbb{E}_{Code}\left[P_{Noise}\left(d(C(z), Y) \leq np'\right)\right] \leq \varepsilon/2$ it suffices to prove that

$$P_{Noise}(\mathbb{E}_{Code}[d(C(z), Y)] \leq np') < \varepsilon/2.$$

Indeed, this is sufficient since we can now prove that

$$\mathbb{E}_{Code}\left[P_{Noise}\left(d(C(z), Y) \leq np'\right)\right] \leq P_{Noise}(\mathbb{E}_{Code}[d(C(z), Y)] \leq np').$$

Summing over all sample outcomes of the of the randomness of the noise, and applying linearity of expectation, we have

$$\mathbb{E}_{Code}\left[\sum_{w\,:\,\text{DF2}} d(C(z), C(x) + w)\right] \leq \sum_{w\,:\,\mathbb{E}_{Code}[\text{DF2}]} \mathbb{E}_{Code}\left[d(C(z), C(x) + w)\right],$$

$$\sum_{w\,:\,\text{DF2}} \mathbb{E}_{Code}[d(C(z), C(x) + w)] \leq \sum_{w\,:\,\mathbb{E}_{Code}[\text{DF2}]} \mathbb{E}_{Code}\left[d(C(z), C(x) + w)\right].$$

The notation w : DF2 is saying that assuming no other channel noise, sum over all w which cause a decoder to have a DF2 error—that is, we will sum over all w which make $d(C(z), C(x) + w \leq np'$, similarly with $\mathbb{E}_{Code}[\text{DF2}]$. Expectations are less restrictive than singular values: as a simple example, if a random variable $X \leq 391$ always, then $\mathbb{E}[X] \leq 391$ always. But if $\mathbb{E}[X] \leq 391$, this does not necessarily mean that $X \leq 391$. Therefore, there are more terms in the right summation than the left and thus the inequality holds.

Now consider $P_{Noise}\Big(\mathbb{E}_{Code}\left[d(C(z), Y)\right] \leq np'\Big)$. Since C is random, averaging over all possible codes allows us to model $C(z)$ as a random string of n *iid* Bernoulli random variables each with probability of $1/2$. Then, compute that

$$P_{Noise}\Big(\mathbb{E}_{Code}[d(C(z), Y)] \leq np'\Big) \propto \sum_{i=0}^{np'} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} = \sum_{i=0}^{np'} \binom{n}{i} \frac{1}{2^n}$$

by summing over the i which would make this true. However the above expression is for a particular fixed z; we want general $z \in \{0,1\}^k$. Note that $z \neq x$ so we instead need to compute

$$P_{Noise}\Big(\mathbb{E}_{Code}[d(C(z), Y)] \leq np'\Big) = (2^k - 1) \sum_{i=0}^{np'} \binom{n}{i} \frac{1}{2^n} = \frac{(2^k - 1)}{2^n} \sum_{i=0}^{np'} \binom{n}{i}.$$

We cannot easily compute this summation so we will upper bound it. Observe that $np' < \frac{n}{2}$ since $p' < \frac{1}{2}$. This means that the largest binomial coefficient in our sum is $\binom{n}{np'}$. Therefore

$$< \left((np' + 1)\frac{2^k - 1}{2^n}\right)\binom{n}{np'}, \qquad \text{(Upper Bounding sum)}$$

$$< \left((np' + 1)\frac{2^k - 1}{2^n}\right) r^{-np'}(1 - r)^{-n(1-p')}, \quad \text{(Lemma 11)}$$

$$= \Big((np' + 1)\frac{2^k - 1}{2^n}\Big)p'^{-np'}(1-p')^{-n(1-p')}, \quad (\text{set } r = p')$$

Now since 2^x and $\log_2(x)$ are inverses, we may put the right term in an exponent:

$$= \Big((np' + 1)\frac{2^k - 1}{2^n}\Big)2^{-n(p'\log_2(p)+(1-p')\log_2(1-p'))},$$

($2^{\log_2}$ is the identity)

$$= (np' + 1)\left(2^k - 1\right)2^{n(H(p')-1)}. \qquad (\text{Recognizing } H(p'))$$

Now, we simplify the expression and substitute in variables to arrive at

$$\leq \Big(np' + 1\Big)2^{n(H(p')-1)+k}, \qquad ((2^k - 1) < 2^k)$$

$$< \Big(np' + 1\Big)2^{n(H(p)+\varepsilon/2)-n+k}. \qquad (H(p') - H(p) < \varepsilon/2)$$

Recall that we have defined k by the relation $k = n\lceil 1 - H(p) - \varepsilon\rceil$, so substitute this into the above expression to get

$$= \Big(np' + 1\Big)2^{(n(H(p)+\varepsilon/2))-n+\lceil n(1-H(p)-\varepsilon)\rceil+1},$$

$$\leq \Big(np' + 1\Big)2^{\lceil -n\varepsilon/2\rceil+2}, \qquad (\lceil a\rceil + b \leq \lceil a + b\rceil + 1)$$

$$= 8\Big(np' + 1\Big)2^{\lceil -n\varepsilon/2\rceil-1}.$$

$$\leq 8(np' + 1)2^{-n\varepsilon/2}. \qquad (\lceil -a\rceil - 1 \leq -a)$$

This goes to zero with n since exponential decay dominates linear growth, so make sure n is large enough such that this is less than $\varepsilon/2$. Since $P_{Noise}\Big(\mathbb{E}_{Code}\left[d(C(z), Y)\right] \leq np'\Big) < \varepsilon/2$, by our above work we can claim that $\mathbb{E}_{Code}\Big[P\left(d(C(z), Y) \leq np'\right)\Big] < \varepsilon/2$.

We have that $\mathbb{E}_{Code}[P\left(d(C(z), Y) \leq np'\right)] < \varepsilon/2$ and $\mathbb{E}_{Code}[P\left(np' \leq d(C(x), Y)\right)] < \varepsilon/2$. So applying linearity of expectation and the logic after definition 7, we can claim that $\mathbb{E}_{Code}[P(DF1 \text{ or } DF2)] < \varepsilon$. Thus by the good score principle (Lemma 10), there is a code with rate at least $\lceil 1 - H(p) - \varepsilon\rceil$ so that the probability of failure is less than ε for x, as desired. Further since we used generic x, we have proved our result.

□

In the context of the probabalistic method, we average across codes, the sample space of which are extremely large; calling the sample space Ω, we have that

$$|\Omega| = |\{0,1\}^n|^{|\{0,1\}^k|} = (2^n)^{(2^k)}.$$

For large n and k, there are too many possible codes to consider any individual code, thus necessitating the use of the probabilistic method. The wording of this proof is also important, we show that for rates greater than $1 - H(p) - \varepsilon$, we can have probability of

error less than ε. However, this does not mean that any rate above this threshold has such an error rate. Indeed, there is something called the *Channel Capacity* which is the largest communication rate in a given channel. I will define it qualitatively from McKean's "Probability: The classical limit theorems"[8].

Definition 12: (Channel Capacity) In a noisy channel—not necessarily communicating strings of binary messages—the maximum communication rate at which messages can be sent with probability of decryption error less than ε is the channel capacity C. Furthermore, communicating at a transmission rate higher than C results in necessary nontrivial probabilities of communication error.

∇

I will not introduce how to find C as this requires much more machinery than this project has time for. In Shannon's paper, channel capacity is defined as a construct and then the above definition is a theorem, however since this is out of the scope of this project, I introduce this as fact. McKean computes that the channel capacity for the binary channel—the one we have been using thus far—is $1 - H(p)$ meaning that Theorem 8 is stating that we may get as close as we wish to the channel capacity and still communicate with arbitrarily small probability of error, however once we exceed $1 - H(p)$, we cannot have such efficient communication.

3 Concluding Thoughts:

In closing, communication is noisy, and we are able to communicate through this noise with low probabilities of error, but this takes work. Even though we may send messages a fixed communication rate, this rate requires that our n get longer (and our k correspondingly), meaning that we still have to send longer messages across channels. Of course, in reality we are alright with some error, we will not demand that we have an astronomically low probability of error, however that we could mathematically demand this is of our code is satisfying.

One of the ideas I will certainly take from this project is the probabilistic method as a method for proving existence. While it takes some setting up, this is a very useful tool, and I'm surprised that it only got a slight mention in Blitzstein & Hwang and nothing in [ASV]. Also, while I did not present it much here, I learned a lot about how to use entropy to inform statistical intuition. For example, if we are choosing a prior distribution of a parameter and we are unsure on the structure of such a parameter, it is a good idea to choose a distribution in a family which has the highest entropy because it allows for the most uncertainty.

Finally, this project, mostly through working with actual sources, has taught me a respect for how we treat random variables and randomness in this class. Many of the sources I read had a much easier time working with these ideas because they brushed off some of their dealings with the randomness. However, working through this randomness gives a much better understanding of the mechanics of the proof and why the claim works the way it does. Alex certainly talked at length about how in the "real world" we will have to deal with people not quite respecting randomness, but I don't think I quite internalized it until now.

4 References

1. Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

2. Shannon, C. E., & Weaver, W. (1998). The mathematical theory of communication. Univ. of Illinois Press.

3. Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. Nature Reviews Neuroscience, 9(4), 292–303. https://doi.org/10.1038/nrn2258

4. Zhang, P. (2008). Data Communications in Distributed Control System. In Industrial Control Technology (pp. 675–774). Elsevier. https://doi.org/10.1016/B978-0815515771-5.50007-4

5. Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. Physical Review, 106(4), 620–630. https://doi.org/10.1103/PhysRev.106.620

6. Loy, D.G. (2017). Music, Expectation, and Information Theory. In: Pareyon, G., Pina-Romero, S., Agustín-Aquino, O., Lluis-Puebla, E. (eds) The Musical-Mathematical Mind. Computational Music Science. Springer, Cham. https://doi.org/10.1007/978-3-319-47337-6_17

7. Blitzstein, J. K., & Hwang, J. (2019). Introduction to probability (Second edition). crc Press/Taylor & Francis Group.

8. McKean, H. P. (2014). Probability: The classical limit theorems. Cambridge University Press.

9. MacKay, D. J. C. (2019). Information theory, inference, and learning algorithms (22nd printing). Cambridge University Press.

10. Google. (n.d.). The Telephone Game—tele-net America. Retrieved December 9, 2024, from https://fity.club/lists/suggestions/telephone-game/