6
Noisy channel coding

Intended learning outcomes:

- You can compute the channel mutual information, suitably simplifying the calculation if the channel exhibits symmetry.
- You understand the formal setup of the noisy channel coding problem, and are familiar with the binary symmetric channel (BSC), binary erasure channel (BEC) and additive white Gaussian noise (AWGN) channel.
- You know the difference between asymptotic and one-shot bounds and can derive the former from the latter.
- You can determine the type of a sequence and compute the empirical distribution.
- You understand the concept of random codes and derandomization.
- You can determine if a discrete memoryless source (DMS) can be transmitted through a discrete memoryless channel (DMC) using the source-channel separation theorem.

Book reference: Chapter 7 in Cover & Thomas[1].

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991. ISBN 9780471748823. DOI: 10.1002/047174882X

6.1 Channel mutual information

To quote Shannon from his pivotal paper "A Mathematical theory of communication":

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

The basic setup of the communication problem consists of a source that generates digital information which is to be reliably communicated to a destination through a channel, preferably in the most efficient manner possible. The destination could be spatially or temporally separated.

In this chapter we will first learn how to transmit a source that produces messages with uniform probability from some set of messages

and then argue that the optimal strategy to transmit an arbitrary source is to first compress it (which makes it approximately uniform) and then send it over the channel. The latter is called the source-channel separation theorem since it allows to treat channel coding and source coding independently as two separate tasks, without loss of efficiency — at least in the asymptotic limit of large block lengths.

Before we state the main theorem of this chapter we want to explore the following quantity:

We fix alphabets of input symbols, $\mathcal{X}$, and output symbols, $\mathcal{Y}$. Let W be a *channel*, a stochastic map represented as a conditional probability distribution $W_{Y|X}(y|x)$. The *channel mutual information* of W is defined as

$$I(W) := \max_{P_X \in \mathcal{P}(\mathcal{X})} I(X:Y)\,, \tag{6.1}$$

where $P_{XY}(x,y) = P_X(x)W_{Y|X}(y|x)$ is the joint distribution of channel input and output.

This is the maximal mutual information between channel input and output. The quantity is often called "channel capacity" in the literature, and we will see that it in fact corresponds to the maximal rate at which information can be transmitted over the channel in a later section. However, we prefer to keep a semantic difference between information quantities, like the channel mutual information, and operational quantities, like the channel capacity. Only through the study of information theory do we actually establish their equivalence, and usually only in special cases, e.g. for discrete memoryless channels in this case.

The optimisation is well-behaved since the underlying function is concave in P_X, as the following lemma shows.

Lemma 6.1. *For a channel W, the mutual information between channel input and output, $I(X:Y)$, is concave in the marginal pmf P_X.*

Proof. We have $I(X:Y) = H(Y) - H(Y|X)$, which we may write as

$$I(X:Y) = H(Y) - \sum_x P_X(x) H(Y|X=x), \tag{6.2}$$

where $P_Y(y) = \sum_x P_X(x)P_{Y|X}(y|x)$. By concavity of the entropy function we now see that the first term is concave in P_X. The second term is linear and thus concave in P_X as well. □

One very important consequence of this concavity is that we can simplify the optimisation for symmetric channels. The strongest symmetry we consider here is one where every permutation of input

symbols can be "undone" be doing a respective permutation of the output symbols of the channel.

The permutation group $S_{\mathcal{X}}$ contains all bijective functions from $\mathcal{X} \to \mathcal{X}$ and each $\pi \in S_{\mathcal{X}}$ corresponds to a relabelling of the different input symbols in $\mathcal{X}$. The total number of permutations is given by $|S_{\mathcal{X}}| = |\mathcal{X}|!$.

Proposition 6.2. *Consider a channel W such that for some subgroup of permutation $S \subseteq S_{\mathcal{X}}$ and every $\pi \in S$, there exists a permutation $\widetilde{\pi} \in S_{\mathcal{Y}}$ such that $W(y|x) = W(\widetilde{\pi}(y)|\pi(x))$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, there is an input distribution Q_X achieving the channel mutual information and satisfies $Q_X(x) = Q_X(\pi(x))$ for all $\pi \in S$ and $x \in \mathcal{X}$.*

Notably, when S is the cyclic group on $\mathcal{X}$ (or if $S = S_{\mathcal{X}}$), then the only input distribution that satisfies the above property is the uniform distribution, $P_X(x) = \frac{1}{|\mathcal{X}|}$. In this case the channel mutual information is simply given by $I(X : Y)$ where X is uniformly distributed, and no maximisation is needed.

Proof. Let $d = |\mathcal{X}|$. Assume P_X is a pmf that achieves the channel mutual information. In a first step, we want to show that $I(X : Y)_P = I(X : Y)_{P^\pi}$ for any permutation $\pi \in S$ and $P_X^\pi(x) = P_X(\pi^{-1}(x))$. For this purpose we introduce the random variable $\pi(X)$ and note that

$$X \longleftrightarrow \pi(X) \longleftrightarrow \widetilde{\pi}(Y) \longleftrightarrow Y \tag{6.3}$$

form a Markov chain. Hence, by data-processing $I(X : Y)_{P^\pi} \geq I(X : Y)_P$ —but since we started with the assumption that P_X is a maximiser the two mutual informations must in fact be equal.

Next we use this identity to write

$$I(W) = I(X : Y)_P = \sum_{\pi \in S} \frac{1}{|S|} I(X : Y)_{P^\pi} \,. \tag{6.4}$$

This can be interpreted as the expectation value of $I(X : Y)$, where the pmf is chosen uniformly at random from amongst the permuted pmfs P^π. However, since the mutual information is concave in the pmf, we have that

$$\sum_{\pi \in S} \frac{1}{|S|} I(X : Y)_{P_k} \leq I(X : Y)_Q, \tag{6.5}$$

where

$$Q_X(x) = \sum_{\pi \in S} \frac{1}{|S|} P_X^\pi(x) = \sum_{\pi \in S} \frac{1}{|S|} P_X(\pi^{-1}(x)) \tag{6.6}$$

is the expected pmf of X. Hence, the pmf Q_X performs at least as good as P_X, and thus must also achieve the channel mutual information. Finally, we show that $Q_X(\pi(x)) = Q_X(x)$. To see this, we use

that S is a group and

$$Q_X(\pi(x)) = \sum_{\bar{\pi}\in S} \frac{1}{|S|} P_X(\bar{\pi}^{-1}\cdot\pi(x)) \tag{6.7}$$

$$= \sum_{\bar{\pi}\in S} \frac{1}{|S|} P_X(\bar{\pi}^{-1}(x)) = Q_X(x) \tag{6.8}$$

by a change of variable. □

We will now consider two very prominent examples of communication channels and compute their channel mutual information.

1. The *binary symmetric channel* (BSC) takes a binary input to a binary output. The bit is flipped with a certain probability, here denoted ϵ, and otherwise left intact:

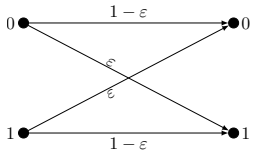


Figure 6.1: Binary symmetric channel with cross-over probability ϵ.

   The conditional probability distribution of the channel is given by

$$W_{\mathrm{BSC}(\epsilon)}(y|x) = (1-\epsilon)1\{x=y\} + \epsilon\, 1\{x\neq y\}\,. \tag{6.9}$$

   The channel mutual information for the BSC is easy to evaluate, even without invoking Proposition 6.2 — which does clearly apply here. Let us simply note that $H(Y|X=x) = h(\epsilon)$, the binary entropy evaluated for ϵ, and this is independent of $x\in\{0,1\}$. Hence the mutual information is given by $I(X:Y) = H(Y) - h(\epsilon)$, which is maximised when Y is uniformly distributed. This is achieved when X is uniformly distributed itself. Hence, the channel mutual information is given by

$$I(W_{\mathrm{BSC}(\epsilon)}) = 1 - h(\epsilon)\,. \tag{6.10}$$

2. The *binary erasure channel* (BEC) takes a binary input to a ternary output, $\{0,1,\perp\}$. The output $\perp$ has probability ϵ on either input, and otherwise the input symbol remains unaffected. Essentially this is a channel that flags errors:

   The conditional probability distribution of the channel is given by

$$W_{\mathrm{BEC}(\epsilon)}(y|x) = (1-\epsilon)1\{x=y\} + \epsilon\, 1\{y=\perp\}\,. \tag{6.11}$$

   By Proposition 6.2 we can again argue that the maximising input distribution is the uniform distribution. And we get the output

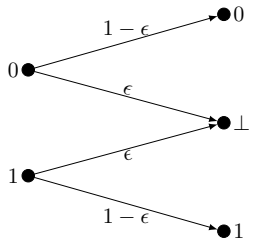


Figure 6.2: Binary erasure channel with error probability ϵ.

distribution

$$P_Y(y) = \begin{cases} \frac{1}{2}(1-\epsilon) & \text{if } y \in \{0,1\} \\ \epsilon & \text{if } y = \perp \end{cases} \tag{6.12}$$

We again have $H(Y|X=x) = h(\epsilon)$ independent of x and can then compute

$$\begin{aligned} I(W_{\mathrm{BEC}(\epsilon)}) &= H(Y) - h(\epsilon) && (6.13)\\ &= -2 \cdot \frac{1}{2}(1-\epsilon)\log\frac{1}{2}(1-\epsilon) - \epsilon\log\epsilon - h(\epsilon) && (6.14)\\ &= (1-\epsilon) + h(\epsilon) - h(\epsilon) && (6.15)\\ &= 1-\epsilon && (6.16) \end{aligned}$$

For the general case the problem is a bit more difficult, but concavity ensures that if we find a local maximum for the mutual information then that maximum is in fact global. Based on this, there are algorithms that can compute the channel mutual information efficiently for any stochastic map.

The following expression for the channel mutual information is useful to know, and expresses it as the information radius of the channel. We will need it later to prove the converse.

Proposition 6.3. *For any stochastic map W, we have*

$$I(W) = \min_{Q \in \mathcal{P}(\mathcal{Y})} \max_{x \in \mathcal{X}} D(W_{Y|X}(\cdot|x) \| Q_Y)\,. \tag{6.17}$$

In particular, there exists a distribution $Q^* \in \mathcal{P}(\mathcal{Y})$ such that $D(W_{Y|X}(\cdot|x)\|Q_Y) \leq I(W)$ for all $x \in \mathcal{X}$.

Proof. We first write, using the definition of the channel mutual information,

$$\begin{aligned} I(W) &= \max_{P_X \in \mathcal{P}(\mathcal{X})} D(P_{XY} \| P_X \times P_Y) && (6.18)\\ &= \max_{P_X \in \mathcal{P}(\mathcal{X})} \min_{Q_Y \in \mathcal{P}(\mathcal{Y})} D(P_{XY} \| P_X \times Q_Y), && (6.19) \end{aligned}$$

where the second equality comes from the fact that $D(P_{XY}\|P_X \times Q_Y) = D(P_{XY}\|P_X \times P_Y) + D(P_Y\|Q_Y)$ and the minimum is thus achieved for $Q_Y = P_Y$. If we further rewrite

$$D(P_{XY}\|P_X \times Q_Y) = \sum_{x\in\mathcal{X}} P_X(x) D(W_{Y|X}(\cdot|x)\|Q_Y) \tag{6.20}$$

we realise that this quantity is linear in P_X and convex in Q_Y. The idea then is to use Sion's minimax theorem[2], which states that the minimum and maximum in the above expressions can be interchanged. Hence, we get

$$I(W) = \min_{Q_Y\in\mathcal{P}(\mathcal{Y})} \max_{P_X\in\mathcal{P}(\mathcal{X})} \sum_{x\in\mathcal{X}} P_X(x) D(W_{Y|X}(\cdot|x)\|Q_Y)\,. \tag{6.21}$$

Finally note that the maximum in the above expression is taken for a P_X that is concentrated on a single point. This yields the expression in (6.17). □

[2] Maurice Sion. On General Minimax Theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958

6.2 *The channel coding theorem*

Let us now move on to a more operational description of the channel coding problem. As we have seen a noisy channel can be described by a conditional probability distribution $W_{Y|X}$. If such a channel can be used multiple times, without any memory effects, we speak of a *discrete memoryless channel* (DMC). We will not consider more complicated channels that change over time or have memory effects here, and thus our definition is restricted to the discrete memoryless case.

A *discrete memoryless channel* $\boldsymbol{W}$ is fully characterised by a stochastic map $W = W_{Y|X}$. For any $n \in \mathbb{N}$, the stochastic map W^n takes a sequence of input symbols $x^n \in \mathcal{X}^n$ to a sequence of output symbols $y^n \in \mathcal{Y}^n$ such that

$$P\big[Y^n = y^n \big| X^n = x^n\big] = W^n(y^n|x^n) = \prod_{i=1}^{n} W_{Y|X}(y_i|x_i)\,. \tag{6.22}$$

For our definition of codes, we can without loss of generality assume that the messages we are interested to send are in fact bit strings, i.e. $M \in \{0,1\}^L$. Any message set can obviously be encoded using bit strings, for example by using a compression algorithm!

An $(\epsilon, 2^L, n)$-*channel code* for a DMC $\boldsymbol{W}$ is comprised of

- an encoder function $e : \{0,1\}^L \to \mathcal{X}^n$ and
- a decoder function $d : \mathcal{Y}^n \to \{0,1\}^L$.

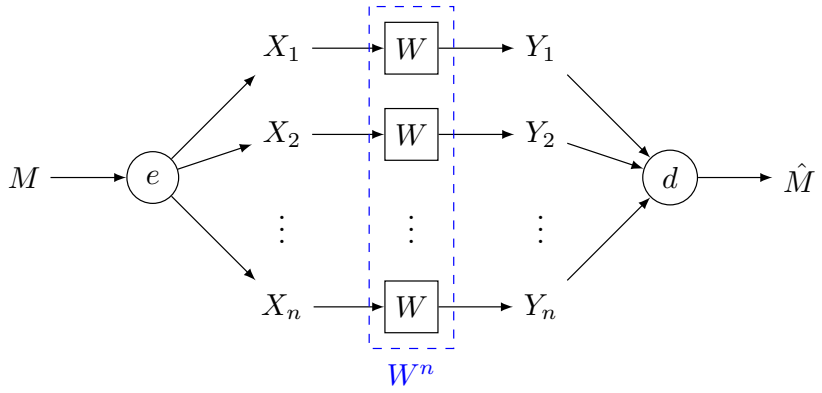


Figure 6.3: **The channel coding setup.** The figure depicts the setup for block length n. The encoder, e, takes a message M and encodes it into n channel input symbols, $X_1, X_2, \ldots, X_n$. The channels act independently on these input symbols. The channel outputs $Y_1, Y_2, \ldots, Y_n$ are decoded using the decoder, d, to an estimate $\hat{M}$ of M.

Consider the Markov chain $M \leftrightarrow X^n \leftrightarrow Y^n \leftrightarrow \hat{M}$ where M follows the uniform distribution on $\{0,1\}^L$, the channel input is $X = e(M)$, the channel output Y^n follows the distribution in (6.22), and $\hat{M} = d(Y)$. For an $(\epsilon, 2^L, n)$-channel code we require that these random variables satisfy $P[M \neq \hat{M}] \leq \epsilon$.

Here n is called the *block length*, L is the message length in bits, and ϵ is the allowed *average probability of error*.

This allows us to define the concept of achievable rates and capacity of a DMC.

We say that a rate R is a achievable for a DMC $\boldsymbol{W}$ if there exists a sequence of $(\epsilon_n, 2^{\lceil nR \rceil}, n)$ codes for all $n \in \mathbb{N}$ such that

$$\lim_{n\to\infty} \epsilon_n = 0 \tag{6.23}$$

The channel capacity of $\boldsymbol{W}$, denoted $C(\boldsymbol{W})$, is the supremum over all achievable rates R.

The main theorem of this chapter now relates the channel capacity of a DMC with the maximal mutual information of the underlying stochastic map.

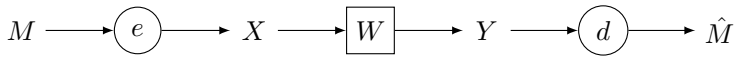
Theorem 6.4 (Channel coding theorem)**.** *For a DMC $\boldsymbol{W}$ with stochastic map W, we have*

$$C(\boldsymbol{W}) = I(W)\,. \tag{6.24}$$

We will prove this theorem in several steps. First we will derive an upper bound on the cardinality of the message set that holds even for a single use of the channel, the so-called meta-converse. From this we will then prove the converse (upper bound on the rate) and finally show how this rate can be achieved.

6.2.1 The meta-converse

The task in noisy channel coding is to transmit a message reliably over a DMC. We will for now assume that the message is uniformly distributed over some set of messages, but this assumption will be relaxed in the next section. Same as with source coding, we will eventually consider this problem in an asymptotic scenario where the the number of times the channel can be used, n, is taken to infinity. However, some results can conveniently be stated in a *one-shot setting*, without such a limit in mind, and we will do this here. Let us first define the notion of a code in the one-shot setting.



The condition on the distribution of M implies that ϵ enforces an *average error* criterion. We could alternatively also require that the probability of error is small for any distribution of M, which would enforce a *maximum error* criterion.

Our first result is a bound on the cardinality of the message set. This result is called the *meta-converse* as it can be used to derive various different fundamental limits (or converse bounds). This result was only established (relatively) recently by Polyanskiy-Poor-Verdú[3]. So even though information theory (and in particular channel coding) is by now a very well-established discipline, some progress can still be made when it comes to simplifying mathematical proofs and presenting them in a unified way.

[3] Yury Polyanskiy, H. Vincent Poor, and Sergio Verdú. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, may 2010. DOI: 10.1109/TIT.2010.2043769

Proposition 6.5. *For any $(\epsilon, 2^L, 1)$-channel code for a stochastic map W, we have*

$$L \leq \max_{P_X \in \mathcal{P}(\mathcal{X})} \min_{Q_Y \in \mathcal{P}(Y)} \log \frac{1}{\beta^*_\epsilon(P_{XY} \| P_X \times Q_Y)}, \tag{6.25}$$

where $P_{XY}(x,y) = P_X(x)W_{Y|X}(y|x)$ is the joint distribution of channel input and output and $\beta^*_\epsilon(P_{XY} \| P_X \times Q_Y)$ is the minimal error of the second kind, as defined as in Eq. (4.6), for the hypothesis testing problem where H_0 is P_{XY} and H_1 is $P_X \times Q_Y$.

The way to think about this hypothesis test is the following. The null hypothesis is that P_{XY} are in fact the channel input and output of our channel W, and the alternative hypothesis is that the output Q_Y has been produced independently of the input P_X, i.e. that it is the output of a channel that is completely useless for information transmission.

We provide the proof here for the special case where the encoder and decoder are deterministic and the encoder is furthermore injec-

tive, i.e. the function e uniquely maps messages to channel inputs. These assumptions make the proof a bit simpler but are not really restrictive. It is easy to verify that for any code using randomness there is a deterministic one performing at least equally well, and non-injective codes give up on distinguishing certain messages from the start, which can only give an advantage in extreme regimes where we tolerate large error and want to transmit more messages than fit into the channel alphabet.

Proof of Proposition 6.5. We first consider a fixed $(\epsilon, 2^L, 1)$-channel code that induces a distribution

$$P_X(x) = 2^{-L} \sum_{m \in \{0,1\}^L} \mathbf{1}\{e(m) = x\}\,. \tag{6.26}$$

on the channel input. By assumption, such a code must satisfy $P[M \neq \hat{M}] \leq \epsilon$. We now consider the hypothesis testing problem at hand, where H_0 is P_{XY} and H_1 is $P_X \times Q_Y$ for some arbitrary distribution $Q_Y \in \mathcal{P}(\mathcal{Y})$. For this problem we take the test

$$\mathcal{A} = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : x \neq e(d(y))\} \tag{6.27}$$

We can then compute the errors of the first and second kind for this test. This yields

$$\begin{aligned} \alpha(\mathcal{A}) = P_{XY}(\mathcal{A}) &= P[X \neq e(d(Y))] = P[e(M) \neq e(\hat{M})] && (6.28)\\ &\leq P[M \neq \hat{M}] \leq \epsilon\,. && (6.29) \end{aligned}$$

Furthermore,

$$\begin{aligned} \beta(\mathcal{A}) &= P_X \times Q_Y(\mathcal{A}^c) && (6.30)\\ &= 2^{-L} \sum_{x,y} \sum_{m \in \{0,1\}^L} Q_Y(y) \mathbf{1}\{e(m) = x\}\, \mathbf{1}\{x = e(d(y))\} && (6.31)\\ &= 2^{-L} \sum_{y} \sum_{m \in \{0,1\}^L} Q_Y(y)\, \mathbf{1}\{e(m) = e(d(y))\} && (6.32)\\ &= 2^{-L} \sum_{y} Q_Y(y) = 2^{-L}\,, && (6.33) \end{aligned}$$

where in the penultimate equality we used that e is injective, and thus there exists exactly one m for which $e(m) = e(d(y))$ holds. Hence, we can deduce that $\beta^*_\epsilon(P_{XY} \| P_X \times Q_Y) \leq \frac{1}{|M|}$, and, optimising over Q_Y, we find

$$L \leq \min_{Q_Y \in \mathcal{P}(\mathcal{Y})} \log \frac{1}{\beta^*_\epsilon(P_{XY} \| P_X \times Q_Y)}\,. \tag{6.34}$$

Finally, since we do not know the distribution P_X that the code induces — it depends on the specific encoding function e used — we maximise over P_X to get a bound that holds for all $(\epsilon, 2^L, 1)$-channel codes. □

6.2.2 *Proof of strong converse and method of types*

In the following we will show that $C(\boldsymbol{W}) \leq I(W)$. In fact, we will show something much stronger. We will show that for any sequence of $(\epsilon, 2^{\lceil Rn \rceil}, n)$-channel codes for a DMC $\boldsymbol{W}$ with $\epsilon \in (0,1)$ fixed, we must have $R \leq I(W)$. Hence, even if we allow for a nonzero error asymptotically, the maximal rate is still bounded by the channel mutual information. This is what is called a strong converse for channel coding. The proof strategy we follow here is inspired by some of my own work[4], and if taken to its conclusion can yield tight higher-order expansions of the channel capacity. However, here we are only interested in the first order, which allows us to simplify the argument quite a bit.

[4] Marco Tomamichel and Vincent Y. F. Tan. A Tight Upper Bound for the Third-Order Asymptotics for Most Discrete Memoryless Channels. *IEEE Transactions on Information Theory*, 59(11):7041–7051, nov 2013. DOI: 10.1109/TIT.2013.2276077

Proposition 6.6 (Strong converse for channel coding)**.** *Let $\boldsymbol{W}$ be a DMC with a stochastic map W. Then, for any sequence of $(\epsilon_n, 2^{\lceil nR \rceil}, n)$-codes with $\limsup_{n\to\infty} \epsilon_n < 1$, we must have $R \leq I(W)$.*

This implies that even if we allow for an error $\epsilon < 1$, we still cannot achieve any rate exceeding the channel mutual information. It also directly implies the converse part of the noisy channel coding theorem, Theorem 6.4, since it ensures that $C(\mathbb{W}) \leq I(W)$.

Before we prove this statement we first derive a relaxation of the meta-converse in terms of the information spectrum relative entropy. For this we will need the following lemma:

Lemma 6.7. *For any pmfs P_X and Q_Y and channel $W_{Y|X}$ it holds that*

$$D_s^\epsilon\big(P_{XY}\big\|P_X \times Q_Y\big) \leq \max_{x\in\mathcal{X}} D_s^\epsilon\big(W(\cdot|x)\big\|Q_Y(\cdot)\big)\,. \tag{6.35}$$

Proof. We first introduce the random variable $Z_x = \log \frac{W_{Y|X}(Y|x)}{Q_Y(Y)}$ where Y is distributed according to the law $W_{Y|X=x}$. Using it, we can write

$$D_s^\epsilon(W(\cdot|x)\big\|Q_Y(\cdot)) = \sup\{R \in \mathbb{R} : P[Z_x \leq R] \leq \epsilon\}\,. \tag{6.36}$$

Moreover, using the law of total probability, we find

$$\begin{aligned} &P\left[\log \frac{P_{XY}(X,Y)}{P_X(X)Q_Y(Y)} \leq R\right] \\ &\quad= \sum_x P_X(x) W_{Y|X=x}\left[\log \frac{W_{Y|X}(Y|x)}{Q_Y(Y)} \leq R\right] &(6.37)\\ &\quad= \sum_x P_X(x) P\left[Z_x \leq R\right]\,. &(6.38)\end{aligned}$$

Plugging this into the definition of the information spectrum relative entropy, we find

$$D_s^\epsilon(P_{XY}\|P_X\times Q_Y) = \sup\left\{R\in\mathbb{R} : \sum_x P_X(x)P\left[Z_x\le R\right]\le\epsilon\right\} \tag{6.39}$$

$$\le \sup\left\{R\in\mathbb{R} : \min_{x\in\mathcal{X}}\left\{P\left[Z_x\le R\right]\right\}\le\epsilon\right\} \tag{6.40}$$

$$\le \max_{x\in\mathcal{X}}\sup\left\{R\in\mathbb{R} : P\left[Z_x\le R\right]\le\epsilon\right\}. \tag{6.41}$$

To establish the first inequality we used that there must (at least) exist one x for which the probability does not exceed its expectation over X. Thus, we have relaxed the condition. The final inequality can be verified as follows. Let us say that the supremum in (6.40) is R^*, and, thus, $P[Z_x \le R^*-\mu]\le\epsilon$ for every $\mu>0$ and some $x\in\mathcal{X}$. Hence, $\max_{x\in\mathcal{X}}\sup\{R\in\mathbb{R} : P\left[Z_x\le R\right]\le\epsilon\}\ge R^*-\mu$. Since this holds for all $\mu>0$ we are done. □

The inequality from (6.40) to (6.41) is in fact an equality. Can you see why?

Proof of Proposition 6.6. Consider a sequence of $(\epsilon_n, 2^{\lceil Rn\rceil}, n)$-codes as in the statement of the result. Since we have that $\epsilon_\infty := \limsup_{n\to\infty}\epsilon_n < 1$ we may choose $\epsilon\in(\epsilon_\infty,1)$ and by definition of the limit, there exists an N such that for any $n\ge N$ there exists an $(\epsilon, 2^{\lceil nR\rceil}, n)$-code in the sequence. Recall that

$$P_{X^nY^n}(x^n,y^n) = P_{X^n}(x^n)\prod_{i=1}^n W_{Y|X}(y_i|x_i) \tag{6.42}$$

is the joint distribution of channel inputs and outputs when the stochastic map W^n is applied to an arbitrary input distributions P_{X^n}. By the meta-converse as well as Lemma 4.5 and Lemma 6.7 for the channel W^n, we have

$$nR \le \lceil nR\rceil \tag{6.43}$$

$$\le \max_{P_{X^n}\in\mathcal{P}(\mathcal{X}^n)}\min_{Q_{Y^n}\in\mathcal{P}(\mathcal{Y}^n)}\log\frac{1}{\beta_\epsilon^*(P_{X^nY^n}\|P_{X^n}\times Q_{Y^n})} \tag{6.44}$$

$$\le \max_{P_{X^n}\in\mathcal{P}(\mathcal{X}^n)}\log\frac{1}{\beta_\epsilon^*(P_{X^nY^n}\|P_{X^n}\times Q_Y^n)} \tag{6.45}$$

$$\le \max_{P_{X^n}\in\mathcal{P}(\mathcal{X}^n)} D_s^{\epsilon+\delta}(P_{X^nY^n}\|P_{X^n}\times Q_Y^n) + \log\frac{1}{\delta} \tag{6.46}$$

$$\le \max_{x^n\in\mathcal{X}^n} D_s^{\epsilon+\delta}\left(W^n(\cdot|x^n)\middle\|Q_Y^n(\cdot)\right) + \log\frac{1}{\delta}, \tag{6.47}$$

where we have chosen δ such that $\epsilon+\delta<1$. In the last step two steps we simply noted that the expression no longer depends on the input distribution P_{X^n} and chose Q_{Y^n} to be an i.i.d. distribution Q_Y^n, where Q_Y is the output distribution that minimises the expression in (6.17), i.e., it satisfies

$$\max_x D(W(\cdot|x)\|Q) = I(W). \tag{6.48}$$

It thus remains to analyse the information spectrum relative entropy in (6.47) as a function of $x^n \in \mathcal{X}^n$. In particular, the probability

$$P\left[\sum_{i=1}^{n} Z^i_{x_i} \leq R\right], \quad \text{where } Z^i_x \text{ are independent rvs,} \tag{6.49}$$

given by $Z^i_x = \log \frac{W_{Y|X}(Y|x)}{Q_Y(Y)}$ with Y distributed according to $W_{Y|X=x}$. Hence, the $Z^i_{x_i}$ are independent but not identically distributed. Using this, we can write

$$D_s^{\epsilon+\delta}\big(W^n(\cdot|x^n)\big\|Q_Y^n(\cdot)\big) = \sup\left\{R \in \mathbb{R} : P\left[\sum_{i=1}^{n} Z^i_{x_i} \leq R\right] \leq \epsilon + \delta\right\}. \tag{6.50}$$

The first thing we note is that this expression actually does not depend on all the properties of x^n, but only the frequency in which the letters of $\mathcal{X}$ occur in x^n. As a consequence, in this proof we will employ the so-called "Method of Types"[5], of which we however will only be able to scratch the surface.

[5] Imre Csiszár. The Method of Types. *IEEE Transactions on Information Theory*, 44(6):2505–2523, oct 1998. DOI: 10.1109/18.720546

Consider a sequence x^n where each symbol x_i for $i \in [n]$ is taken from a finite set $\mathcal{X}$. There are obviously $\mathcal{X}^n$ different such sequences. But in many circumstances it suffices to classify these sequences simply by how many times each of the elements of $\mathcal{X}$ appear in it. This is called the *type* of the sequence x^n. We will also introduce the *empirical distribution* of a sequence x^n, which also only depends on the number of times each symbol appears, and is defined as

$$P_X^{x^n}(x) = \frac{1}{n}\big|\{i \in [n] : x_i = x\}\big|. \tag{6.51}$$

for all $x \in \mathcal{X}$. Note that this is a pmf on $\mathcal{X}$ where each symbol has probability proportional to the number of times it appears in x^n.

We have already seen that in the homework that the empirical distribution of a i.i.d. sequence X^n will concentrate around the underlying pmf P_X with which the X_i are chosen.

One of the other interesting observations we can make here is that there are not too many types, or at least not exponentially many. The number of different types of sequences in $\mathcal{X}^n$ is simply given by the number of partitions of n into $|\mathcal{X}|$ segments. A simple upper bound on the number of types can be given as $(n+1)^{|\mathcal{X}|-1}$, which grows only polynomially in n. This upper bound can be shown by counting the number of possible values the probability $P_X^{x^n}(x)$ in (6.51) can take for each symbol $x \in \mathcal{X}$.

More precisely, the argument given here leads to an upper bound $(n+1)^{|\mathcal{X}|}$. Can you see how this can be improved to the bound $(n+1)^{|\mathcal{X}|-1}$ claimed?

The random variable in (6.49) is a sum of independent (but not identical!) random variables. Let us compute its expectation and

variance, which are

$$\mathbb{E}\left[\sum_{i=1}^{n} Z_{x_i}^i\right] = \sum_{i=1}^{n} D\left(W(\cdot|x_i)\|Q\right) \tag{6.52}$$

$$= n \underbrace{\sum_{x\in\mathcal{X}} P_X^{x^n}(x) D\left(W(\cdot|x)\|Q\right)}_{=:J} \tag{6.53}$$

$$\mathrm{Var}\left[\sum_{i=1}^{n} Z_{x_i}^i\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\log \frac{W_{Y|X}(Y_i|x)}{Q_Y(Y_i)}\right] \tag{6.54}$$

$$\leq n \underbrace{\max_{x\in\mathcal{X}} \mathrm{Var}\left[\log \frac{W_{Y|X}(Y_i|x)}{Q_Y(Y_i)}\right]}_{=:\,\sigma^2}, \tag{6.55}$$

where σ^2 is some constant. So if we set $K = n(J+\nu)$ for any small $\nu > 0$, Chebyshev's inequality yields

$$P\left[\sum_{i=1}^{n} Z_{x_i}^i \geq K\right] \leq \frac{\sigma^2}{\nu^2 n}, \tag{6.56}$$

or, equivalently,

$$P\left[\sum_{i=1}^{n} Z_{x_i}^i \leq K\right] \geq 1 - \frac{\sigma^2}{\nu^2 n}, \tag{6.57}$$

If we choose n large enough, then $1 - \frac{\sigma^2}{\nu^2 n}$ is always larger than $\epsilon + \delta$. Hence, we can deduce that, for sufficiently large n,

$$D_s^{\epsilon+\delta}\left(W^n(\cdot|x^n)\middle\|Q_Y^n(\cdot)\right) \leq n(J+\nu) \tag{6.58}$$

$$\leq n(I(W)+\nu)\,. \tag{6.59}$$

This expression now no longer depends on x^n. Plugging it into Eq. (6.47) and dividing both sides by n, we find

$$R \leq I(W) + \nu + \frac{1}{n}\log\frac{1}{\delta} \tag{6.60}$$

Since $\frac{1}{n}\log\frac{1}{\delta} \leq \nu$ for sufficiently large n and furthermore ν can be chosen arbitrarily small, we can conclude that the inequality $R \leq I(W)$ must hold for any such sequence of codes. □

6.2.3 *Proof of achievability and random codes*

We will need the following technical lemma. (Our proof is inspired by the analysis of Hayashi and Nagaoka in the domain of quantum information theory[6].)

[6] Masahito Hayashi and Hiroshi Nagaoka. General Formulas for Capacity of Classical-Quantum Channels. *IEEE Transactions on Information Theory*, 49(7):1753–1768, jul 2003. DOI: 10.1109/TIT.2003.813556

Lemma 6.8. *Let $t \geq 0$ and $s \in [0,1]$. Then, $1 - \frac{s}{s+t} \leq 1 - s + t$.*

Proof. We may rewrite the statement as

$$0 \leq t - s + \frac{s}{s+t}\,. \tag{6.61}$$

If $t \geq s$ this is trivially true. If $t < s$ we use the convexity of the function $f(t) = \frac{s}{s+t}$ to bound it with its tangent at $t = 0$. This yields

$$\frac{s}{s+t} \geq 1 - \frac{t}{s} = \frac{s-t}{s} \geq s - t, \tag{6.62}$$

where the second inequality follows since $s \in (0,1]$ and $s - t \geq 0$. □

We again first analyse the channel coding problem in the one-shot setting where the channel is only used once.

Proposition 6.9. *For any $\epsilon, \delta \in (0,1)$ such that $\epsilon + \delta < 1$ there exists an $(\epsilon + \delta, 2^L, 1)$-channel code for a stochastic map $W_{Y|X}$ as long as the code parameters satisfy*

$$L \leq \log \frac{\delta}{\beta_\epsilon^*(P_{XY} \| P_X \times P_Y)} \tag{6.63}$$

for some pmf $P_X \in \mathcal{P}(\mathcal{X})$.

Proof. We now construct a random code for a single use of the channel. First, we fix any distribution $P_X \in \mathcal{P}(\mathcal{X})$. From this we generate $|M|$ codewords independently by picking them from the distribution P_X, i.e. the output of the decoder, $E(m)$, is itself a random variable following the distributions P_X for each message m. The decoder is constructed as follows. Consider the binary hypothesis testing problem between $H_0 : P_{XY}$ and $H_1 : P_X \times P_Y$. By definition of $\beta_\epsilon^*(P_{XY} \| P_X \times P_Y)$, there exists a subset $\mathcal{A} \subset \mathcal{X} \times \mathcal{Y}$ that satisfies

$$P_{XY}(\mathcal{A}^c) \leq \epsilon \quad \text{and} \quad (P_X \times P_Y)(\mathcal{A}) = \beta_\epsilon^*(P_{XY} \| P_X \times P_Y)\,. \tag{6.64}$$

From this we construct the sets $\mathcal{A}_x = \{y \in \mathcal{Y} : (x,y) \in \mathcal{A}\}$ for all $x \in \mathcal{X}$. For a fixed encoder $E = e$, the decoder is probabilistic. Given a channel output y it assigns $\hat{M} = m$ with probability

$$P[\hat{M} = m | Y = y, E = e] = \frac{\mathbf{1}\{y \in \mathcal{A}_{e(m)}\}}{\sum_{m'} \mathbf{1}\{y \in \mathcal{A}_{e(m')}\}} \tag{6.65}$$

$$= \frac{\mathbf{1}\{y \in \mathcal{A}_{e(m)}\}}{\mathbf{1}\{y \in \mathcal{A}_{e(m)}\} + \sum_{m' \neq m} \mathbf{1}\{y \in \mathcal{A}_{e(m')}\}}\,. \tag{6.66}$$

Let us now analyse the probability of error for this code, first for a fixed set of codewords (or fixed encoder, e) and fixed message m.

$$\begin{aligned} &P[M \neq \hat{M} | M = m, E = e] \\ &\quad = 1 - \sum_y W(y|e(m)) P[\hat{M} = m | Y = y, E = e] && (6.67) \\ &\quad = \sum_y W(y|e(m)) \left(1 - \frac{\mathbf{1}\{y \in \mathcal{A}_{e(m)}\}}{\mathbf{1}\{y \in \mathcal{A}_m\} + \sum_{m' \neq m} \mathbf{1}\{y \in \mathcal{A}_{e(m')}\}} \right) && (6.68) \end{aligned}$$

We can now use Lemma 6.8 to bound this as

$$\begin{aligned} &P[M \neq \hat{M} | M = m, E = e] && (6.69) \\ &\quad \leq \sum_y W(y|e(m)) \left(\mathbf{1}\{y \notin \mathcal{A}_{e(m)}\} + \sum_{m' \neq m} \mathbf{1}\{y \in \mathcal{A}_{e(m')}\} \right) . && (6.70) \end{aligned}$$

We may now take the average over all encoders e, so that $e(m)$ and $e(m')$ for $m \neq m'$ are independent and follow the distribution P_X. This gives the following bound

$$\begin{aligned} &P[M \neq \hat{M} | M = m] && (6.71) \\ &\leq \sum_{x,y} P_X(x) W(y|x) \left(\mathbf{1}\{y \notin \mathcal{A}_x\} + \underbrace{(2^L - 1)}_{\leq 2^L} \sum_{x' \in \mathcal{X}} P_X(x') \mathbf{1}\{y \in \mathcal{A}_{x'}\} \right) && (6.72) \end{aligned}$$

and we note that the bound no longer depends on the choice of m, i.e. Eq. (6.72) is in fact an upper bound on $P[M \neq \hat{M}]$. Let us now investigate the two summands in (6.72) individually. We first observe that

$$\sum_{x,y} P_X(x) W(y|x) \mathbf{1}\{y \notin \mathcal{A}_x\} = P_{XY}\left[(x,y) \notin \mathcal{A}\right] = P_{XY}[A^c] \leq \epsilon \quad (6.73)$$

by definition of the sets $\mathcal{A}_x$ and $\mathcal{A}$. We can also evaluate

$$\begin{aligned} &\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_X(x) W(y|x) \sum_{x' \in \mathcal{X}} P_X(x') \mathbf{1}\{y \in \mathcal{A}_{x'}\} \\ &\quad = \sum_{y \in \mathcal{Y}} P_Y(x) \sum_{x' \in \mathcal{X}} P_X(x') \mathbf{1}\{(x', y) \in \mathcal{A}\} && (6.74) \\ &\quad = (P_X \times P_Y)[\mathcal{A}] && (6.75) \\ &\quad = \beta_\epsilon^*(P_{XY} \| P_X \times P_Y) . && (6.76) \end{aligned}$$

Summarising this, we find that

$$P[M \neq \hat{M}] \leq \epsilon + 2^L \beta_\epsilon^*(P_{XY} \| P_X \times P_Y) . \quad (6.77)$$

So, in particular, as long as we choose $2^L \leq \delta \cdot \beta_\epsilon^*(P_{XY} \| P_X \times P_Y)^{-1}$, we achieve $P[M \neq \hat{M}] \leq \epsilon + \delta$, as required.

Finally, since this bound holds on average over all choices of encoders e, there exists (at least) one encoder that satisfies $P[M \neq \hat{M}|E = e] \leq \epsilon + \delta$ as well. This is the code we are looking for. □

Finally, we can complete the proof of Theorem 6.4, showing that any rate $R < I(W)$ is achievable.

Achievability of Theorem 6.4. We consider the one-shot results applied to the super-channel W^n. Proposition 6.9 stipulates that there exists a code with $2^{\lceil nR \rceil}$ codewords and error 2ϵ as long as

$$\lceil nR \rceil \leq \log \frac{\epsilon}{\beta^*_\epsilon(P_{X^nY^n} \| P_{X^n} \times P_{Y^n})} \tag{6.78}$$

for some input distribution $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$. Or, removing the ceiling function and deviding by n on both sides, as long as

$$R \leq \frac{1}{n}\left(\log \frac{1}{\beta^*_\epsilon(P_{X^nY^n} \| P_{X^n} \times P_{Y^n})} + \log \epsilon - 1\right). \tag{6.79}$$

We further choose P_{X^n} to be i.i.d. and P_X the maximiser in the definition of the channel mutual information, i.e. $I(X:Y)_P = I(W)$ to make our analysis simpler. We can then use the Chernoff-Stein's Lemma (cf. Theorem 4.4) to show that

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta^*_\epsilon(P^n_{XY} \| P^n_X \times P^n_Y)} = D(P_{XY} \| P_X \times P_Y) \tag{6.80}$$

$$= I(X:Y) = C(W). \tag{6.81}$$

And thus, we can conclude that if $R < I(W)$ is strictly smaller than the channel mutual information, then for any error $\epsilon > 0$ the condition (6.79) will be satisfied for sufficiently large n. Hence, there exists a sequence of codes with vanishing probability of error at this rate, and it is thus achievable. Taking the supremum over all such rates yields our achievability bound $C(\boldsymbol{W}) \geq I(W)$. □

6.2.4 *Maximum probability of error*

Consider a code with $|M|$ codewords. So far we have used the *average probability of error* as a metric for our codes, namely we required that

$$P[\hat{M} \neq M] = \sum_{m \in \{0,1\}^L} 2^{-L} P[\hat{M} \neq m | M = m] \tag{6.82}$$

vanishes asymptotically. Sometimes we would however like to impose an even stricter condition, namely that the *maximum probability of error*, given by

$$\max_{m \in \{0,1\}^L} P[\hat{M} \neq m | M = m], \tag{6.83}$$

vanishes asymptotically. Because the condition is stricter our converse bounds still hold even with this new definition of error; however, the random codes we constructed so far do not necessarily lead to a small maximum probability of error.

The following lemma allows us to construct codes that overcome this.

Proposition 6.10. *Given an $(\epsilon, 2^L, 1)$-average error channel code, we can construct a $(2\epsilon, 2^{L-1}, 1)$-maximum error channel code.*

Proof. The proof uses expurgation of bad codewords. By definition of the $(\epsilon, 2^L, 1)$-average error channel code, we have

$$\sum_{m \in \{0,1\}^L} 2^{-L} P[\hat{M} \neq m | M = m] \leq \epsilon \tag{6.84}$$

Hence, there must be a subset $M_g \subseteq \{0,1\}^L$ of size at least 2^{L-1} with

$$P[\hat{M} \neq m | M = m] \leq 2\epsilon \qquad \forall\, m \in M_g \tag{6.85}$$

as otherwise the inequality in Eq. (6.84) cannot hold. The codewords in M_g constitute an $(2\epsilon, \frac{|M|}{2}, 1)$-maximum error channel code. □

6.3 *Source-channel separation theorem*

We have until now covered the case where a message that is uniformly chosen from a set needs to be transmitted through the noisy channel. Does anything change when instead we want to transmit a general source? The setting is the same as with channel coding, except that now for each block length n we want to transmit a memoryless source given by i.i.d. $Z^n = (Z_1, Z_2, \ldots, Z_n)$. A code for block length n is given by an encoder $e_n : \mathcal{Z}^n \to \mathcal{X}^n$ and a decoder $d_n : \mathcal{Y}^n \to \mathcal{Z}^n$ and our goal is to find a sequence of such codes that satisfy

$$\lim_{n \to \infty} P[\hat{Z}^n \neq Z^n] \to 0 \tag{6.86}$$

Here we want to show the following theorem:

Theorem 6.11. *Given a DMS $\mathbf{Z}$ and DMC $\mathbf{W}$, there exists a sequence of codes satisfying with asymptotically vanishing error if $H(Z) < I(W)$. Moreover, if $H(Z) > I(W)$ such a sequence of codes cannot exist.*

When $H(Z) < I(W)$ we can simply compress the source at a rate $R = H(Z) + \mu$ and then transmit it over the channel at the same rate $R = I(W) - \mu$, where we choose $\mu = \frac{1}{2}(I(W) - H(Z))$. That is, we first

apply the encoder for source compression, transmit the compressed source through the channel using a maximum probability of error channel code, and finally decompress the source at the receiver. The error of such a scheme is simply the sum of the individual errors of the source compression code and the channel code, both of which vanish asymptotically as shown in the source and channel coding theorems, respectively.

The second statement of this theorem, which is conceptually more interesting, shows that such a separate treatment of compression and channel coding is in fact optimal (at least when we only look at the first order asymptotics). We will only give a formal proof of the second statement. We will need the following lemma for this purpose, which ensures the additivity of the channel mutual information.

Lemma 6.12. *Let W_1 and W_2 be two channels. Then,*

$$I(W_1 \times W_2) = I(W_1) + I(W_2)\,. \tag{6.87}$$

Proof. We first note that for two channel inputs X_1 and X_2 following any distribution $P_{X_1X_2}$ and two channel outputs Y_1 and Y_2 produced by two channels W_1 and W_2 applied to X_1 and X_2, respectively, we have

$$\begin{aligned}
I(X_1X_2 : Y_1Y_2) &= \\
&= H(Y_1Y_2) - H(Y_1Y_2|X_1X_2) && (6.88)\\
&\leq H(Y_1) + H(Y_2) \\
&\quad - \sum_{x_1,x_2\in\mathcal{X}} P_{X_1X_2}(x_1,x_2)\, \underbrace{H(Y_1Y_2|X_1=x_1, X_2=x_2)}_{=H(Y_1|X_1=x_1)+H(Y_2|X_2=x_2)} && (6.89)\\
&= H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (6.90)\\
&= I(X_1 : Y_1) + I(X_2 : Y_2) && (6.91)\\
&\leq I(W_1) + I(W_2)\,. && (6.92)
\end{aligned}$$

Here we used sub-additivity of entropy and the fact that Y_1 and Y_2 are independent once we condition on X_1 and X_2. We conclude that this inequality holds in particular also for the the distribution $P_{X_1X_2}$ achieving $I(W_1 \times W_2)$. The other direction and thus equality clearly holds since we can always just restrict ourselves to product distributions when optimising $I(W_1 \times W_2)$, and, thus

$$\begin{aligned}
I(W_1 \times W_2) &\geq \max_{P_{X_1},P_{X_2}} I(X_1X_2 : Y_1Y_2) && (6.93)\\
&= \max_{P_{X_1},P_{X_2}} I(X_1 : Y_1) + I(X_2 : Y_2) = I(W_1) + I(W_2)\,. && (6.94)
\end{aligned}$$

□

Proof of Theorem 6.11. Assume $H(Z) - I(W) = \nu > 0$. If there is a sequence of codes with asymptotically vanishing error then for every $\epsilon > 0$ there must be a block length n such that $P[\hat{Z}^n \neq Z^n] \leq \epsilon$. For such a code, by Fano's inequality, we have

$$H(Z^n) - I(Z^n : Y^n) = H(Z^n|Y^n) \leq H(Z^n|\hat{Z}^n) \leq 1 + \epsilon\, n \log |Z| \quad (6.95)$$

We can now evaluate $H(Z^n) = nH(Z)$ since the source is i.d.d., and furthermore

$$I(Z^n : Y^n) \leq I(X^n : Y^n) \leq I(W^n) = nI(W)\,. \quad (6.96)$$

To verify the first inequality we simply note that $Z^n \to X^n \to Y^n$ form a Markov chain, which implies that $I(Z^n : Y^n) \leq I(X^n : Y^n)$. Maximising $I(X^n : Y^n)$ over all input distributions then yields the inequality. The last inequality follows from Lemma 6.12, which can be used to verify that $I(W^n) = nI(W)$ by induction.

Finally, combining Eqs. (6.95) and (6.96) yields

$$\epsilon \log |Z| \geq H(Z) - I(W) - \frac{1}{n} = \nu - \frac{1}{n} \quad (6.97)$$

but since for large enough n the term on the right-hand side is strictly positive, ϵ is bounded away from zero, leading to a contradiction. □

6.4 *Gaussian channels*

We have not discussed continuous variables in any detail and indeed all our proofs so far have assumed that the random variables take values in a finite alphabet. We will now however explore one very important channel that is continuous, the additive white Gaussian noise (AWGN) channel. This channel takes an input $X \in \mathbb{R}$ and outputs

$$Y = X + Z, \quad (6.98)$$

where Z follows a Gaussian distribution with mean 0 and standard deviation σ, and is indpendent of X. The channel behaviour can thus be characterised by the conditional pdf

$$w_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}}, \quad (6.99)$$

which is the Gaussian pdf with mean x and standard deviation σ. We can now ask the usual question about this channel — at what rate can we transmit information over it? It turns out that without further restrictions the answer to this that we can transmit as much information as we want, even through a single use of the channel.

We simply map the messages to a lattice of values $x_m \in \mathbb{R}$ that are sufficiently separated so that even after the noise is added $w(y|x_m)$ and $w(y|x_{m'})$ only have small overlap for distinct messages m and m'. If the grid distance is chosen to be 6σ, for example, we will get a decoding error that is lower than 0.5% by such a construction. And we can get an arbitrarily small error by spreading the lattice even further.

Formally construct such an encoder and decoder and compute the probability of error.

In practical applications an AWGN for example arises when we encode information in an electromagnetic field, and X_i and Y_i for each channel use $i \in [n]$ are then simply amplitudes of the field. On the other hand, the energy stored in the field grows with the square of the amplitude and needs to be invested by the sender of the electromagnetic pulse. It is natural to restrict how much energy per channel-use, or power, is available at the source.[7] Formally, this is done by requiring that every codeword $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P \tag{6.100}$$

[7] The nomenclature makes sense since power is energy per time unit, and channel uses are temporally separated in this context.

This communication channel models many practical channels, including wireless and satellite links. The noise may be due to a variety of (independent) microscopic reasons; however, the central limit ensures that collectively these noise sources resemble an additive noise with a Gaussian distribution.

We will now analyse the channel capacity of the AWGN channel under the above constraint. To do this, we however need to first introduce the notion of differential entropy.

6.4.1 *Differential entropy and mutual information*

Let X be a real-valued continuous random variable with support on S and pdf p_X. The *differential entropy* of X is defined as

$$h(X) = -\int_S p_X(x) \log p_X(x) \mathrm{d}x\,. \tag{6.101}$$

It is worth noting that this integral does not always exist and might in fact be infinite in many cases.

A class of distributions for which it is relatively well-behaved is the uniform distribution, where $p_X(x) = \frac{1}{a}$ in an interval $[0, a]$ and zero elsewhere. In this case it is easy to verify that we have $h(X) = \log a$. An other interesting case is the case of Gaussian distribution with

For maths enthusiasts: Construct an example with a valid pdf for which the integral diverges and one for which it becomes negative.

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,. \tag{6.102}$$

In this case we can evaluate the differential entropy as follows

$$h(X) = -\int p_X(x) \log p_X(x) \mathrm{d}x \tag{6.103}$$

$$= \int p_X(x) \left(\frac{(x-\mu)^2}{2\sigma^2} \log e + \log \sqrt{2\pi\sigma^2} \right) \mathrm{d}x \tag{6.104}$$

$$= \mathbb{E}[(x-\mu)^2] \frac{\log e}{2\sigma^2} + \frac{1}{2} \log (2\pi\sigma^2) \tag{6.105}$$

$$= \frac{1}{2} \log (2e\pi\sigma^2) . \tag{6.106}$$

But note that since the differential entropy can get negative for small σ it is hard to give it operational meaning.

One thing we can immediately observe is that the differential entropy is independent of the mean of X. This is true more generally: $h(X) = h(X + a)$ for any constant a, which can be verified by a simple change of variable. Note, however, that $h(cX) = h(X) + \log |c|$, so the entropy is not invariant under rescaling. This can be seen in contrast to the invariance of the entropy of discrete random variables under relabelings.

We can define conditional entropy and mutual information analogously to the discrete case.

Let X and Y be real-valued continuous random variables with joint pdf p_{XY}. The *conditional differential entropy* of X given Y is defined as

$$h(X|Y) = -\int_S p_{XY}(x,y) \log p_{X|Y}(x|y) \, \mathrm{d}x \, \mathrm{d}y , \tag{6.107}$$

where $p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$ is the conditional pdf and $p_Y(y)$ the marginal pdf on Y. Moreover, the *mutual information* between X and Y is defined as

$$I(X:Y) = \int p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \, \mathrm{d}x \, \mathrm{d}y . \tag{6.108}$$

As expected, we can again decompose $I(X:Y) = h(X) - h(X|Y)$, assuming that all the expressions are finite.

The mutual information for continuous variables is very naturally linked to mutual information for discrete variables. To see this, consider the discrete random variables X^Δ that takes values $x_r = r\Delta$ for $r \in \mathbb{Z}$ with probability

$$P_{X^\Delta}(x_r) = \int_{x_r - \Delta/2}^{x_r + \Delta/2} p_X(x) \, \mathrm{d}x . \tag{6.109}$$

This is simply a discretised version of X, where everything in an interval of length Δ is course-grained into a single discrete value.

For sufficiently small Δ continuity of $p_X(x)$ implies that $P_{X^\Delta}(x_r) \to \Delta \cdot p_X(x_r)$, and thus we find, under some mild regularity assumptions,

$$I(X:Y) = \int p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \mathrm{d}x\, \mathrm{d}y \tag{6.110}$$

$$= \lim_{\Delta\to 0} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta^2\, p_{XY}(x_i,y_j) \log \frac{p_{XY}(x_i,y_j)}{p_X(x_i)p_Y(y_j)} \tag{6.111}$$

$$= \lim_{\Delta\to 0} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} P_{X^\Delta Y^\Delta}(x_i,y_j) \log \frac{P_{X^\Delta Y^\Delta}(x_i,y_j)}{P_{X^\Delta}(x_i)P_{Y^\Delta}(y_j)} \tag{6.112}$$

$$= \lim_{\Delta\to 0} I(X^\Delta : Y^\Delta), \tag{6.113}$$

where the first equality simply follows by the definition of the Riemann integral. A corresponding result does not hold for differential entropy, instead a different normalisation is required. This is evident simply from the fact that fine-graining a random variable will generally strictly increase its entropy, and thus the entropy would always diverge to infinity in the above limit.

Argue that this implies that the mutual information is always non-negative even for continuous variables.

Finally, we can define the relative entropy between two pdfs p_X and q_X as

$$D(p_X \| q_X) = \int p_X(x) \log \frac{p_X(x)}{q_X(x)}\, \mathrm{d}x\,. \tag{6.114}$$

The same argument we used in Chapter 1, based on Jensen's inequality, reveals that

$$D(p_X \| q_X) \geq 0 \tag{6.115}$$

for all pairs of pdfs.

6.4.2 *Channel coding theorem for the AWGN channel*

Theorem 6.13. *The capacity of the AWGN channel $\boldsymbol{W}$ with variance σ^2 and power constraint P is given by*

$$C(\boldsymbol{W}) = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right). \tag{6.116}$$

The expression $\frac{P}{\sigma^2}$ is called the *signal-to-noise ratio* (SNR).

This expression does not remind us of the usual channel coding theorem, but this is only because it is already simplified for the channel at hand. Let us thus first show the following identity.

Lemma 6.14. *For an AWGN channel W with variance σ^2 and power con-*

straint P, we have

$$\max_{\substack{p_X \in \mathcal{P}(\mathbb{R}) \\ \mathbb{E}[X^2] \leq P}} I(X:Y) = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right) \tag{6.117}$$

where $p_{XY}(x,y) = p_X(x) w_{Y|X}(y|x)$ as usual.

Due to the constraint on the codewords the optimisation is now not over all input distributions but only such distributions that satisfy the required bound on the expectation of X^2.

Proof. We may rewrite $I(X:Y) = h(Y) - h(Y|X)$ where

$$h(Y|X) = h(X+Z|X) = h(Z) = \frac{1}{2}\log(2\pi e\sigma^2) \tag{6.118}$$

can be simplified immediately. We now make the following observations. We have

$$\mathbb{E}[Y^2] = \mathbb{E}[X^2 + 2XZ + Z^2] \tag{6.119}$$

$$= \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Z] + \mathbb{E}[Z^2] \leq P + \sigma^2 \tag{6.120}$$

since X and Z are independent and $\mathbb{E}[Z] = 0$. So we have a bound on the variance of Y — does this allow us to conclude anything about its differential entropy?

We now argue that $h(Y)_p$ cannot exceed the entropy of a Gaussian ϕ_Y with the same variance as p_Y, let us call it σ^2. To see this, we first note that we can assume without loss of generality that both p_Y and ϕ_Y have mean zero as the entropy is independent of constant shifts. Using this, we can write

$$0 \leq D(p_Y \| \phi_Y) = -h(Y)_p + \int p_Y(y) \log \frac{1}{\phi_Y(y)} \mathrm{d}y\,. \tag{6.121}$$

Now we note that since ϕ_Y is Gaussian the expression $\log \phi_Y(y)$ is of the form $A + By^2$ for two constants A and B. Hence, we may replace $p_Y(y)$ with $\phi_Y(y)$ in the latter integral in (6.121) since they both yield $A + B\sigma^2$. Therefore, we have shown that $h(Y)_p \leq h(Y)_\phi$.

Thus, we can in particular write

$$h(y)_p \leq \frac{1}{2}\log(2\pi e(P+\sigma^2))\,. \tag{6.122}$$

Hence, we can conclude that

$$I(X:Y) \leq \frac{1}{2}\log(2\pi e(P+\sigma^2)) - \frac{1}{2}\log(2\pi e\sigma^2) = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) \tag{6.123}$$

Finally, we can see that equality can be achieved by choosing p_X Gaussian with standard deviation P (and zero mean) as the input distribution. □

It now remains to show that the channel capacity equals the power-restricted channel mutual information, i.e.,

$$C(\boldsymbol{W}) = \max_{\substack{p_X \in \mathcal{P}(\mathbb{R}) \\ \mathbb{E}[X^2] \leq P}} I(X:Y) \tag{6.124}$$

For the converse, we can actually largely build on the proof we already have — we will thus only sketch the argument here. We will ignore all technical aspects that come from the fact that we are now dealing with pdfs instead of pmfs and focus on the main idea. In the meta-converse, in the last step we introduced a maximisation over all channel input distributions: if we have restrictions on which codewords are allowed, we can also restrict the distribution there. Thus, when we apply the meta-converse for n channels, we now get

$$R \leq \min_{Q_{Y^n} \in \mathcal{P}(\mathcal{Y}^n)} \max_{P_{X^n}} \frac{1}{n} \log \frac{1}{\beta_\epsilon^*(P_{X^nY^n} \| P_{X^n} \times Q_{Y^n})}\,, \tag{6.125}$$

where we optimise over pdfs P_{X^n} has support only on codewords x^n that satisfy the constraint $\frac{1}{n}\sum_{i=1}^n x_i^2 \leq P$. This will help us in Eq. (6.35), where we can now restrict the optimisation over sequences x^n with the above property as well. It remains now only to note that the empirical distributions corresponding to these sequences satisfy

$$\mathbb{E}[X^2] = \sum_x P_X^{x^k}(x) x^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 \leq P\,. \tag{6.126}$$

The converse can thus be generalised along these lines to the continuous case, although significant care has to be taken when we go from discrete to continuous variables.

A very rough sketch can also be drawn up for achievability. The critical part here is that we choose codewords X^n at random using the i.i.d. law $P_X(x)^n$ and for some distribution with $\mathbb{E}[X^2] \leq P - \epsilon$, their power consumptions satisfies

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \to \mathbb{E}[X^2] \leq P - \epsilon, \tag{6.127}$$

as $n \to \infty$ by the weak law of large numbers, and, thus

$$P\left[\frac{1}{n}\sum_{i=1}^n X_i^2 \leq P\right] \to 1\,. \tag{6.128}$$

Thus, with high probability the codewords constructed in our random coding scheme satisfy the power constraint. In case an invalid codeword is chosen by the random process we simply discard it and count this as an error.