EE2340/EE5847: Information Sciences/Information Theory **2020**

Handout 3: Channel coding and Classification

Instructor: Shashank Vatedka

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. Please email the course instructor in case of any errors.*

Recap

- Memoryless sources
- Data compression: rate, probability of error
- Source coding theorem: The optimal rate of compression for iid sources is $H(X)$
- Entropy:

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x).$$

3.1 Discrete memoryless channels

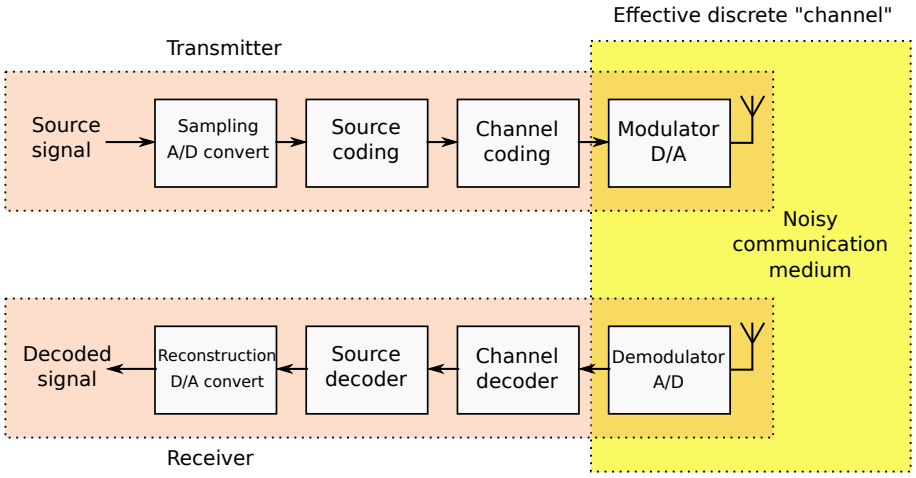

Figure 3.1: Basic block diagram of a digital communication system.

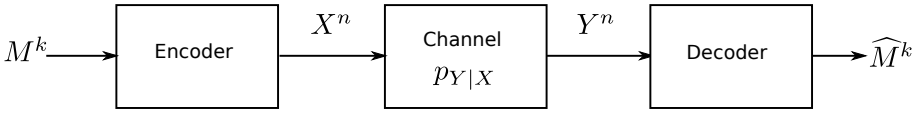

Figure 3.2: Basic block diagram for the channel coding problem.

Fig. 3.1 shows the block diagram of a a typical digital communication system. The system has a modular structure. If we combine the physical channel and the modulator/demodulator blocks, then effectively from the modulator input at the transmitter to the demodulator output at the receiver we have a discrete channel with a finite input and output alphabet.

Formally, a discrete memoryless channel is defined by an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ (here, $\mathcal{X}$ and $\mathcal{Y}$ are discrete but potentially could be infinite) and a transition probability law $p_{Y|X}$ (a conditional probability) on $\mathcal{Y} \times \mathcal{X}$. When used for n channel uses, the probability law is

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

In other words, we assume that the noise induced by the channel has no 'memory', i.e., acts independently across time.

Since the input to the channel decoder is compressed data, we will assume that the "message," or the input to the channel encoder, is a uniformly distributed k-bit sequence. Our goal is to ensure that the receiver (the decoder in this case) is able to recover the message correctly.

3.1.1 Channel code

A channel code is formally defined as a pair of maps, an encoder $f : \{0,1\}^k \to \mathcal{X}^n$, and a decoder $g : \mathcal{Y}^n \to \{0,1\}^k$. The *throughput*, or *rate* of the code is

$$R \stackrel{\text{def}}{=} \frac{k}{n}$$

while the probability of error is

$$P_e \stackrel{\text{def}}{=} \sum_{m^k \in \{0,1\}^k} \sum_{x^n, y^n} \frac{1}{2^k} \Pr[f(m^k) = x^n] p_{Y^n|X^n}(y^n|x^n) \Pr[g(y^n) = m^k].$$

Our goal is to maximize R for a desired probability of error.

We say that a rate R is *achievable* for a channel if there exist codes of rate R that achieve $\lim_{n\to 0} P_e$ over this channel. The *capacity* of the channel is the maximum R which is achievable.

Unlike the source coding problem, finding the optimal channel code is harder.

3.1.1.1 Common channels

- Binary symmetric channel (BSC): The BSC with crossover probability p, denoted BSC(p), has $\mathcal{X} = \mathcal{Y} = \{0,1\}$, and
$$p_{Y|X}(y|x) = \begin{cases} 1-p & \text{if } x = y \\ p & \text{if } x \neq y. \end{cases}$$
- Binary erasure channel (BEC): The BEC with erasure probability p, denoted BEC(p), has input alphabet $\mathcal{X} = \{0,1\}$, and output alphabet $\mathcal{Y} = \{0,1,e\}$, where e is called the erasure symbol. The transition probabilities are
$$p_{Y|X}(y|x) = \begin{cases} p & \text{if } y = e \\ 1-p & \text{if } x = y \\ 0 & \text{otherwise.} \end{cases}$$

- Additive white Gaussian noise (AWGN) channel: This is not a discrete channel, as the input and output alphabet are continuous. The input and output alphabets are $\mathbb{R}$. The channel can be described by the rule:
$$Y_i = x_i + Z_i, \qquad i = 1, 2, \ldots, n$$
where $(Z_1, \ldots, Z_n)$ are iid with $\mathcal{N}(0, \sigma^2)$ components. The transition law is:
$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}}.$$
Additionally, there is usually a power constraint imposed: the transmitted vector x^n must satisfy
$$\|x^n\|^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} x_i^2 \leqslant nP$$
for some $P > 0$ called the power constraint.
- Complex slow/quasi-static fading channel: Another continuous time channel. The input and output alphabets are $\mathbb{C}$. This can be described by the rule
$$Y_i = hX_i + Z_i,$$
where h is called the fading coefficient and is random, while Z_i is circularly symmetric complex Gaussian[1] with zero mean and variance σ^2. The transmitted vector $\underline{x} \in \mathbb{C}^n$ must also satisfy a power constraint of P:
$$\|\underline{x}\|^2 = \sum_{i=1}^{n} |x_i|^2 \leqslant nP.$$
In the simplest case, h is assumed to be complex Gaussian (Rayleigh fading).
- Fast fading channel:
$$Y_i = h_i X_i + Z_i,$$
where $h^n = (h_1, \ldots, h_n)$ has iid components.
- Multiple antenna/multi-input multi-output (MIMO) channels: Here, the input alphabet is $\mathbb{R}^{t_s}$ while the output alphabet is $\mathbb{R}^{t_r}$.
$$\mathscr{Y}_i = \mathscr{H}_i \mathscr{X}_i + \mathscr{Z}_i, \qquad i = 1, \ldots, n$$
where $\mathscr{Y}_i \in \mathbb{R}^{t_r}$, $\mathscr{X}_i \in \mathbb{R}^{t_s}$. The channel coefficient matrix $\mathscr{H}_i$ is a random $t_r \times t_s$ matrix, while $\mathscr{Z}_i \in \mathbb{R}^{t_s}$ has AWGN components. The input $(\mathscr{Y}_1, \ldots, \mathscr{Y}_n)$ must satisfy a power constraint.

3.1.1.2 More complicated channels

The following channels are *not* memoryless, and not dealt with in this course:

- A simple channel with memory: These are reasonable models for most wireless communication systems.
$$Y_i = a_0 X_i + a_1 X_{i-1} + \ldots + a_k X_{i-k} + Z_i$$
where $Z_i \sim \mathcal{N}(0, \sigma^2)$ is iid.
- Insertion/deletion channels: These model synchronization errors. In this setup, the input to the channel is an n-length vector (corresponding to the n channel uses). If the clocks between the sender and receiver are not synchronized, then the length of the vector seen at the output (after sampling) is different from n. See fig. 3.5.

[1] In other words, $h = h_R + \sqrt{-1} h_I$, where h_R and h_I are independent $\mathcal{N}(0,)$

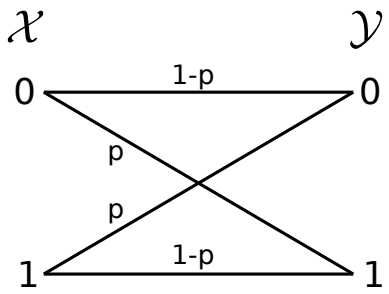

Figure 3.3: The binary symmetric channel with crossover probability p.

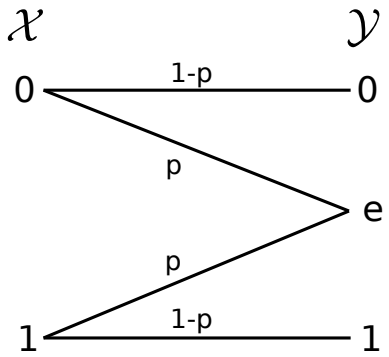

Figure 3.4: The binary erasure channel with crossover probability p.

3.1.1.3 Simple codes

Assume that $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and the following binary symmetric channel (BSC) with crossover probability p.

- **Uncoded transmission:** If we transmit the message directly, then the rate is $R = 1$, and the probability of error is $1 - (1 - p)^n$.
- **Repetition code:** Suppose that we resend each bit l times. Decoding follows the majority rule: declare the message bit to be 1 if at least $l/2$ code bits are equal to 1. Then, $n = kl$, the rate is $1/l$.

  Observe that if we want $\lim_{n\to\infty} P_e = 0$, then the asymptotic rate is zero if we use a repetition code.

3.1.2 Mutual information

If X, Y are discrete random variables with joint distributiom p_{XY}, then the mutual information between X and Y is defined as

$$I(X;Y) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) \log_2 \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)},$$

where $p_X(x) = \sum_{y\in\mathcal{Y}} p_{XY}(x,y)$ and $p_Y(y) = \sum_{x\in\mathcal{X}} p_{XY}(x,y)$ are the marginals of X and Y respectively.

Some notes and observations:

- The points we made about entropy in the previous handout also holds for mutual information. $I(X;Y)$

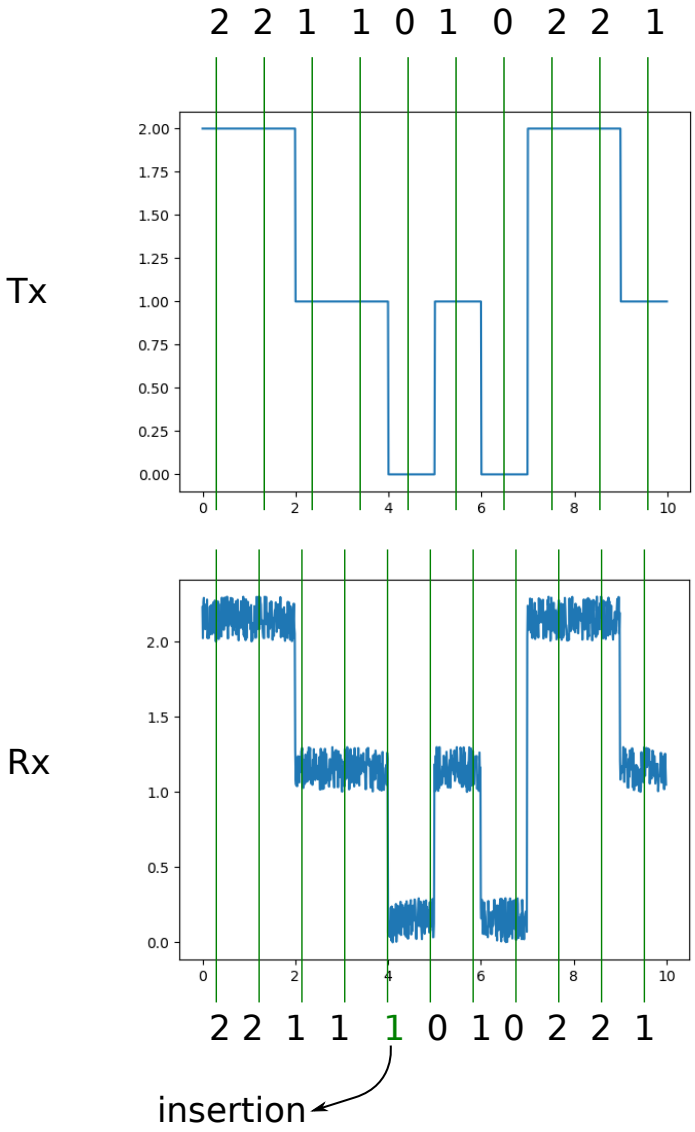

Figure 3.5: Insertion channel. Due to the difference in clock frequencies, additional symbols may get "inserted". The length of the output is different from the length of the input.

is not a function of X, Y but rather a function of p_{XY}. Moreover, the sum in $I(X;Y)$ is taken over only those (x, y) for which $p_{XY}(x, y) > 0$.

- Mutual information is symmetric in X and Y. In other words, $I(X;Y) = I(Y;X)$. The following is an example of a function which is *not symmetric* in X, Y:

$$\widetilde{I}(X;Y) = \sum_{(x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) \log_2 \frac{p_{XY}(x,y)}{p_X(x)}.$$

- The mutual information is a measure of "information" that X gives about Y, or (due to symmetry) that Y gives about X.

We will briefly give an operational meaning to mutual information before studying the properties of entropy and mutual information.

3.2 The channel coding theorem

Theorem 3.1 (Shannon, 1948)**.** *The capacity of a discrete memoryless channel with transition probabilities $p_{Y|X}$ is equal to*

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

Let us see what the channel coding theorem says for the binary symmetric channel.

Claim: It suffices to consider $0 < p < 1/2$.

- What is the optimal channel code for $p = 0$?
- Why do you think the capacity is zero when $p = 1/2$? What can you do?
- What if $p > 1/2$?

For the BSC(p), the capacity $C = 1 - H_2(p) = 1 + p\log_2 p + (1-p)\log_2(1-p)$. The p_X that maximizes $I(X;Y)$ is Bernoulli(1/2).

Definition 3.2. *Given two binary sequences x^n and y^n, the* Hamming distance *between x^n and y^n, denoted $d_H(x^n, y^n)$, is the number of locations in which x^n and y^n differ.*

The Hamming distance is a *metric*. A metric generalizes the notion of distance (such as the Euclidean distance you are familiar with). A metric on a set $\mathcal{X}$ is a function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

- (Non-negativity) $f(x, y) \geqslant 0$ for all $x, y \in \mathcal{X}$, and equality holds if and only if $x = y$,
- (Symmetry) $f(x, y) = f(y, x)$ for all x, y, and
- (Triangle inequality) $f(x, y) \leqslant f(x, z) + f(z, y)$ for all x, y, z.

You should be able to prove the following lemma:

Lemma 3.3. *For any binary sequence x^n, let Y^n be the sequence that the receiver observes when x^n is passed through a BSC(p). Then,*

$$\Pr[d_H(x^n, Y^n) > np(1+\epsilon)] \leqslant 2^{-\alpha\epsilon^2 n}$$

for some universal constant $\alpha > 0$.

3.2.0.1 Channel coding as a packing problem

Lemma 3.3 says that the received vector is at a Hamming distance of $\approx np$ from the transmitted codeword. This gives a nice way of constructing a code: The codebook must be a set of points in $\{0,1\}^n$ such that every pair of points are at a distance of at least $2np(1+2\epsilon)$ away from each other.

The above construction guarantees that the probability of error is at most $2^{-\alpha\epsilon^2 n}$.

However, it turns out that this method gives us a rate less than Theorem 3.1.

3.3 Relationship between entropy and mutual information

We can write,

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} p_{XY}(x,y) \log_2 \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \\
&= \sum_{x,y} p_{XY}(x,y) \log_2 \frac{p_{X|Y}(x,y)}{p_X(x)} \\
&= \sum_{x,y} p_{XY}(x,y) \left(\log_2 p_{X|Y}(x,y) - \log_2 p_X(x)\right) \\
&= -\sum_x p_X(x) \log_2 p_X(x) + -\sum_x p_{XY}(x,y) \log_2 p_{X|Y}(x|y) \\
&= H(X) - H(X|Y),
\end{aligned}$$

where

$$H(X|Y) \stackrel{\text{def}}{=} -\sum_x p_{XY}(x,y) \log_2 p_{X|Y}(x|y)$$

is called the conditional entropy of X given Y. Similarly,

$$I(X;Y) = H(Y) - H(Y|X)$$

Interpretation: Mutual information is a measure of the information that X gives about Y (or by symmetry, Y gives about X). It is equal to the difference in the randomness that was originally there about X and the amount of randomness that remains about X after observing Y (and the other way around, by symmetry).

3.4 The problem of classification

Consider the following problem: you observe n iid samples $X_1, \ldots, X_n$, where each sample may have distribution p_s or p_g. In the context of email spam classification, p_s is the distribution of the email if it is spam, and p_g is the distribution if it is not spam (good).

We want a rule such that

$$\Pr[\text{declare spam}|\text{email is not spam}] \leqslant \epsilon$$

and

$$\Pr[\text{declare not spam}|\text{email is spam}] \text{ is minimized.}$$

The optimal test turns out to be the likelihood ratio test:

$$\text{Output} \begin{cases} \text{spam} & \text{if } \log_2 \frac{p_s(x^n)}{p_g(x^n)} > \alpha \\ \text{not spam} & \text{if } \log_2 \frac{p_s(x^n)}{p_g(x^n)} \leqslant \alpha \end{cases}$$

The value of α is chosen to satisfy

$$\Pr\left[\log_2 \frac{p_s(X^n)}{p_g(X^n)} > \alpha \middle| \text{not spam}\right] = \epsilon$$

Then,

$$\lim_{n\to\infty} \frac{1}{n} \Pr[\text{declare not spam}|\text{email is spam}] = -D(p_s \| p_g),$$

where

$$D(p_s \| p_g) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} p_s(x) \log_2 \frac{p_s(x)}{p_g(x)}$$

is called the Kullback-Liebler (KL) divergence (or the relative entropy) between p_s and p_g.

- Relationship between KL divergence and mutual information:

$$I(X;Y) = D(p_{XY} \| p_X p_Y).$$

3.5 Continuous alphabet

Suppose that we have continuous random variables X, Y with a joint pdf f_{XY} and marginals f_X, f_Y.

The differential entropy of X,

$$h(X) \stackrel{\text{def}}{=} -\int_{-\infty}^{\infty} f_X(x) \log_2 f_X(x) dx.$$

Recall that we use the convention $0 \log_2 0 = 0$.

The conditional differential entropy of X given Y is

$$h(X|Y) \stackrel{\text{def}}{=} -\int_{x,y} f_{XY}(x,y) \log_2 f_{X|Y}(x|y) dx dy.$$

The mutual information between X and Y is

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = \int_{x,y} f_{XY}(x,y) \log_2 \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dx dy$$

Shannon's channel coding theorem holds for continuous alphabet as well: The capacity of any channel with power constraint P and transition law $f_{Y|X}$ is

$$C = \max_{f_X : \text{Var}(X) \leqslant P} I(X;Y) \tag{3.1}$$

The Gaussian random variable is very important as we encounter it frequently in communications and signal processing. You can compute the differential entropy of the Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$: It is equal to

$$h(X) = \frac{1}{2} \log_2(2\pi e \sigma^2).$$

For the power constrained AWGN channel $Y_i = X_i + Z_i$ with $Z_i \sim \mathcal{N}(0, \sigma^2)$, (3.1) is maximized when f_X is $\mathcal{N}(0, P)$. Then,

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z).$$

Since X, Z are independent Gaussians, $Y \sim \mathcal{N}(0, P + \sigma^2)$. Using the formula for the entropy of a Gaussian and simplifying, we get

$$C = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right).$$