GI12/4C59: Information Theory

Lectures 19–21

Massimiliano Pontil

About these lectures

Theme of these lectures: We continue the discussion of discrete noisy channels and consider channels with memory (a.k.a. channels with feedback). We then turn our attention to continuous channels and analyze the important Gaussian channel.

Outline

1. Review of the channel coding theorem
2. Channels with feedback
3. Continuous channel
4. The Gaussian channel

Review of channel coding theorem

Memoryless channel:

$$W \longrightarrow \boxed{\text{Encoder}} \rightarrow X^n \rightarrow \boxed{\textit{Channel}} \rightarrow Y^n \rightarrow \boxed{\text{Decoder}} \rightarrow \hat{W}$$

The received signal y^n has conditional probability distribution

$$p(y^n|x^n) = p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n).$$

A code $C^{(n)}$ is identified by the codewords $x^n(w) \in \mathcal{X}^n,\ w \in \mathcal{W}$ and a decoding function $g : \mathcal{Y}^n \rightarrow \mathcal{W}$.

The rate R of the code $\mathcal{C}^{(n)}$ is defined to be $R := \frac{\log|\mathcal{W}|}{n}$.

Review of channel coding theorem (cont.)

- The probability of error of $\mathcal{C}^{(n)}$ is defined as

$$\lambda(E|\mathcal{C}^{(n)}) = \max_{w \in \mathcal{W}} P(g(Y^n) \neq w).$$

- A rate R is said achievable if there exists a sequence of codes $\mathcal{C}^{(n)}, n \in \mathbb{N}$ such that $\lambda(E|\mathcal{C}^{(n)})$ converges to zero as $n \to \infty$.

- The capacity C of the channel is the supremum of all possible achievable rates.

Channel coding theorem (main part): $C = \max_{p(x)} I(X;Y)$.

Review of channel coding theorem (cont.)

The idea in proving the theorem consisted in generating a code $C^{(n)}$ at random, i.e. $x^n \sim p(x_1) \cdots p(x_n)$, with the decoding function based on "joint typicality verification".

The main steps in proving that $R < C$ is achievable were

- The average probability of error $P^{(n)}(E)$ with respect to $C^{(n)}$ and a random choice of the sent message equals, by symmetry, $P^{(n)}(E|W = 1)$.

- If $R < I(X;Y)$, then $P^{(n)}(E|W = 1)$ converges to zero as $n \to \infty$.

- There exists a sequence of codes $C^{*(n)}, n \in \mathbb{N}$ whose maximal probability of error converges goes to zero as $n \to \infty$.

Then, the result follows by optimizing with respect to $p(x)$.

Decoding rule (informal argument)

Informally, if the input codewords have long length, the channel looks like the noisy typewriter channel: there is a subset of inputs which produces disjoint output subsets.

- Each input x^n will likely produce about $2^{nH(Y|X)}$ output sequences which are equally likely.
- The total number of typical output sequences is about $2^{nH(Y)}$.
- Thus, there is a subset of about $2^{nH(Y)}/2^{nH(Y|X)} = 2^{nI(X;Y)}$ inputs which produces disjoint subsets of typical output sequences.

The above argument suggests that a rate $R = I(X;Y)$ is achievable and, so, the capacity is $C = \max_{p(x)} I$.

Channels with feedback

In this case the codewords associated to message w is a computed with feedback, that is

$$x_1 = x_1(w), \quad x_i = x_i(w, y^{i-1}) \ , i = 2, \ldots, n.$$

The rate of a $\mathcal{C}^{(n)}$ feedback code is still defined as

$$R = \frac{\log |\mathcal{W}|}{n}$$

Remark: In a memoryless channel the codewords are determined before sending the message. Here each message has multiple possible codes associated to it. The code which is actually used depends on the measured outputs.

Feedback capacity

We define the capacity with feedback, C_F, of a discrete memoryless channel as the supremum of all rates achievable by feedback codes.

Let $C = \max_{p(x)} I(X;Y)$ be the capacity without feedback. By construction $C_F \geq C$, and we may expect that $C_F > C$ as feedback allows for more reliable transmission. Instead we have the rather surprising result.

Theorem: $C_F = C$ (Feedback does not increase the capacity).

Proof of the theorem

Let $P^{(n)}$ be the average probability of error when $p(w)$ is uniform. Then, by Fano inequality, $H(W|Y^n) \leq 1 + nRP^{(n)}$ and

$$nR = H(W) = H(W|Y^n) + I(W;Y^n) \leq 1 + P^{(n)}nR + I(W;Y^n)$$

and

$$\begin{aligned} I(W;Y^n) &= H(Y^n) - H(Y^n|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{i-1},\ldots,Y_1,W) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{i-1},\ldots,Y_1,W,X_i) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i,Y_i) \leq nC. \end{aligned}$$

(we used the fact that $X_i = X_i(W, Y_1, \ldots, Y_{i-1})$ and, conditional on X_i, Y_i is independent of W and $Y_1, \ldots, Y_{i-1}$). Putting all together, we conclude that

$$nR \leq 1 + nRP^{(n)} + nC \quad \Rightarrow \quad R \leq \frac{1}{n(1-P^{(n)})} + \frac{C}{1-P^{(n)}} \to C.$$

Entropy of a continuous random variable

Let X be a continuous r.v. with c.d.f. $F(x) = P(\{X \leq x\})$ and assume that $f(x) = F'(x)$ is well defined and $\int_{-\infty}^{\infty} f(x) = 1$.

The entropy of a r.v. X with density f (also called differential entropy) is defined as

$$H(X) = H(f) = -\int_{-a}^{a} f(x) \log f(x) dx = -E[\log f(x)]$$

(Remember that we use the convention $0 \log 0 = 0$.

Note: Unlike in the case of a discrete r.v., now $H(X)$ can be negative.

Entropy of a continuous random variable (cont)

Example 1: If $f(x)$ is the uniform distribution on the interval $[0, a]$, that is, $f(x) = \frac{1}{a}$ if $x \in [0, a]$ and zero otherwise, we have

$$H(X) = -\int_{-\infty}^{\infty} \frac{1}{a} \log \frac{1}{a} dx = \log a.$$

Example 2: If $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$, the Gaussian distribution with mean μ and variance σ,

$$H_e(X) = -\int_{-\infty}^{\infty} f(x) \left(-\frac{(x-\mu)^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right) dx = \frac{1}{2} + \frac{1}{2} \ln 2\pi\sigma^2 = \frac{1}{2} \ln 2\pi e\sigma^2.$$

Thus, $H(X) = H_2(X) = \frac{1}{2} \log 2\pi e\sigma^2$. (Remember $\log_b x = \log_b a \log_a x$, in particular $\log x = \log e \ln x$)

Remark: The extension of the entropy (and relative entropy and mutual information – see below) to the continuous case needs some care since sums are replaced by integrals which may not exist or be infinite.

Joint entropy

If $X_1, \ldots, X_n$ is a sequence of continuous r.v., their *joint entropy* is defined by

$$H(X_1, \ldots, X_n) = -E[\log f(X_1, \ldots, X_n)].$$

This satisfies the chain rule:

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1).$$

In particular $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$.

Entropy of multivariate Normal distribution

Let $\mathbf{x} = (x_1, \ldots, x_n)$, $f(\mathbf{x}) \equiv G(\mu, K) = \frac{1}{\sqrt{2\pi}^n \sqrt{det(K)}} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top K^{-1}(\mathbf{x}-\mu))$, the multivariate normal distribution with mean $\mu \in \mathbb{R}^n$ and $n \times n$ covariance matrix K. We have

$$H_e(X_1, \ldots, X_n) = \frac{n}{2} \log(2\pi e \det(K))$$

Proof:

$$\begin{aligned} H(X_1, \ldots, X_n) &= -\int f(\mathbf{x}) \left(-\frac{1}{2}(\mathbf{x}-\mu)^\top K^{-1}(x-\mu)) + \frac{1}{2} \ln(2\pi)^n \det(K) \right) \\ &= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_j) K_{ij}^{-1} (x_j - \mu_j) \right] + \frac{1}{2} \ln(2\pi)^n \det(K) \end{aligned}$$

and the result follows by noting that

$$E\left[\sum_{i,j} (x_i - \mu_i) K_{ij}^{-1} (x_j - \mu_j) \right] = \sum_{ij} K_{ji} K_{ji}^{-1} = n$$

Relative entropy / mutual information

The relative entropy of the ordered density pair f, g is defined by

$$D(f \parallel g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \ .$$

This is finite only if $supp(f) \subseteq supp(g)$. Here X can be a vector of r.v., $X = (X_1, \ldots, X_n)$, in which case $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

If $(X, Y) \sim f(X, Y)$ are jointly continuous r.v., their mutual information is defined by

$$I(X;Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy = D(f(x,y) \parallel f(x)f(y))$$

and we have $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Quantization of a continuous r.v.

H, D and I are the limit of their quantized version.

If we "approximate" X by a infinite discrete r.v. X_δ which takes the values $f(x_i)\delta$, $i \in \mathbb{N}$, where the x_i (bins) are chosen to satisfy

$$p_i := \int_{i\delta}^{(i+1)\delta} f(x)dx = f(x_i)\delta$$

we have

$$H(X_\delta) = -\sum_i p_i \log p_i = -\sum_i (\delta f(x_i) \log f(x_i) + f(x_i)\delta \log \delta) \approx H(X) - \log \delta$$

In particular, the entropy of an $n-$bit quantization of X (that is $\delta = 2^{-n}$) is $H(X) + n$.

Examples

Example 1: If $f(X)$ is uniform in $[0,1]$, $H(X) = 0$ and n bits suffice to describe X to $n-$bit accuracy.

Example 2: If $f(X)$ has support $[0, \frac{1}{8}]$ and we choose $\delta = 2^{-n}$, that is X_δ takes $\frac{1}{8}2^n = 2^{n-3}$ possible values. Thus, $n-3$ bits suffice to describe X to $n-$bit accuracy.

Remarks:

- $\frac{X_\delta}{\delta}$ is a step function approximating X.
- In the above examples, each value of X_δ is described by the same number of bits. In general $H(X) + n$ is the average number of bits required to describe X to n bit accuracy.
- In general the bins do not need to have the same length δ. A more efficient quantization method is to use a variable parameter δ where δ is smaller in regions where $f(x)$ is large.

Mutual information

Let X, Y be continuous r.v. and X^δ, Y^δ their quantized versions.

We have

$$\begin{aligned} I(X^\delta; Y^\delta) &= H(X^\delta) - H(X^\delta|Y^\delta) \\ &\approx H(X) - \log\delta - (H(X|Y) - \log\delta) \\ &= I(X;Y) \end{aligned}$$

Thus, the mutual information of X and Y is the limit of the mutual information of their quantized version.

Properties of D and I

1. $D(f \parallel g) \geq 0$, with equality if and only if $f = g$ with probability 1.
2. $D(f \parallel g)$ is convex in the pair (f, g).
3. $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.
4. $I(X;Y)$ is a concave function of $p(x)$ for $p(y|x)$ fixed and a convex function of $p(y|x)$ for $p(x)$ fixed.

Note: Here $D(f \parallel g)$ is a *functional* of the functions f and g.
A functional $J(f)$ is convex if for every f_1, f_2 and $\lambda \in [0,1]$,

$$J(\lambda f_1 + (1-\lambda) f_2) \leq \lambda J(f_1) + (1-\lambda) J(f_2).$$

The convexity of a functional of two functions f and g is defined similarly.

Proof of property 1

Let S be the support of f. Then

$$\begin{aligned} -D(f \parallel g) &= \int_S f \log \frac{g}{f} \\ &\leq \log \int_S f \frac{g}{f} \\ &= \log \int_S g \leq \log 1 = 0 \end{aligned}$$

The first inequality follows by Jensen inequality. In that inequality, equality holds if and only if $f = g$ almost everywhere.

Properties of the entropy

a) $H(X|Y) \leq H(X)$, with equality iff X and Y are indep.

- Follows from above property 3 and $I(X;Y) = H(X) - H(X|Y)$.

b) $H(X^n) \leq \sum_{i=1}^{n} H(X_i)$ with equality iff the X_i are indep.

- Follows from chain rule for entropy,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$$

and above property 3.

Properties of the entropy (cont.)

For every $a \in \mathbb{R}, c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$, we have that

c) $H(X^n + c) = H(X^n)$.

d) $H(AX^n) = H(X^n) + \log|\det(A)|$.

- In particular, $H(aX) = H(X) + \log|a|$, $a \in \mathbb{R}$. In fact, if $Y = aX$ we have

$$f_Y(y) = \frac{1}{|a|} f\left(\frac{y}{a}\right)$$

and so

$$\begin{aligned} H(Y) &= -\int f_Y(y) \log f_Y(y) = -\int \frac{1}{|a|} f\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f\left(\frac{y}{|a|}\right)\right) dy \\ &= -\int f(x) \log f(x) + \log|a| = H(X) + \log|a| \end{aligned}$$

One more important property of H

Lemma: If $\mathbf{X} = (X_1, \ldots, X_n)$ has zero mean and covariance $K = E[\mathbf{X}\mathbf{X}^\top]$, then

$$H(\mathbf{X}) \leq \frac{n}{2} \log(2\pi e \det(K))$$

with equality if and only if $\mathbf{X} \sim G(0, K)$. In particular, if $n = 1$ and $E[X^2] \leq \sigma^2$, then $H(X) \leq \frac{n}{2} \log(2\pi e \sigma^2)$.

Proof: Let $g = G(0, K)$. We have

$$\begin{aligned} 0 &\leq D(f \parallel g) = \int f \log \frac{f}{g} = -H(f) - \int f \log g \\ &= -H(f) - \int g \log g = -H(f) + H(g) \quad \Rightarrow \quad H(f) \leq H(g) \end{aligned}$$

where we use the fact that $\log g$ is a quadratic form in $\mathbf{X}$ and f satisfies the zero mean and covariance constraint.

Typical sequences

If $X_1, \ldots, X_n$ are **i.i.d.** continuous r.v. with density f, then, by the weak law of large numbers,

$$\lim_{n \to \infty} -\frac{\log f(X_1, \ldots, X_n)}{n} = -E[\log f(X)] = H(X) \quad \text{in probability}$$

The $\epsilon-$typical set is defined as

$$A_\epsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \left| \frac{\log f(x_1, \ldots, x_n)}{n} + H(X) \right| \leq \epsilon \right\}$$

Remark: The above definition, the below definition of $\epsilon-$jointly typical sequences, and the properties of typical sets follow closely the discrete case.

Property of the typical sequences

- $P(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.

This follows from the law of large number (see above)

- $Vol(A_\epsilon^{(n)}) := \int_{A_\epsilon^{(n)}} dx_1 \dots dx_n \leq 2^{n(H(X)+\epsilon)}$ for every n.

In fact, we have:

$$\begin{aligned} 1 &= \int f(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} f(x_1, \dots, x_n) dx_1 \cdots dx_n \geq \int_{A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} dx_1 \cdots dx_n \\ &= 2^{-n(H(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 \cdots dx_n = 2^{-n(H(X)+\epsilon)} Vol(A_\epsilon^{(n)}) \end{aligned}$$

Property of the typical sequences (cont.)

- $Vol(A_\epsilon^{(n)}) \geq (1 - \epsilon) 2^{n(H(X)-\epsilon)}$ for n sufficiently large.

In fact, if n is sufficiently large so that $P(A_\epsilon^{(n)}) > 1 - \epsilon$, we have

$$\begin{aligned} 1 - \epsilon &= \int_{A_\epsilon^{(n)}} f(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &\leq \int_{A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} dx_1 \cdots dx_n \\ &= 2^{-n(H(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 \cdots dx_n = 2^{-n(H(X)-\epsilon)} Vol(A_\epsilon^{(n)}) \end{aligned}$$

Jointly typical sequences

Given i.i.d. r.v. $(X_i, Y_i), i = 1, \ldots, n$, the set of jointly $\epsilon-$typical sequences is defined by

$$\mathcal{A}_\epsilon^{(n)} := \left\{ (x^n, y^n) : x^n, y^n \text{ are } \epsilon - \text{typical, and } \left| \frac{\log f(x^n, y^n)}{n} + H(X^n, Y^n) \right| \leq \epsilon \right\}$$

For every $\epsilon > 0$, we have that

1. $P\left(\left\{(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\}\right) \to 1$ when $n \to \infty$.
2. $Vol(\mathcal{A}_\epsilon^{(n)}) \in \left[(1-\epsilon)2^{n(H(X,Y)-\epsilon)}, 2^{n(H(X,Y)+\epsilon)}\right]$.
3. If S^n and T^n are independent with the same marginal distributions as X^n and Y^n respectively, then

$$P\left(\left\{(S^n, T^n) \in \mathcal{A}_\epsilon^{(n)}\right\}\right) \in \left[(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}, 2^{-n(I(X;Y)-3\epsilon)}\right].$$

The Gaussian channel

The Gaussian channel is a time discrete memoryless channel specified by the conditional density

$$p(y|x) = G(x, K)$$

that is

$$Y = X + Z$$

where X and Y are independent and $Z \sim G(0, K)$.

Remark: like in the discrete case, the memoryless assumption says that if $x^n = (x_1, \ldots, x_n)$ is the input to the channel and $y^n = (y_1, \ldots, y_n)$ the output,

$$p(y^n|x^n) = p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n)$$

The Gaussian channel (cont.)

We wish to send $M = |\mathcal{W}|$ possible messages $w \in \mathcal{W}$ over the channel. The set-up is like in the discrete case except that we require that each message codeword $x^n(w),\ w \in \mathcal{W}$ satisfies the constraint

$$\sum_{i=1}^{n} x_i^2(w) \leq n\rho, \quad \rho > 0.$$

Without this constraint the capacity of the channel would be infinite (verify that we could code infinitely many messages just using codewords of length 1...).

Gaussian channel vs. discrete channel

Assume we have two messages ($M = 2$) which we wish to code with 1 bit. Given the above constraint, the best code is $x(1) = \sqrt{\rho}$, $x(2) = -\sqrt{\rho}$. If $p(W = 1) = p(W = 2) = \frac{1}{2}$ we have

$$\begin{aligned} P(error) &= \frac{1}{2}P(Y < 0|X = \sqrt{\rho}) + \frac{1}{2}P(Y > 0|X = -\sqrt{\rho}) \\ &= P(Z > \sqrt{\rho}) = 1 - \Phi(\sqrt{\frac{\rho}{K}}) \end{aligned}$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})dx$$

Thus, we can use a Gaussian channel as a discrete binary symmetric channel with crossover probability $p = P(error)$

Capacity

A code $C^{(n)}$ is specified by the codewords $x^n(w),\ w \in \mathcal{W}$ satisfying the above constraint and a decoding function $g : \mathcal{Y}^n \to \mathcal{W}$.

The probability of error of $C^{(n)}$ is defined by

$$\lambda(E|\mathcal{C}^{(n)}) := \max\{P(\{g(Y^n) \neq w\}|\{X^n = x^n(w)\}) : w \in \mathcal{W}\}$$

A transmission rate R is said to be *achievable* if there exists a sequence of $(\lceil 2^{nR} \rceil, n), n \in \mathbb{N}$ codes such that,

$$\lim_{n\to\infty} \lambda(E|\mathcal{C}^{(n)}) = 0$$

The **capacity** C of the channel is the supremum of all achievable rates.

The channel coding theorem for Gaussian channel

Theorem (Shannon) The capacity of the Gaussian channel is

$$C = \max_{p(x)}\{I(X;Y) : E[X^2] \leq \rho\} = \frac{1}{2}\log\left(1 + \frac{\rho}{K}\right)$$

that is, for every rate $R < C$, there exits a sequence of $\mathcal{C}^{(n)} = (\lceil 2^{nR} \rceil, n)$ codes such that $\lambda(E|\mathcal{C}^{(n)}) \to 0$ as $n \to \infty$. Conversely, any sequence of codes for which $\lambda(E|\mathcal{C}^{(n)}) \to 0$ must have a rate $R \leq C$.

Computation of the capacity

We first show that $\max_{p(x)}\{I(X;Y) : E[X^2] \leq \rho\} = \frac{1}{2}\log\left(1+\frac{\rho}{K}\right)$.

Since $Y = X + Z$, $Z \sim G(0,K)$, and X and Z independent, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(X+Z|X) \\ &= H(Y) - H(Z|X) = H(Y) - H(Z) \end{aligned}$$

We also have

$$H(Z) = \frac{1}{2}\log 2\pi e K, \quad E[Y^2] = E[X^2] + 2E[X]E[Z] + E[Z^2] = \rho + K.$$

Consequently, by the property above $H(Y) \leq \frac{1}{2}\log 2\pi e(\rho+K)$ and

$$I(X;Y) \leq -\frac{1}{2}\log 2\pi e K + \frac{1}{2}\log 2\pi e(\rho+K) = \frac{1}{2}\log\left(1+\frac{\rho}{K}\right).$$

Proof of part 1

The proof of the theorem parallels the proof for the discrete case with the minor difference that, here, we need to take into account the codeword constraint. The key steps are

1. Random code generation.
2. Decoding by jointly typicality.
3. Computation of the average probability of error.
4. Extraction of a good code.

Part 1: code generation

We generate a $(\lceil 2^{nR} \rceil, n)$ code $\mathcal{C}$ at random according to $p(x) = G(0, \rho - \epsilon)$. Each codeword $x^n(w), w = 1, \ldots, M := \lceil 2^{nR} \rceil$ is generated with probability

$$p(x^n(w)) = \prod_{i=1}^{n} p(x_i(w)) \quad \Rightarrow \quad P(C) = \prod_{w=1}^{M} \prod_{i=1}^{n} p(x_i(w))$$

When n grows the probability that a generated codeword does not satisfy the power constraint goes to zero. In fact, if we set $E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2 > \rho \right\}$, for every $\epsilon > 0$, by the weak law of large numbers

$$P(E_0) \to 0 \quad \text{as } n \to \infty$$

Part 1: decoding function

Let $x^n(w)$ and y^n be the sent and received signals.

We set $g(y^n) = \widehat{w}$ if i) $x^n(\widehat{w})$ is the only codeword which is jointly typical with y^n and ii) $x^n(\widehat{w})$ satisfies the power constraint. Otherwise we declare an error.

We also define the events

$$E_w = \left\{ (X^n(w), Y^n) \in \mathcal{A}_\epsilon^{(n)} \right\}, \ w = 1, \ldots, M.$$

Part 1: probability of error

By the symmetry of the code generation process we have that

$$P(E) = P(E|W = 1).$$

The events which contribute to the occurrence of an error are E_0, $\bar{E}_1$, E_w, $w \geq 2$. Then, using the union bound,

$$\begin{aligned} P(E|W=1) &= P(E_0 \cup \bar{E}_1 \cup E_2 \cup E_3 \ldots \cup E_M) \qquad (1) \\ &\leq P(E_0) + P(\bar{E}_1) + \sum_{w=2}^{M} P(E_w) \\ &\leq 2\epsilon + \sum_{w=2}^{M} P(E_w). \end{aligned}$$

Part 1: probability of error (cont.)

If $w \geq 2$ $X^n(1)$ and $X^n(w)$ and independent and, so, Y^n and $X^n(w)$ are also independent. By the property of typical sequences, the probability that Y^n and $X^n(w)$ are jointed typical is bounded as

$$P(E_w) \leq 2^{-n(I(X;Y)-3\epsilon)}, \; w = 2, 3, \ldots, M = 2^{nR}.$$

We conclude that

$$\begin{aligned} P(E|W=1) &\leq 2\epsilon + \sum_{w=2}^{2^{nR}} P(E_w) \qquad (2) \\ &\leq 2\epsilon + \sum_{w=1}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= 2\epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \leq 2\epsilon + 2^{3n\epsilon}2^{-n(I(X;Y)-R)} \leq 3\epsilon \end{aligned}$$

where the last step holds if n is sufficiently large and $R < I(X;Y) - 3\epsilon$.

Like in the discrete case, we conclude that, for n large enough, $P(E) < \epsilon$.

Part 1: extracting a good code

The above argument shows that, if n is large enough,

$$P(E) = \sum_{\mathcal{C}^{(n)}} P(\mathcal{C}^{(n)}) P(E|\mathcal{C}^{(n)}) \leq \epsilon.$$

Then, there must exist a code $P(\mathcal{C}^{*(n)})$ such that

$$P(E|\mathcal{C}^{*(n)}) \leq \epsilon.$$

Finally, if we remove the worst half codewords (all the codewords which do not satisfy the power constraint are included in this set), we have a new code with rate $R - \frac{1}{n}$ whose codewords satisfy the power constraint and has maximal probability of error $\leq 2\epsilon$.

Proof of part 2

We now show the converse of the theorem: a rate $R > C = \frac{1}{2}\log\left(1 + \frac{\rho}{K}\right)$, is not achievable or, equivalently,

$$\lambda(E|C^{(n)}) \to 0 \quad \Rightarrow \quad R \leq C.$$

Proof: We assume that a message w is selected at random and denote by $P^{(n)}$ the average probability of error. By Fano inequality we have

$$H(W|Y^n) \leq 1 + nRP^{(n)}$$

and, as we shown for the discrete channel, we have

$$H(W) = nR \leq 1 + nRP^{(n)} + \sum_{i=1}^{n} I(X_i; Y_i).$$

Proof of part 2 (cont.)

We set $\kappa_i = \frac{1}{M}\sum_w x_i^2(w)$, $M = 2^{nR}$. Since $Y_i = X_i + Z_i$ with X_i and Z_i independent, the second order momentum of Y_i is $\kappa_i + K$.

Thus, by the property of the entropy with covariance constraint, we have

$$H(Y_i) \leq \frac{1}{2}\log 2\pi e(\kappa_i + K)$$

and

$$\begin{aligned} I(X_i;Y_i) &= H(Y_i) - H(Y_i|X_i) = H(Y_i) - H(Z_i) \\ &\leq \frac{1}{2}\log 2\pi e(\kappa_i + K) - \frac{1}{2}\log 2\pi e K = \frac{1}{2}\log(1 + \kappa_i/K) \end{aligned}$$

We conclude that

$$nR \leq 1 + nRP^{(n)} + \sum_{i=1}^{n} I(X_i;Y_i) \leq 1 + nRP^{(n)} + \frac{1}{2}\sum_{i=1}^{M}\log(1 + \kappa_i/K)$$

Proof of part 2 (cont.)

We rewrite last equation as

$$R \leq \frac{1}{2n}\sum_{i=1}^{n}\log(1 + \kappa_i/K) + RP^{(n)} + \frac{1}{n}.$$

and observe that, since $\log(1+x)$ is concave in x, by Jensen inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\log(1 + \kappa_i/K) \leq \log\left(1 + \frac{1}{n}\sum_{i=1}^{n}\kappa_i/K\right) \leq \log\left(1 + \frac{\rho}{K}\right)$$

where the last inequality we use $\frac{1}{n}\sum_{i=1}^{n} k_i \leq \rho$ which is because of the power constraint of each codeword. Thus,

$$R \leq \frac{1}{2}\log\left(1 + \frac{\rho}{K}\right) + RP^{(n)} + \frac{1}{n} \quad \longrightarrow \quad \frac{1}{2}\log\left(1 + \frac{\rho}{K}\right).$$

Bibliography

This lectures are based on Sec. 8.12-13, Chapters 9, and Sec. 10.1-2 of Cover and Thomas's.