10-704: Information Processing and Learning **Spring 2012**

Lecture 18: Gaussian channel, Parallel channels and Rate-distortion theory

Lecturer: Aarti Singh *Scribe: Danai Koutra*

18.1 Joint Source-Channel Coding (cont'd)

Clarification: Last time we mentioned that we should be able to transfer the source over the channel only if $H(src) < Capacity(ch)$. Why not design a very low rate code by repeating some bits of the source (which would of course overcome any errors in the long run)? This design is not valid, because we cannot accumulate bits from the source into a buffer; we have to transmit them immediately as they are generated.

18.2 Continuous Alphabet (discrete-time, memoryless) Channel

The most important continuous alphabet channel is the Gaussian channel. We assume that we have a signal X_i with gaussian noise $Z_i \sim \mathcal{N}(0, \sigma^2)$ (which is independent from X_i). Then, $Y_i = X_i + Z_i$. Without any

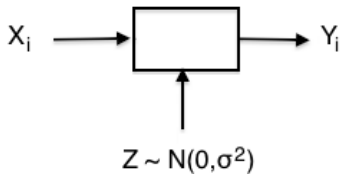


Figure 18.1: Gaussian channel

further constraints, the capacity of the channel can be infinite as there exists an inifinte subset of the inputs which can be perfectly recovery (as discussed in last lecture). Therefore, we are interested in the case where we have a power constraint on the input. For every transmitted codeword $(x_1, x_2, \ldots, x_n)$ the following inequality should hold:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P \;\; (deterministic) \qquad or \qquad \frac{1}{n}\sum_{i=1}^{n} E[x_i^2] = E[X^2] \leq P \;\; (randomly \;\; generated \;\; codebits)$$

Last time we computed the information capacity of a Gaussian channel with power constraint.

Definition 18.1 (Information Capacity) *The information capacity C of the Gaussian channel with power constraint P is*

$$C = \max_{p(x):E[X^2]\leq P} I(X,Y) = \frac{1}{2}\log(\frac{P+\sigma^2}{\sigma^2}).$$

Definition 18.2 (Operational Capacity) *The operational capacity of the Gaussian channel with power constraint P is the supremum of all achievable rates. A rate is achievable if there exists a sequence of $(2^{nR}, n)$ codes with codewords that satisfy the power constraint and the maximal probability of error $\lambda^{(n)} \to 0$.*

Similar to the DMC (Discrete Memoryless Channel), the operational capacity of a Gaussian channel is same as the information capacity, as we discuss next.

Sphere-packing argument Here we try to give an informal argument why we can construct $(2^{\frac{n}{2}\log\frac{P+\sigma^2}{\sigma^2}}, n)$ codes with low probability of error. For any codeword of length n, the received vector is:

$$Y = [Y_1 \dots Y_n]^T \sim \mathcal{N}([X_1 \dots X_n]^T, \sigma^2\mathbf{I}).$$

Each received codeword lives with high probability in a sphere (because of the gaussian noise) around the sent vector, $\mathbf{X}$; the radius of the sphere is $\sqrt{n\sigma^2}$. Moreover, the norm of any codeword Y is $\sqrt{n(P+\sigma^2)}$, so we expect that every Y will lie in a sphere with radius $\sqrt{n(P+\sigma^2)}$. The number of distinguishable codewords is equal to the number of small spheres that can be packed in the large sphere:

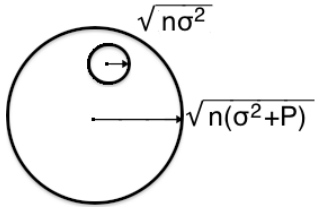


Figure 18.2: Sphere packing.

$$number\ of\ codewords = \frac{c_n\sqrt{n(P+\sigma^2)}^n}{c_n\sqrt{n\sigma^2}^n} = 2^{\frac{n}{2}\log\frac{P+\sigma^2}{\sigma^2}} = 2^{nC},$$

where c_n is the volume of a unit sphere in n dimensions. The maximum achievable rate can be obtained by sending the maximum number of codewords that are distinguishable.

Remark: The rigorous proof, which we will not give here, differs from the discrete case we saw last time in two main points (for details refer to Thomas-Cover Sections 9.1, 9.2):

- random coding: we want the random codewords to satisfy the power constraint with high probability, so choose $X_i(w) \sim \mathcal{N}(0, P-\epsilon)$, where $i = 1, \dots, n$ and $w = 1, \dots, 2^{nR}$.
- jointly typical decoding: when considering the probability of error we take into account both the case when the transmitted and received codewords are not jointly typical (or a different received codeword is jointly typical with the transmitted codeword) and when the power constraint is violated. Also, all the arguments for joint typicality we saw in the discrete case hold true for continuous random variables if we replace the size of the typical sets with volume.

18.3 Continuous Time Channel

In this section, we consider continuous-time channels which are band-limited to $[-W, W]$. According to the Nyquist theorem, sampling a bandlimited signal at a rate $2W$ Hz (or one sample every $1/(2W)$ sec) is sufficient to reconstruct the signal from the samples. Thus, the *Nyquist rate* is $2W$ samples per second. We will use this result to understand the capacity of the continuous channel in terms of the corresponding discrete-time channel obtained by sampling at Nyquist rate.

Recall that the constraint is the discrete-time channel discussed above is really an average energy constraint on the codeword (though we loosely call it power constraint). For continuous time signals, their power is measured in Watts and

$$Power\ P\ [Watts] = \frac{energy}{time} = \frac{energy}{sec}.$$

This translates into average energy of $P/2W$ per sample. Moreover, the noise power is typically specified in terms of Power Spectrum Density (PSD), say $\frac{N_0}{2}\frac{Watts}{Hz}$ which translates to average noise energy of $\frac{N_0}{2}$ per sample. We can now write the capacity of continuous time channel, with average input signal energy of $P/2W$ per sample and average noise energy of $\frac{N_0}{2}$ per sample, as

$$Capacity = \frac{1}{2}\log\left(1+\frac{\frac{P}{2W}}{\frac{N_0}{2}}\right) = \frac{1}{2}\log\left(1+\frac{P}{N_0W}\right) bits/sample$$

$$= W\log\left(1+\frac{P}{N_0W}\right) bits/sec.$$

18.4 Parallel Gaussian Channels

Now we consider k independent discrete-time Gaussian channels in parallel, with a common power constraint $\sum_{j=1}^{k} P_j \le P$. How can we distribute the total power over all channels so that we maximize the capacity?

For channel j, $Y_j = X_j + Z_j$, where $Z_j \sim \mathcal{N}(0,\sigma_j^2)$ is the noise. We assume that the noise is independent from channel to channel. The information capacity of the channel is

$$\begin{aligned} C &= \max_{p(x_1,\ldots,x_n):\sum_{j=1}^{k} P_j \le P} I(X_1,\ldots,X_k\ ;\ Y_1,\ldots,Y_k) \\ &= \max_{\{P_j\} s.t. \sum P_j \le P} \sum_{j=1}^{k} \frac{1}{2}\log\left(1+\frac{P_j}{\sigma_j^2}\right) \end{aligned}$$

which is achieved if $(X_1,\ldots,X_k) \sim \mathcal{N}(0, diag(P_1,\ldots,P_k))$.

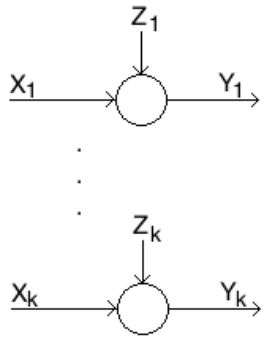


Figure 18.3: Parallel Gaussian channels.

The power allotment problem reduces to a standard optimization problem which can be solved using Lagrange multipliers

$$J(P_1,\ldots,P_k) = \sum_{j=1}^{k} \frac{1}{2}\log\left(1+\frac{P_j}{\sigma_j^2}\right) + \lambda\left(\sum_{j=1}^{k} P_j - P\right).$$

We take the derivative w.r.t. P_j

$$\frac{\partial J}{\partial P_j} = \frac{1}{2}\frac{1}{\sigma_j^2+P_j} + \lambda$$

and by setting it to 0 and using the Karush-Kuhn-Tucker conditions, we find the solution is of the form

$$P_j = (U-\sigma_j^2)_+$$

where U is chosen s.t. $\sum_j P_j \le P$. The above notation means that if we allocate power to channel j, then its power will be $P_j = U - \sigma_j^2$; otherwise, $P_j = 0$. The solution has an interesting form which is related to the so-called *water-filling* process (Fig. 18.4). We start allocating power to the channel with smallest noise variance, and as the power budget increases, we fill up the channels in order of increasing noise variances. At any point, the sum of noise variance and signal power in the channels being used is constant (equal to U).

This case also describes time-varying channels since instead of parallel channels with different noise variances, one may view these as sequential transmission through a time-varying channel.

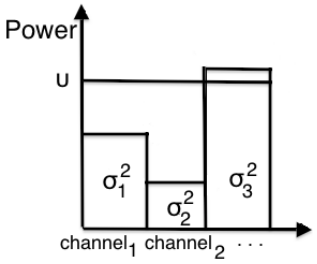


Figure 18.4: Water-filling process for parallel channels.

18.5 Gaussian channels with dependent noise

Now we assume that the noise in the channels is correlated, $Z_i \sim \mathcal{N}(0, \Sigma_Z)$, where Σ_Z can be decomposed to $U\Lambda U^T$. This case describes also channels with memory since instead of parallel correlated channels, one may view these as sequential tranmissions with correlations with previous inputs aka "memory". In this case, the optimal power allocation is obtained by water-filling in the spectral domain.

We can alterantively think of the channel as

$$U^T Y_i = U^T X_i + U^T Z_i$$

where $U^T Z_i \sim \mathcal{N}(0, \Lambda)$. This is akin to the parallel independent Gaussian channels case discussed in previous section, and capacity is achieved if $U^T X_i \sim \mathcal{N}(0, \begin{pmatrix} P_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & P_k \end{pmatrix})$, where P_j are the solutions described by water-filling as before, except that now the water-filling is done in the spectral domain (eigendomain). Equivalently, $X_i \sim \mathcal{N}(0, U \begin{pmatrix} P_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & P_k \end{pmatrix} U^T)$, where $\Sigma_X = U \begin{pmatrix} P_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & P_k \end{pmatrix} U^T$ is the power matrix. The capacity C can be written as

$$C = \max_{tr(\Sigma_X) \le nP} \frac{1}{2} \log \frac{|\Sigma_X + \Sigma_Z|}{|\Sigma_Z|}.$$

Remark: If we also exploit feedback, the capacity becomes (see Thomas-Cover Sec 9.6)

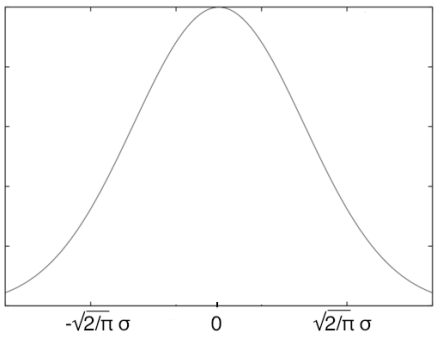
$$C_{FB} = \max_{tr(\Sigma_X) \le nP} \frac{1}{2} \log \frac{|\Sigma_{X+Z}|}{|\Sigma_Z|}.$$

which can be shown to be larger than capacity without feedback. Thus, in channels with *memory*, feedback *can* help. Recall that this does not hold in *memoryless* cases.

18.6 Rate Distortion Theory (lossy compression)

After considering continuous alphabet channels, lets also talk about continuous alphabet source coding. So far we have only talked about discrete alphabet lossless source coding. Observe that we cannot hope to represent a continuous alphabet (real value) without any loss or distortion. Even in discrete case, we might wonder how much distortion we must suffer if we compress the source to fewer than entropy number of bits. Thus, we talk about lossy compression, also known as rate distortion theoy.

Lets start with a simple example. Consider $X \sim \mathcal{N}(0, \sigma^2)$. If we have 1 bit for the encoding, then we can send the conditional mean of the positive or negative region so as to distinguish whether X is positive or negative. This one-bit quantization process of a Gaussian random variable is depicted in Fig. 18.6. If we are allowed more than 1 bit, we can map source symbols to the nearest representation where the representation points are chosen to minimize conditional expected distortion over the regions that are mapped to them. This idea of dividing the input domain into different regions and allocating them to bits is common and known with different names, such as vector quantization, and Lloyd's algorithm or k-means.



Now we are interested in compressing the input more than we are "allowed" (for lossless compression) and recovering the signal up to a certain distortion. We will study distortion functions under the following setting:

$$\text{src } X \to \text{encoder} \to_{<<H(X)} \text{decoder} \to \hat{X}$$

Definition 18.3 (Distortion Function) *A distortion function, $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by $\hat{x}$.*

Examples of distortion functions are:

- Squared error distortion: $(x - \hat{x})^2$
- Hamming distortion: $1_{x \neq \hat{x}}$.

Definition 18.4 *Distortion between sequences x^n and $\hat{x}^n$ is defined as $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$.*

Definition 18.5 (Rate distortion code) *A $(2^{nR}, n)$-rate distortion code consists of an encoder*

$$f_n : X^n \to \{1, 2, \ldots, 2^{nR}\}$$

and a decoder

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{X}^n.$$

The distortion of this code is $E_X[d(X^n, g_n(f_n(X^n))]$.

A rate distortion pair (R, D) is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$ rate disortion code (f_n, g_n) with $\lim_{n\to\infty} E[d(X^n, g_n(f_n(X^n))] \leq D$.

Definition 18.6 *The rate distortion function $R(D)$ is defined as: $R(D) = \inf(R : (R, D)$ pair is achievable$)$.*

An example of such a pair is $(\geq H(x), 0)$.

Definition 18.7 *The information rate distortion function is defined by:*

$$R^{(I)}(D) = \min_{p(\hat{x}|x): E[d(X,\hat{X})] \leq D} I(X; \hat{X}).$$

If $D = 0$ i.e. lossless compression is desired, the information rate distortion function would be the entropy, since in this case $\hat{X} = X$ and $I(X, X) = H(X)$ (it is not possible to compress X with less bits than the entropy without loss).

Definition 18.8 (Rate distortion theorem) *The rate distortion function $R(D)$, the minimum achievable rate at distortion D, for an iid source with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to $R^{(I)}(D)$.*

Before we discuss the proof sketch, lets look at some examples of information rate distortion functions.

Example 1: Bernoulli(p) source with $p < 1/2$ and Hamming distortion

We want to find the rate distortion function.

$$\begin{aligned} I(X, \hat{X}) &= H(X) - H(X|\hat{X}) \\ &= H(Ber(p)) - H(X \oplus \hat{X}|\hat{X}) \\ &\geq H(Ber(p)) - H(X \oplus \hat{X}) \\ &\geq H(Ber(p)) - H(Ber(D)) \end{aligned}$$

Note that in the second equality we used the fact that introducing an invertible function of the conditioning variable does not change the entropy. The inequality in the third step follows by follows since conditioning does not increase entropy, and the last step follows since $X \oplus \hat{X}$ is a Bernoulli random variable with expected hamming distortion $Pr(X \oplus \hat{X} = 1) \leq D$.

Now we want to see if we can design $\hat{X}$ so that the lower bound that we found above is achievable; if we can design it, then $H(Ber(p)) - H(Ber(D))$ will be the information rate distortion function. First consider the case $D \leq p \leq 1/2$. The binary symmetric channel in Fig. 18.6 is a simple construction which gives $H(X|\hat{X}) = H(Ber(D))$ while $Pr(X \neq \hat{X}) = D$, and, thus, achieves the lower bound. The idea behind the

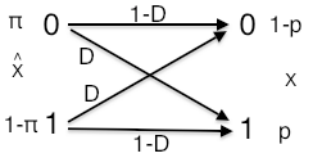


Figure 18.5: Joint distribution for binary source.

design is to think of $\hat{X}$ as the input of the channel and X as the output, and define π so that the output X appears to be generated by $Bernoulli(p)$. To find such a π, lets consider $Pr(X = 1)$:

$$p = Pr(X = 1) = \pi D + (1 - \pi)(1 - D)$$

which yields $\pi = \frac{1-p-D}{1-2D}$. Thus, the information rate distortion function is $H(Ber(p)) - H(Ber(D))$ if $D \leq p \leq 1/2$. For $D > p$, i.e. we are allowed distortion larger than p, then if we don't encode X at all, i.e. $R = 0$, we can achieve distortion D trivially.

Thus, the rate distortion function for the Bernoulli(p) source is:

$$R(D) = \begin{cases} H(Ber(p)) - H(Ber(D)), & \text{if } D \leq \min(p, 1-p) \\ 0, & \text{if } D > \min(p, 1-p) \end{cases}$$

This holds for all p, though we only worked it out for $p \leq 1/2$.

Example 2: Gaussian source $\mathcal{N}(0, \sigma^2)$ and squared-error distortion

As before, we first find a lower bound for the rate distortion function

$$\begin{aligned} I(X, \hat{X}) &= H(X) - H(X|\hat{X}) \\ &\geq \frac{1}{2}\log(2\pi e)\sigma^2 - H(X - \hat{X}) \\ &\geq \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D \\ &= \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned}$$

where $\sigma^2 \geq D$. Note that we used the fact that conditioning reduces entropy (1st inequality), and the maximum entropy under the second moment constraint $E[(X - \hat{X})^2] \leq D$ is achieved by the Gaussian distribution (2nd inequality). The lower bound can be achieved by the simple construction in Fig. 18.6 with $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$ and $Z \perp \hat{X}$. It is easy to verify that $H(X|\hat{X}) = \frac{1}{2}\log(2\pi e)D$ and $E[(X - \hat{X})^2] = D$.

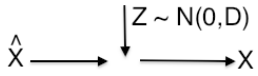


Figure 18.6: Joint distribution for Gaussian source.

As before if $\sigma^2 < D$, i.e. we are allowed mean square distortion larger than σ^2, then if we don't encode X at all, i.e. $R = 0$, we can achieve mean square distortion D trivially.

The rate distortion function is

$$R^{(I)}(D) = \min_{p(\hat{x}|x):E[d(X,\hat{X})]\leq D} I(X;\hat{X}) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & \text{if } D \leq \sigma^2 \\ 0, & \text{if } D > \sigma^2 \end{cases}$$

Equivalently, the distortion we can get for a given a rate is $D(R) = \sigma^2 2^{-2R}$, i.e. each bit of description reduces expected distortion by a factor of 4. If we are allowed to encode with 1 bit, then the best distortion we can get is $D(R) = \frac{\sigma^2}{4}$. Vector quantization or Lloyd's method give expected distortion to the centers $\frac{\pi-2}{\pi}\sigma^2 = 0.36\sigma^2$, which is not optimal. The best rate distortion limit is achieved by encoding a stream of symbols instead of one symbol at a time (even if the symbols are independent!).