

## Summary

So far, we have discussed only discrete channels; we shall now have a look at *continuous* channels.

If the discrete case is mastered, it is not difficult to handle the analogous definitions and results of the continuous case. For example, sums are replaced by integrals, etc. There are some important differences, however, and some care is needed. Calculations, for example, are generally more involved in the continuous case.

*Entropy* in the discrete case corresponds to *differential entropy* in the continuous case.

## Differential Entropy (2)

Two remarks:

1. As in the discrete case, the differential entropy depends only on the probability function of the random variable. Hence the differential entropy is sometimes written  $h(f)$  instead of  $h(X)$ .
2. Note that the differential entropy is defined only if the aforementioned integral and density function exist!

## Differential Entropy (1)

The **cumulative distribution function** of a random variable  $X$  is  $F(x) = \Pr(X \leq x)$ . If  $F(x)$  is continuous, the random variable is said to be continuous. Let  $f(x) = F'(x)$  when the derivative is defined. If  $\int_{-\infty}^{\infty} f(x) = 1$ , then  $f(x)$  is called a **probability density function** for  $X$ . The set where  $f(x) > 0$  is called the *support set* of  $X$ .

The **differential entropy**  $h(X)$  of a continuous random variable  $X$  with density  $f(x)$  is defined as

$$h(X) = - \int_S f(x) \log f(x) dx,$$

where  $S$  is the support set of the random variable.

## Example: Uniform Distribution

We consider a random variable that is uniformly distributed from 0 to  $a$ , so  $\int_0^a f(x) = 1$  gives that  $f(x) = 1/a$  when  $0 \leq x \leq a$  and 0 otherwise. The differential entropy is

$$h(X) = - \int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a.$$

**Note:** If  $a < 1$  in this example, then  $h(X) < 0$ . Unlike discrete entropy, differential entropy can be negative!

### Example: Normal Distribution

For  $X \sim \phi(x) = (1/\sqrt{2\pi\sigma^2}) \times e^{-x^2/2\sigma^2}$ , the entropy in *nats* is

$$\begin{aligned} h(\phi) &= - \int \phi(x) \ln \phi(x) dx = - \int \phi(x) \left( -\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right) dx \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2 = \frac{1}{2} + \frac{1}{2} \ln 2\pi\sigma^2 \\ &= \frac{1}{2} \ln e + \frac{1}{2} \ln 2\pi\sigma^2 = \frac{1}{2} \ln 2\pi e\sigma^2. \end{aligned}$$

Changing the base, we get

$$h(\phi) = \frac{1}{2} \log 2\pi e\sigma^2 \text{ bits.}$$

©Patric Östergård

### Typical Sets for Continuous Variables

For  $\epsilon > 0$  and any positive integer  $n$ , the **typical set**  $A_\epsilon^{(n)}$  with respect to  $f(x)$  is defined as follows:

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \dots, x_n) : \left| -\frac{1}{n} \log f(x_1, x_2, \dots, x_n) - h(X) \right| \leq \epsilon \right\},$$

where  $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$ .

©Patric Östergård

### The AEP for Continuous Random Variables

For a sequence of i.i.d. discrete random variables, the AEP states that  $p(x_1, x_2, \dots, x_n)$  is close to  $2^{-nH(X)}$  with high probability. This enables us to define a *typical set*. For the continuous case, we have analogous definitions and results.

**Theorem 9.2.1.** Let  $X_1, X_2, \dots, X_n$  be a sequence of random variables drawn i.i.d. according to the density  $f(x)$ . Then

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow h(X) \text{ in probability.}$$

©Patric Östergård

### Differential Entropy vs. Discrete Entropy (1)

We consider a random variable  $X$  with density  $f(x)$ , and divide the range of  $X$  into intervals  $I_i$  of length  $\Delta$ ; see [Cov, Fig. 9.1]. We define the quantized random variable  $X^\Delta$  by

$$X^\Delta := x_i,$$

where  $x_i \in I_i$  and

$$f(x_i)\Delta = \int_{I_i} f(x) dx.$$

©Patric Östergård

## Differential Entropy vs. Discrete Entropy (2)

**Theorem 9.3.1.** If the density  $f(x)$  of the random variable  $X$  is Riemann integrable, then

$$H(X^\Delta) + \log \Delta \rightarrow h(X) \text{ as } \Delta \rightarrow 0.$$

Thus the entropy of an  $n$ -bit quantization of a continuous random variable  $X$  is approximately  $h(X) + n$ .

**Example 1.** If  $X$  has a uniform distribution on  $[0, 1]$  and we let  $\Delta = 2^{-n}$ , then  $h(X) = 0$  and  $\log \Delta = -n$ , so  $H(X^\Delta) = n$  and  $n$  bits suffice to describe  $X$  to  $n$  bit accuracy.

©Patric Östergård

## Joint and Conditional Differential Entropy

The **joint differential entropy**  $h(X_1, X_2, \dots, X_n)$  of a set  $X_1, X_2, \dots, X_n$  of random variables with density  $f(x_1, x_2, \dots, x_n)$  is defined as

$$-\int f(x_1, x_2, \dots, x_n) \log f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n.$$

The **conditional differential entropy** is defined as

$$h(X|Y) = -\int f(x, y) \log f(x|y) dx dy.$$

©Patric Östergård

## Differential Entropy vs. Discrete Entropy (3)

**Example 2.** If  $X$  is uniformly distributed on  $[0, \frac{1}{8}]$  and  $\Delta = 2^{-n}$ , then  $h(X) = -3$  and  $\log \Delta = -n$ . Now we get that  $H(X^\Delta) = n - 3$ . Indeed, to describe  $X$  to  $n$  bit accuracy requires only  $n - 3$  bits, as the first 3 bits to the right of the decimal point must be 0.

In these examples, every value of  $X$  requires the same number of bits. In general, however,  $h(X) + n$  is the number of bits *on the average* required to describe  $X$  to  $n$  bit accuracy.

The differential entropy of a discrete random variable can be considered to be  $-\infty$ .

©Patric Östergård

## Relative Entropy and Mutual Information

The **relative entropy** between two densities  $f$  and  $g$  is defined by

$$D(f||g) = \int f \log \frac{f}{g}.$$

The relative entropy is finite only if the support set of  $f$  is contained in the support set of  $g$ .

The **mutual information** between two random variables is defined by

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy.$$

©Patric Östergård

## Mutual Information

From the definition of mutual information it is clear that

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

and

$$I(X; Y) = D(f(x, y) \| f(x)f(y)).$$

The mutual information between two random variables is the limit of the mutual information between their quantized versions, as

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

© Patric Östergård

## The Gaussian Channel (1)

The most important continuous alphabet channel is the **Gaussian channel**, which is depicted in [Cov, Fig. 10.1]. This is a *time discrete* channel with output  $Y_i$  at time  $i$ , where  $Y_i$  is the sum of the input  $X_i$  and the noise  $Z_i$ . The noise is drawn i.i.d. from a Gaussian distribution with variance  $N$ , and  $Z_i$  is therefore independent of the input  $X_i$ . Summing up,

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, N).$$

© Patric Östergård

## Properties of Differential Entropy

All but the last of the following properties are analogous to the those in the discrete case.

- $D(f \| g) \geq 0$ .
- $I(X; Y) \geq 0$  with equality iff  $X$  and  $Y$  are independent.
- $h(X|Y) \leq h(X)$  with equality iff  $X$  and  $Y$  are independent.
- $h(X_1, X_2, \dots, X_n) = \sum_{i=1}^n h(X_i | X_{i-1}, X_{i-2}, \dots, X_1)$ .
- $h(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n h(X_i)$ .
- $h(X + c) = h(X)$ .
- $h(aX) = h(X) + \log |a|$ .

© Patric Östergård

## The Gaussian Channel (2)

Without further conditions, the capacity of the Gaussian channel may be infinite:

- If the noise variance  $N$  is zero,  $Y_i = X_i$ . Since  $X_i$  can take any real value, the channel can then transmit an arbitrary real number with no error.  $\Rightarrow$  We must have  $N > 0$ .
- If there is no constraint in the input, we can choose an infinite subset of inputs arbitrarily far apart so that they are distinguishable at the output with arbitrarily small probability of error, and we get infinite capacity.  $\Rightarrow$  We must have a (usually, energy or power) constraint on the input.

© Patric Östergård

### The Gaussian Channel (3)

With an average power constraint, we require for any codeword  $(x_1, x_2, \dots, x_n)$  transmitted over the channel that

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

This communication channel models many practical channels, including radio and satellite links.

We shall now look at the simplest (and suboptimal) case when only one bit is transmitted over the channel. Given the power constraint, we have  $-\sqrt{P} \leq X \leq \sqrt{P}$ , so the best we can do is obviously to use the values  $-\sqrt{P}$  and  $\sqrt{P}$ .

©Patric Östergård

### The Gaussian Channel (5)

▷ The previous scheme converts the Gaussian channel into a discrete BSC with crossover probability  $P_e$ .

By using more than two levels in an analogous way, other discrete channels can be obtained. Similar ideas are used in some practical modulation schemes.

*Advantage:* Ease of processing of the output signal for error correction.

*Drawback:* The capacity of the channel is not reached.

©Patric Östergård

### The Gaussian Channel (4)

If both values are equally likely, the optimum decoding rule is to check whether  $Y > 0$  (then,  $\hat{X} = \sqrt{P}$ ) or  $Y < 0$  (then,  $\hat{X} = -\sqrt{P}$ ).

The probability of error with this decoding scheme is

$$\begin{aligned} P_e &= \frac{1}{2} \Pr(Y < 0 | X = \sqrt{P}) + \frac{1}{2} \Pr(Y > 0 | X = -\sqrt{P}) \\ &= \frac{1}{2} \Pr(Z < -\sqrt{P} | X = \sqrt{P}) + \frac{1}{2} \Pr(Z > \sqrt{P} | X = -\sqrt{P}) \\ &= \Pr(Z > \sqrt{P}) = 1 - \Phi\left(\sqrt{\frac{P}{N}}\right), \end{aligned}$$

where  $\Phi(x)$  is the cumulative normal function

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

©Patric Östergård

### Information Capacity for Gaussian Channels (1)

The **information capacity** of the Gaussian channel with power constraint  $P$  is

$$C = \max_{p(x): EX^2 \leq P} I(X; Y).$$

By expanding  $I(X; Y)$ , we get

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

since  $Z$  is independent of  $X$ .

©Patric Östergård

### Information Capacity for Gaussian Channels (2)

We get

$$EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2 \leq P + N,$$

since  $X$  and  $Z$  are independent and  $EZ = 0$ . By [Cov, Theorem 9.6.5] (omitted here), we have that

$$h(Y) \leq \frac{1}{2} \log 2\pi e(EY^2) \leq \frac{1}{2} \log 2\pi e(P + N).$$

©Patric Östergård

### Coding Theorem for Gaussian Channels

By using arguments analogous to those of the discrete case, the channel coding theorem for the Gaussian channel can be proved; see [Cov, pp. 242–247]. A rate  $R$  is said to be *achievable* for a Gaussian channel with a power constraint  $P$  if there exists a sequence of  $(n, 2^{nR})$  codes satisfying the power constraint such that the maximal probability of error tends to zero. The **capacity** of the channel is the supremum of the achievable rates.

**Theorem 10.1.1.** The capacity of the Gaussian channel with power constraint  $P$  and noise variance  $N$  is

$$C = \frac{1}{2} \log \left( 1 + \frac{P}{N} \right) \text{ bits per transmission.}$$

©Patric Östergård

### Information Capacity for Gaussian Channels (3)

Since we know  $h(Z)$  from an earlier example, we are now able to finish the calculations.

$$\begin{aligned} I(X; Y) &= h(Y) - h(Z) \leq \frac{1}{2} \log 2\pi e(P + N) - \frac{1}{2} \log 2\pi eN \\ &= \frac{1}{2} \log \left( 1 + \frac{P}{N} \right). \end{aligned}$$

Hence the information capacity of the Gaussian channel is

$$C = \max_{EX^2 \leq P} I(X; Y) = \frac{1}{2} \log \left( 1 + \frac{P}{N} \right),$$

and the maximum is attained when  $X \sim \mathcal{N}(0, P)$ .

©Patric Östergård