Kim Boström, July 2007

Random Variables

1 Basics

A *random variable* is an abstract conception. Let us endow random variables with a hat, as in $\hat{x}$, to distinguish them from ordinary variables. A continuous real-valued random variable $\hat{x}$ takes real values $x \in \mathbb{R}$ which are called the *realizations* of $\hat{x}$. The random variable $\hat{x}$ is then completely defined by a nonnegative probability distribution $p : \mathbb{R} \to [0, \infty)$ with $\int dx\, p(x) = 1$ (completeness). The probability distribution is also called the *probability density function* (pdf). The subset $\mathcal{X} \subset \mathbb{R}$ where $p(x) > 0$ is the set of possible realizations of $\hat{x}$.

The probability distribution determines the probability of *events*. For continuous random variables, an *event* is a subset $X \subset \mathbb{R}$ so that the probability that $\hat{x}$ takes a value in the set X is given by

$$\mathcal{P}(X) := \int_X dx\, p(x). \tag{1}$$

We mostly refer to an event in a more intuitive way. For example, we refer to an event $X \subset \mathbb{R}$ as "$\hat{x} \in X$" in order to make explicit that we consider the event that the random variable $\hat{x}$ takes a value in X. The probability that $\hat{x}$ takes a value below x is given by the *cumulative probability distribution* (cdf)

$$P(x) := \mathcal{P}(\hat{x} \leq x), \tag{2}$$

which is equal to $\mathcal{P}((-\infty, x])$ or, more explicitely,

$$P(x) = \int_{-\infty}^{x} dx'\, p(x'), \tag{3}$$

which implies

$$P'(x) = p(x). \tag{4}$$

The most important property of a random variable $\hat{x}$ is its *expectation value*, also called the *mean* or the *first moment*,

$$\mu(\hat{x}) := \int dx\, p(x)\, x. \tag{5}$$

Because the mean is of such a central meaning in stochastics, a more convenient notation using brackets is used, namely

$$\langle \hat{x} \rangle := \mu(\hat{x}). \tag{6}$$

Any function $f(\hat{x})$ of a random variable $\hat{x}$ is also a random variable and its expectation value reads

$$\langle f(\hat{x}) \rangle = \int dx\, p(x) f(x). \tag{7}$$

For two real numbers α, β we have

$$\langle \alpha\hat{x} + \beta \rangle = \alpha\langle \hat{x} \rangle + \beta. \tag{8}$$

The second most important property of a random variable $\hat{x}$ is the *variance*

$$\sigma^2(\hat{x}) := \langle (\hat{x} - \langle \hat{x} \rangle)^2 \rangle = \langle \hat{x}^2 \rangle - \langle \hat{x} \rangle^2, \tag{9}$$

which is identical to the *second central moment*, where the n-th moment of $\hat{x}$ about a value c is defined by

$$\mu_{n,c}(\hat{x}) := \langle (\hat{x} - c)^n \rangle. \tag{10}$$

The moments about the mean are called the *central moments*. The square root of the variance is called the *standard deviation*,

$$\sigma(\hat{x}) := \sqrt{\sigma^2(\hat{x})}. \tag{11}$$

For two real numbers α, β we have

$$\sigma^2(\alpha\hat{x} + \beta) = \alpha^2\sigma^2(\hat{x}). \tag{12}$$

A *discrete real-valued random variable* $\hat{x}$ is defined by a countable set $\mathcal{X} = \{x_i\}$ of realizations and a probability distribution $p_i = p(x_i)$, so that the mean of any function $f(\hat{x})$ is given by

$$\langle f(\hat{x}) \rangle := \sum_i p_i\, f(x_i), \tag{13}$$

and the probability to find $\hat{x}$ realized in the set $X \subset \mathcal{X}$ is given by

$$P\{\hat{x} \in X\} := \sum_{x_i \in X} p_i. \tag{14}$$

Particularly, the probability that $\hat{x}$ is realized as x_i reads

$$P\{\hat{x} = x_i\} = p(x_i) = p_i. \tag{15}$$

A discrete random variable is equivalent to a continuous random variable having a δ-peaked probability distribution

$$p(x) = \sum_i \delta(x - x_i) p_i. \tag{16}$$

The use of singular distributions like the δ-function requires special treatment and care.

2 Characteristic function

The characteristic function χ of a continuous random variable $\hat{x}$ is the Fourier transform of its probability function,

$$\chi(k) := \langle e^{ik\hat{x}} \rangle = \int dx\, p(x)\, e^{ikx}. \tag{17}$$

The characteristic function can be used to easily get the moments of a probability distribution. With $\chi^{(n)}(k) \equiv \frac{\partial^n}{\partial k^n}\chi(k)$ being the n-th derivative of the characteristic function, we have

$$\langle \hat{x}^n \rangle = \frac{1}{i^n}\, \chi^{(n)}(0). \tag{18}$$

3 Gaussian variables

A Gaussian random variable $\hat{x}$ has a probability distribution $p(x) = G(x|\mu, \sigma)$ with only two parameters, the mean μ and the variance σ, and it takes the form

$$G(x|\mu, \sigma) := \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{19}$$

The *half-width* η is the distance between those two points on the x-axis where the Gaussian function equals $1/2$, whereas the *Gauss-width* γ is the distance between those two points on the x-axis where the Gaussian function equals $1/\mathrm{e}$. Both values are connected to the standard deviation σ via

$$\eta = (2\sqrt{2\ln 2}) \cdot \sigma \approx 2.3548 \cdot \sigma \tag{20}$$

$$\gamma = (2\sqrt{2}) \cdot \sigma \approx 2.8284 \cdot \sigma. \tag{21}$$

If $\hat{x}$ is Gaussian, then the random variable

$$\hat{z} = \frac{\hat{x} - \mu}{\sigma} \tag{22}$$

is *normally distributed*, that is, it has a Gaussian distribution $p(z) = G(z|0, 1)$ being centered at $\mu_z = 0$ and having a standard deviation of $\sigma_z = 1$,

$$G(z) := G(z|0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{23}$$

Hence, $\hat{z}$ indicates the *distance from the mean in standard deviations*, often referred to as the *z-score*. The transformation between $\hat{x}$ and $\hat{z}$ are accomplished by

$$\hat{z} = \frac{\hat{x} - \mu}{\sigma} \tag{24}$$

$$\text{and} \quad \hat{x} = \sigma\hat{z} + \mu. \tag{25}$$

The Gaussian cumulative distribution function (cdf) is given by

$$\Phi(x|\mu, \sigma) := \int_{-\infty}^{x} dx'\, G(x'|\mu, \sigma), \tag{26}$$

and the *standard normal* cdf by

$$\Phi(x) := \Phi(x|0, 1). \tag{27}$$

Between the Gaussian cdf and the normal cdf there is the useful relationship

$$\Phi(x|\mu,\sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{28}$$

The *error function* $\mathrm{erf}(x)$ is defined as

$$\mathrm{erf}(x) := 2\Phi(\sqrt{2}x) - 1, \tag{29}$$

or explicitely

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x dx'\, e^{-x'^2}, \tag{30}$$

so that

$$\Phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{31}$$

The inverse Gaussian cdf is called the *Quantile function*,

$$Q(p) := \Phi^{-1}(p), \tag{32}$$

where the value $q = Q(p)$ is called the *p-quantile*, and it represents a threshold for $\hat{x}$ so that $\hat{x}$ falls below q with probability $p \in [0,1]$. We have

$$Q(p) = \sqrt{2} \cdot \mathrm{erf}^{-1}(2p-1). \tag{33}$$

4 Entropy and information

The concept of information is subtle and there are many equivalent approaches which not all appear equivalent at first (and maybe second) sight. One should always remember that information does not exist *per se* but is always information *about* something. If someone says "The amount of information encoded in this spike train" without indicating what this piece of information could possibly *tell* us, he is simply talking nonsense. A stream of bits, symbols, amino acids or action potentials does not contain any information in itself. Only if the stream *reduces our ignorance* about a certain quality, we may say that it *conveys* information about this quality. Now, there is a well-known measure of *ignorance* (or *uncertainty*) about a particular random variable $\hat{x}$, the famous *Shannon entropy*. In the continuous case, the Shannon entropy is defined as

$$H(\hat{x}) := -\int dx\, p(x) \log p(x), \tag{34}$$

and has values on the entire real line,

$$H(\hat{x}) \in \mathbb{R}. \tag{35}$$

In the case of a *discrete* random variable, the Shannon entropy is defined as

$$H(\hat{x}) := -\sum_i p_i \log p_i, \tag{36}$$

and is always bigger or equal to zero,

$$H(\hat{x}) \geq 0, \tag{37}$$

with $H = 0$ exactly if $p_j = 1$ for a certain j and $p_i = 0$ for $i \neq j$. Thus, we can interpret the situation $H = 0$ as if our ignorance with respect to the value of $\hat{x}$ is zero because we know that with certainty $\hat{x}$ is realized as x_j.

Unfortunately, such straightforward interpretation is not possible for the continuous case, because the Shannon entropy of a continuous random variable can also be negative. In the extreme case where we know with certainty that $\hat{x}$ is realized as some real number x_0, then the corresponding probability distribution reads $p(x) = \delta(x - x_0)$. The Shannon entropy of such a distribution equals minus infinity,

$$\begin{aligned} H(\hat{x}) &= -\int dx\, p(x) \log p(x) && (38)\\ &= -\int dx\, \delta(x - x_0) \log \delta(x - x_0) && (39)\\ &= -\log \delta(0) = -\infty. && (40) \end{aligned}$$

A possible approach to the concept of information content is to consider the *difference in entropy* before and after a piece of information about $\hat{x}$ has been received. Such "piece of information" can be a measurement, or some additional insights that have not been recognized before, or somebody telling us something about $\hat{x}$, or whatever. Let us call this piece of information simply a *message*. Before receipt of the message, the unknown value is described by the random variable $\hat{x}$, after the receipt, the unknown value is described by the random variable $\hat{x}'$. The probability distribution changes from $p(x)$ to $p'(x)$ (where the prime does not indicate differentiation!). The differential entropy between $\hat{x}$ and $\hat{x}'$ reads

$$\Delta H = H(\hat{x}) - H(\hat{x}'). \tag{41}$$

If $\Delta H > 0$ then we interpret this as our knowledge about the unknown value having been *increased* by the piece of information, in other words, we have *gained information*. Similarly, if $\Delta H < 0$ then we have *lost information*. This interpretation is independent of H being positive or negative. Furthermore, the interpretation also goes for discrete random variables. Hence, the differential entropy is a much better measure for the amount of information than the Shannon entropy alone. It does not refer to the random variable $\hat{x}$ but rather to a *message about* $\hat{x}$. The information content of the message with respect to the unknown value $\hat{x}$ equals the differential entropy ΔH.

It should be noted that there is a useful and meaningful measure of the information content of a message *without* reference to an external random variable. It is the *maximal amount of*

information that a message may convey. This maximal amount is connected to the *number of possible states* the message can be in. Let $M = \{m_i\}$ be the finite set of possible message states, then

$$I = \log |M| \tag{42}$$

is the maximal amount of information the message can carry. This concept of information content is realized as the *file size* on a computer. A record of N zeroes and ones can be in one out of 2^N possible states, thus, the file size of the record is $I = \log(2^N) = N$ bits.

If the message is in a continuum of possible states, then one may replace the size with the *volume* of the set of all possible realizations of $\hat{x}$,

$$I = \log \left[\int\limits_{p(x)>0} dx \right]. \tag{43}$$

In our example, a spike train does not contain information *per se*, it rather *conveys* information about something else, for example an external stimulus or an internal brain state. Moreover, since spike trains are represented by *continuous* random variables (because the instantaneous spike rates take on a continuum of possible values), the Shannon entropy of the spike train is no meaningful measure for the information content anyway. Also, since the set of possible spike trains is infinite, there is not even a meaningful amount of maximal information a spike train may convey. Transforming the continuous variable into a discrete one by particular "time binning" of the spike train is strongly dependent on the binning and therefore highly artificial. Also, it neglects information possibly conveyed in the interspike interval and in the graded potential.

Examples Consider the random variable $\hat{x}$ with

$$p(x) = \begin{cases} \frac{1}{2} & -a \leq x \leq a \\ 0 & \text{else} \end{cases} \tag{44}$$

where a is some positive real number. The entropy reads

$$\begin{aligned} H(\hat{x}) &= -\int dx\, p(x) \log p(x) && (45)\\ &= -\int_{-a}^{a} dx\, \frac{1}{2} \log \frac{1}{2} = \frac{1}{2} \int_{-a}^{a} dx && (46)\\ &= \frac{1}{2} \big[x\big]_{-a}^{a} = \frac{1}{2}\{a + a\} && (47)\\ &= a. && (48) \end{aligned}$$

Thus, the entropy grows with the length of the interval where $\hat{x}$ is realized. There is a straightforward interpretation: The uncertainty about the value of $\hat{x}$ grows when the range of possible values becomes larger.

***

Intuitively, we would expect a message about $\hat{x}$ to always *increase* our knowledge. However, there can be also *negative* values of ΔH which would mean that the message contains "negative information" about $\hat{x}$, although we know more about $\hat{x}$ than before. Impossible? Consider the following example: "The key is either in my pocket or somewhere in the room." Given that there are, say, 128 places in the room where the key can be, the probability distribution reads

$$p_1 = 0.5 \tag{49}$$

$$p_2 = \ldots = p_{129} = \frac{0.5}{128} = \frac{1}{256}, \tag{50}$$

whose entropy is

$$H = -\frac{1}{2}\log\frac{1}{2} - \sum_{i=1}^{128}\frac{1}{256}\log\frac{1}{256} \tag{51}$$

$$= \frac{1}{2} + \frac{1}{2}\log(256) = 4.5 \tag{52}$$

bits. Now I check my pocket to see that *the key is not there*. This piece of information updates my probability distribution to

$$p_1' = 0 \tag{53}$$

$$p_2' = \ldots = p_{129}' = \frac{1}{128}, \tag{54}$$

whose entropy reads

$$H' = \sum_{i=1}^{128}\frac{1}{128}\log\frac{1}{128} = \log(128) = 7 \tag{55}$$

bits. Therefore, the message "the key is not in my pocket" contained an amount of

$$\Delta H = H - H' = -2.5 \tag{56}$$

bits, which is *negative*! It definitely was a *bad* message. My knowledge about the true location of the key has become poorer. Before, there was a 50% chance that the key is in my pocket, which was a quite comfortable situation. Now, I lost this possibility and have to look in roughly half of the 128 places in the room before I find the key.

***

How much information do I obtain when I learn the exact value of a continuous random variable $\hat{x}$? Initially, my amount of ignorance equals the usual Shannon entropy

$$H = -\int dx\, p(x)\log p(x). \tag{57}$$

After getting to know the exact value of $\hat{x}$, say $\hat{x} = x_0$, the probability distribution is updated to

$$p'(x) = \delta(x - x_0), \tag{58}$$

whose Shannon entropy equals minus infinity,

$$H' = -\int dx\, \delta(x - x_0) \log \delta(x - x_0) = -\log \delta(0) = -\infty. \tag{59}$$

Thus, the differential entropy, whatever my knowledge about the initial random variable $\hat{x}$, reads

$$\Delta H = H - H' = H + \infty = \infty. \tag{60}$$

Consequently, getting to know the *exact value* of a random variable requires an infinite amount of information to be transmitted! This is insofar plausible as a real number has infinitely many digits. Knowing the exact value of $\hat{x}$ means knowing all of these infinitely many digits, which requires an infinite amount of information to be transmitted and stored.

But then, how come I ever get to know the exact value of a continuous random variable? Obviously never, because my brain is not infinitely large. But are there really real numbers in the real world out there *at all*? Does any of the physical systems out there *really have an exact state*, which is only one out of a continuum of possible states? And if so, then I as well as any other finite information-processing system will never, not even in principle, have the ability to get to know this exact physical state. And do we have the *chance* at least? The probability to find a continuous random variable having the exact value, say, x_0 is *zero*,

$$P\{\hat{x} = x_0\} = \lim_{\epsilon \to 0} \int_{x_0 - \epsilon}^{x_0 + \epsilon} dx\, p(x) = 0, \tag{61}$$

for any probability distribution that is non-singular at x_0. Thus it will *never* happen that the random variable takes the exact value x_0. I can repeat this argument for arbitrary x_0, so I have just shown that $\hat{x}$ cannot take *any* value at all, right?

Altogether, continuous random variables are pretty much different from ordinary numbers and have to be handled with care.

5 Functions of random variables

5.1 Probability distribution

The function of a random variable is also a random variable. Consider the continuous real-valued random variable $\hat{x}$ and some function $f : \mathbb{R} \to \mathbb{R}$ then the random variable

$$\hat{y} = f(\hat{x}) \tag{62}$$

has the probability distribution

$$q(y) = \int dx\, \delta(y - f(x))p(x). \tag{63}$$

This is reasonable because then we have for any other function $g : \mathbb{R} \to \mathbb{R}$

$$\begin{aligned}
\langle g(\hat{y})\rangle &= \int dy\, q(y)g(y) && (64)\\
&= \int dx\, dy\, \delta(y - f(x))p(x)g(y) && (65)\\
&= \int dx\, p(x)g(f(x)) && (66)\\
&= \langle g(f(\hat{x}))\rangle, && (67)
\end{aligned}$$

as desired.

Examples Consider the random variable $\hat{y} = \hat{x}^2$ where $\hat{x}$ has the probability distribution (44). Then $\hat{y}$ is distributed by

$$\begin{aligned}
q(y) &= \int dx\, \delta(y - f(x))p(x) && (68)\\
&= \int_{-a}^{a} dx\, \frac{1}{2}\delta(y - x^2) && (69)\\
&= \int_{-a}^{a} dx\, \frac{1}{|2x|}\{\delta(x - \sqrt{y}) + \delta(x + \sqrt{y})\}p(x) && (70)\\
&= \begin{cases} \frac{1}{2\sqrt{y}}\{p(\sqrt{y}) + p(-\sqrt{y})\} & ; 0 \le y \le a^2 \\ 0 & ; \text{else} \end{cases} && (71)\\
&= \begin{cases} \frac{1}{2\sqrt{y}} & ; 0 \le y \le a^2 \\ 0 & ; \text{else,} \end{cases} && (72)
\end{aligned}$$

where we have used the relation

$$\delta(f(x)) = \frac{1}{|f'(x)|} \sum_{f(x_i)=0} \delta(x - x_i). \tag{73}$$

The mean of $\hat{y} = \hat{x}^2$ yields

$$\begin{aligned}
\langle \hat{y}\rangle &= \int dy\, q(y)y = \int_0^{a^2} dy\, \frac{y}{2\sqrt{y}} = \int_0^{a^2} dy\, \frac{1}{2}\sqrt{y} && (74)\\
&= \frac{1}{2}\Big[\frac{2}{3}y^{\frac{3}{2}}\Big]_0^{a^2} = \frac{1}{3}a^3. && (75)
\end{aligned}$$

5.2 Entropy

Let us calculate the entropy $H(\hat{y})$ for $\hat{y} = f(\hat{x})$,

$$H(\hat{y}) = -\int dy\, q(y) \log q(y) \tag{76}$$

$$= -\int dy\, dx\, p(x)\delta(y - f(x)) \log \left[\int dx'\, p(x')\delta(y - f(x')) \right] \tag{77}$$

$$= -\int dx\, p(x) \log \left[\int dx'\, p(x')\delta(f(x) - f(x')) \right]. \tag{78}$$

In order for this expression to exist, we have to assume that $f'(x) \neq 0$ on the set of realizations of $\hat{x}$, i.e. where $p(x) > 0$, then it follows that f can be inverted on that domain, and then the condition $f(x') = f(x)$ is equivalent to $x' = x$ and hence

$$H(\hat{y}) = -\int dx\, p(x) \log \frac{p(x)}{|f'(x)|} \tag{79}$$

$$= -\int dx\, p(x) \log p(x) + \int dx\, p(x) \log |f'(x)|. \tag{80}$$

Recalling that $\hat{y} = f(\hat{x})$, and that the first term above equals the Shannon entropy of $\hat{x}$, and that the second term has the form of an expectation value of the random variable $|f'(\hat{x})|$, we arrive at the compact formula

$$H(f(\hat{x})) = H(\hat{x}) + \langle \log |f'(\hat{x})| \rangle. \tag{81}$$

Examples Consider the trivial function $f(\hat{x}) = \hat{x}$, then we expect that the random variable $\hat{y} = \hat{x}$ has the same entropy as $\hat{x}$. Indeed, using (81) we find

$$H(\hat{y}) = H(\hat{x}) + \langle \log |f'(\hat{x})| \rangle \tag{82}$$

$$= H(\hat{x}) + \langle \log 1 \rangle = H(\hat{x}). \tag{83}$$

Next, consider the quadratic function $f(\hat{x}) = \hat{x}^2$, with $\hat{x}$ being defined by the probability distribution (44). Using (81), the entropy of $\hat{y} = \hat{x}^2$ yields

$$H(\hat{x}^2) = H(\hat{x}) + \langle \log |2\hat{x}| \rangle \tag{84}$$

$$= a + \int_{-a}^{a} dx\, \frac{1}{2} \log |2x| \tag{85}$$

$$= a + 2\int_{0}^{a} dx\, \log x \tag{86}$$

$$= a + 2 \cdot \big[x(\log x - 1) \big]_0^a \tag{87}$$

$$= a + 2a(\log a - 1). \tag{88}$$

Interestingly, if $a < 2$ then the entropy of $\hat{x}^2$ is *smaller* than the entropy of $\hat{x}$ because then the factor $(\log a - 1)$ above is negative. Straightforward interpretation: The interval

$[-a, a]$ of possible values for $\hat{x}$ is *compressed* to the interval $[0, a^2]$ of possible values for $\hat{x}^2$, hence, the uncertainty about the value of $\hat{x}^2$ is smaller than the uncertainty about the value of $\hat{x}$. Similarly, for $a > 2$ the interval is *extended* to $[0, a^2]$ so that the uncertainty is increased.

6 Correlation

In order to understand how information is conveyed by one variable about another variable, we need the notion of *correlation*.

A pair of continuous real-valued random variables $\hat{x}, \hat{y}$ take values in $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. Since the two random variables are not necessarily independent from each other, the realization of both variables is conceived as a *joint event* whose probability is determined by the *joint probability distribution* $p(x, y)$. The probability to find $\hat{x}$ in X and $\hat{y}$ in Y, where $X, Y \subset \mathbb{R}$, reads

$$P\{(x \in X) \wedge (y \in Y)\} = \int_{X \times Y} dx\, dy\, p(x, y). \tag{89}$$

Note that $P\{(x \in X) \wedge (y \in Y)\} \equiv P\{(x, y) \in X \times Y\}$. The expectation value of any function of $\hat{x}, \hat{y}$ is given by

$$\langle f(\hat{x}, \hat{y}) \rangle := \int dx\, dy\, p(x, y) f(x, y). \tag{90}$$

The two random variables $\hat{x}, \hat{y}$ are mutually *independent* exactly if the joint probability distribution factorizes,

$$p(x, y) = p(x) q(y), \tag{91}$$

where

$$p(x) := \int dy\, p(x, y), \quad q(y) := \int dx\, p(x, y) \tag{92}$$

are the *marginal probabilities*.

If $\hat{y}$ is a function of $\hat{x}$, thus $\hat{y} = f(\hat{x})$ then the joint probability distribution reads

$$p(x, y) = p(x) \delta(y - f(x)), \tag{93}$$

which is demonstrated by showing that

$$\int dy\, p(x, y) = \int dy\, p(x) \delta(y - f(x)) = p(x) \tag{94}$$

and, by considering (63),

$$\int dx\, p(x, y) = \int dx\, p(x) \delta(y - f(x)) = q(y), \tag{95}$$

as desired.

It can be shown that the mean and the variance of independent random variables are additive,

$$\langle \hat{x} + \hat{y} \rangle = \langle \hat{x} \rangle + \langle \hat{y} \rangle \tag{96}$$
$$\sigma^2(\hat{x} + \hat{y}) = \sigma^2(\hat{x}) + \sigma^2(\hat{y}). \tag{97}$$

6.1 Mutual information

Two random variables that are not mutually independent are *correlated*. The degree of correlation can measured in different ways. The best measure for correlation is the *mutual information*

$$I(\hat{x}, \hat{y}) := \int dx\, dy\, p(x, y) \log \frac{p(x, y)}{p(x)q(y)}, \tag{98}$$

where it is common to take the binary logarithm, so that the resulting unit is the *bit*. (A *bit* is the binary logarithm of a dimensionless number.)

We can think of the mutual information as the amount of information I learn about one variable when getting to know the other variable. The mutual information is the best measure of correlation between $\hat{x}$ and $\hat{y}$, because if and only if $I(\hat{x}, \hat{y}) = 0$ then the two random variables are mutually independent. Moreover, we have

$$I(\hat{x}, \hat{y}) \geq 0 \quad \text{(positivity)} \tag{99}$$
$$I(\hat{x}, \hat{y}) = I(\hat{y}, \hat{x}) \quad \text{(symmetry)}. \tag{100}$$

In the *discrete case*, the maximal value of the mutual information is the minimum of the two individual entropies,

$$I(\hat{x}, \hat{y}) \leq \min\big\{(H(\hat{x}), H(\hat{y})\big\}. \tag{101}$$

In the continuous case there is no such upper bound, because the information needed to learn the exact value of a continuous variable is infinite. Hence, if two variables fully depend on each other (one is completely redundant with the other), their mutual information is infinite.

Other equivalent expressions for the mutual information are

$$I(\hat{x}, \hat{y}) = H(\hat{x}) + H(\hat{y}) - H(\hat{x}, \hat{y}) \tag{102}$$
$$I(\hat{x}, \hat{y}) = H(\hat{x}) - H(\hat{x}|\hat{y}) \tag{103}$$
$$I(\hat{x}, \hat{y}) = H(\hat{y}) - H(\hat{y}|\hat{x}), \tag{104}$$

where

$$H(\hat{x}|\hat{y}) := -\int dx\, dy\, p(x, y) \log p(x|y) \tag{105}$$

is the *conditional entropy*, and

$$H(\hat{x},\hat{y}) := -\int dx\, dy\, p(x,y) \log p(x,y) \tag{106}$$

is the *joint entropy*.

6.2 Pearson correlation coefficient

Another, often used, measure of correlation is the *Pearson correlation coefficient*,

$$\mathrm{Corr}(\hat{x},\hat{y}) := \frac{\mathrm{Cov}(\hat{x},\hat{y})}{\sigma(\hat{x})\sigma(\hat{y})}, \tag{107}$$

where

$$\mathrm{Cov}(\hat{x},\hat{y}) := \langle (\hat{x} - \langle \hat{x} \rangle)(\hat{y} - \langle \hat{y} \rangle) \rangle \tag{108}$$

is the *covariance* of $\hat{x}$ and $\hat{y}$. In fact, the Pearson correlation coefficient only measures the amount of *linear* correlation between two variables. A positive or negative coefficient indicates positive or negative linear correlation, respectively. When the two variables are independent, then the coreelation coefficient is zero. Unfortunately, the converse is not true.

Examples Consider a continuous random variable $\hat{x}$ and a fully dependent random variable $\hat{y} = f(\hat{x})$. Using (93), the joint entropy reads

$$\begin{aligned} H(\hat{x},\hat{y}) &= -\int dx\, dy\, p(x,y) \log p(x,y) && (109)\\ &= -\int dx\, dy\, p(x)\delta(y - f(x)) \log \big[p(x)\delta(y - f(x))\big] && (110)\\ &= -\int dx\, p(x) \log \big(p(x)\delta(0)\big) && (111)\\ &= -\infty. && (112) \end{aligned}$$

Thus for any continuous random variable $\hat{x}$

$$H(\hat{x}, f(\hat{x})) = -\infty. \tag{113}$$

***

Consider the random variables $\hat{x}$ given by (44) and the random variable $\hat{y} = \hat{x}^2$. The joint probability according to (93) reads

$$\begin{aligned} p(x,y) &= p(x)\delta(y - f(x)) && (114)\\ &= \begin{cases} \frac{1}{2}\delta(y - x^2) & ; -a \le x \le a \\ 0 & ; \text{else.} \end{cases} && (115) \end{aligned}$$

Therefore, the covariance of $\hat{x}$ and $\hat{y}$ yields

$$\begin{aligned}
\mathrm{Cov}(\hat{x},\hat{y}) &= \int dx\, dy\, p(x,y)(x - \langle\hat{x}\rangle)(y - \langle\hat{y}\rangle)\rangle && (116)\\
&= \int dx\, dy\, p(x)\delta(y - x^2)(x - 0)(y - \frac{1}{3}) && (117)\\
&= \int dx\, p(x)x(x^2 - \frac{1}{3}) && (118)\\
&= \int_{-a}^{a} dx\, \frac{1}{2}\left(x^3 - \frac{x}{3}\right) = 0, && (119)
\end{aligned}$$

where in the last step we have used that the integration interval is symmetric around zero and that the integrand is an odd function. Thus, the Pearson correlation coefficient equals *zero* for all values of a,

$$\mathrm{Corr}(\hat{x},\hat{y}) = \frac{\mathrm{Cov}(\hat{x},\hat{y})}{\sigma(\hat{x})\sigma(\hat{y})} = 0. \quad (120)$$

The example shows that there are quite simple cases where the Pearson correlation coefficient completely vanishes *although* the variables are strongly correlated. In contrast to that, the mutual information between $\hat{x}$ and $\hat{y}$ is infinite because the two variables are continuous and are completely dependent on each other. The information obtained about one variable when learning the exact value of the other is infinite, because, as we have shown above, the joint entropy $H(\hat{x},\hat{y})$ equals minus infinity. Collecting the terms we obtain

$$\begin{aligned}
I(\hat{x}, f(\hat{x})) &= H(\hat{x}) + H(f(\hat{x})) - H(\hat{x}, f(\hat{x})) && (121)\\
&= 2 \cdot H(\hat{x}) + \langle \log|f'(\hat{x})|\rangle - (-\infty) = \infty. && (122)
\end{aligned}$$

Concluding, *the mutual information between two completely dependent continuous random numbers is infinite*.