C PROBABILITY THEORY

C.1 *Kolmogorov axioms for probability*

Probabilities serve as a quantification of the chances of success in a game, when an element of randomness is involved. A random experiment consists in a random selection from a **finite** and non-empty set of events Ω consisting of individual outcomes. The power set $\mathcal{P}(\Omega)$ is the set of all possible subsets (including the empty set) of Ω. The probability measure is now assigning a probability to each of the possible selections $\mathcal{P}(\Omega)$ from the set Ω, and this probability is required to be a real number between 0 and 1.

⚠ *If Ω has n elements, the power set $\mathcal{P}(\Omega)$ contains 2^n elements.*

The probability measure is a probability if it fulfils Kolmogorov's axioms:

1. $p(\Omega) = 1$
2. $p(\mathrm{A}) \geq 0$ for all $\mathrm{A} \subset \Omega$
3. $p(\mathrm{A} \cup \mathrm{B}) = p(\mathrm{A}) + p(\mathrm{B})$ if $\mathrm{A} \cap \mathrm{B} = \emptyset$, otherwise $p(\mathrm{A} \cup \mathrm{B}) = p(\mathrm{A}) + p(\mathrm{B}) - p(\mathrm{A} \cap \mathrm{B})$

The first axiom says that the probability of some event in Ω is certainly to come up, and the second axiom makes sure that the probabilities are always positive. Probabilities of mutually exclusive events add, as stated by the third axiom. From these axioms one can draw a number of important conclusions:

1. $p(\mathrm{A}) + \mathrm{P}(\mathcal{C}(\mathrm{A})) = p(\Omega) = 1$, because $\Omega = \mathrm{A} \cup \mathcal{C}(\mathrm{A})$
2. $p(\mathrm{A}) \leq p(\mathrm{B})$ for $\mathrm{A} \subset \mathrm{B}$
3. $p(\mathrm{A} \cup \mathrm{B}) + p(\mathrm{A} \cap \mathrm{B}) = p(\mathrm{A}) + p(\mathrm{B})$
4. $p(\mathrm{A} \backslash \mathrm{B}) = p(\mathrm{A}) - p(\mathrm{A} \cup \mathrm{B})$
5. $p(\emptyset) = 0$, from the previous statement with $\emptyset = \mathrm{A} \backslash \mathrm{A}$

Generalisations to *infinite* sets are possible by replacing the set of Kolmogorov-axioms with a Borel-σ algebra.

C.2 *Laplace probability*

We can narrow down the set A to contain a single element, $\omega \in \Omega$ in the set of possible outcomes Ω. Then, the probability for a single element is given by

$$p(\mathrm{A}) = p\left(\bigcup_i \mathrm{A}_i\right) = \sum_i p(\mathrm{A}_i) = \sum_i p(\omega_i) \quad \text{for} \quad \mathrm{A} \subset \Omega \tag{C.67}$$

where $\mathrm{A}_i \cap \mathrm{A}_j = \emptyset$ for $i \neq j$, such that straightforward additivity is given. $\omega \to p(\omega)$ is the probability function which assigns a probability to each ω.

If the elements ω_i are equally likely to be selected, like identical lottery tickets, the probability of selecting an individual one must be the inverse of how many tickets are available, i.e. the cardinality $\#(\Omega)$ of the set Ω

$$p(\omega) = \frac{1}{\#\Omega} \quad \text{and therefore} \quad p(\mathrm{A}) = \frac{\#\mathrm{A}}{\#\Omega} \tag{C.68}$$

for any set A grouping a couple of elements ω_i into a set. This is exactly Laplace's idea about a probability being the number of favourable cases divided by the number of possible cases.

C.3 *Conditional probabilities and Bayes' law*

Conditional probabilities refer to a random experiment that is carried out in two steps: Firstly, a subset $\mathrm{A} \subset \Omega$ is selected in the first step, so that the events in the complement $\omega \in \mathcal{C}(\mathrm{A})$ have been assigned a probability $= 0$ in the successive second step of the random experiment: Then, from these preselected objects a new random selection is made, $\omega \in \mathrm{B}$ under the condition $\omega \in \mathrm{A}$. The conditional probability of selecting objects from B under the condition that they have been members of the selection of A is given by

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{\#(\mathrm{A} \cap \mathrm{B})}{\#(\mathrm{A})} \quad \text{with the Laplacian probability} \quad \mathrm{P}(\mathrm{A}) = \frac{\#(\mathrm{A})}{\#(\Omega)} \tag{C.69}$$

Extending the expression by the cardinality $\#(\Omega)$ of Ω gives

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{\#(\mathrm{A} \cap \mathrm{B})}{\#(\Omega)} \cdot \frac{\#(\Omega)}{\#(\mathrm{A})} = \frac{\mathrm{P}(\mathrm{A} \cap \mathrm{B})}{\mathrm{P}(\mathrm{A})} \tag{C.70}$$

Then, Bayes' law appears naturally from the realisation that $\mathrm{P}(\mathrm{A} \cap \mathrm{B})$ is symmetric

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{B} \cap \mathrm{A}) \tag{C.71}$$

so that one obtains:

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) \cdot \mathrm{P}(\mathrm{A}) = \mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{B} \cap \mathrm{A}) = \mathrm{P}(\mathrm{A}|\mathrm{B}) \cdot \mathrm{P}(\mathrm{B}) \tag{C.72}$$

implying in particular that $\mathrm{P}(\mathrm{A}|\mathrm{B}) \neq \mathrm{P}(\mathrm{B}|\mathrm{A})$. A classic example to remember this result is the following idea: If A corresponds to a person being female (in a biological or medical sense) and B corresponds to a person being pregnant, $\mathrm{P}(\mathrm{B}|\mathrm{A}) \simeq 10^{-2}$ (which can be easily estimated from the number of children per woman, and the duration of a pregnancy in relation to the life expectancy). On the contrary, $\mathrm{P}(\mathrm{A}|\mathrm{B})$ is essentially unity. So the gist of Bayes' law is that switching condition and random outcome of a conditional random process needs to be corrected by the ratio of the so-called prior probabilities $p(\mathrm{A})$ and $p(\mathrm{B})$,

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) = \mathrm{P}(\mathrm{A}|\mathrm{B})\frac{p(\mathrm{B})}{p(\mathrm{A})} \tag{C.73}$$

C.4 *Random variables*

Up to this point, the outcome of a random experiment was a selection of events from the set Ω, all contained in the power set $\mathcal{P}(\Omega)$. The idea of a random variable x now is to assign a value $x(\omega)$ to each of the possible individual outcomes, and to think of the probability $p(x)$ in terms of the value rather than the randomly selected elements. A straightforward example would be the value assigned to lottery tickets: The ticket that are drawn in a lottery would form the elements in Ω and the random variable x would be the money that is paid to the winner.

$$p(x) = \mathrm{P}(\omega \in \Omega | x(\omega) = x) \tag{C.74}$$

In this case, the probabilities $p(x)$ as a function of x are called a distribution. Clearly, the same value of the random variable x could correspond to different elements in Ω, so the probability $p(x)$ collects up the contribution from each element ω which is assigned the value x.

The expectation value $\langle x\rangle$ or the first moment of the random variable x following the distribution $p(x)$ is given by

$$\langle x\rangle = \sum_{\omega\in\Omega} \mathrm{P}(\omega)\cdot x(\omega) = \sum_i x_i p(x_i) = \int \mathrm{d}x\, p(x)\cdot x \tag{C.75}$$

glossing over a fundamental difference between finite and infinite sets Ω. Similarly, the variance $\langle x^2\rangle$ or the second moment is defined

$$\langle x^2\rangle = \sum_{\omega\in\Omega} \mathrm{P}(\omega)\cdot x^2(\omega) = \sum_i x_i^2 p(x_i) = \int \mathrm{d}x\, p(x)\cdot x^2 \tag{C.76}$$

which immediately generalises to moments $\langle x^n\rangle$ of arbitrary order,

$$\langle x^n\rangle = \sum_{\omega\in\Omega} \mathrm{P}(\omega)\cdot x^n(\omega) = \sum_i x_i^n p(x_i) = \int \mathrm{d}x\, p(x)\cdot x^n \tag{C.77}$$

where it is interesting to note that the moments can be defined by summing over the set of possible events ω or by integrating over the possible range of values for x, as the probabilities $\mathrm{P}(\omega)$ and $p(x)$ are not identical.

The normalisation required by the Kolmogorov-axioms suggests a transformation law for continuous probabilities,

$$1 = \int \mathrm{d}x\, p_x(x) = \int \mathrm{d}y\, \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|\cdot p_x(x(y)) = \int \mathrm{d}y\, p_y(y) \quad \text{such that} \quad p(x)\mathrm{d}x = p(y)\mathrm{d}y \tag{C.78}$$

from the Jacobian appearing in the variable change when integrating by substitution.

Summing random numbers $z = x + y$ from two distributions $p_x(x)$ and $p_y(y)$ leads to a distribution of the sum z which is given by convolution of the two original distributions,

$$p_z(z) = \int \mathrm{d}x p_x(x) \int \mathrm{d}y p_y(y)\delta_\mathrm{D}(z-(x+y)) = \int \mathrm{d}x p_x(x) p_y(z-x) = \int \mathrm{d}y p_x(z-y) p_y(y) \tag{C.79}$$

where the δ_D-distribution selects from all possible values x and y the ones that make the sum $x + y$ equal to a predefined z. Similarly, the distribution of differences, products and ratios of random numbers can be computed. Of course everybody knows that convolutions are most practically computed in Fourier-space, so would the Fourier-transform of a distribution be a sensible mathematical object?

C.5 *Characteristic function and moment generating function*

The characteristic function $\varphi(t)$ of a distribution $p(x)$ is defined as the Fourier-transform,

$$\varphi(t) = \int \mathrm{d}x\, p(x)\exp(-\mathrm{i}tx) = \langle\exp(-\mathrm{i}tx)\rangle \tag{C.80}$$

Substituting the series expansion of the exponential then yields

$$\varphi(t) = \int \mathrm{d}x p(x) \sum_n \frac{(-\mathrm{i}tx)^n}{n!} = \sum_n \frac{(-\mathrm{i}t)^n}{n!} \cdot \int \mathrm{d}x\, p(x) \cdot x^n = \sum_n \frac{(-\mathrm{i}t)^n}{n!} \langle x^n \rangle \tag{C.81}$$

That actually implies that the moments $\langle x^n \rangle$ can be computed by a differentiation

$$\langle x^n \rangle = \frac{1}{(-\mathrm{i})^n} \cdot \frac{\mathrm{d}^n}{\mathrm{d}t^n} \varphi(t) \bigg|_{t=0} \tag{C.82}$$

instead by an integration process: The n-fold differentiation isolates the nth moment $\langle x^n \rangle$ in the series, because the differentiation of the lower powers in t vanish and the higher order powers of t are set to zero, leaving just $\langle x^n \rangle$. Related to the characteristic function is the moment generating function, defined as the Laplace- instead of the Fourier-transform,

$$\mathrm{M}(t) = \int \mathrm{d}x\, p(x) \exp(-tx) = \langle \exp(-tx) \rangle \tag{C.83}$$

such that

$$\langle x^n \rangle = \frac{1}{(-1)^n} \cdot \frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathrm{M}(t) \bigg|_{t=0} \tag{C.84}$$

without having to worry about i. The above result about convolving distributions is now particularly simple,

$$\varphi_z(t) = \varphi_x(t) \cdot \varphi_y(t) \tag{C.85}$$

for the sum $z = x + y$ of two random variables. The Taylor-expansion of $\ln \varphi(t)$ yields the cumulants κ_n as coefficients,

$$\ln \varphi(t) = \sum_n \kappa_n \cdot \frac{t^n}{n!} \tag{C.86}$$

which are different compared to the moments, $\langle x^n \rangle \neq \kappa_n$ in general! First of all, cumulants add when random numbers are added, because $\ln \varphi_z(t) = \ln \varphi_x(t) + \ln \varphi_y(t)$, and they serve as a quantification, how close a distribution is to a Gauß-distribution.

The Gauß-distribution has the specific functional form

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{C.87}$$

and the characteristic function follows straight away to be of equal Gaußian shape, $\varphi(t) = \exp(-\frac{1}{2}\sigma^2 t^2)$ with the corresponding logarithm $\ln \varphi(t) \sim -it\mu - t^2\sigma^2$. Consequently, only the first three cumulants are nonzero $\kappa_0 = 1$ as a reflection of normalisation, the mean $\kappa_1 = \mu$ and the variance $\kappa_2 = \sigma^2$. One should be very cautious at this point: Cumulants and moments are not identical, and in general one needs Faa' di Bruno's formula to convert between them. Hence, the cumulant series truncates after κ_2, so that any higher-order cumulant must contain information about the non-Gaußian shape of a distribution.

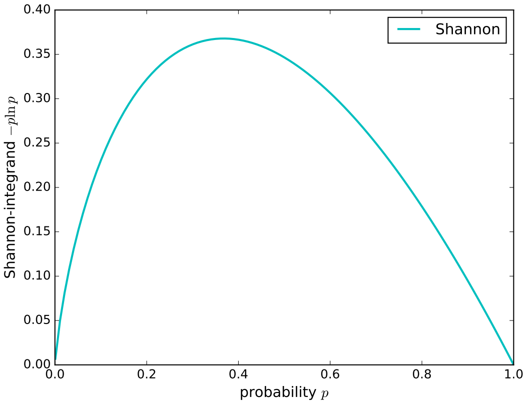

Figure 5: Integrand $-p \ln p$ of the Shannon entropy measure. Clearly, $p = 0$ and $p = 1$ are certain events, with no entropy S at all.

C.6 *Information entropies*

It is an abstract but very interesting question how much randomness is contained in a random process with probabilities p_i in the discrete and $p(x)\mathrm{d}x$ in the continuous case. For that quantification one computes Shannon's information entropy S

$$\mathrm{S} = -\sum_i p_i \ln p_i = -\int \mathrm{d}x\, p(x) \ln p(x) \tag{C.88}$$

which has the properties

1. $\mathrm{S} \geq 0$ for $0 < p_i \leq 1$
2. $\mathrm{S} = 0$ for $p_i = 1$ (certain outcome)
3. $p_i = \frac{1}{\#\Omega}$ for equally probable outcomes according to Laplace. Then,

$$\sum_i p_i = \frac{1}{\#\Omega} \sum_i 1 = \frac{\#\Omega}{\#\Omega} = 1 \rightarrow \mathrm{S} = -\sum_i \frac{1}{\#\Omega} \ln \frac{1}{\#\Omega} = \ln \#\Omega \tag{C.89}$$

Clearly, the first requirement is chosen to have S as a positive number, while the second and third requirement make sure that the information entropy increases (logarithmically) as there are more possible outcomes, starting from 0 if there is no randomness at all. This is illustrated in Fig. 5.

Information entropy in this definition is additive for independent subsystems. Having a factorising probability for the events i and j from two different sets $p_{ij} = p_i \cdot q_j$ and therefore statistical independence,

$$\mathrm{S} = -\sum_{ij} p_{ij} \ln p_{ij} = -\sum_{ij} p_i q_j \left(\ln p_i + \ln q_j\right) = -\sum_{ij} p_1 q_j \ln p_i + p_i q_j \ln q_j \tag{C.90}$$

such that separation of the terms yields

$$\mathrm{S} = -\sum_i\left(\sum_j q_j\right)p_i\ln p_i + \sum_j\left(\sum_i p_i\right)q_j\ln q_j = -\sum_i p_i\ln p_i - \sum_j q_j\ln q_j = \mathrm{S}_p + \mathrm{S}_q \tag{C.91}$$

i.e. the entropies of independent random processes are additive. In the discrete case, they are bounded from below by 0, which is reached for a certain event, where one of the p_i is one and the others zero, as normalisation is respected, $\sum_i p_i = 1$.

Contrary to common belief, Shannon's entropy measure is not the only quantity with these properties. A wider class of entropy measures, for instance the Rényi-entropy has the same properties and is defined as

$$\mathrm{S}_\alpha = \frac{1}{1-\alpha}\ln\sum_i p_i^\alpha = \frac{1}{1-\alpha}\ln\int \mathrm{d}x\, p(x)^\alpha \tag{C.92}$$

with a positive parameter α. Rényi's entropy falls back onto Shannon's expression for $\alpha = 1$ (by application of de l'Hôpital's rule), and the particular case of $\alpha = 1/2$ is referred to as Bhattacharyya-entropy,

$$\mathrm{S}_{\alpha=1/2} = 2\ln\sum_i\sqrt{p_i} = 2\ln\int\mathrm{d}x\,\sqrt{p(x).} \tag{C.93}$$

It should be emphasised that information entropies defined for a continuous distribution, i.e. a probability density $p(x)\mathrm{d}x$ is not invariant under changes of the random variable, which is not an issue at all for the discrete probabilties. In fact, $p(x)\mathrm{d}x = p(y)\mathrm{d}y$ as the transformation law gives

$$\mathrm{S} = -\int\mathrm{d}x\, p(x)\ln p(x) = -\langle\ln p(x)\rangle \quad\rightarrow\quad \mathrm{S} = -\int\mathrm{d}y\, p(y)\ln\left(p(y)\frac{\mathrm{d}y}{\mathrm{d}x}\right) \tag{C.94}$$

with an additional Jacobian $\mathrm{d}y/\mathrm{d}x$, and neither are they necessarily positive. In order to remedy this, relative entropies such as the Kullback-Leibler divergence have been introduced

$$\Delta\mathrm{S} = \sum_i p_i\ln\left(\frac{p_i}{q_i}\right) = \int\mathrm{d}x\, p(x)\ln\left(\frac{p(x)}{q(x)}\right) = \left\langle\ln\left(\frac{p}{q}\right)\right\rangle \tag{C.95}$$

which measures the relative amount of randomness between two distributions $p(x)\mathrm{d}x$ and $q(x)\mathrm{d}x$. In fact, the same transformation Jacobian $\mathrm{d}y/\mathrm{d}x$ is introduced for both $p(x)$ and $q(x)$, thus canceling out. The Rényi-equivalents of the Kullback-Leibler divergence are called α-divergence and reads:

$$\Delta\mathrm{S}_\alpha = \frac{1}{\alpha-1}\ln\sum_i\frac{p_i^\alpha}{q_i^{\alpha-1}} = \frac{1}{\alpha-1}\ln\int\mathrm{d}x\,\frac{p(x)^\alpha}{q(x)^{\alpha-1}} = \frac{1}{\alpha-1}\ln\left\langle\left(\frac{p}{q}\right)^{\alpha-1}\right\rangle \tag{C.96}$$

Neither Rényi- nor Shannon-measures are symmetric comparisons between the two distributions p and q, with one exception: The choice $\alpha = 1/2$ leads to the Bhattacharyya-entropy,

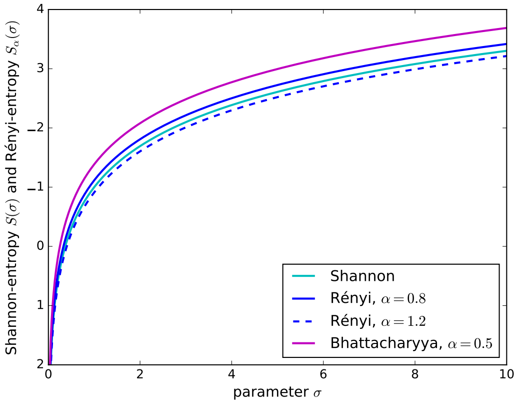

Figure 6: Shannon-, Rényi- and Bhattacharyya-entropies for the exponential distribution $p(x) = \exp(-x/\sigma)/\sigma$ on $[0, +\infty)$.

$$\Delta S_{\alpha=1/2} = -2 \ln \sum_i \sqrt{p_i q_i} = -2 \ln \int \mathrm{d}x \, \sqrt{p(x) q(x)} \tag{C.97}$$

The entropies are depicted in Fig. 13 for the exponential distribution

$$p(x) = \frac{1}{\sigma} \exp\left(-\frac{x}{\sigma}\right) \tag{C.98}$$

on the interval $[0 \ldots + \infty)$ as an example. They have the explicit expressions

$$S = \ln \sigma + 1 \quad \text{and} \quad S_\alpha = \frac{1}{\alpha - 1} \ln \left(\alpha \sigma^{\alpha - 1}\right) = \frac{\ln \alpha}{\alpha - 1} + \ln \sigma. \tag{C.99}$$

The amount of randomness of the distribution is larger for large σ, as large values for the outcome x occur with higher probability and effectively, a wider range of likely outcomes is covered. Consequently, the entropies as measures of the randomness of a distribution, increase with σ. The limit $\lim_{\alpha \to 1} \ln \alpha / (\alpha - 1)$ is given by 1, so the Shannon-case is recovered.

It is a very interesting thought to consider Shannon's entropy as a functional for the distribution p_i or $p(x)\mathrm{d}x$ and ask for which distribution the information entropy as a functional is maximised. For instance, the variation of S would be

$$\delta S = -\sum_i (\ln p_i + 1) \delta p_i = 0 \tag{C.100}$$

which needs to be augmented by a boundary condition making sure that the resulting probabilities add up to one, as required by Kolmogorov's first axiom:

$$\sum_i p_i = 1 \rightarrow \delta \sum_i p_i = \sum_i \delta p_i = 0 \tag{C.101}$$

such that

$$\delta S + \lambda \sum_i \delta p_i = 0 \tag{C.102}$$

implying that $\sum(\ln p_i + 1 + \lambda)\delta p_i = 0$ and therefore

$$p_i = \exp(-(1+\lambda)) \tag{C.103}$$

i.e. a constant probability: Information entropy is maximal for the uniform distribution, which defines the microcanonical ensemble in statistical physics.

Maximising Shannon's entropy with additional constraint

$$U = \sum_i p_i E_i = \langle E \rangle \tag{C.104}$$

with a fixed expectation value U, where we have already chosen suggestive variable names, alongside the normalisation. Formulating both constraints as Lagrange multipliers for the variation δS entropy

$$\delta S = -\sum_i (\ln p_i + 1)\delta p_i = 0 \tag{C.105}$$

would require

1. $\sum_i p_i = 1 \to \delta \sum_i p_i = \sum_i \delta p_i = 0$

2. $\sum_i p_i E_i = U \to \delta \sum_i p_i E_i = \sum_i E_i \delta p_i = 0$

leading to

$$\delta S + \lambda \sum_i \delta p_i + \mu \sum_i E_i \delta p_i = 0 \tag{C.106}$$

which can be computed to yield $\sum(\ln p_i + 1 + \lambda + \mu E_i)\delta p_i = 0$ and solved for the probabilties to give

$$p_i = \exp(-(1+\lambda+\mu E_i)) \tag{C.107}$$

The two Lagrange multipliers can be determined by resubstituting p_i into the two boundary conditions:

$$\sum_i p_i = \sum_i \exp(-(1+\lambda+\mu E_i)) = 1 \quad \to \quad \exp(-(1+\lambda)) = \frac{1}{\sum_i \exp(-\mu E_i)} = \frac{1}{Z} \tag{C.108}$$

with the partition sum $Z = \sum_i \exp(-\mu E_i)$ as well as

$$\frac{\sum_i E_i \exp(-\mu E_i)}{\sum_i \exp(-\mu E_i)} = U = \frac{1}{Z}\sum_i E_i \cdot \exp(-\mu E_i) \tag{C.109}$$

with the probabitliy

$$p_i = \frac{1}{Z}\exp(-\mu E) \tag{C.110}$$

which looks a bit reminiscent of the Boltzmann-probability,

$$p_i \sim \exp\left(-\frac{E_i}{k_B T}\right) \tag{C.111}$$

if the identification $\mu = 1/(k_B T)$ is valid. This is the basis of the so-called canonical ensemble, where states at higher energy are less likely according to the Boltzmann probability. In summary I would like to point out that the realisation of entropy maximising probabilities replaces the fundamental postulates of statistical physics: It is superfluous to define the equipartition of states or the Boltzmann-factor in an axiomatic way when in fact the two distributions are the ones that maximise Shannon's entropy, subject to boundary conditions. Perhaps it is much more intuitive to imagine that the equipartition of states is a condition with makes the least assumption about the system.