Large deviation theory

Having discussed typical events, let us discuss a-typical events. There are two reasons (at least) why a-typical events may be of interest. First we may be interested in *rare* events that involve fluctuations of quantities that are larger than the typically expected ones. For example, the credit rating of an insurance company is based on its estimated default probability. This occurs if an unexpectedly large number of contracts in its portfolio demand claims that exceed the equity[245] A of the insurance company. The claims X_i from contracts $i = 1, \dots, n$ can be modeled as random variables and the default corresponds to the event

$$D = \{S_n \geq A\}, \qquad S_n = \sum_{i=1}^{n} X_i.$$

If $n \gg 1$, which is the case in this example, we know that as long $\mathbb{E}[S_n] < A$ this even does not typically occur. So default D is an a-typical event.

Communications engineers face a similar problem: they need to calculate safe buffer and bandwidth sizes for network traffic which arises from a population of many users. This entails estimating the probability of traffic overflow, making sure that these will be very rare events. In both cases, we want to estimate how small is the probability of the large deviation and how do we expect it to occur.

In a stylised picture, biological evolution occurs through random mutations. Most of them have neutral or deleterious effects but some rare mutation bring some advantages that increase the reproduction probability – the fitness – of individuals carrying them. Even if average fitness decreases, because cells with larger fitness are selected, the fitness of the population as a whole does not decrease. Evolution is propelled by rare events.

More in general, when we study a phenomenon we might represent our current state of knowledge with a distribution $Q(\omega)$ defined on the sample space of all possible realisations $\omega \in \Omega$ of that phenomenon. You may think of ω as a complete description of that phenomenon and of Q as the distribution encoding all known (experimental) facts. The distribution Q is the *theory* that allows us to predict the value

There are several textbooks devoted to Large Deviation Theory, as e.g.

Richard S Ellis. *Entropy, large deviations, and statistical mechanics*. Springer Verlag, New York, Berlin, 1985

[245] The equity is a measure of the value of the company, and it equals the amount of money that would result if all of the assets of the company were liquidated and all debts were paid off.

$\mu_Q = \mathbb{E}_Q[X]$ of a quantity $X(\omega)$. Clearly, we're interested in predictions of the theory Q going beyond the range of events that have been used to derive it.

This prediction can be tested in a repeated series of independent experiments $\underline{X} = (X_1, \dots, X_n)$ and, if $\mathbb{E}_Q[X] < +\infty$ we expect that $S_n/n \to \mathbb{E}_Q[X]$. If this expectation is confirmed by the experiment, then the experiment brings no new information. But if S_n/n is very different from $\mathbb{E}_Q[X]$, then the experimental result calls for a revised theory P that can accommodate all existing knowledge and the new observation. In this case, the experiment is an a-typical event because the theory Q is wrong[246]. How should we revise the theory $Q \to P$ in order to incorporate the new information? And how much did we learn?

[246] This logic is routinely applied in statistics, when we want to test an hypothesis. Then $Q(\omega)$ stands for the distribution that we expect if a certain hypothesis H_0 is satisfied. A practical example is that of subjects that receive a treatment for a certain disease. Then one wants to rule out the *null hypothesis* H_0 that the treatment is completely ineffective, on the basis of a sample $\underline{X}$ of measurement of a quantity X that is known to be relevant. In hypothesis testing, we take $Q(x)$ as the distribution that X would follow in untreated patients. In this case, if the treatment is effective then the sample $\underline{X}$ is a-typical.

THE STUDY OF RARE (a-typical) events is the domain of *Large Deviation Theory*. Let us start by formalising the main questions and concepts in the case of sequences $\underline{X} = (X_1, \dots, X_n)$ of i.i.d. random variables. Let us assume that the variance $\mathbb{V}[X_i] = \sigma^2 < \infty$ is finite, so that both the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) hold. Then, for large n, the mean S_n/n will be very close to $\mu = \mathbb{E}[X]$ (LLN) and the sum S_n is well approximated by $S_n \simeq n\mu + \sigma\sqrt{n}\zeta$ where ζ is a Gaussian random variable with zero mean and unit variance (CLT). This is what we *typically* expect. Yet it may happen to observe *large deviations*[247], i.e. events such that, for some $\epsilon > 0$,

$$A_n(\bar{x}) = \left\{ \underline{X} : \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \bar{x} \right| < \epsilon \right\} \tag{309}$$

with $\bar{x} \neq \mu$. These are clearly *a-typical* events that we expect to occur with a vanishingly small probability, as $n \to \infty$.

[247] NY Times reports on Dec. 11, 2021 that Kentuky "was hit by four tornadoes [...] including one that stayed on the ground for more than 200 miles." The Governor of Kentuky said "This has been the most devastating tornado event in our state's history, [...] The level of devastation is unlike anything I have ever seen." This is a very unlikely event according to the distribution of past events. Scientists suspect that this suggests that the distribution of severity of these events has changed because of climate change.

The questions that we shall focus on are:

1. what is the probability $P\{A_n(\bar{x})\}$ of the large deviation? More specifically, since $P\{A_n(\bar{x})\} \to 0$ as $n \to \infty$, we shall be interested in the leading behaviour of $P\{A_n(\bar{x})\}$ with n.
2. Conditional on the fact that $A_n(\bar{x})$ occurs, what is the distribution of the X_i? In other words, how are large deviations typically realised?

The answers to these questions depend on the distribution from which the sample $\underline{X}$ is drawn. We shall discuss separately the different cases.

Large deviations for i.i.d. variables with finite support

Consider a sequence of n i.i.d. random variables $\underline{X} = (X_1, \dots, X_n)$ drawn from a distribution $Q(x)$ over a finite alphabet $x \in \mathcal{X}$ (i.e. $|\mathcal{X}| < +\infty$). The probability of a sample $\underline{X}$ is given by

This part is discussed in COVER Chapter 11.

Remember that

$$\mathcal{H}[P] = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$$

is the entropy as a functional of $P(x)$.

$$P\{\underline{X}\} = \prod_{i=1}^{n} Q(X_i) = \prod_{x\in\mathcal{X}} Q(x)^{nP_{\underline{X}}(x)} = e^{-n\mathcal{H}[P_{\underline{X}}]-nD_{KL}(P_{\underline{X}}||Q)}. \quad (310)$$

where

$$P_{\underline{X}}(x) = \frac{1}{n}|\{i:\ X_i = x\}| \quad (311)$$

is the empirical distribution[248], which is the fraction of points in the sample that are equal to x. In particular, the probability of a sample $P\{\underline{X}\}$ only depends on its type $P_{\underline{X}}$. The event A_n also can be defined in terms of types, as a subset in the space of distributions[249] $\mathcal{P}$ or of types $\mathcal{P}_n \subseteq \mathcal{P}$ of samples of n points. More precisely, the event defined in Eq. (309) can be rewritten as $A_n = \{P_{\underline{X}} \in \mathcal{A}_n\} \subseteq \mathcal{P}_n$ where

$$\mathcal{A}_n = \{P \in \mathcal{P}_n:\ |\mathbb{E}_P[X] - \bar{x}| < \epsilon\}, \qquad \mathbb{E}_P[X] = \sum_{x\in\mathcal{X}} P(x)x \quad (312)$$

is a subset of the space of distributions defined on $\mathcal{X}$.

The probability that an event A_n occurs can be written as

$$\begin{aligned} P\{A_n\} &= \sum_{\underline{X}\in A_n} P\{\underline{X}\} = \sum_{\underline{X}\in A_n} e^{-n\mathcal{H}[P_{\underline{X}}]-nD_{KL}(P_{\underline{X}}||Q)} && (313)\\ &= \sum_{P\in\mathcal{A}_n} \sum_{\underline{X}:\ P_{\underline{X}}=P} e^{-n\mathcal{H}[P]-nD_{KL}(P||Q)} && (314)\\ &= \sum_{P\in\mathcal{A}_n} e^{-n\mathcal{H}[P]-nD_{KL}(P||Q)} \left|\{\underline{X}:\ P_{\underline{X}} = P\}\right| && (315)\\ &\backsim \sum_{P\in\mathcal{A}_n} e^{-nD_{KL}(P||Q)} && (316) \end{aligned}$$

where we used the fact that, by Eq. (310) $P\{\underline{X}\}$ only depends on $P_{\underline{X}}$ in the first equation, and the fact that the number $\left|\{\underline{X}:\ P_{\underline{X}} = P\}\right|$ of samples with $P_{\underline{X}} = P$ is $\backsim e^{n\mathcal{H}[P]}$, by the Asymptotic Equipartition Property[250].

If $|\bar{x} - \mathbb{E}_Q[X]| < \epsilon$ then the event A_n is typical, which means that there is at least one distribution $P \in \mathcal{A}_n$ that is very close to Q, and that asymptotically converges to it. Therefore for these distributions $D_{KL}(P||Q) \to 0$ as $n \to \infty$ and, as a consequence, $P\{A_n\} \to 1$. If $\bar{x}$ is significantly different from $\mathbb{E}_Q[X]$ then Q is "far" from any $P \in \mathcal{A}_n$. Then A_n is an *a-typical* event and its probability vanishes as $n \to \infty$. Every type $P \in \mathcal{A}_n$ contributes with a term which is exponentially small in n, with a coefficient that is proportional to $D_{KL}(P||Q)$. Then for n large, we expect that the sum will be dominated by the type

$$P^* = \arg\min_{P\in\mathcal{A}_n} D_{KL}(P||Q) \quad (317)$$

that is "closest" to Q, in terms of D_{KL} divergence. Indeed, taking only the term $P = P^*$ in the sum over $\mathcal{A}_n$ in Eq. (316), one gets $P\{A_n\} \geq e^{-nD_{KL}(P^*||Q)}$. On the other hand, $e^{-nD_{KL}(P||Q)} \leq e^{-nD_{KL}(P^*||Q)}$

[248] $P_{\underline{X}}(x)$ is called the type of $\underline{X}$. We refer to Cover Chapter 11 for a detailed discussion of types.

[249] The space of disrtibutions is defined as

$$\mathcal{P} = \left\{P : \mathcal{X} \to \mathbb{R},\ P(x) \geq 0,\ \sum_{x\in\mathcal{X}} P(x) = 1\right\}.$$

The set $\mathcal{P}_n$ of types is a subset of $\mathcal{P}$ of distributions where, for all $x \in \mathcal{X}$, $p(x) = k_x/n$ with $k_x = 0, 1, \ldots, n$ and $\sum_{x\in\mathcal{X}} k_x = n$. $\mathcal{P}_n$ is a discrete set of points in $\mathcal{P}$. For each $x \in \mathcal{X}$, k_x can take $n+1$ values, so the number of points in $\mathcal{P}_n$ can be at most $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$. As n increases, the number of points in $\mathcal{P}_n$ becomes denser and denser, so that each $P \in \mathcal{P}$ can be approximated to arbitrary precision by a $P \in \mathcal{P}_n$ if n is sufficiently large.

[250] Let us remind that $a_n \backsim e^{cn}$, where c is a constant, means that $\frac{1}{n}\log a_n \to c$ as $n \to \infty$.

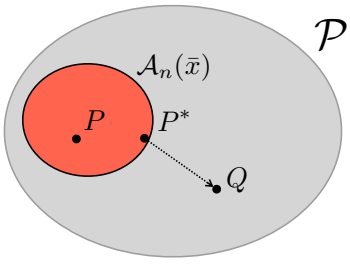


Figure 35: Sketch of the minimisation problem in large deviation theory. We note in passing that the relative entropy $D_{KL}(P||Q) \geq D_{KL}(P||P^*) + D_{KL}(P^*||Q)$ satisfies the opposite of the triangle inequality (see Cover Theorem 11.6.1).

that provides an upper bound

$$P\{A_n\} \leq e^{-nD_{KL}(P^*||Q)}|\mathcal{A}_n| \tag{318}$$
$$\leq (1+n)^{|\mathcal{X}|}e^{-nD_{KL}(P^*||Q)} \tag{319}$$

where we first used the fact that the number $|\mathcal{A}_n|$ of types $P \in \mathcal{A}_n$ is upper bounded by the total number of types $|\mathcal{P}_n|$, which is less than $(n+1)^{|\mathcal{X}|}$. This means that $P\{A_n\}$ decays exponentially with a rate which is equal to $D_{KL}(P^*||Q)$. This is the content of *Sanov's theorem*, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \log P\{A_n\} = -D_{KL}(P^*||Q). \tag{320}$$

Summarising, the leading order in the behaviour of the probability of an a-typical event A_n when $n \to \infty$, is given by $P\{A_n\} \backsim e^{-nD_{KL}(P^*||Q)}$ where P^* is the solution of Eq. (317).

Let us illustrate this for the case

$$\mathcal{A}_n = \left\{ P : \sum_{x\in\mathcal{X}} P(x)f(x) \geq \bar{f} \right\}$$

that corresponds to event A_n where the average of $f(X)$ over a sample $\underline{X}$ of points drawn independently from $Q(x)$, is larger than $\bar{f}$. If

$$\mathbb{E}_Q[f(X)] \equiv \sum_{x\in\mathcal{X}} Q(x)f(x) \geq \bar{f}$$

then $Q \in \mathcal{A}_n$ and the event is typical. The interesting case is when $\mathbb{E}_Q[f(X)] < \bar{f}$ because then A_n is an a-typical event where the sample average

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \bar{f}$$

does not satisfies the law of large numbers. In order to compute $P\{A_n\}$ we should first solve the problem Eq. (317). This is done introducing Lagrange multipliers and solving the problem

$$\min_{P,\beta,\lambda} \left[D_{KL}(P||Q) + \beta \left(\sum_{x\in\mathcal{X}} P(x)f(x) - f_0 \right) + \lambda \left(\sum_{x\in\mathcal{X}} P(x) - 1 \right) \right],$$

where $f_0 \geq \bar{f}$ has to be chosen so as to satisfy Eq. (317). The solution has the form

$$P_\beta(x) = \frac{Q(x)e^{-\beta f(x)}}{Z(\beta)} \tag{321}$$

where

$$Z(\beta) = \mathbb{E}_Q\left[e^{-\beta f(x)}\right] = \sum_{x\in\mathcal{X}} Q(x)e^{-\beta f(x)} \tag{322}$$

is the normalisation constant[251]. The parameter β has to be fixed so that

$$\mathbb{E}_\beta[f(X)] = \sum_{x\in\mathcal{X}} P_\beta(x)f(x) = -\frac{d}{d\beta}\log Z(\beta) \tag{323}$$

[251] $Z(\beta)$ is often called the *partition function*. Note that the derivatives of $\log Z(\beta)$ is closely related to the cumulant generating function of X, $\phi(h) = \log Z(-h)$. We use this property to relate the derivatives of $\log Z(\beta)$ to the cumulants of X under the distribution P_β.

where we used $\mathbb{E}_\beta[\dots]$ for expectations over the distribution P_β. Notice that when $\beta = 0$ then $P_\beta(x) = Q(x)$ is the original distribution. For this reason, the curve $\mathbb{E}_\beta[f(X)]$ takes the value $\mathbb{E}_Q[f(X)]$ for $\beta = 0$. In other words, the point $\beta = 0$ corresponds to typical events, where the law of large numbers holds. Varying β one "explores" rare events with large fluctuations of the sample mean of f. In particular, $\mathbb{E}_\beta[f(X)]$ is a decreasing function of β (see Fig. 36), because

$$\frac{d\mathbb{E}_\beta[f(X)]}{d\beta} = -\left\{\mathbb{E}_\beta\left[f^2(X)\right] - \mathbb{E}_\beta[f(X)]^2\right\} = -\mathbb{V}_\beta[f(X)] \leq 0.$$

So the event $\mathcal{A}_n$ corresponds to all those β for which $\mathbb{E}_\beta[f(X)] \geq \bar{f}$, i.e. to the region $\beta \leq \beta^*$ where β^* is such that $\mathbb{E}_{\beta^*}[f(X)] = \bar{f}$.

Among all the distributions P_β with $\beta \leq \beta^*$ we should chose that one with the smallest $D_{KL}(\cdot||Q)$. Now

$$D_{KL}(P_\beta||Q) = -\beta\mathbb{E}_\beta[f(X)] - \log Z(\beta)$$

and

$$\frac{dD_{KL}(P_\beta||Q)}{d\beta} = \beta\mathbb{V}_\beta[f(X)]$$

has the same sign of β. Therefore, $D_{KL}(P_\beta||Q)$ has a minimum at $\beta = 0$ and its minimum for $\beta \leq \beta^* \leq 0$ is attained at β^*. Summarizing,

$$P\{A_n\} \sim e^{-nD_{KL}(P^*||Q)}, \qquad D_{KL}(P^*||Q) = -\beta^*\bar{f} - \log Z(\beta^*)$$

where β^* satisfies $\mathbb{E}_{\beta^*}[f(X)] = \bar{f}$ and $P^* = P_{\beta^*}$.

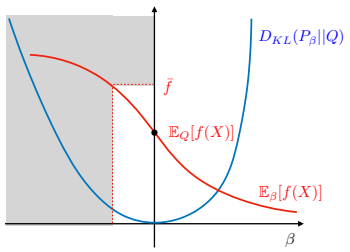


Figure 36: The shaded region corresponds to the event E.

What is the meaning of the distribution P_{β^*}? In order to address this question, let us compute the marginal distribution of the first m variables $\vec{X} = (X_1, \dots, X_m)$

$$P(\vec{x}|A_n(\bar{x})) = P\{X_1 = x_1, \dots, X_m = x_m|A_n(\bar{x})\}$$

when $n \to \infty$ with m finite, conditional on the occurrence of the large deviation $A_n(\bar{x})$. We observe that[252]

$$P(\vec{x}|A_n(\bar{x})) = \frac{P\{\{X_1 = x_1, \dots, X_m = x_m\} \bigcap A_{n-m}(\bar{x}')\}}{P\{A_n(\bar{x})\}}, \quad (324)$$

$$= \frac{Q(x_1)\cdots Q(x_m)P\{A_{n-m}(\bar{x}')\}}{P\{A_n(\bar{x})\}} \quad (325)$$

$$\simeq P_{\beta^*}(x_1)\cdots P_{\beta^*}(x_m) \quad (326)$$

where in the first equality, the event $A_{n-m}(\bar{x}')$ is the event that the $n-m$ variables $(X_{m+1}, \dots, X_n)$ sum up to

$$\sum_{i=m+1}^{n} X_i = n\bar{x} - \sum_{i=1}^{m} x_i \equiv (n-m)\bar{x}'. \quad (327)$$

[252] We use the previous results with $f(x) = x$ for simplicity.

In Eq. (325) we use the fact that the variables X_i are independent and they are drawn from Q. Finally, Eq. (326) holds because[253]

$$\frac{P\{A_{n-m}(\bar{x}')\}}{P\{A_n(\bar{x})\}} \simeq e^{-(n-m)D_{KL}(P_{\beta'}||Q)+nD_{KL}(P_\beta||Q)} \quad (328)$$

$$\simeq e^{(n-m)\beta'\bar{x}'-n\beta\bar{x}-n[\log Z(\beta)-\log Z(\beta')]-m\log Z(\beta')}$$

$$\simeq \frac{1}{Z(\beta^*)^m} e^{-\beta^*\sum_{i=1}^m x_i} \quad (329)$$

for $n \to \infty$. Eq. (325) shows that in the limit $n \to \infty$ the joint distribution of $\vec{X}$ coincides with the distribution of m variables $X_1, \dots, X_m$ which are drawn independently from the same distribution $P_\beta(x)$. In loose words, *the large deviation is realised as a typical sample of independently drawn variables from a distribution $P_\beta(x)$, which is different from Q.*

There are in principle many other ways in which a sample that satisfies $A_n(\bar{x})$ could be realised. Any other distribution $P \in \mathcal{A}_n(\bar{x})$ such that $\mathbb{E}_P[X] = \bar{x}$ would generate samples that satisfies $A_n(\bar{x})$, typically. However, the probability to generate samples with type $P_{\underline{X}} = P$ is $e^{-nD_{KL}(P||Q)}$, which is exponentially smaller (in n) than the probability of typical samples generated as i.i.d. draws from P^* in Eq. (317). The distribution that is most likely to be observed is the "closest" to Q in terms of the KL divergence[254].

Large deviations for i.i.d. continuous variables with thin tails

The same solution can be derived by a direct calculation for the cases where $X_i \in \mathbb{R}$ are continuous i.i.d. random variables whose common pdf $q(x)$ decays at least exponentially fast[255]. We refer to this case by saying that $q(x)$ has *thin tails*. The case of *fat tails*, where $q(x)$ decays slower than an exponential, will be discussed later.

Let $A_n(\bar{x})$ be the event that the mean falls in an interval $[\bar{x}, \bar{x} + d\bar{x})$ for an infinitesimal $d\bar{x}$. Then $P\{A_n(\bar{x})\} = p_n(\bar{x})d\bar{x}$ where $p_n(\bar{x})$ is the pdf of $\bar{x}$. This can be computed using the integral representation of the delta function[256]

$$p_n(\bar{x}) = n\int_{-\infty}^{\infty}\prod_{i=1}^n dx_i q(x_i)\delta\left(\sum_i x_i - n\bar{x}\right) \quad (330)$$

$$= n\int_{-\infty}^{\infty}\frac{dk}{2\pi}e^{ikn\bar{x}}\left[\int dx q(x)e^{-ikx}\right]^n \quad (331)$$

$$= n\int_{-\infty}^{\infty}\frac{dk}{2\pi}e^{ng(ik)} \quad (332)$$

where the function $g(\beta)$ is defined as.

$$g(\beta) = \beta\bar{x} + \log\int dx q(x)e^{-\beta x}$$

[253] Here we use the shorthand $\beta = \beta^*(\bar{x})$ and $\beta' = \beta^*(\bar{x}')$. The second line follows from the fact that

$$D_{KL}(P^*||Q) = -\beta\bar{x} - \log Z(\beta).$$

In the first term of the exponent we use Eq. (327) so that

$$(n-m)\beta'\bar{x}' - n\beta\bar{x} = n(\beta'-\beta)\bar{x} - \sum_{i=1}^m x_i$$

The first term cancels with

$$\log Z(\beta') - \log Z(\beta) \simeq -(\beta'-\beta)\bar{x} + \dots$$

that is obtained expanding $\log Z(\beta')$ around β (note that $\beta - \beta' \sim \bar{x} - \bar{x}'$ is of order $1/n$) using $\bar{x} = -\frac{d}{d\beta}\log Z(\beta)$.

[254] This is because the type $P_{\underline{X}}$ of a random sample of i.i.d. draws from a distribution is not random at all, by the Glivenko-Cantelli theorem.

This derivation can be found also in the appendix of

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009

[255] i.e. distributions such that for some $\lambda, K > 0$

$$\lim_{x\to\pm\infty} q(x)e^{\lambda|x|} \le K.$$

[256] The Dirac's $\delta(x)$ function is defined as that (generalized) function such that for any function $f(x)$

$$\int_{-\infty}^{\infty} dx f(x)\delta(x-x_0) = f(x_0)$$

In particular with $f(x) = 1$ this shows that $\delta(x-x_0)$ is a pdf whose mass is concentrated in x_0. With $f(x) = e^{ikx}$ the relation above shows that the Fourier transform of $\delta(x)$ is 1. Hence

$$\delta(x) = \int_{-\infty}^{\infty}\frac{dk}{2\pi}e^{-ikx}.$$

Also note that $\delta(ax) = \delta(x)/a$.

The integral in Eq. (332) can be evaluated by the saddle point method. This entails looking at the stationary point of $g(\beta)$ and expanding around it. The maximum of $g(\beta)$ is attained at $\beta^*(\bar{x})$ that satisfies the equation $g'(\beta) = 0$, i.e.

$$\bar{x} = \frac{1}{Z(\beta)} \int dx x q(x) e^{-\beta x}, \qquad Z(\beta) = \int dx q(x) e^{-\beta x} = \mathbb{E}_Q \left[e^{-\beta X} \right] \tag{333}$$

Then one can perform the integral in Eq. (332) by substituting

$$g(ik) = g(\beta^*) + \frac{g''(\beta^*)}{2}(ik - \beta^*)^2 + O(ik - \beta^*)^3$$

Upon changing variables to $y = \sqrt{ng''(\beta)}(k + i\beta^*)$ one can check that higher order terms in the expansion of g beyond the second one are small for n large and can be neglected. Therefore one can compute the Gaussian integral with the result

$$p_n(\bar{x}) \simeq \sqrt{\frac{n}{2\pi g''(\beta^*)}} e^{ng(\beta^*)} \sim e^{ng(\beta^*)} \tag{334}$$

where the leading order behavior in n is retained in the last equation.

There are few things to observe in this result:

1. The form of Eq. (333) that fixes β^* is of the form $\bar{x} = \mathbb{E}_\beta [X]$ where the expectation is taken on the modified distribution

$$p_\beta(x) = \frac{q(x) e^{-\beta x}}{Z(\beta)} \tag{335}$$

   This is not a coincidence, as we're going to see.

2. The second derivative of g is positive as it is the variance of a random variable X with pdf $p_\beta(x)$

$$g''(\beta) = \mathbb{E}_\beta \left[X^2 \right] - \mathbb{E}_\beta [X]^2 = \mathbb{V}_\beta [X]$$

3. The marginal joint distribution of a finite number m of variables, say $\vec{X} = (X_1, \ldots, X_m)$ conditional on the occurrence of $A_n(\bar{x})$, defined as

$$p(\vec{x}|A_n(\bar{x}))\, dx_1 \cdots dx_m = P\{X_1 \in [x_1, x_1 + dx_1), \ldots, X_m \in [x_m + dx_m)|A_n(\bar{x})\}$$

   can be estimated as before, and

$$\lim_{n\to\infty} p(\vec{x}|A_n(\bar{x})) = p_\beta(x_1) \cdots p_\beta(x_m)$$

   This shows that the large deviation is realised as an independent draw of variables from the distribution $p_\beta(x)$.

Exercise: There are other ways in which a large deviation $\bar{x}$ can be realised. Imagine that a large deviation $\bar{x} = \mathbb{E}_Q [X] + a$ is observed, with $a > 0$. The "explanation" of large deviation theory is that the event $A_n(\bar{x})$ occurs because X_i are actually not drawn from $q(x)$ but from $p_\beta(x)$ of Eq. (335), with β determined by the condition $\bar{x} = \mathbb{E}_\beta [X]$. A different explanation is that, instead, the X_i are drawn i.i.d. from a "shifted" distribution $p_a(x) = q(x - a)$. Show, for the specific example of exponential random variables, $q(x) = e^{-x}$ for $x \geq 0$ and $q(x) = 0$ for $x < 0$, that the "shifted" distribution hypothesis is much less plausible than the one offered by large deviation theory.

4. The expression of the rate of exponential decay of the probability $P\{A_n(\bar{x})\}$ can be written as

$$\lim_{n\to\infty} \frac{1}{n} \log P\{A_n(\bar{x})\} = g(\beta^*) = -D_{KL}(p_{\beta^*}||q)$$

as shown by a direct calculation. This is the same content of Sanov's theorem Eq. (320). The fact that $P\{A_n(\bar{x})\}$ is related to a relative entropy is not accidental, as we discussed earlier.

Large deviations and the Legendre transform

The function

$$I(\bar{x}) = -\lim_{n\to\infty} \frac{1}{n} \log P\{A_n(\bar{x})\} \tag{336}$$

is called the Cramer's function or the large deviation (rate) function. As shown above, $I(\bar{x}) = D_{KL}(P_{\beta^*(\bar{x})}||Q)$ is a relative entropy. Rephrasing the steps we did above, the practical recipe to compute the Cramer's function is condensed in the following steps[257]:

1. Compute the cumulant generating function

$$\phi(h) = \log \int dx q(x) e^{hx} = \log \mathbb{E}_Q \left[e^{hX}\right]$$

2. Take a derivative of ϕ and compute

$$\bar{x}(h) = \frac{d\phi}{dh} \tag{337}$$

3. invert this function and compute $h(\bar{x})$

4. compute

$$I(\bar{x}) = \bar{x}h(\bar{x}) - \phi[h(\bar{x})]$$

The variables h and $\bar{x}$ are called *conjugate variables*. Notice that the function $\phi(h)$ has to be concave, i.e. its second derivative must be positive. This is always true in the present case, because $\phi''(h) = \mathbb{V}_\beta[X] > 0$ is given by the variance of X on the distribution P_β (with $\beta = -h$). Indeed the steps above "map" a concave function $\phi(h)$ into another concave function $I(\bar{x})$, because you can easily check that $I''(\bar{x}) = 1/\phi''(h) > 0$.

As a general remark, note that the function $I(\bar{x})$ contains (and it has to be consistent with) both the law of large numbers and the central limit theorem. The first implies that $I(\bar{x}) = 0$ when $\bar{x} = \mathbb{E}_Q[X]$. The second that the pdf of $\bar{x}$ is well approximated by a Gaussian for $\bar{x} \simeq \mathbb{E}_Q[X]$, i.e.

$$p_n(\bar{x}) \simeq \sqrt{\frac{n}{2\pi \mathbb{V}_Q[X]}} e^{-\frac{n(\bar{x}-\mathbb{E}_Q[X])^2}{2\mathbb{V}_Q[X]}}.$$

[257] In the derivation above we had

$$I(\bar{x}) = -g(\beta^*), \quad h = -\beta$$

and $\phi(h) = \log Z(\beta)$. The reason for this change of notation will become clear in what follows.

Therefore, $I(\bar{x}) \simeq \frac{(\bar{x}-\mathbb{E}_Q[X])^2}{2\mathbb{V}_Q[X]} + \ldots$ is well approximated by a quadratic function for $\bar{x} \simeq \mathbb{E}_Q[X]$. This can indeed be checked explicitly, because the second derivative of $I(\bar{x})$ for $\bar{x} = \mathbb{E}_Q[X]$ is the inverse of the second derivative of the cumulant generating function $\phi(h)$, which is the variance $\mathbb{V}_Q[X]$.

Exercise: Compute the Cramer function $I(\bar{x})$ for the exponential distribution $p(x) = e^{-x}$, $x \geq 0$.

THE MATHEMATICS described here is that of Legendre transforms[258]. This mathematical construction does not arise accidentally. Consider the following constrained optimisation problem

$$I(\bar{x}) = \min_{P:\ x(P)=\bar{x}} U(P) \tag{338}$$

where $P \in \mathbb{R}^d$ is a d-dimensional vector and the function $U(P)$ is concave[259]. In the case of large deviations for distributions with finite support, P is a distribution, $U(P) = D_{KL}(P||Q)$ and $x(P) = \sum_x P(x)x$ is a linear function of P (an expected value). P identifies a point in the (x, U) plane, with $x = x(P)$, and the solution of the problem lies on the boundary in the (x, U) plane between points that can be achieved for some value of P and points that cannot be achieved. This boundary is the function $I(\bar{x})$ that we want to characterise (see Fig. 37).

[258] A warmly suggested reading on the Legendre transform, which discusses its geometric interpretation and gives much intuition on its nature, can be found in

Royce KP Zia, Edward F Redish, and Susan R McKay. Making sense of the legendre transform. *American Journal of Physics*, 77(7):614–622, 2009

[259] i.e.

$$U(\lambda P_1 + (1-\lambda)P_0) \leq \lambda U(P_1) + (1-\lambda)U(P_0).$$

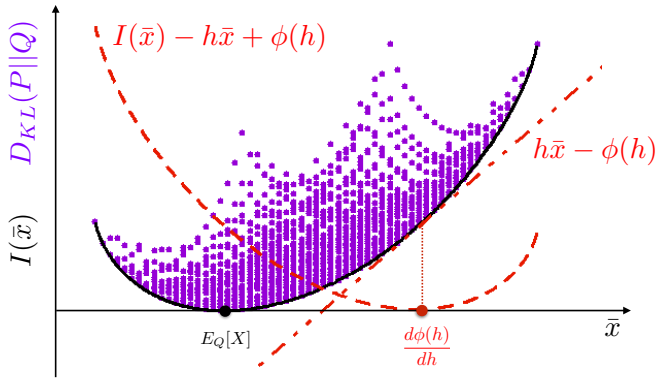


Figure 37: Construction of the large deviation function for $Q(x)$ defined for $x \in \chi = \{1, 2, 3, 4\}$ and $Q(1) = 2Q(2) = 4Q(3) = 4Q(4) = 1/2$ and $n = 20$ points. The red curves show the construction implied by the Legendre transform for $h = 1$.

With the introduction of Lagrange multipliers, we transform the problem in Eq. (338) into[260]

$$I(\bar{x}) = \min_P \max_h \{U(P) - h[x(P) - \bar{x}]\} \tag{339}$$

$$= \max_h \{\bar{x}h - \phi(h)\} \tag{340}$$

$$\phi(h) = \max_P \{hx(P) - U(P)\}. \tag{341}$$

In this way we relate the original optimisation problem Eq. (338) to a *dual* problem Eq. (341).

[260] The fact that the optimisation over h is a maximisation derives from the fact that it is the solution of the optimisation of a concave function $hx(P) - U(P)$. As $I(\bar{x})$ inherits its concavity from $U(P)$, $\phi(h)$ inherits its convexity from $hx(P) - U(P)$.

The meaning of h is clear if we consider the same problem, but for a value $\bar{x} + d\bar{x}$ of the constraint. Then if $P^*(\bar{x})$ is the point where the extreme is achieved,

$$I(\bar{x} + d\bar{x}) = U\left(P^*(\bar{x} + d\bar{x})\right) = U\left(P^*(\bar{x})\right) + \nabla_P U \cdot \delta P^* + \dots$$

where $\delta P^* = P^*(\bar{x} + d\bar{x}) - P^*(\bar{x})$. The first order conditions of the optimisation in Eq. (339) on P imply that $\nabla_P U = h \nabla_P x$. Hence the equation above reads $I(\bar{x} + d\bar{x}) = I(\bar{x}) + h \nabla_P x \delta P^* + \dots$. The equation $x(P(\bar{x})) = \bar{x}$, on the other hand, implies that $\nabla_P x \delta P^* = d\bar{x}$. These, taken together, show that

$$h = \frac{dI}{d\bar{x}}$$

is the slope of the tangent of the curve that is the locus of the set of solutions of the optimisation in the $(\bar{x}, U)$ plane. This set can equivalently be described by the coordinate h. Indeed, because of the concavity of $U(P)$, the function $h(\bar{x})$ is an increasing function. Furthermore, this description is totally equivalent to the one in terms of $\bar{x}$. If we let $P(h)$ be the solution of the problem in Eq. (341), then one has

$$\begin{aligned} \phi(h + dh) &= x(P(h + dh))(h + dh) - U(P(h + dh)) \\ &= \phi(h) + \bar{x}(h) dh + \left[h \nabla_P x - \nabla_P U\right] \delta P + \dots \end{aligned}$$

The term in braces vanishes because of the first order conditions of the problem in Eq. (341). Therefore one concludes that

$$\bar{x} = \frac{d\phi}{dh}.$$

Indeed the relation between $I(\bar{x})$ and $\phi(h)$ is completely symmetric, i.e.

$$I(\bar{x}) + \phi(h) = \bar{x} h,$$

so I is the Legendre transform of ϕ and ϕ is the Legendre transform of I. Indeed, notice that Eq. (341) can be rewritten as

$$\phi(h) = \max_{\bar{x}} \left[h\bar{x} - \min_{P:\ x(P) = \bar{x}} U(P) \right] = \max_{\bar{x}} \left[h\bar{x} - I(\bar{x}) \right]. \tag{342}$$

The Legendre transform is not a mere change of variables. Rather it is a mapping of the solution $(\bar{x}, I)$ of a constrained optimisation problem Eq. (338) into the solution (h, ϕ) of a dual unconstrained optimisation problem (Eq. 341). The Legendre transform provides a precise prescription for identifying the *conjugate* variable h that should be used in the transformed problem[261].

LET US ILLUSTRATE the properties of $I(\bar{x})$ for sums $S_n = \sum_{k=1}^n X_k$ of binary variables that take values $X_k = \pm 1$ with equal probability.

[261] The Legendre transform is the bread and butter of statistical mechanics. As we shall see, the thermodynamics of an isolated system is described by distributions of maximal entropy, which is called the *microcanonical ensemble*. In an isolated system the energy E is a constant of the motion and hence it is fixed, as well as the volume V and the number of particles. This problem can be related to the description of a system in equilibrium with its environment (the *heat bath*) removing the constraint on E. In this description, which is the *canonical ensemble*, the new variable is the temperature T and the objective function is the free energy $F = \langle E \rangle - TS$. Likewise, the constraint on fixed volume V can be removed with a Legendre transform that maps the problem in one where the pressure P is fixed, and the constraint on N can be removed introducing the chemical potential μ. As an **Exercise**, identify in each of these cases what are the variables $\bar{x}$ and h and what are the functions $I(\bar{x})$ and $\phi(h)$.

Then, both the recipe above and a direct calculation using Stirling's approximation of the binomial coefficient, show that[262]

$$I(\bar{x}) = \frac{1-\bar{x}}{2}\ln(1-\bar{x}) + \frac{1+\bar{x}}{2}\ln(1+\bar{x})$$

which is just the relative entropy between the distribution $P_\beta = (\frac{1-\bar{x}}{2}, \frac{1+\bar{x}}{2})$ and the uniform distribution $Q = (1/2, 1/2)$, as it should.

[262] **Exercise**: Compute $I(\bar{x})$ in both ways.

The expansion for $|\bar{x}| \ll 1$ yields $I(\bar{x}) \simeq \frac{1}{2}\bar{x}^2 + O(\bar{x}^4)$ for $\bar{x} \ll 1$, which is consistent with the law of large number and the central limit theorem for $|\bar{x}| \sim 1/\sqrt{n} \ll 1$. For larger values of $\bar{x}$ the function $I(\bar{x})$ provides much more informations on the large deviation properties of the mean S_n/n. Note that $I(\bar{x})$ is defined only for $\bar{x} \in [-1, 1]$. Indeed also $|S_n/n| \leq 1$ by definition in this case. Next note that $I(\pm 1) = \ln 2$, and indeed the probability that $S_n = \pm n$ is exactly 2^{-n}.

How much do we learn?

This section is a side remark, and it should be taken as a curiosity driven digression.

Let us go back to our discussion where the distribution Q encodes our current state of knowledge, i.e. our *theory*. The theory Q predicts that an observable X should take a value $\approx \mathbb{E}_Q[X]$. When we perform and experiment and measure X, the measurement may be consistent with this prediction or not. In the latter case we need to revise our theory Q and replace it by P_β, depending on the observed value $\bar{x}$ of X. How much do we learn?

The uncertainty is reduced from $\mathcal{H}[Q]$ to $\mathcal{H}[P_\beta]$. Hence the acquired information is

$$\begin{aligned} -\Delta\mathcal{H} &= \mathcal{H}[Q] - \mathcal{H}[P_\beta] && (343) \\ &= I(\bar{x}) + \mathbb{E}_Q\left[\left(e^{hX-\phi(h)} - 1\right)\log Q\right], && (344) \end{aligned}$$

where the second line results from a trite calculation using the results in previous sections[263]. The first term $I(\bar{x}) = D_{KL}(P_\beta||Q)$ quantifies how surprising the result of the experiment is. The second instead, has the form of a covariance[264] between $e^{hX-\phi(h)}$ and $\log Q$. Hence it depends on what observable X has been probed in the experiment. This allows us to ask, given Q, what quantity X should be probed in order for the experiment to be as informative as possible? Yet $\Delta\mathcal{H}$ also depends on h, i.e. on the observed value $\bar{x}$ of X. One way to address this question is to "explore the neighbourhood" of Q, searching for "directions" X where the reduction in uncertainty $\Delta\mathcal{H}$ increases faster. Hence we expand $\Delta\mathcal{H}$ for small values of h and, after some work, we find[265]

$$\begin{aligned} \Delta\mathcal{H} \simeq\ & -h\,\mathrm{Cov}_Q(X, \log Q) \\ & -\frac{1}{2}h^2\left\{\mathbb{V}_Q[X] + \mathrm{Cov}_Q\left[\left(X - \mathbb{E}_Q[X]\right)^2, \log Q\right]\right\} + O(h^3) \end{aligned}$$

[263] And it is left as an **Exercise**.

[264] Note that $\mathbb{E}_Q\left[e^{hX-\phi(h)}\right] = 1$.

[265] We remind that the covariance is defined as

$$\mathrm{Cov}_Q(X, Y) = \mathbb{E}_Q\left[(X - \mathbb{E}_Q[X])(Y - \mathbb{E}_Q[Y])\right]$$

where the index specifies that the expectation is taken with respect to Q.

which is an interesting result.

The leading linear term implies that the largest change in $\Delta\mathcal{H}$ occurs when $X = \log Q$, which is the X that maximises the covariance with $\log Q$. Note indeed that, by the Asymptotic Equipartition Property, the value of $-\log Q \approx \mathcal{H}[Q]$ permits to identify the set of typical outcomes.

The choice $X = \log Q$ explores the space of distributions along the curve of parametric distributions[266]

$$P_h(x) = \frac{1}{\mathbb{E}_Q\left[Q^h\right]} Q^{1+h}(x).$$

The change $\Delta\mathcal{H}$ can however be either positive or negative, depending on whether $h < 0$ or $h > 0$. In order to make sure that the measurement reduces the uncertainty on the system, the measured quantity X should be such that $\mathrm{Cov}_Q(X, \log Q) = 0$, so that the linear term vanishes.

The first term of order h^2 is $I(\bar{x}) \simeq \frac{1}{2}h^2\mathbb{V}_Q[X]$, which suggests that the most potentially surprising experiments, are those that probe quantities with large fluctuations. This is indeed a well established recipe in experimental design.

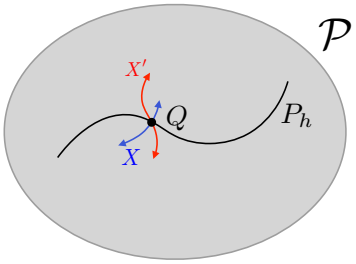


Figure 38: Probing the space of distributions around Q. Each experiments X explores the space along a different trajectory P_h.

[266] In a statistical mechanics analogy, as we shall see Q takes the form $Q(x) = \frac{1}{Z}e^{-E(x)/T}$, where T is the temperature. Then also $P_h(x)$ has the same form, with $T' = T/(1+h)$. In addition

$$\begin{aligned}\Delta\mathcal{H} &= -\frac{h}{T^2}\mathbb{V}_Q[E] \\ &- \frac{h^2}{2T^2}\left[\mathbb{V}_Q[E] + \frac{1}{T}\mathbb{E}_Q\left[(E - \mathbb{E}_Q[E])^3\right]\right] + \dots\end{aligned}$$

and the coefficient of the linear term in h is the specific heat.

Weakly correlated variables: Phase transitions and the Gartner-Ellis theorem

The results we have derived so far for large deviations extend to the case where the random variables X_i are weakly dependent. How weak the dependence will be clarified below[267].

Consider the following situation: we have a sample $X_1, \dots, X_n$ drawn i.i.d. from a distribution, but we're not sure what the distribution is. With probability a the sample comes from the distribution P and with probability $1 - a$ it comes form the distribution Q. Both P and Q have either finite support or thin tails. What is the probability $P\{A_n(\bar{x})\}$ in this case? Clearly

$$\mathbb{E}[X] = a\mathbb{E}_P[X] + (1-a)\mathbb{E}_Q[X],$$

where $\mathbb{E}_P[\dots]$ and $\mathbb{E}_Q[\dots]$ stand for expectations on the distributions P and Q, respectively. Do we expect that the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to a\mathbb{E}_P[X] + (1-a)\mathbb{E}_Q[X]$$

holds?

The answer can be found by a direct calculation:

$$\begin{aligned}P\{A_n(\bar{x})\} &= aP\{A_n(\bar{x})|P\} + (1-a)P\{A_n(\bar{x})|Q\} &(345)\\ &\backsim ae^{-nI_P(\bar{x})} + (1-a)e^{-nI_Q(\bar{x})} &(346)\end{aligned}$$

[267] To give an idea, one example where the theory applies is when random variables interact only "locally". This means that for each X_i there is a finite subset $\partial_i \subset \{1, \dots, n\}$ of indices such that, conditional on the values of the variables X_j for $j \in \partial_i$, X_i is independent of all the other variables $k \notin \partial_i$, i.e.

$$P\{X_i|X_j,\ \forall j \neq i\} = P\{X_i|X_j,\ \forall j \in \partial_i\}.$$

A Markov process, where X_i only depends on X_{i-1} and X_{i+1} (i.e. $\partial_i = \{i-1, i+1\}$) is a sequence of weakly dependent random variables.

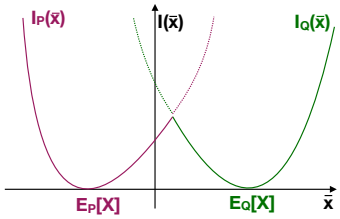


Figure 39: The construction of the Cramer function $I(\bar{x})$ for the example discussed in the text.

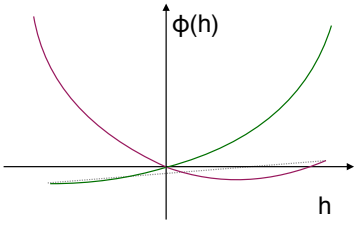


Figure 40: The functions ϕ_P and ϕ_Q for the example discussed in the text.

where $P\{A_n(\bar{x})|W\}$ is the probability of the large deviation, conditional on the assumption that the variables X_i are drawn i.i.d. from the distribution $W = P$ or Q, and

$$I_W(\bar{x}) = -\lim_{n\to\infty} \frac{1}{n} \log P\{A_n(\bar{x})|W\} = \min_{P\in\mathcal{A}_n(\bar{x})} D_{KL}(P||W)\,.$$

It is now clear that

$$I(\bar{x}) = -\lim_{n\to\infty} \frac{1}{n} \log P\{A_n(\bar{x})\} = \min[I_P(\bar{x}), I_Q(\bar{x})]. \tag{347}$$

Notice that:

- The curve $I(\bar{x})$ touches the $\bar{x}$ axis in *two* point $\bar{x} = \mathbb{E}_P[X]$ and $\bar{x} = \mathbb{E}_Q[X]$. This means that, typically we expect that the sample mean converges to either $\mathbb{E}_P[X]$ or to $\mathbb{E}_Q[X]$, but *not* to $\mathbb{E}[X]$. This violation of the law of large numbers occurs because the variables $X_1, \ldots, X_n$ are *not* independent. Indeed, knowledge of a subset k of the X_i allows us to infer whether the right distribution is P or Q, and hence informs us on the values of the remaining $n - k$.
- The curve $I(\bar{x})$ is not convex. Locally it is convex, apart from the point $\bar{x}_c$ where $I_P(\bar{x}_c) = I_Q(\bar{x}_c)$, where it has a cusp.
- The derivative h of $I(\bar{x})$ is no longer a continuous function of $\bar{x}$. Rather it has a jump at the point $\bar{x}_c$, i.e. $\lim_{x\to\bar{x}_c^{\pm}} I'(x) = h_{\pm}$.
- Following the geometric construction of the function $\phi(h)$, one finds that the function $\phi(h)$ is not single valued in the interval $h \in [h_+, h_-]$ and that it is not continuous.

The Maxwell construction and the Gärtner-Ellis theorem The fact that $I(\bar{x})$ derived above is non convex makes the recipe based on the Legendre transform, that we discussed for i.i.d. variables inapplicable. The *Gärtner-Ellis theorem* describes what happens if we apply this recipe anyhow. Suppose that the function

$$\bar{\phi}(h) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{h(X_1+\ldots+X_n)}\right] \tag{348}$$

exists and is finite, for h in a neighbourhood of the origin. Then the convex hull $\bar{I}(\bar{x})$ of the large deviation function is given by the Legendre transform of $\bar{\phi}(h)$.

Let us see how this works for the problem we discussed above, of a sequence $\underline{X}$ of variables which is drawn i.i.d. from either P or Q. It is easy to see that $\mathbb{E}\left[e^{h(X_1+\ldots+X_n)}|W\right] = e^{n\phi_W(h)}$, where $\phi_W(h)$ is drawn in Fig. 40 for $W = P$ or Q. Then

$$\begin{aligned} \bar{\phi}(h) &= \lim_{n\to\infty} \frac{1}{n} \log\left[a e^{n\phi_P(h)} + (1-a) e^{n\phi_Q(h)}\right] \\ &= \max\left[\phi_P(h), \phi_Q(h)\right] \end{aligned} \tag{349}$$

Exercise: Let $\underline{X} = (X_1, \ldots, X_n)$ where $X_i = Y_0 Y_i$, with $Y_0 = \pm 1$ with equal probability, and $Y_i \in \{0, 1\}$ are i.i.d. random variables with $P\{Y_i = 1\} = p = 1 - P\{Y_i = 0\}$, and they are all independent of Y_0. Compute the large deviation function for the random variables X_i, i.e.

$$I(\bar{x}) = -\lim_{n\to\infty} \frac{1}{n} \log P\left\{\sum_{i=1}^{n} Y_i \in [\bar{x}, \bar{x} + \epsilon)\right\}$$

for some $\epsilon > 0$. Compute the function $\bar{I}(\bar{x})$ by Gärtner-Ellis theorem, i.e. as the Legendre transform of $\bar{\phi}$.

as shown in Fig. 41(bottom).

Notice that $\bar{\phi}(h)$ has a cusp – i.e. a discontinuity in its first derivative – for $h = 0$. The derivative of $\bar{\phi}(h)$ as $h \to 0^+$ equals $\mathbb{E}_Q[X]$ whereas when $h \to 0^-$ one finds $\bar{\phi}'(h) = \mathbb{E}_P[X]$.

The Legendre transform $\bar{I}(\bar{x})$ of $\bar{\phi}(h)$ is shown in Fig. 41(top). This function $\bar{I}(\bar{x})$ is identical to $I(\bar{x})$, except for the part in the interval $\bar{x} \in [\mathbb{E}_P[x], \mathbb{E}_Q[x]]$, where $I(\bar{x})$ is replaced by a straight line.

The Gärtner-Ellis theorem provides the solution to a different yet related problem, which is the case where an unknown fraction of the variables are drawn from P and the rest from Q. Specifically, let X_i be drawn from P if $i \leq \nu n$ and from Q if $i > \nu n$, with $\nu \in [0,1]$ which is unknown.

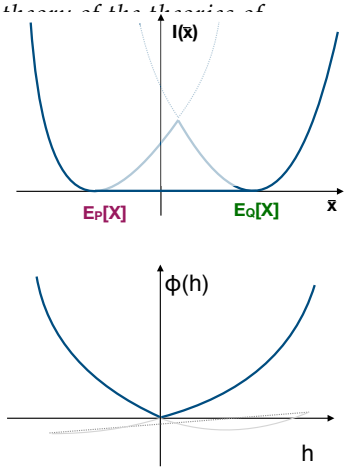


Figure 41: The Gärtner-Ellis theorem applied to the problem of a sequence $\underline{X}$ drawn i.i.d. from either P or Q.

Again we consider the event $A_n(\bar{x})$, i.e. that the mean of a sample $X_1, \dots X_n$ of points obtained in this way equals $\bar{x}$, and we want to compute the probability of $A_n(\bar{x})$. The probability of finding a large deviation with a sample mean equal to $\bar{x}$ is

$$\begin{aligned} P\{A_n(\bar{x})\} &= \int_0^1 d\nu \int d\bar{x}_P \int d\bar{x}_Q P\{A_{\nu n}(\bar{x}_P)|P\} P\{A_{(1-\nu)n}(\bar{x}_Q)|Q\} \delta(\bar{x} - \nu\bar{x}_P - (1-\nu)\bar{x}_Q) \\ &\sim \int_0^1 d\nu \int d\bar{x}_P \int d\bar{x}_Q e^{-n\left[\nu I_P(\bar{x}_P) + (1-\nu) I_Q(\bar{x}_Q)\right]} \delta(\bar{x} - \nu\bar{x}_P - (1-\nu)\bar{x}_Q) \end{aligned}$$

where we assume a uniform prior on ν. For all values of $\bar{x} \in [\mathbb{E}_P[X], \mathbb{E}_Q[X]]$ this multiple integral is dominated by the values $\bar{x}_P = \mathbb{E}_P[X]$ and $\bar{x}_Q = \mathbb{E}_Q[X]$, and ν such that $\bar{x} = \nu\mathbb{E}_P[X] + (1-\nu)\mathbb{E}_Q[X]$, because then $I_P(\bar{x}_P) = I_Q(\bar{x}_Q) = 0$, and one finds that

$$I_\nu(\bar{x}) = -\lim_{n\to\infty} \frac{1}{n} \log P\{A_n^{(\nu)}(\bar{x})\} = 0.$$

Put differently, for every $\bar{x} \in [\mathbb{E}_P[X], \mathbb{E}_Q[X]]$ it is possible to find a value

$$\nu = \frac{\mathbb{E}_Q[X] - \bar{x}}{\mathbb{E}_Q[X] - \mathbb{E}_P[X]} \in [0,1] \tag{350}$$

such that the above construction allows us to realise the large deviation $\bar{x}$ as a typical event (i.e. with $I_\nu(\bar{x}) = 0$).

As we're going to discuss (see footnote 313) the replacement of non-concave part of $I(\bar{x})$ with a straight line is conceptually identical to the Maxwell's construction in thermodynamics. In physics this construction relates the thermodynamics of homogenous but unstable states that that of inhomogeneous states, which are a mixture of two pure states. Here, it relates the (large deviation) properties of a system which is either in one *pure* state (P) or in another (Q), to one which is a *mixture* $P_\nu = \nu P + (1-\nu)Q$ of the two states. Mathematically, the first case is described by the Cramer function $I(\bar{x})$ while the mixture is described by its *convex hull* $\bar{I}(\bar{x})$, defined in Eq. (347), which is the Legendre transform of $\bar{\phi}(h)$ in Eq. (349)[268].

[268] **Exercise:** Consider yet a different problem where each of the variables X_i is drawn from P, with probability ν, or from Q with probability $1-\nu$. What is the large deviation function $I(\bar{x})$ in this case when ν is known and when ν is unknown?

Large deviations for Markov Chains

A further example of a sequence of weakly dependent random variables is given by Markov Chains. Let us recall that a Markov Chain $Z_0, Z_1, \ldots, Z_t, \ldots$ is a sequence of random variables that take values in a discrete set $\mathcal{S}$, and which is defined by a transition matrix

$$W_{s,s'} = P\{Z_t = s | Z_{t-1} = s'\}, \qquad s, s' \in \mathcal{S}. \tag{351}$$

We restrict our attention to irreducible Markov Chains for which the distribution $p\{Z_t = s\}$ converges, as $t \to \infty$, to the unique invariant measure μ_s which satisfies the equation $\mu_s = \sum_{s'} W_{s,s'} \mu_{s'}$.

For an observable X_t with a distribution $P\{X_t = x | Z_t = s\} = q(x|s)$ that depends only on the state Z_t at time t, we expect that its time average between times $\tau + 1$ and $\tau + N$ converges as $N \to \infty$ to the expected value of X_t on μ_s, for $\tau \to \infty$, i.e.

$$\lim_{\tau \to \infty} \lim_{N \to \infty} \frac{1}{N} \sum_{t=\tau+1}^{\tau+N} X_t \to \mathbb{E}_\mu [X_t] \equiv \sum_{x,s} x q(x|s) \mu_s.$$

What is the probability to observe instead a value $\bar{x}$ different from $\mathbb{E}_\mu [X_t]$? In order to apply Eq. (348) we need to compute the expected value

$$\mathbb{E}\left[e^{h \sum_t X_t}\right] = \sum_{s_\tau, s_{\tau+1}, \ldots, s_{\tau+N}} \prod_{t=\tau+1}^{\tau+N} W_{s_t, s_{t-1}} \mathbb{E}\left[e^{h X_t} | s_t\right] P_0(s_\tau) \tag{352}$$

$$= \sum_{s_\tau, s_{\tau+N}} \{\hat{U}^N\}_{s_{\tau+N}, s_\tau} P_0(s_\tau), \tag{353}$$

where $\{\hat{U}^N\}_{s_{\tau+N}, s_\tau}$ is the $s_{\tau+N}, s_\tau$ element of the N^{th} power of the matrix $U_{s,s'} = \mathbb{E}\left[e^{h X_t} | s\right] W_{s,s'}$. In the repeated matrix multiplication, the dominant component is the one corresponding to the largest eigenvalue of $\hat{U}$, corresponding to the right eigenvector

$$\lambda v_s = \sum_{s'} U_{s,s'} v_{s'} = \sum_{s'} \mathbb{E}\left[e^{h X_t} | s\right] W_{s,s'} v_{s'} \tag{354}$$

which leads to $\mathbb{E}\left[e^{h \sum_t X_t}\right] \sim \lambda^N$. Note that, by virtue of the Perron-Frobenius theorem, λ and all components of v_s are positive, because $U_{s,s'} \geq 0$ for all s, s'. Hence the limit in Eq. (348) leads to $\bar{\phi}(h) = \log \lambda$.

Summarising, the recipe of large deviations for a Markov Chain is *i)* compute the matrix $\hat{U}$, *ii)* compute its largest eigenvalue λ as a function of h, *iii)* compute the rate function $\bar{I}(\bar{x})$ from the Legendre transform of $\bar{\phi}(h) = \log \lambda$. The distribution of Z_t conditional on the large deviation is given by the normalised right eigenvector

$$P\{Z_t = s | A_n(\bar{x})\} = \frac{v_s}{\sum_{s'} v_{s'}}$$

(where h is the solution of $\frac{d\bar{\phi}}{dh} = \bar{x}$). Note that when $h \to 0$, this distribution reverts back to the invariant measure μ_s.

Evolution as a large deviation Let us take a branching process as a simple model of evolution. In order to allow for mutations, let us assume that the probability $p(k|f)$ that an individual generates k offsprings depends on its fitness f and that each offspring has a different fitness f'. We define the fitness as the logarithm of the expected number of offsprings, i.e. $f = \log \mathbb{E}[X_i|f]$. If the fitness is drawn at random independently from the same distribution $p(f)$ for each offspring, then it is easy to see that[269] that the total population after n generations grows as $\mathbb{E}[Z_n] = e^{n\mathbb{E}[f]}$ where $\mathbb{E}[f] = \sum_f f p(f)$ is the expected fitness. However, if the fitness is inherited, in part, from the parents then the growth rate of the population is larger than $\mathbb{E}[f]$. In order to see this, consider the case where the fitness is a function $f(s)$ of the "type" s of individual and assume that the offsprings of an individual of type s' end up of type s with probability $W_{s,s'}$. In this way, their fitness will depend on the fitness of the parent. Let $Z_s^{(n)}$ be the number of individuals of type s at generation n. Then

[269] **Exercise**: do that.

$$\mathbb{E}\left[Z_s^{(n)}\right] = \sum_{s'} W_{s,s'} e^{f(s')} \mathbb{E}\left[Z_{s'}^{(n-1)}\right].$$

The vector $y_s^{(n)} = e^{f(s)}\mathbb{E}\left[Z_s^{(n)}\right]$ of the expected number of offsprings from parents of type s at generation n, satisfies the equation

$$\vec{y}^{(n)} = \hat{U}\vec{y}^{(n-1)} = \hat{U}^N \vec{y}^{(0)}$$

where the matrix $\hat{U}$ has elements $U_{s,s'} = e^{f(s)} W_{s,s'}$. This implies that the growth rate of the population is controlled by the largest eigenvalue λ of the matrix $\hat{U}$, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[Z^{(n)}\right] = \lambda.$$

The corresponding eigenvector satisfies the analogue equation to (354)

$$\lambda v_s = \sum_{s'} U_{s,s'} v_{s'} = \sum_{s'} e^{f(s)} W_{s,s'} v_{s'}$$

This is equivalent to a large deviation principle for the variable $X(s_t) = f(s)$ for the Markov chain with transition probability $W_{s,s'}$ (with $h = 1$). In particular, when the offsprings of fit individuals are likely fit, i.e. when $W_{s,s'}$ is close to diagonal, we expect that the growth rate to be larger than $\mathbb{E}[f(s)]$. Also, we expect that the fraction of individuals of type s at generation $n \gg 1$ is asymptotically given by

$$\tilde{\mu}_s = \frac{\mathbb{E}\left[Z_s^{(n)}\right]}{\sum_{s'} \mathbb{E}\left[Z_{s'}^{(n)}\right]} \simeq \frac{v_s e^{-f(s)}}{\sum_{s'} v_{s'} e^{-f(s')}}.$$

Notice also that this is also the fraction of individuals of type s that we expect to find tracing back the types of ancestors of an individual

at generation $n \gg 1$. This is in general different from the invariant measure μ_s, which corresponds to the fraction of descendants of type s of an individual[270].

[270] **Exercise:** Compute λ and $\vec{v}$ in a branching process with two types $s = \pm 1$ and $f(s) = f + gs$, in the case $W_{s,s} = 1 - \epsilon$ and $W_{s,-s} = \epsilon$.

Large deviations for fat tailed distributions

The Cramer function $I(\bar{x})$ has the property that it is positive and it vanishes for $\bar{x} = E[X]$, which corresponds to the point $h = 0$. The machinery above works if $\phi(h)$ exists at least for h in an open neighbourhood of the origin. This requires that the pdf of X decays at least as an exponential for $|x| \to \infty$. What happens if this is not true?

We shall call fat tailed distribution any distribution $Q(x)$ for which

$$\lim_{|x|\to\infty} \frac{1}{|x|} \log Q(x) = 0 \tag{355}$$

for $x \to +\infty$ or $x \to +\infty$, or both. In this limit, $e^{hx}Q(x)$ diverges for at least one value of h in the neighbourhood of $h = 0$ as $x \to \pm\infty$.

For simplicity, we focus on the right tail of the pdf, and assume that $Q(x)$ vanishes at least exponentially fast as $x \to -\infty$. This includes stretched exponential distributions $Q(x) \sim e^{-ax^\alpha}$ with $\alpha < 1$ and power law distributions $Q(x) \sim Ax^{-\gamma}$ for $x \gg 1$. Again we focus on the event

$$A_n(\bar{x}) = \left\{ \underline{X} : \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \bar{x} \right| < \epsilon \right\}$$

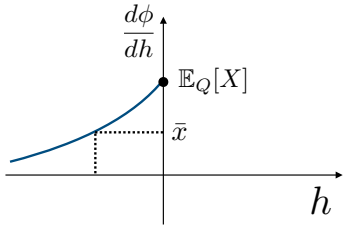
for some arbitrarily small $\epsilon > 0$ and our goal is to compute the Cramer's function $I(\bar{x})$ in Eq. (336). For $h \leq 0$ we can follow the recipe outlined in the previous sections because $\mathbb{E}\left[e^{hX}\right]$, and hence $\phi(h)$, is finite. This allows us to define the Cramer function $I(\bar{x})$ for all $\bar{x} \leq \mathbb{E}_Q[X]$, which is expected to vanish as $\bar{x} \to \mathbb{E}_Q[X]$ with a quadratic behaviour $I(\bar{x}) \simeq \frac{1}{2\mathbb{V}_Q[X]} \left(\bar{x} - \mathbb{E}_Q[X]\right)^2 + \dots$ for $\bar{x} \lesssim \mathbb{E}_Q[X]$.

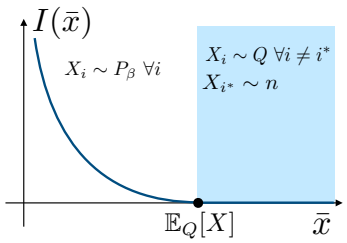
In order to explore the behaviour of $I(\bar{x})$ for $\bar{x} > \mathbb{E}_Q[X]$, let us consider the event

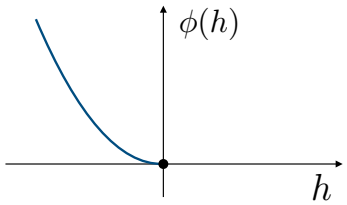
$$\begin{aligned} \tilde{A}_n(\bar{x}) &= \bigcup_{i^*=1}^{n} \left\{ \underline{X} : \left| \frac{1}{n-1} \sum_{i \neq i^*} X_i - \mathbb{E}_Q[X] \right| < \epsilon,\ X_{i^*} = x_n^* \right\} \\ x_n^* &= n\bar{x} - (n-1)\mathbb{E}_Q[X] \end{aligned} \tag{356}$$

In words, the event $\tilde{A}_n(\bar{x})$ describes a large deviation where the mean $\frac{1}{n}\sum_i X_i = \bar{x}$ deviates from the expected value $\mathbb{E}_Q[X]$, but all the excess of the mean is concentrated on only one variable $X_{i^*} = x_n^*$, which is proportional to n, whereas all the other variables are "typical", i.e. $X_i \approx \mathbb{E}_Q[X]$. The probability of this event is

$$P\{\tilde{A}_n(\bar{x})\} \geq (1-\epsilon)nQ\left(n\bar{x} - (n-1)\mathbb{E}_Q[X]\right)$$







where the factor $1-\epsilon$ comes from the fact that the $n-1$ variables $i \neq i^*$ take typical values, the factor n accounts for the fact that i^* can take n values, and the last factor is the probability of X_{i^*}.

The event $\tilde{A}_n(\bar{x})$ is only one way in which the large deviation can occur, therefore $\tilde{A}_n(\bar{x}) \subseteq A_n(\bar{x})$. As a consequence $P\{A_n(\bar{x})\} \geq P\{\tilde{A}_n(\bar{x})\}$ and

$$I(\bar{x}) = -\lim_{n\to\infty} \frac{1}{n} \log P\{A_n(\bar{x})\} \tag{357}$$

$$\leq -\lim_{n\to\infty} \frac{1}{n} \log P\{\tilde{A}_n(\bar{x})\} = 0 \tag{358}$$

where the last equality is a consequence of Eq. (355). Therefore, for all $\bar{x} \geq \mathbb{E}_Q[X]$ the Cramer function vanishes, $I(\bar{x}) = 0$.

In loose words, "democratic" ways to realise large deviations, where $\bar{x}$ is obtained as the average of i.i.d. draws from a modified distribution, are not typical. For fat tailed distributions, large deviations typically concentrate on a single variable X_{i^*} which is responsible for the whole excess of the mean $\bar{x}$. The symmetry between the variables, which are identically distributed *a priori*, is *broken spontaneously*, because one of them takes an *extensive* value (i.e. a value proportional to n). *Spontaneously* refers to the fact that, *a priori*, any variable X_{i^*} can carry the excess deviation.

The fact that $I(\bar{x}) = 0$ for all $\bar{x} \geq \mathbb{E}_Q[X]$ implies that $I(\bar{x})$ has a singularity at $\bar{x} = \mathbb{E}_Q[X]$ in the second derivative. This is the analogue of a second order phase transition in statistical physics[271], that generally occur when a symmetry of the system is spontaneously broken[272], precisely as in the current situation where the *a priori* (permutation) symmetry between the variables X_i is broken.

Note that when the pdf of X decays slower than $|x|^{-2}$ the expected value of X diverges. Then "large deviations" occur typically, and, as we have seen, they occur with one variable being of the same order of the whole sum. So also in that case, the large deviation concentrates on few variables[273].

[271] In thermodynamics, the order of a transition is defined as the order of the derivative that develops a singularity at the critical point. As we shall see, $I(\bar{x})$ is the analog of the entropy in statistical mechanics.

[272] The typical example is the spontaneous magnetisation of metals when the temperature is decreased below the Curie temperature.

[273] In the special case where X are Cauchy variables $p(x) = \pi^{-1}(1+x^2)^{-1}$, you can check that the $\sum_i X_i/n$ is itself a Cauchy variable. Therefore the probability of a large deviation

$$P\{A_n(\bar{x})\} = \frac{1}{\pi}\frac{1}{1+\bar{x}^2}$$

does not decay exponentially with n. Actually it does not decay at all.

States of knowledge

Now that we have a quantitative notion of information, we can address the problem of finding distributions that are consistent with a given state of knowledge. Just like Socrates has been claimed to say that

> The only true wisdom is in knowing you know nothing

it seems the only state of knowledge we can precisely identify is the one where we "know nothing". If lack of information can be measured by the entropy, the state where we know nothing corresponds to a probability distribution of maximal entropy. In addition, as we shall see, large deviation theory allows us to be precise in understanding how new information can be incorporated in our current state of knowledge (i.e. in probability distributions). This "becomes a methodology for a very general type of scientific reasoning", as claimed by E. T. Jaynes[274]. We shall discuss this general approach and then, statistical mechanics as one of its particular applications.

[274] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957

Maximum entropy

Consider the case of a discrete random variable $X \in \chi$ drawn from a finite set χ. The state of maximal ignorance corresponds to a distribution $p(x) = P\{X = x\}$ of maximal entropy[275]

$$p(x) = \frac{1}{|\chi|}. \tag{359}$$

[275] You can show this by studying the maximisation of $\mathcal{H}[p]$ with the normalisation constraint $\sum_x p(x) = 1$.

Indeed, in order to dispel uncertainty the number of binary questions we need to ask is as large as possible[276], i.e. $H[X] = \log_2 |\chi|$. The state of maximal ignorance is also such that the distribution of X is invariant under any permutation of the possible values $x \in \chi$. This is consistent with a state of knowledge where we don't know anything that can distinguish event $\{X = x\}$ from event $\{X = x'\}$.

[276] In this case, the optimal way to elicit information is to ask questions that split the number of possible alternatives in half each time. If $|\chi| = 2^H$, there are $\binom{2^H}{2^{H-1}}$ ways to choose how to make the first question, $\binom{2^{H-1}}{2^{H-2}}^2$ ways to pose the second and so on. In total there are

$$\mathcal{N} = \prod_{k=0}^{H-k} \binom{2^{H-k-1}}{2^{H-2}}^{2^k}$$

ways to ask the H questions. Which of these ways one choses to ask questions is irrelevant. The fact that this number is so large means that we have no clue of how to pose questions in a smart way.

Now assume that we know that[277]

$$\mathbb{E}\left[F(X)\right] = \sum_{x\in\chi} p(x)F(x) = f \tag{360}$$

[277] This knowledge may come from the fact that, in a series of $N \gg 1$ independent experiments where we measure the variables $Y_i = F(X_i)$ for $i = 1, \ldots, N$, we observe that

$$\frac{1}{N}\sum_{i=1}^{N} F(X_i) \simeq f\,,$$

and that we expect the Law of Large Numbers to hold. It may also come from the fact that we expect that $\mathbb{E}\left[F(X)\right] = f$ based on theoretical grounds.

for a function $F(X)$. Then the distribution that encodes this and only this information, is given by the one that maximises the entropy, subject to these constraints. This implies that we have to solve the problem:

$$\max_{p,\lambda,\nu} \left\{ -\sum_{x\in\mathcal{X}} p(x)\log p(x) + \lambda\left[\sum_{x\in\mathcal{X}} p(x)F(x) - f\right] + \nu\left[\sum_{x\in\mathcal{X}} p(x) - 1\right]\right\}.$$

The solution is

$$p_\lambda(x) = \frac{1}{Z(\lambda)} e^{\lambda F(x)} \tag{361}$$

where $Z(\lambda)$ ensures normalisation, and the value of λ should be adjusted in such a way that Eq. (360) is satisfied, i.e.

$$\mathbb{E}\left[F(X)\right] = \frac{d\log Z}{d\lambda} = f\,. \tag{362}$$

Note that the solution to this problem is unique. The way to show this is to observe that λ is the solution of a convex optimisation problem. Indeed Eqs. (362) correspond to the first order condition of the maximisation of the entropy as a function of λ

$$\Sigma(\lambda) = \mathcal{H}[p_\lambda] = \log Z(\lambda) - \lambda\mathbb{E}\left[F(X)\right]\,.$$

where $\mathbb{E}\left[F(X)\right]$ is a function of λ. Note that

$$\frac{d\Sigma}{d\lambda} = -\lambda\frac{d\mathbb{E}\left[F(X)\right]}{d\lambda} = -\lambda\mathbb{V}\left[F(X)\right]$$

has the opposite sign of λ, where $\mathbb{V}\left[F(X)\right] \geq 0$ is the variance of $F(X)$ under the distribution p_λ. So $\Sigma(\lambda)$ has a unique maximum at $\lambda = 0$, because it increases for $\lambda < 0$ and it decreases for $\lambda > 0$.

YET IT IS IMPORTANT to stress that the entropy

$$S(f) = \max_{p:\mathbb{E}[F(X)]=f} \mathcal{H}[p] \tag{363}$$

is a function of f, which is the independent variable. The variables f and λ are conjugate under the Legendre transform that maps the problem Eq. (363) into the conjugate problem[278]

$$\psi(\lambda) = \min_p\left[-\mathcal{H}[p] - \lambda\mathbb{E}\left[F\right]\right] \tag{364}$$

The solution of Eq. (363) is given by $S(f) = \log Z(\lambda) - \lambda f$, where $\lambda = \lambda(f)$ is given by the solution of Eq. (362), whereas the solution of Eq. (364) is given by

$$\psi(\lambda) = \min_f[-S(f) - \lambda f] = -\log Z(\lambda). \tag{365}$$

The function ψ is not an entropy[279]. It is called a *free energy*.

[278] This follows from

$$\begin{aligned} S(f) &= \min_\lambda \max_p \left\{\mathcal{H}[p] + \lambda(\mathbb{E}\left[F\right] - f)\right\} \\ &= \min_\lambda\left\{-\lambda f - \min_p\left[-\mathcal{H}[p] - \lambda\mathbb{E}\left[F\right]\right]\right\} \\ &= \min_\lambda\left\{-\lambda f - \psi(\lambda)\right\} \end{aligned}$$

[279] Note that $-\psi(\lambda)$ is the cumulant generating function of the random variable $F(X)$.

Summarising, the maximisation of the entropy at a fixed value of $f = \mathbb{E}[F]$ corresponds to the minimisation of the free energy ψ at a fixed value of the conjugate parameter λ. Because of this

$$\lambda(f) = -\frac{dS}{df} \qquad \text{and} \qquad f(\lambda) = -\frac{d\psi}{d\lambda} \tag{366}$$

and the functions S and ψ stand in the relation[280] $S + \psi = -\lambda f$.

This construction generalises in a straightforward manner to the case where $F(X) = (F_1(X), \dots, F_K(X))$ is a vector of K observables and $f = (f_1, \dots, f_K)$ is a vector of measurements. The solution of the maximisation of the entropy is again given by Eq. (361) with $\lambda = (\lambda_1, \dots, \lambda_K)$ being a vector of parameters, fixed by Eqs. (362), where the derivative is replaced by the gradient, and $\lambda F(x) = \sum_k \lambda_k F_k(x)$ is given by the dot product.

[280] **Exercise:** When $Q(x) = 1/|\chi|$ the construction discussed in this section is identical to the one we have followed in large deviation theory. What is the relation between the parameters $h, \bar{x}$ and λ, f, and between the functions I, ϕ and S, ψ?

There are several ways to see that Eq . (361) is the correct choice for the probability of X that encodes only the information that $\mathbb{E}[F(X)] = f$, as discussed in [281]. Let us discuss one of them. Imagine the situation where you have a sample of $n \gg 1$ values of X, that you think are drawn from a distribution $p(x)$. Then the analogous of Eq. (360) is

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} F(X_i) = \sum_{x \in \chi} P_{\underline{X}}(x) F(x) \,. \tag{367}$$

[281] Y Tikochinsky, NZ Tishby, and Raphael David Levine. Alternative approach to maximum-entropy inference. *Physical Review A*, 30(5):2638, 1984

where $P_{\underline{X}}(x)$ is the fraction of times that the outcome x occurs in the sample $\underline{X} = (X_1, \dots, X_n)$. The number of samples $\underline{X}$ that correspond to a given $P_{\underline{X}} = P$ is

$$\left| \{ \underline{X} : \; P_{\underline{X}} = P \} \right| = \frac{n!}{\prod_x [nP(x)]!} \simeq e^{n\mathcal{H}[P]}$$

where the second relation is a trite application of Stirling's formula. Then it is clear that, among all the possible distributions P that are consistent with Eq. (367) those for which $H[P]$ is maximal correspond to an overwhelmingly larger number of samples. So the probability that the observed sample is not one of these, is negligibly small as $n \to \infty$.

Distributions of maximal entropy are special because the probability of a sample $\underline{X} = (X_1, \dots, X_n)$

$$p^{(k)}(\underline{X}) = \frac{1}{Z^n} \exp\left\{ \sum_k \lambda_k \sum_{i=1}^{n} F_k(X_i) \right\}$$

depends on the data *only* through the empirical averages

$$\hat{f}_k(\underline{X}) = \frac{1}{n} \sum_{i=1}^{n} F_k(X_i)$$

of F_k. Therefore these averages contain all the information that is needed to identify the parameters λ of the distribution $p^{(k)}$. All other information in the sample is uninformative noise. This is why the empirical averages $\hat{f}_k$ are called *sufficient statistics*. This should not be surprising. Indeed, the distribution $p^{(k)}$ has been derived precisely as the one that encodes the state of knowledge in which the values of F, and only these, are known.

Generalised thermodynamics

EQUILIBRIUM: The principle of maximum entropy can also be applied to a system composed of two or more parts, of which we know the value of an aggregate quantity. More precisely, let X_1 and X_2 be two random variables and imagine that the value

$$f = \mathbb{E}\left[\nu_1 F(X_1) + \nu_2 F(X_2)\right]$$

is known. Here $\nu_\ell \geq 0$ can be thought of as the size of system X_ℓ, for $\ell = 1, 2$[282]. The same analysis as before, implies that the maximum entropy prediction for the combined system is

$$p(x_1, x_2) = \frac{1}{Z(\lambda)} e^{\lambda[\nu_1 F(x_1) + \nu_2 F(x_2)]} = p_{\nu_1 \lambda}(x_1) p_{\nu_2 \lambda}(x_2).$$

In other words, in the absence of further information, we need to assume that X_1 and X_2 are independent and that their distributions are given by Eq. (361). In order to check that this is consistent with the construction above, let $S_\ell(f_\ell)$ be the maximum entropy of X_ℓ given that $\mathbb{E}\left[F(X_\ell)\right] = f_\ell$, for $\ell = 1, 2$. Then it must be that

$$S(f) = \max_{f_1} \left[S_1(f_1) + S_2\left(\frac{f - \nu_1 f_1}{\nu_2}\right) \right],$$

where the argument $f_2 = \frac{f - \nu_1 f_1}{\nu_2}$ of S_2 is determined by the constraint $f = \nu_1 f_1 + \nu_2 f_2$. The first order condition of this maximisation, implies that

$$\frac{1}{\nu_1}\frac{dS_1}{df_1} \equiv \frac{\lambda_1}{\nu_1} = \frac{1}{\nu_2}\frac{dS_2}{df_2} \equiv \frac{\lambda_2}{\nu_2} = \lambda\,. \tag{368}$$

In words, the maximum entropy principle is associated to a notion of *equilibrium* where each of the parts has the same value of λ_ℓ / ν_ℓ. This generalises to systems composed of many parts X_ℓ, $\ell = 1, \ldots, L$ in a straightforward manner[283].

THE FIRST LAW OF THERMODYNAMICS: Consider now a slightly different problem where the observables $F_k(X)$ change slightly, i.e. $F_k \to F_k + \delta F_k$ and the measurement also changes $f_k(x) \to f_k(x) + \delta f_k(x)$.

[282] Variables that are proportional to the size of the system ν_ℓ are called *extensive*. Examples in physics include the entropy, the volume, the energy and the number of particles. Variables that are independent of the system's size – such as the temperature, the pressure and the particle density – are called *intensive*.

[283] In words, in equilibrium all the intensive variables take the same value in each part of the subsystem.

This transformation involves an arbitrary (infinitesimal) change of both the "internal" parameters F_k and of the "external" variables f_k, and it can be regarded as a generalised infinitesimal "thermodynamic" transformation. The new system is described by new parameters $\lambda'_k = \lambda_k + \delta\lambda_k$, which are again given by the solution of Eqs. (362). The change in the entropy, to leading order, can be written as[284]

$$\delta\mathcal{H} = \mathcal{H}[p_{\lambda+\delta\lambda}] - \mathcal{H}[p_\lambda] \simeq \sum_{k=1}^{K} \lambda_k \delta Q_k \tag{369}$$

where

$$\delta Q_k = -\delta f_k + \mathbb{E}\left[\delta F_k(X)\right] \tag{370}$$

is a generalised "heat", that is composed of two parts. The first is due to to the action δf_k of the external variables on the system and the second is the internal change of the observables. Put differently, the change δf_k of f_k in any transformation between maximum entropy states is given by two terms, one is the "work" $\mathbb{E}\left[\delta F_k(X)\right]$ done on the system and the other is due to the change δQ_k in the information content. Eq. (370) is the analog of the *first law of thermodynamics* in physics.

[284] The entropy at the maximum is given by

$$\mathcal{H}[p_\lambda] = -\lambda f + \log Z(\lambda)$$

where $\lambda f = \sum_k \lambda_k F_k$ stands for the scalar product. The change in the first term is given by $\delta(\lambda f) = \delta\lambda\, f + \lambda\, \delta f$. The change in the second term instead is given by $\delta \log Z = \delta\lambda \mathbb{E}\left[F\right] + \lambda \mathbb{E}\left[\delta F\right]$, where expected values are taken with respect to p_λ, and hence $\mathbb{E}\left[F\right] = f$ so that the terms proportional to $\delta\lambda$ cancel.

Maxent learning

Maximal entropy – sometimes called *maxent* – provides a procedure to learn theories from data. Imagine we're interested to acquire knowledge about an unknown quantity X, that we know takes values in a finite set $X \in \chi$. Our goal is to learn the distribution $p(x) = P\{X = x\}$ and to reduce our uncertainty about X. If we're in a state of total ignorance about X then our starting point is the maximum entropy distribution $p^{(0)}(x) = 1/|\chi|$. Imagine that we make an experiment and measure[285] the observable $\mathbb{E}\left[Y_1\right] = \mathbb{E}\left[f_1(X)\right]$. If the value $f_1 = \mathbb{E}\left[Y_1\right]$ that we obtain is consistent with the theory, i.e. if

$$f_1 = \sum_{x\in\chi} p^{(0)}(x) F_1(x)$$

then the experiment confirms the theory. If it does not, then, in order to include this observation, the theory has to be modified as

$$p^{(1)}(x) = \frac{1}{Z^{(1)}(\lambda_1)} e^{\lambda_1 F_1(x)},$$

where λ_1 has to be fixed so that $\sum_{x\in\chi} p^{(1)}(x) F_1(x) = \frac{\partial \log Z^{(1)}}{\partial \lambda_1} = f_1$. This procedure can be repeated by performing further experiments on other observables $Y_k = F_k(X)$, for $k = 2, 3, \ldots$. At each step, if the prediction of the current theory $p^{(k-1)}$ does not match the outcome f_k of the experiment, i.e. if $f_k \neq \sum_{x\in\chi} p^{(k-1)}(x) F_k(x)$, then the theory has to be refined $p^{(k-1)} \mapsto p^{(k)}$ with the procedure given above. In

[285] For example, we can take a sample $\underline{Y}_1 = (Y_1^{(1)}, \ldots, Y_1^{(N)})$ and estimate

$$\mathbb{E}\left[Y\right] \simeq \frac{1}{N} \sum_{i=1}^{N} Y_1^{(i)},$$

if N is very large.

this way the theory $p^{(k)}$ encodes, at each step, all the knowledge that has been accumulated in past experiments. Notice that if $\lambda_k = 0$ then $p^{(k)} = p^{(k-1)}$.

The entropy $\mathcal{H}[p^{(k)}]$ is clearly a non-increasing function of k, so it generally decreases in the process of refining the theory[286]. The difference $H[p^{(k-1)}] - H[p^{(k)}]$ is the amount of information that is learned in the k^{th} step.

[286] Remember our discussion on the mutual information: the knowledge of a random variable Y decreases our uncertainty on X *a priori*, but *a posteriori* there may be values of Y such that the entropy of X is actually larger. Why is this not the case in the situation we're discussing here?

Continuous variables

It seems natural to generalise the discussion above to continuous variables X with pdf $p(x)$, by adopting the differential entropy $h[X]$ instead of $H[X]$ and replacing partial with functional derivatives. So, for example, the distribution of maximal (differential) entropy for $X \in [0, \infty)$ with $E[X] = \mu$ is the exponential $p(x) = \mu^{-1} e^{-x/\mu}$ and the distribution of maximal entropy for $X \in \mathbb{R}$ with $E[X] = \mu$ and $V[X] = \sigma^2$ is the Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

The main problem with this approach is that re-parametrisation invariance is lost. Imagine two observers that want to make inference on the same system and measure the same quantity ϕ. Yet the first observer represent the observables $\phi(x)$ as a function of x and the second as a function of y, where $y = f(x)$, with $f(x)$ a strictly increasing function of x. Hence, the second observer represents the same quantity with a different function $\tilde{\phi}(y) = \phi(f^{-1}(y))$. On the basis of the same data $\underline{\phi} = (\phi_1, \ldots, \phi_n)$ and the same measurement

$$\bar{\phi} = \frac{1}{n} \sum_{i=1}^{n} \phi_i$$

their states of knowledge would be encoded in the two distributions

$$p(x) = \frac{1}{Z} e^{\theta \phi(x)} , \qquad \tilde{p}(y) = \frac{1}{\tilde{Z}} e^{\tilde{\theta} \tilde{\phi}(y)}$$

respectively, where we assume that the two distributions are normalisable (i.e. $Z, \tilde{Z} < +\infty$). Yet these correspond to two different states of knowledge. Indeed, by a change of variable, the pdf $\tilde{p}(y)$ for the second observer would correspond to

$$\tilde{p}(x) = \tilde{p}\left(f(x)\right) \frac{df(x)}{dx} = \frac{1}{\tilde{Z}} e^{\tilde{\theta} \phi(x)} \frac{df(x)}{dx}$$

which is different from $p(x)$. Indeed it is not even a maximum (differential) entropy distribution[287]. Indeed, the two observers maximise two different functions $h[X]$ and $h[Y]$ subject to the same constraint. It is

[287] For discrete variables X this problem does not arise. Both observers assign the same probabilities to corresponding values of X and Y, because f is a bijection between discrete values.

no wonder that their states of knowledge are different. The problem is that for continuous variables the differential entropy does not provide a way to encode a state of complete ignorance, rather it allows us only to quantify changes in our state of knowledge. The issue of how to represent, from first principles, a state of ignorance for continuous variables, corresponds to the problem of choosing the non-informative prior in Bayesian statistics that is discussed in [288]. The bottom line is that, when possible, symmetries of the problem can be used to determine the prior. In order to give a flavour of the argument, imagine we want to find the pdf $p_0(x)$ that encodes the state of complete ignorance for a random variable $X \in \mathbb{R}$. We shall call this a *prior* because this pdf represents what is known on X *before* we make any measurement. Imagine two observers, one that measures the variable X and the other that measures $Y = X + a$, with $a \in \mathbb{R}$ a constant. Because of translation invariance, the state of knowledge of the two observers must be the same, they both have no clue of what the value of X (or Y) is, i.e. they should both use the same prior p_0. They must also assign the same probability $p_0(x)dx = p_0(y)dy$ to the same intervals of X. This means that $p_0(x) = p_0(x + a)$ for all values of a, which means that[289]

$$p_0(x) = c$$

must be a constant. The problem is that in order for this pdf to be normalisable one should have $c \to 0$, i.e. $p_0(x)$ is an *improper prior*. In order to understand the origin of the problem, let's go back to the discrete case. There, the state of complete ignorance is the one which is further away from the state of complete knowledge $X = x$, in terms of the minimal number of binary questions that need to be asked to determine X. If X is continuous, it is clear that the minimal number of binary questions should be infinite. This tallies with the fact that, when symmetries can be used, one finds improper (i.e. non normalisable) priors, i.e. priors for which $h[X] = +\infty$.

Even if it is disturbing, the fact that p_0 is not normalisibale, does not prevents us from using it in learning. Imagine indeed that we collect a sample $\underline{\phi}$ of n independent observations of the variable $\phi(X)$, and we observe that

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) = \bar{\phi}.$$

Then we can use the machinery of large deviation theory to incorporate this information in the state of knowledge p_0. Formally, the updated state of knowledge now would read

$$p(x|\bar{\phi}) = \frac{p_0(x)e^{\lambda\phi(x)}}{Z(\lambda)}, \qquad Z(\lambda) = \int_{-\infty}^{\infty} dx p_0(x) e^{\lambda\phi(x)}. \tag{371}$$

If we substitute $p_0(x) = c$, the constant c cancels in both the numerator

[288] Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241, 1968

[289] **Exercise:** Using the same argument, show that the prior that encodes a state of complete ignorance on a positive real random variable $X > 0$ is $p_0(x) = c/x$. This is again an improper prior.

and the denominator. So the fact that $p_0(x)$ is improper, does not prevent $p(x|\bar{\phi})$ to be a proper pdf, provided that $Z(\lambda) < \infty$[290].

Yet there's another problem with Eq. (361). Take the example where our current state of knowledge $p^{(0)}$ implies that X is a Gaussian random variable with mean μ and variance σ^2. On the basis of this, you would predict that $S = E[X^3]$ should take the value $S = \mu^3 + 3\mu\sigma^2$. Imagine you observe that S is significantly different from this value. What should you conclude?

If you try to incorporate this information in the distribution, you end with a distribution

$$p^{(1)}(x) = \frac{1}{Z} e^{\lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3}$$

that cannot be normalised, so the recipe of maximum entropy fails.

There is a way to accommodate the observation $S \neq \mu^3 + 3\mu\sigma^2$ that requires a minimal modification of the distribution $p^{(0)}$. Take

$$\tilde{p}^{(1)}(x) = \epsilon\delta(x - \Lambda) + (1 - \epsilon)p^{(0)}(x)$$

then, a trite calculation leads to

$$\mathbb{E}[X] = \mu + \epsilon(\Lambda - \mu) \tag{372}$$

$$\mathbb{V}[X] = \sigma^2 + \epsilon[\sigma^2 + (1 - \epsilon)\mu(\mu - 2\Lambda)] \tag{373}$$

$$\mathbb{E}\left[X^3\right] = \mu^3 + 3\mu\sigma^2 + \epsilon\left[\Lambda^3 - 3\mu\sigma^2 - 3\mu^3\right]. \tag{374}$$

If we take

$$\Lambda = \left(S - \mu^3 - 3\mu\sigma^2\right)^{1/3} \epsilon^{-1/3}$$

then in the limit $\epsilon \to 0$ we recover all the three observed moments. At the same time, in this limit, $p^{(1)} \to p^{(0)}$ which is the original distribution. Formally this is correct, but what does it mean?

The fact that $h[\tilde{p}^{(1)}] = h[p^{(0)}]$ implies that the observation on S does not dispel any uncertainty on X. The distribution $p^{(1)}$ can be realised by a sample of $n \sim \epsilon^{-1}$ observations of $S_i = X_i^3$, in $n - 1$ of which, X_i is a typical draw from $p^{(0)}$, and one of them takes value $X_{i^*} = \Lambda \sim n^{1/3}$ which is very large. All this is reminiscent of the discussion we had concerning large deviations of fat tailed distributions.

Indeed the pdf of S, behaves asymptotically as

$$P\{S \in [s, s + ds)\} \sim e^{-c|s|^{2/3}} ds\,, \qquad |s| \to \infty\,.$$

Therefore S has a fat tailed distribution. As we have seen in the previous chapter, we expect that large deviations (or unexpected events) of samples drawn from such distributions occur in a peculiar manner, where one of the points in the sample attains an anomalously large (or small) value, whereas all the others take typical values. In this

[290] A limiting procedure that could be applied is to limit the values of X to the interval $[-1/(2c), 1/(2c)]$, do the calculation, and then let $c \to 0$. This is an example of a *regularisation*, a technique used to remove singularities from a problem. The prior p_0 should be invariant under affine transformation $X' = a + bX$ for all $a \in \mathbb{R}$ and all $b > 0$. This suggests that location and scale of a random variable X both need improper priors and both introduce a singularity that needs to be regularised. An interesting question, which is open to the best of my knowledge, is: are these the only (primitive) singularities or can there be other ones?

situation, the observation on S cannot change the state of knowledge on the variable X.

This indicates what type of observables will bring new information, in the sense that unexpected events allow us to update our state of knowledge on X, and what observables do not. This suggests that it is useless to sample observables which have a fat tailed distribution under the current state of knowledge, if our goal is to test a theory $p^{(0)}$.

What can we learn?

Remember our discussion on complex systems that maximise a complex function $U(\underline{s}, \bar{s})$ over a set of variables $\vec{s} = (\underline{s}, \bar{s})$ which are known only in part, because $\bar{s}$ are *unknown unknowns*. We concluded that the probability to observe a certain value $\underline{s}$ is given by

$$P\{\underline{s}^* = \underline{s}\} = \frac{1}{Z(\beta)} e^{\beta u_{\underline{s}}},$$

where $u_{\underline{s}} = \mathbb{E}\left[U(\underline{s}, \bar{s})|\underline{s}\right]$ is the known part of the function that is optimised and $\beta > 0$ depends on the optimisation over unknown variables.

If we do not know the function $u_{\underline{s}}$, can we use the procedure outlined above to learn it? In other words, can the function $u_{\underline{s}}$ be learned from a series of experiments?

Let $p^{(0)}(\underline{s})$ be the distribution that encodes the current state of knowledge about the system. For a quantity $q_{\underline{s}}$, it is possible to compute its distribution

$$p^{(0)}(q) = \sum_{\underline{s}} p^{(0)}(\underline{s}) \delta(q - q_{\underline{s}})$$

Imagine running an experiment where the value q_{exp} is measured. In particular, for a complex system, we can assume that $\underline{s}$ is a high dimensional vector of weakly dependent variables. So that the distribution of q should be sharply peaked around its expected value $\mathbb{E}^{(0)}[q] = \sum_{\underline{s}} p^{(0)}(\underline{s}) q_{\underline{s}}$, and hence $q_{\text{exp}} \approx \mathbb{E}^{(0)}[q]$.

If $q_{\text{exp}} \approx \mathbb{E}^{(0)}[q]$ within experimental errors, then the state of knowledge $p^{(0)}$ does not need to be updated. Otherwise it has to be revised[291]. In the latter case, the standard recipe to update $p^{(0)}$ is given by Large Deviation Theory. This maintains that the new distribution should be such that $\mathbb{E}^{(1)}[q] = q_{\text{exp}}$, without assuming anything else. More precisely, the amount of information that the measurement gives on the state $\underline{s}$ is given by the mutual information $I(\underline{s}, q) = D_{KL}(p^{(1)}||p^{(0)})$. Hence, $p^{(1)}$ should be the distribution with $\mathbb{E}^{(1)}[q] = q_{\text{exp}}$ for which $D_{KL}(p^{(1)}||p^{(0)})$ is minimal. The distribution that satisfies this requirement is

$$p^{(1)}(\underline{s}) = \frac{1}{Z(g)} p^{(0)}(\underline{s}) e^{g q_{\underline{s}}}, \qquad Z(g) = \int dq p^{(0)}(q) e^{gq} \tag{375}$$

[291] There is a long tradition of experiments designed to test our state of knowledge in physics. For example, until 1964, we expected that the laws of Nature should be invariant under time reversal T. The CPT theorem states that the laws of Nature should be invariant under the combined transformation CPT, where C stands for charge conjugation and P for parity transformations. The discovery of the violation of the CP symmetry in experiments on the decays of neutral kaons, changed our state of knowledge in particle physics.

where g is adjusted in such a way to satisfy $\mathbb{E}^{(1)}[q] = q_{\text{exp}}$. This process can be continued with additional measures of different observables $q'_{\underline{s}}, q_{\underline{s}}'', \ldots$, and, in principle, it leads to infer

$$\beta u_{\underline{s}} = \log p^{(0)}(\underline{s}) + g q_{\underline{s}} + g' q'_{\underline{s}} + g'' q_{\underline{s}}'' + \ldots \tag{376}$$

to the desired accuracy from a series of experiments.

This recipe, however, only works for quantities which have a distribution which falls off faster than exponential as $q \to \pm\infty$. If $-\log p^{(0)}(q) \simeq c|q|^{\gamma}$ for $|q| \to \infty$ with $\gamma < 1$, then the integral defining $Z(g)$ in Eq. (375) is not defined. There is no well defined way to incorporate an observation $q_{\text{exp}} \neq \mathbb{E}[q]$ in the current state of knowledge in this case. This clearly applies to $u_{\underline{s}}$ itself. The only models $u_{\underline{s}}$ that can be learned are those for which the density of states

$$\mathcal{N}(u)du = |\{\underline{s} : \ u_{\underline{s}} \in [u, u + du)\}|$$

has thin tails, i.e. decays like or faster than an exponential as $u \to \infty$. In this sense, systems where $\mathcal{N}(u)$ have an exponential behaviour with u are special, because they separates the region of learnable systems – those for which $\mathcal{N}(u)$ has thin tails – from unlearnable one – those where $u_{\underline{s}}$ has a fat tailed distribution. Interestingly, these are the systems that are best at learning according to [292].

[292] Ryan John Cubero, Junghyo Jo, Matteo Marsili, Yasser Roudi, and Juyong Song. Statistical criticality arises in most informative representations. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (6):063402, jun 2019. DOI: 10.1088/1742-5468/ab16c8; and Matteo Marsili and Yasser Roudi. Quantifying relevance in learning and inference. *Physics Reports*, 963:1–43, 2022