Statistical Methods in Physics
Final project I - One Dimensional Harmonic Oscillator

Antonin Vallelian

21 January, 2023

Contents

1 Introduction

Statistical methods are fundamental tools in both research activity and professional science-related domains. In this report, I will present the one-dimensional Harmonic Oscillator. We consider the system as a large number of Potassium (^{40}K) atoms trapped in a Harmonic well. The mass of each atom is $m = 40m_0$, where m_0 is the atomic unit mass. The energy of each atom follows the probability density function (PDF)

$$f(E_\omega) = \frac{A}{\sqrt{E_\omega}} \exp\{-\beta E_\omega\}\,, \tag{1}$$

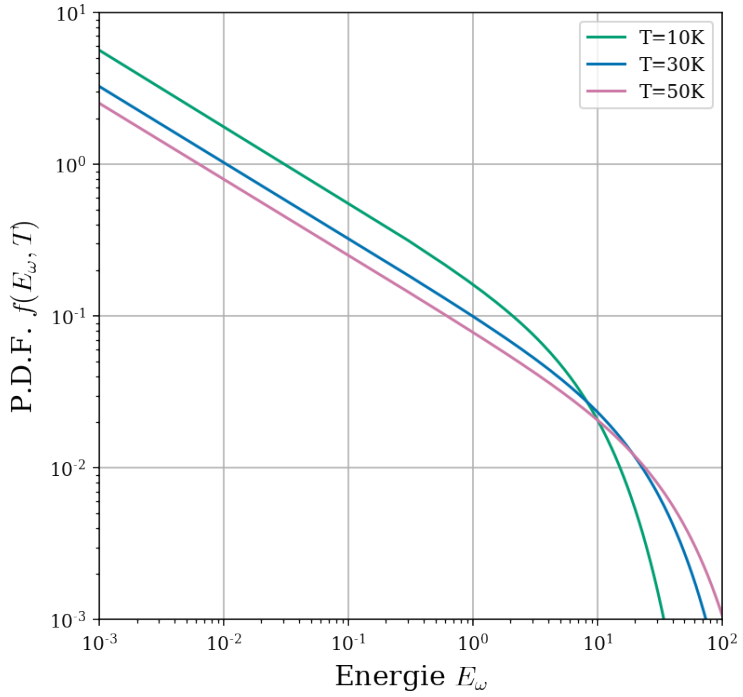


Figure 1: Probability density function given by Eq. (1) for different temperatures, $T = 10$K , $T = 30$K, and $T = 50$K.

where, $\beta = \frac{1}{k_B T}$, k_B being the Boltzmann constant. Since we are working with only simulation, and k_B has a very low value in the SI system, we will set it to 1 without loss of information, which means the unit of the energy is [E] = [T] = K. Also, $A = \sqrt{\frac{\beta}{\pi}}$, and we will consider $T = 30K$. This function is defined for energy in the range $(0, \infty)$.

2 Task-1

2.1 Demonstrate that $f(E_\omega)$ is a PDF

The function provided by Eq. (1), depends on two parameters, E_ω and T. Fig. 1, shows the shape of the function for different temperatures. Since T influences the behavior of f, to be a PDF, Eq. (1) must be positive and its integral over the full energy space (positive and real) must be equal

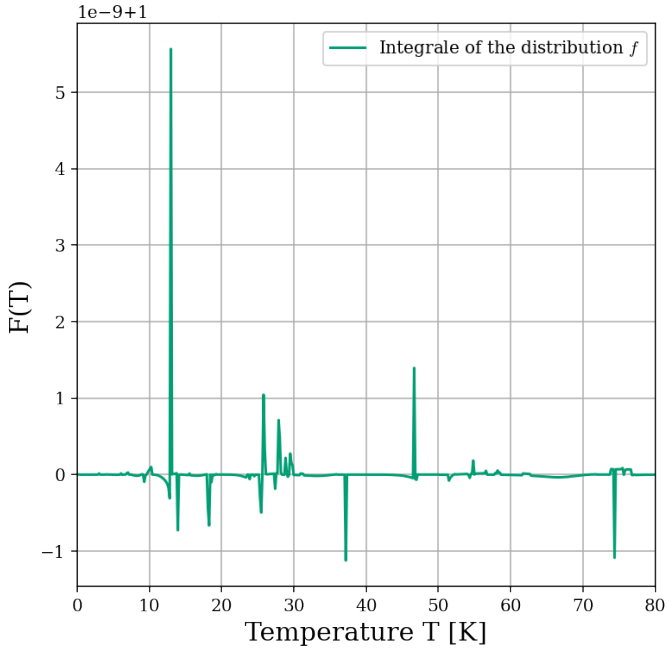


Figure 2: Numerical integral of Eq. (1) over the energy space.

to 1 at any temperature.

The positivity is easy to show. Indeed, the Eq. (1) is a product $\frac{A}{\sqrt{E_\omega}}$ which is positive for any energy or temperature, and $\exp\{-\beta E_\omega\}$ which is also all greater than 0.

The integral of Eq. (1) can be computed analytically or numerically. The integral of f is

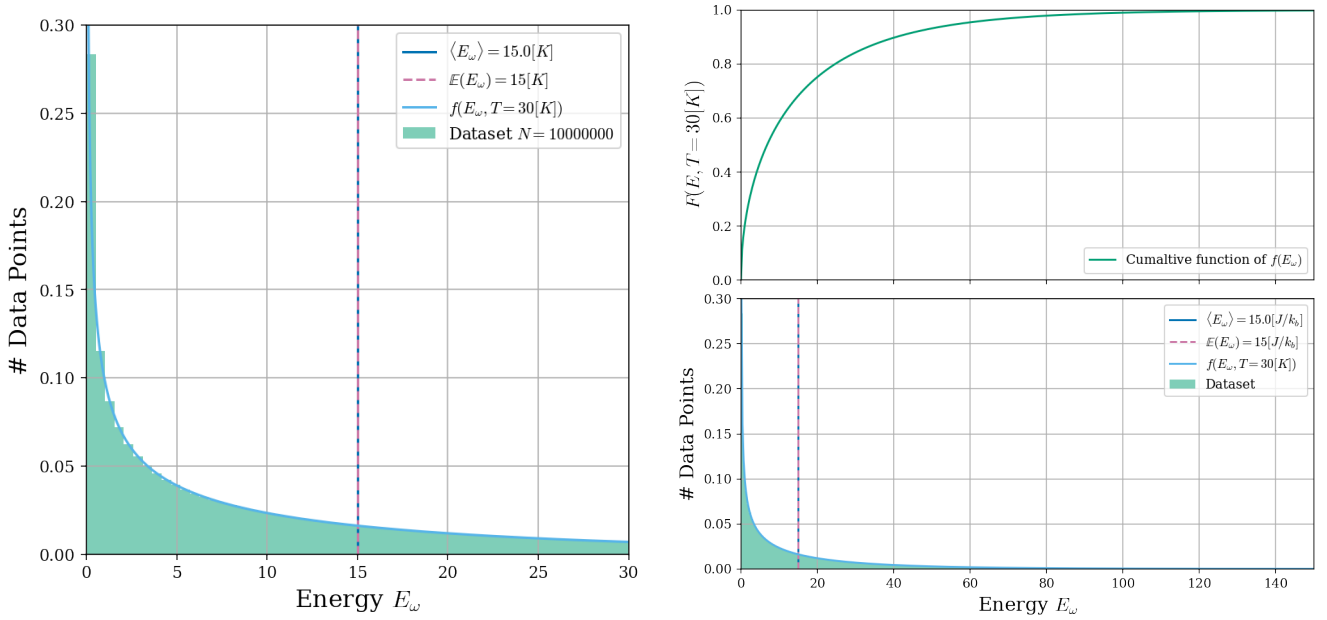
$$\begin{aligned}\int_0^\infty f(E_\omega, T)dE_\omega &= \int_0^\infty \frac{A}{\sqrt{E_\omega}} \exp\{-\beta E_\omega\}dE_\omega \,, \\ &= \frac{E_\omega\sqrt{\frac{\beta}{E_\omega}}\mathrm{Erf}(\sqrt{\beta E_\omega})}{\sqrt{\beta E_\omega}}\Bigg|_0^\infty = \mathrm{Erf}(\sqrt{\beta E_\omega})\Big|_0^\infty = 1\,. \end{aligned} \tag{2}$$

This result is true for all the temperatures. The integral can also be computed numerically with PYTHON using the function SCIPY.INTEGRATE.QUAD which calculates the integral with the Gaussian quadrature rule, which approximates the integral by taking a weighted sum of the function's values at specific points, called the quadrature points. It gives a better approximation than the traditional trapezoid method which can give errors when the variation is big like in the low energy of our PDF. Fig. 2 shows the numerical result for temperature between 0K and 80K. We see that the integral is very close to 1 for all the values of the temperature. Note that to compute this integral the maximum error allowed was 10^{-7}. The fluctuations in the curve are the numerical errors from the function.

Since Eq. (1), fulfills the two required conditions, we can faithfully affirm that it is a PDF.

2.2 Build a Monte Carlo from the PDF $f(E_\omega)$

There are several ways to build a Monte Carlo. The considered PDF is not defined in $E_\omega = 0$ which makes the computation time of an accepted rejected very long as we will discuss in 2.4. This is why we will use the inverse transform method.



(a) Histogram of the N value of energy generated by the PDF with the inverse transform method from 30[K].

(b) *top* Cumulative function of the PDF (1) at $T =$ 30K. *bottom* Histogram of the N value of energy generated by the PDF with the inverse transform method from 0[K] to 150[K].

Figure 3: Dataset generated by the PDF (1). The full blue line represents the numerical mean of the sample and the dashed pink line is the expectation value of the energy. The light blue line corresponds to the PDF.

To do so, we need to calculate the inverse of the cumulative function. We have already calculated the cumulative Eq. (2). To inverse it we can use the inverse error function which can be used in the simulation from the package SCIPY.SPECIAL.ERFINV

$$F(E_\omega) = \text{Erf}(\sqrt{\beta E_\omega}) \,, \tag{3}$$

$$F^{-1}(r) = \frac{1}{\sqrt{\beta}} (\text{Erf}^{-1}(r))^2 \,. \tag{4}$$

Note that this inversion is possible only because is E_ω is real and positive. With this function, we can randomly produce a sample of N data point r from 0 to 1 and with the inverse cumulative will gives us the simulated dataset. The inverse transform method is very fast and thanks to that we can generate many more points than the accepted rejected method. We can generate a sample of $N = 10000000$ and fill them in a histogram. Fig. 3, shows this sample: 3a is a rescaled histogram for energies from c, and 3b is the same graph but from 0[K] to 150[K] with the cumulative to understand the generation process. We can see that the PDF focuses the energy on the low scale but the cumulative becomes close to 1 only above 150[K]. The random set of r between 0 and 1 can

be considered entering the top graph. Once a value of r hits the cumulative green line, it returns the energy generated and can be filled in the histogram below. The histogram is normalized and follows perfectly the PDF, which attests to the validity of the computation of the inverse transform method.

2.3 Comparison between the expectation value and the mean of the sample

The expectation value is

$$\mathbb{E}(E_\omega) = \frac{1}{2}k_b T = 15[\mathrm{K}]\,. \tag{5}$$

We can see in fig. 3a, the expectation value and the mean of the data sample, given by the function NUMPY.MEAN, fit perfectly. The numerical mean gives $\langle E_\omega \rangle = 15.003[\mathrm{K}]$ differs from $\mathbb{E}(E_\omega)$ only to the third digit. This small difference is due to the finite number of events N, bigger is N smaller will it be.

2.4 Another method?

We used the inverse transform method to generate our sample, but we can use the Von Neumann or accepted rejected method. It consists of generating a sample of points in a 2D space with a uniform envelope and selecting those below the PDF. This method is more approximative since it required a cut-off at high energies. Also, the PDF is not defined in $E_\omega = 0$ which requires also adding a cut-off at low energies. These considerations add a lot of error in the computation. We could also take another function in place of the uniform envelope, but this requires rescaling the bin heights of the histogram after the computation which is not always suitable.

Fig. 4 shows a dataset of $N = 10000$ energies. the energy interval considered is $(10^{-3}, 150)$. First, we can see that the area considered for the generation is very compared with the accepted region. In fact, if we compute the ratio $R = \frac{\#\text{ Accepted}}{\#\text{ Total}}$ we find $R = 0.0019$. This means when a point is generated, it has 0.19% of a chance to be accepted in those conditions. But one could ask: why is the uniform envelope this high? Because of the singularity in $E_\omega = 0$. Choosing a lower maximum would suppress a lot of area to the PDF and thus would not be normalized anymore. So we would have to change the normalization, which could have changed the expectation value. Also, the numerical mean represented by the blue line is shifted by the cutoff choices a higher envelope corresponds to a lower mean and a higher cut-off in the energy would give a bigger mean. Finally, the generation time for this method is far bigger than the inverse transform method. Indeed, with this method, we can generate 100000 values in $\sim 0.8s$ but with the inverse transform method, we had 10000000 in $\sim 0.6s$. Comparing the rate of generation $\Gamma = \frac{\#\text{ generated points}}{\text{generation time}}$, we have $\Gamma_{IM} \sim 133\Gamma_{VN}$, the inverse transform method is ~ 133 times faster than the Van Neumann method. This difference in a generation will be very important for the next sections where the computation can be very long. For this PDF, the accepted rejected method is not optimal. It is important to specify "this PDF because not all PDFs have an invertible cumulative function.

2.5 Variance, Skewness, and Kurtosis

Variance The variance of the PDF can be calculated with the formula

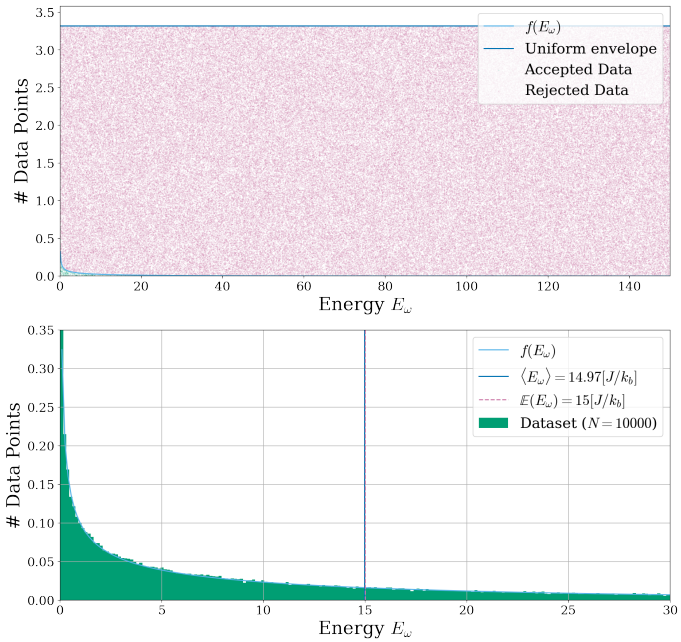


Figure 4: Dataset generated by the Von Neumann methods. *top* The blue line represents the uniform envelope, and the light blue line is the PDF. The scattered points are generated by a uniform law, pinks are rejected and greens are accepted. Note that this plot is only for $N = 500$. For higher N the distinction between the points would have been difficult. *bottom* Histogram of the dataset with the same color code as in Fig. 3.

$$V\left(E_\omega, \text{PDF}\right) = \int_0^\infty \left(E'_\omega - \mu\right) f(E'_\omega) dE'_\omega = \mathbb{E}\left(E_\omega^2\right) - \mathbb{E}\left(E_\omega\right)^2 . \tag{6}$$

We can use once again use the SCIPY.INTEGRATE.QUAD function and we find $V\left(E_\omega\right) = 450.0[K^2]$. This value is directly given by the PDF. We can confirm it with the NUMPY.VAR function which gives us $V_n p\left(E_\omega\right) = 449.6[K^2]$. The two results are similar, the difference comes from the dataset which is not exactly following the PDF. This variance of the mean of a generated sample of random variables will then need to be divided by N since one event is not dependant of another

$$V_N\left(E_\omega\right) = \frac{V_{\text{PDF}}\left(E_\omega\right)}{N} . \tag{7}$$

Skewness The coefficient of skewness is

$$\gamma_1 = \sqrt{\frac{\mathbb{E}\left(E_\omega - \mathbb{E}(E_\omega)^3\right)^2}{\mathbb{E}\left(E_\omega - \mathbb{E}(E_\omega)^2\right)^3}} , \tag{8}$$

which can be directly calculated using the SCIPY.STATS.SKEW function. For this PDF we have $\gamma_1 = 2.82$ which means that the distribution is asymmetric and strongly shifted to the left since it is not 0 ad positive. Looking at Fig. 3 this result is coherent.

Kurtosis Similar to the skewness the kurtosis can be calculated with

$$\gamma_2 = \frac{\mathbb{E}(E_\omega - \mathbb{E}(E_\omega))^4}{\mathbb{E}\left(E_\omega - \mathbb{E}(E_\omega)^2\right)^2} - 3\,, \tag{9}$$

with the function SCIPY.STATS.KURTOSIS function we have $\gamma_2 = 11.99$. The kurtosis reflects the length of the distribution's tail. Indeed, if we had no tail the kurtosis would have been smaller than zero. But in our case, the tail is really long as we can see in Fig. 3. This is why the kurtosis is far from 0.

3 Task-2

3.1 Law of Large Numbers

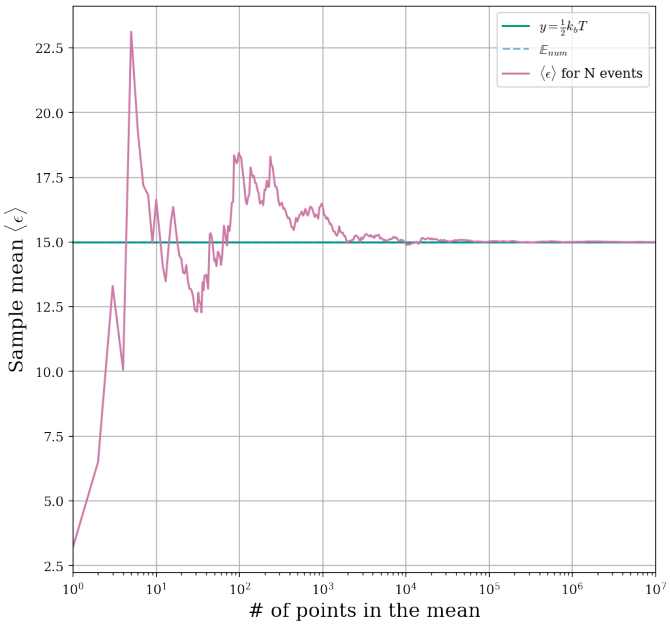


Figure 5: Mean of the sample in terms of the number of points considered.

The strong law of large numbers states that in an experiment we acquired a number N of random variables $\{x_1, x_2, ..., x_N\}$, if

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{\sigma_i^2}{i^2} \neq \infty\,, \tag{10}$$

where σ_i^2 is the variance, then the mean of the sample will converge to the expectation value. There is also the weak law of large numbers which also predicts the convergence but quadratically with the condition

$$\lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \sigma_i^2 = 0\,. \tag{11}$$

In Fig. 5, we can see the mean converging to the expectation value as the law of large numbers predicts.

3.2 Variance of $\sqrt{\frac{2E_\omega}{m\omega^2}}$

The expectation value of $x = \sqrt{\frac{2E_\omega}{m\omega^2}}$ can be calculated as

$$\begin{aligned}\mathbb{E}(x) &= \int_0^\infty x(E_\omega) f(E_\omega) dE_\omega \,, \\ &= \sqrt{\frac{2\beta}{m\omega^2\pi}} \int_0^\infty e^{-\beta E_\omega} dE_\omega \,, \\ \mathbb{E}(x) &= \sqrt{\frac{2}{m\omega^2\pi\beta}} \,, \end{aligned} \tag{12}$$

with ω constant. With the parameter considered, $m = 40m_0$, $T = 30$[K] and we set $\omega = 1[\text{s}^{-1}]$ we have $\mathbb{E}(x) = 0.6910$[m]. If we compute it from the generated variables we have $\langle x \rangle_{num} = 0.6912$[m]. Which is very close to the expectation value. This value corresponds to the mean oscillation of the ^{40}K atoms, if the temperature was higher, the atoms would have been more excited which could have amplified the oscillation and the mean of x would be higher. Note that if we write $\mathbb{E}(x)$ with SI unit we have, $\mathbb{E}(x) \sim 2 \cdot 10^{-12}$[m]. Knowing that the size of an atom is $\sim 10^{-10}$[m], their oscillations are low compared to their size. This result is coherent since 30[K] can be considered as a low temperature, in consequence, the atoms are almost frozen.

3.3 Central Limit Theorem (CLT)

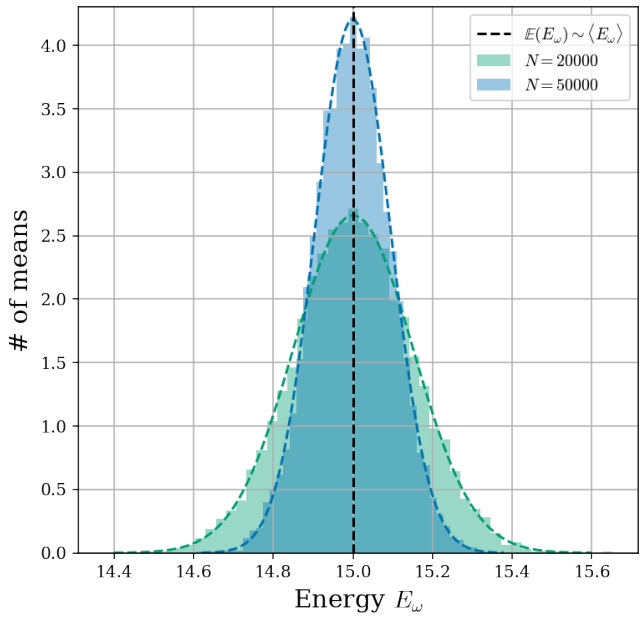
The CLT is useful when we simulate a large number of experiences $N_{\text{exp}} \ll 1$ where a set of N random variables are generated. Indeed, it tells if one computes the means of each generated experiment, they will be distributed around the expectation value as a Gaussian law with a variance $\sigma = \frac{\sigma_{\text{PDF}^2}}{N}$ where σ_{PDF} is the variance calculated in 2.5. To prove it, Fig. 6a shows two sets of 10000 experiments with $N = 20000$ and $N = 50000$. The two distributions follow perfectly their associated Gaussian laws centered around the expectation value which attest to the veracity of the theorem. The numerical means $\langle E \rangle_N$ are $\langle E \rangle_{20000} = 15.00$[K] and $\langle E \rangle_{50000} = 15.00$[K] both means are close to the expectation value. The variances can be calculated from the variance in 2.5 $\sigma^2_{20000} = \frac{\sigma^2_{\text{PDF}}}{N} = \frac{450}{20000} = 0.0225[\text{K}^2]$ and $\sigma_{50000} = 0.009[\text{K}^2]$. Numerically we have $\sigma^2_{20000,\text{Num}} = 0.0225[\text{K}^2]$ and $\sigma^2_{50000,\text{Num}} = 0.00900[\text{K}^2]$. This is a perfect match. The variance for $N = 20000$ is bigger since there are fewer data for the means.

But what if we change N_{exp} instead of N? Fig. 6b, shows the difference in the distribution of $N_{\text{exp}} = 300$ and $N_{\text{exp}} = 10000$. We see that the variance of the normal distribution does not change as expected but we see that the mean of all the means is not exactly the expectation value. Which is coherent since there are less data.

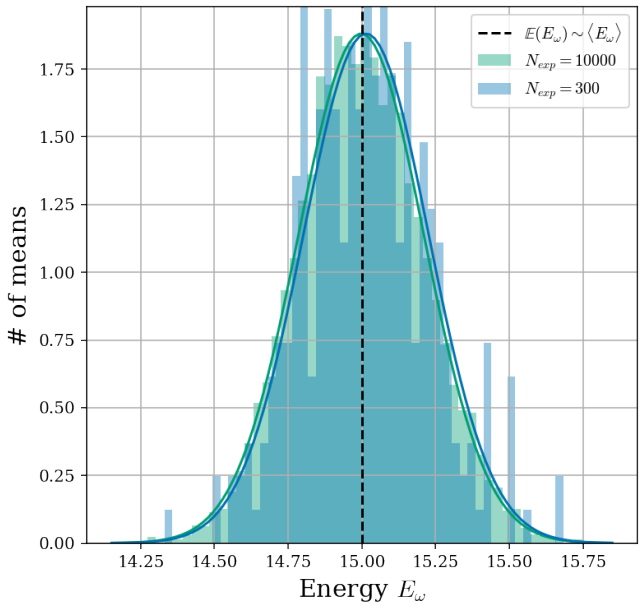
4 Task-3

4.1 The expected PDF for the number of entries per bin

The number of entries per bin should follow a Poisson's distribution, but since we generate $N \gg 1$ set of random variables, the number of entries per bin is really high the distribution will have the same shape as a normal distribution as shown in Fig. 7.



(a) Distribution of the means from two set of $N_{\text{exp}} = 10000$ experiments with $N = 5000$ (green) and $N = 10000$ (blue). The green (blue respectively) line is the Gaussian fit around the expectation value with $\sigma = \sigma_{\text{PDF}}/\sqrt{N}$.



(b) Distribution of the means from two set of $N = 10000$ experiments with $N_{\text{exp}} = 10000$ (green) and $N_{\text{exp}} = 300$ (blue). The green and blue lines are the Gaussian fits around the expectation value with $\sigma = \sigma_{\text{PDF}}/\sqrt{N}$.

Figure 6: Distributions of the means of different sets of random variables.

4.2 Show the the χ^2 of the obtained entries per bins follows a χ^2 distribution.

We generate N_{exp} experiments, that can be distributed in a histogram with n_{bins} bins each. The χ^2 value for each experiment will be given by

$$\chi^2 = \sum_{i=1}^{n_{\text{bins}}} \frac{(\text{bins entry} - \text{bins expectation})^2}{\text{bins expectation}}\,, \tag{13}$$

where the bins entries are the height value for each bin and the bin expectation is the nominal value from the PDF (1) for each bin which is given by

$$\text{bins expectation} = N \int_{E_{\omega,i}-w_b}^{E_{\omega,i}+w_b} f(E_\omega) dE_\omega\,, \tag{14}$$

where w_b is half of the bins width. This value is computed with the SCIPY package's function SCIPY.INTEGRATE.QUAD as before. Then we can distribute the χ^2 values in a histogram to look a the obtained distribution.

Fig. 8a shows the χ^2 distribution for 3000 experiments with 3000 events each for 15 bins. The distribution is expected to follow a χ^2 distribution with $n_{bins} - 1$ degrees of freedom, which is the plot in blue thanks to the SCIPY.STATS.CHI2.PDF function. We can see the distribution fits perfectly the expected shape of the χ^2 PDF.

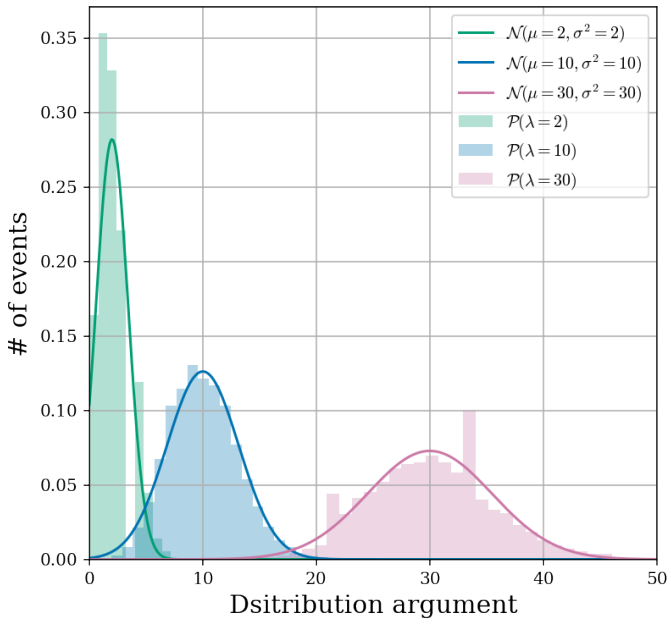


Figure 7: Comparison between the normal distribution and a sample generated by Poisson's distributions.

4.3 How does the χ^2 distribution changes with the number of bins?

We can repeat the process we did in 4.2 but for different numbers of freedom. Fig. 8b shows the χ^2 distribution for different numbers of bins. This PDF is expected to tend to a normal distribution as the number of degrees of freedom increases like the Poisson's distributions in 4.1 when the mean increases. We can easily see the change of behavior in the plot of Fig. 8b. For high degrees of freedom n_{df} the value of χ^2 are distributed around $n_{\text{df}} - 1 \sim n_{\text{df}}$. However, when we increase the number of bins, the number of events per bin decreases. Which makes the assumption that the number of entries per bin follows a normal distribution incoherent. This is why for 32 bins (orange histogram in Fig. 8b), the distribution seems a bit shifted to the left.

5 Task-4

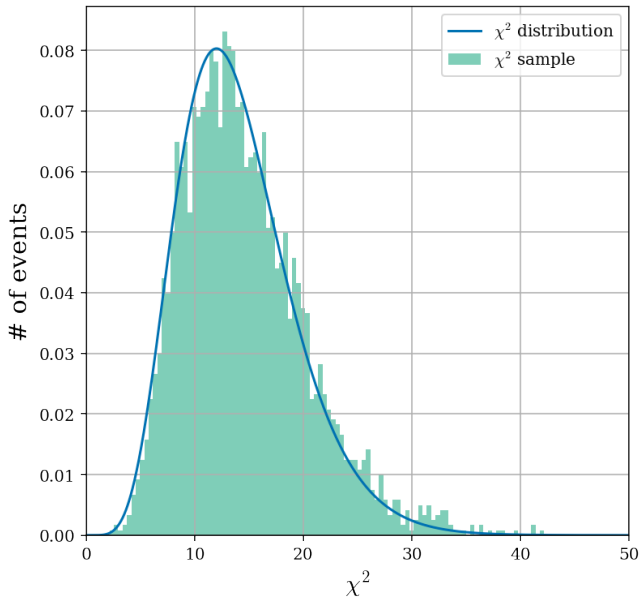
5.1 Maximum likelihood estimation of the temperature

To estimate the parameter of the simulation T, we can use the maximum likelihood method. To do so we create a set of test temperatures for which we will calculate

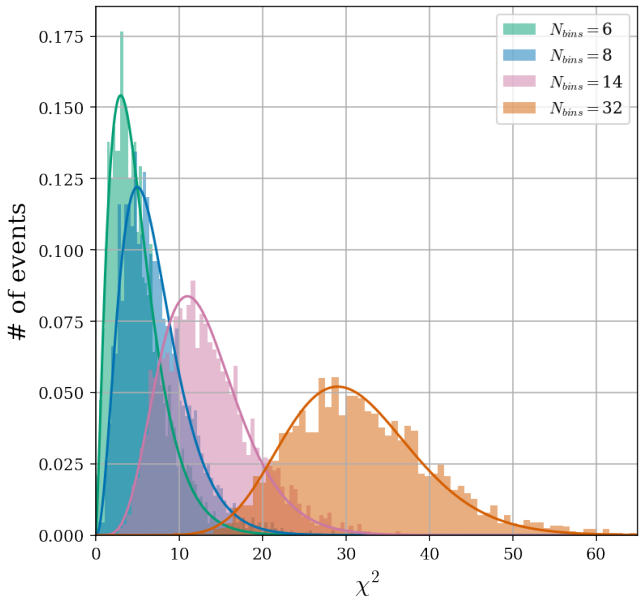
$$-\log L(E_{\omega,1}, E_{\omega,2}, E_{\omega,3}, ..., E_{\omega,N} \mid T) = -\sum_{i=1}^{N} \log f(E_{\omega,i}, T_{\text{Test}})\,. \tag{15}$$

from the energy generated with the Monte Carlo simulation. If we find the minimum of this function, it will return the temperature for which the likelihood $L(E_{\omega,1}, E_{\omega,2}, E_{\omega,3}, ..., E_{\omega,N} \mid T)$ is maximized.

The error on the estimated temperature can be calculated by renormalizing $-\log L$ by its minimum. At this point, the 1σ error will simply be given by the interval



(a) χ^2 distribution for 3000 experiments with 3000 events each for 15 bins. The blue line is the χ^2 function fit and in green is the χ^2 sample.



(b) χ^2 distributions for four different numbers of bins: green 6, blue 8, pink 14 and orange for 32, for 3000 experiments with 3000 random variables. Each distribution has an associated χ^2 PDF with $n_{bins} - 1$ degrees of freedom.

Figure 8: χ^2 for differents experiment.

$$\delta T = \frac{T_2 - T_1}{2}\,, \tag{16}$$

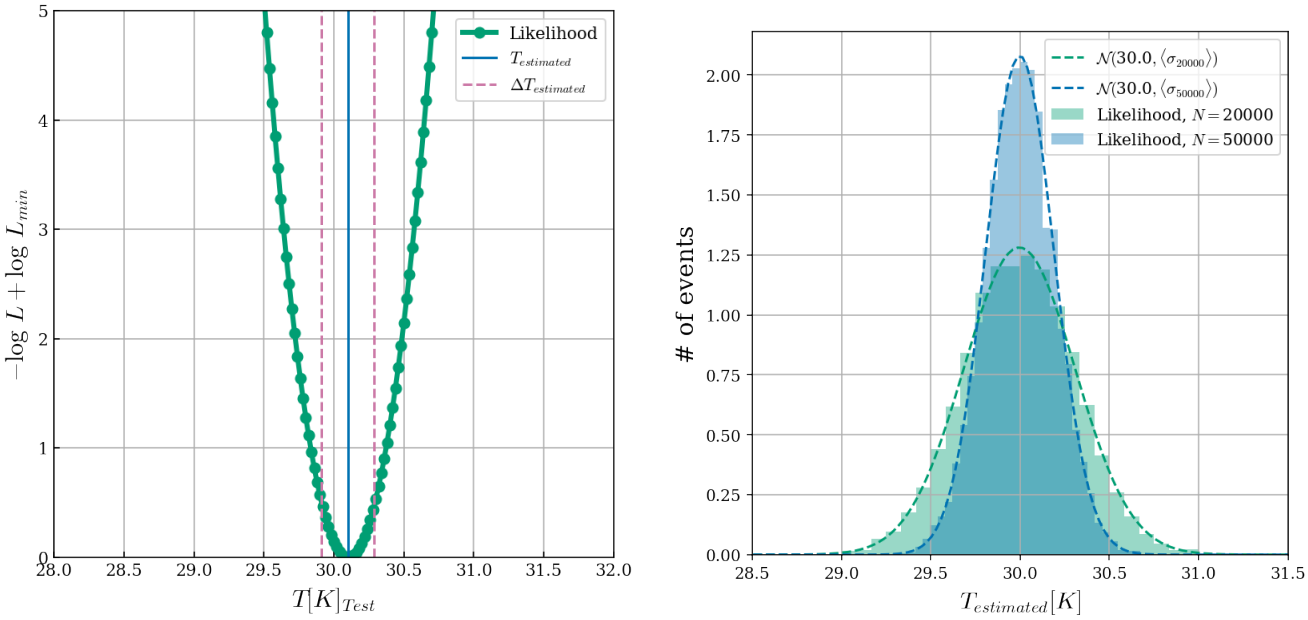
$$-\log L(T_i) - (-\log L)_{\min} = 0.5, i = 1, 2\,,$$

$$T_1 < T_{\text{estimated}} < T_2 \tag{17}$$

Fig. 9a shows $-\log L + (\log L)_{\min}$ for temperatures 201 temperatures taken from 28[K] to 32[K] for 50000 energies. The estimated temperature is $T_{\text{estimated}} = 30.2 + -0.2$[K].
We can repeat this process N_{exp} times and fills the estimated parameter in a histogram. The values will be distributed following the Gaussian shape with the estimated temperatures as the mean and the variance can be approximated by

$$\sigma^2 = \left(\frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} \delta T_i \right)^2 . \tag{18}$$

Fig. 9b shows the distribution of two set of $N_{\text{exp}} = 10000$ parameters for $N = 20000$ and $N = 50000$. The two histograms are fitted with the function SCIPY.OPTIMIZE.CURVEFIT. For $N = 20000$, the fitted variance is $\sigma^2_{20000,\text{fit}} = 0.091[\text{K}^2]$ and the mean is $\langle E_{\omega,20000,\text{fit}} \rangle = 30.0 \pm 0.3$[K]. For $N = 50000$ we have $\sigma^2_{50000,\text{fit}} = 0.037[\text{K}^2]$ and the mean is $\langle E_{\omega,20000,\text{fit}} \rangle = 30.0 \pm 0.2$[K] If we compare these values to the variance given by Eq. (18) $\sigma^2_{20000,\text{num}} = 0.082[\text{K}^2]$, $\sigma^2_{50000,\text{num}} = 0.037[\text{K}^2]$, for $N = 50000$, the variance is the same but for $N = 20000$, the difference can be explained by the lack of events. Nevertheless, the values have a similar magnitude.



(a) Logarithm of Likelihood minus its minimum $-\log L + (\log L)_{\min}$ in terms of the test temperatures is represented with the red dotted line. Each dot represents a test temperature. The blue is the minimum which returns the estimated temperature. The error is represented by the pink dashed lines.

(b) Histograms of 10000 estimated temperature with $N = 20000$ (green) and $N = 50000$ (blue). Both histograms are fitted with a Gaussian distribution (dashed lines).

Figure 9: Result from the maximum likelihood estimation method.

5.2 Goodness and least square estimation method

The goodness of fit is a way to check the simulation by comparing the set of variables $\{E_{\omega,1}, E_{\omega,2}, E_{\omega,3,\ldots E_{\omega,N}}\}$ with the fit. This is a good test to estimate a parameter. Indeed, if the goodness test calculated for a given temperature is not coherent then the fit must be modified. To do so we calculate

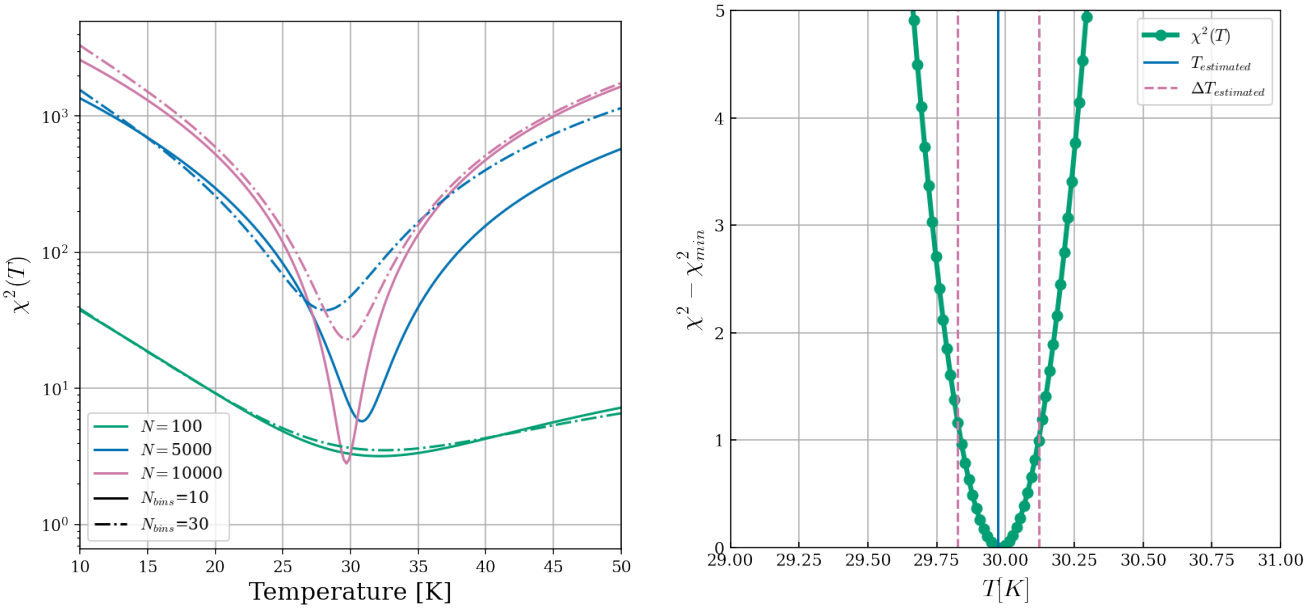
$$\chi^2 = \sum_i \left[\frac{y_i - h\left(E_{\omega,i} \mid T\right)}{\sigma_i} \right]^2 . \tag{19}$$

In 4.3, we see that if the number of degrees of freedom, which corresponds to the number of entries per bin, is high enough the χ^2 distribution will follow a normal distribution, so we can approximate $\sigma_i = \sqrt{y_i}$. the goal is to have the lower χ^2 so if we calculate it for different values of the temperature we can find the minimum to have an estimation as the maximum likelihood method. As in 4.2 the quantity h is the expectation value in each bin given by (14).
As before, to calculate the error we have to rescale the χ^2 function with her minimum. At this point, the 1σ is found with

$$\delta T = \frac{T_2 - T_1}{2} , \tag{20}$$

$$\chi^2_{1\sigma} + \chi^2_{\min} = 1, i = 1, 2 , \tag{21}$$

$$T_1 < T_{\text{estimated}} < T_2 . \tag{22}$$



(a) Different χ^2 in terms of the tested temperatures. The solid lines represent χ^2 for 10 bins and the dot-dashed lines are for $30 bins$. the color green, blue and pink represents $N = 100$, $N = 5000$ and $N = 10000$ respectively.

(b) $\chi^2 - \chi^2_{\text{min}}$ in terms of the test temperatures is represented with the red dotted line. Each dot represents a test temperature. The blue is the minimum which returns the estimated temperature. The error is represented by the pink dashed lines.

Figure 10: Plots of the χ^2 values for different parameters.

Fig. 8a represents different $\chi^2(T)$ in terms of the number of data N and the number of bins. We can see that the error for low N is very big since the curve is flatter. Also Fig. 8b, shows the minimizing process of the χ^2 function for 30 bins, $N = 100000$ and 201 temperatures between 29[K] and 31[K]. The estimated temperature is $T_{\text{estimated}} = 30.0 \pm 0.1$[K] The error seems inferior to the maximum likelihood method.

Let's now repeat this procedure 10000 times and fill in the result in a histogram. Fig. 11 shows the result of the least squares estimation. It gives us $T_{\text{estimated}} = 29.9 \pm 0.2$[K] and the variance $V(\hat{T}) = 0.041[\text{K}^2]$. The estimated parameter is less precise than the maximum likelihood function but is close to the expectation value. In this method, after several computations, the mean value is always shifted below the reel value of the parameter T. This could be explained by the number of bins $b_{\text{bins}} = 30$ which is too high as we can see in Fig. 10a when the number of bins is low the approximation that the number of entries per bin follows a Gaussian distribution is not coherent anymore. However, the computation time (~ 222[min]) makes the test difficult to perform.

5.3 Monte Carlo integral estimation: method of the moment

The last estimation method we will try is the Monte Carlo integration. For a large number N, the momenta of the distribution

$$\hat{E}_\omega = \frac{1}{N} \sum_i h(E_\omega) \,, \tag{23}$$

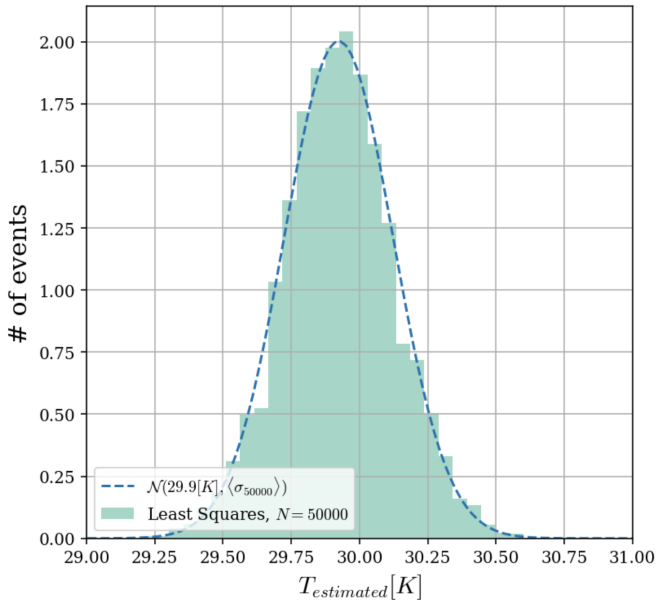


Figure 11: Least Squares estimation of the temperature.

can be equated to the estimator

$$E_\omega(T) = \int E'_\omega f(E'_\omega; T) dE'_\omega \,. \tag{24}$$

Then using the relation between the temperature and the energy it can return the estimation of the temperature

$$\hat{T} = 2 \cdot \hat{E}_\omega = \frac{2}{N} \sum_i h(E_\omega) \,. \tag{25}$$

Then repeat this process N_{exp} times, we can fill a histogram with all the estimated temperatures which will follow a normal distribution around the best estimation. The variance can be estimated by

$$\sigma^2 = \frac{2\left(\frac{1}{N}\sum_i (T_i)^2 - (\frac{1}{N}\sum_i T_i)^2\right)}{N} \,. \tag{26}$$

Fig. 12, shows the result of the method of the moments. We can see that the Gaussian fit with the variance given in Eq. (26) fits perfectly the histogram. The estimated temperature is $T_{\text{estimated}} = 30.0 \pm 0.2[\text{K}]$ and the variance is $\sigma^2 = 0.035[\text{K}^2]$.

5.4 Comparison between the estimation methods

Let's now compare the different estimation methods. For all three procedure we considered $N_{\text{exp}} = 10000$ and $N = 50000$. The different estimated parameters and their variances are resumed in table 1.

We can see that all the variances are on the same scale, but the least square method gives an estimated parameter away from the initial value $T = 30[\text{K}]$. Also, we can see the compilation time is a lot longer for this method. Maybe the used code was not optimized for this computation, for

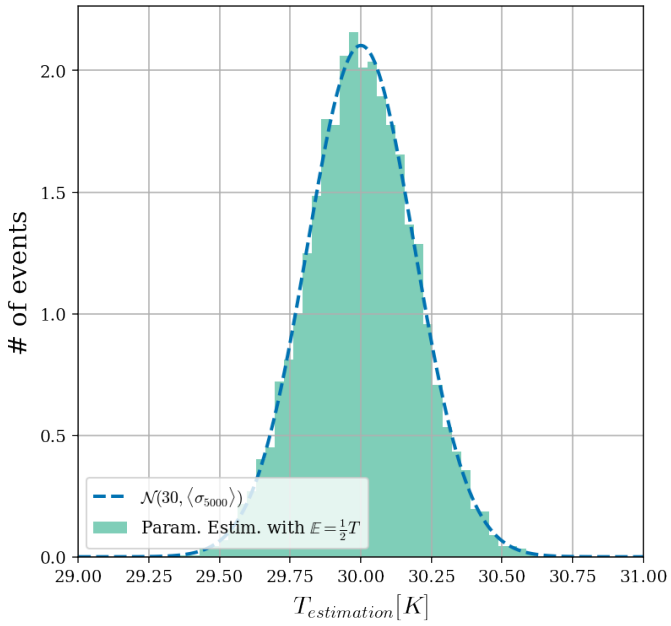


Figure 12: Estimation of the temperature using the method of the moment for 10000 experiments of $N = 50000$ random variables.

Estimation	$\langle T \rangle$[K]	V[K^2]	Time of compilation
Maximum Likelihood	30.0 ± 0.2	0.039	$\sim$ 8[min]
Least Squares	29.9 ± 0.2	0.041	$\sim$ 222[min]
Moment	30.0 ± 0.2	0.037	$\sim$ 25[s]

Table 1: Results of the different estimation methods.

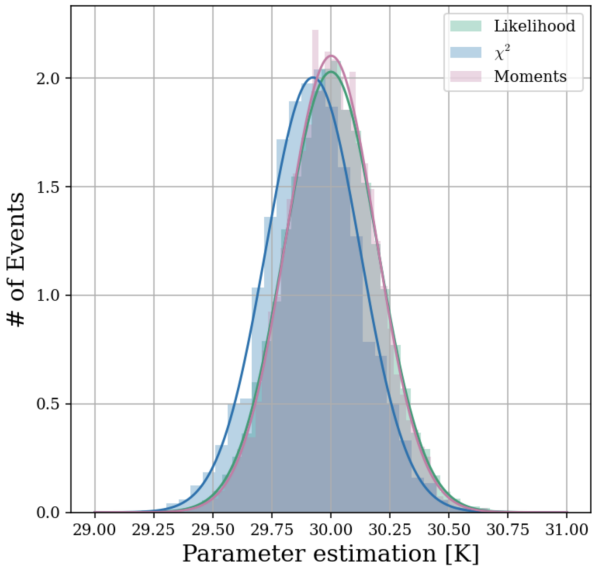
example when computing the quantity (19) we ask the computer to calculate the bins expectation using the function SCIPY.INTEGRATE.QUAD which can be really heavy.

Note that the moment method gives a good result with the lowest variance and computation time. Note also that all the estimations have been done on an 2.6 GHz INTEL CORE I7 6 CORES processor.

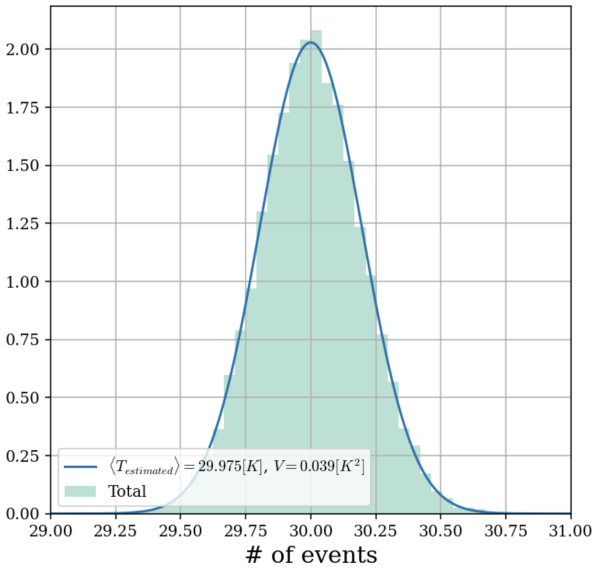
Fig. 13a shows all the histograms for the different estimation methods. We can see that they are all similar with the same initial parameter N_{exp} and N. The Least Squares histogram is a bit shifted giving a less precise mean value. The sum of all the created data, shown in Fig. 13b, gives the mean estimation $T_{\text{estimated}} = 29.975$[K] with a variance $V = 0.039$[K^2]. It is just the mean between all three estimated results. The moment method seems to be the most efficient method for this PDF.

6 Task-5

We will test the hypothesis using the Kolmogorov procedure. To do so, we generate 200 datasets that will be matched per pair. For each pair, we can calculate the test statistic by comparing the cumulative distribution functions of the two datasets with the distribution function f(E). Which can be calculated by



(a) Comparison of the histogram of the different estimation methods. Green for the Maximum Likelihood, blue for the Least Squares method, and pink for the moments' method.



(b) Histogram of all the estimated temperature.

Figure 13: Comparison of the estimation methods

$$\text{cum}_k = \frac{1}{N} \sum_i^k E_{\omega,i} \forall k < n_{\text{bins}} \,, \tag{27}$$

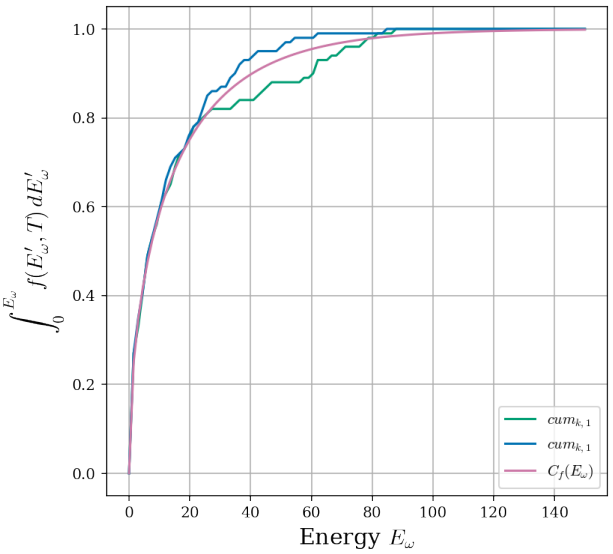
We set then a threshold for the hypothesis testing refers to the level of significance at which you will reject or accept the null hypothesis that the two datasets come from the same distribution. This threshold is typically set to a small value such as 5%, meaning that if the calculated test statistic is greater than the threshold, you would reject the null hypothesis and conclude that the two datasets do not come from the same distribution. We can next calculate the cumulative C_k of the PDF given by

$$C_f(E_\omega) = \int_{-\infty}^{E_\omega} f\left(E'_\omega\right) dE'_\omega \,, \tag{28}$$

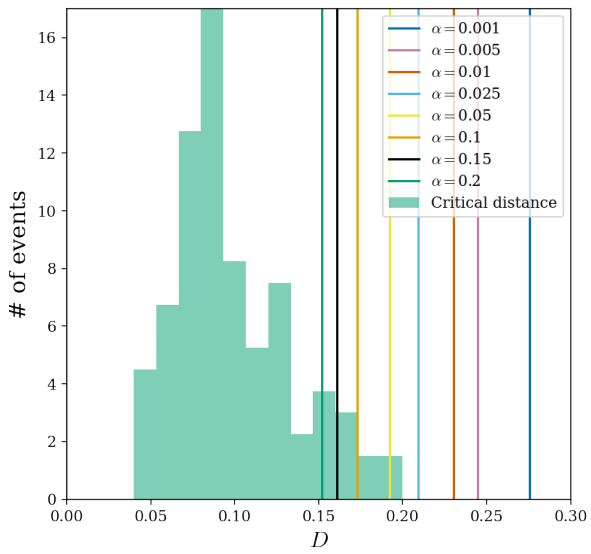
and calculated the largest distance between C_f and cum_k given by

$$d = \sqrt{N} \max\left(|\text{cum}_k - C_f(x)|\right) \,. \tag{29}$$

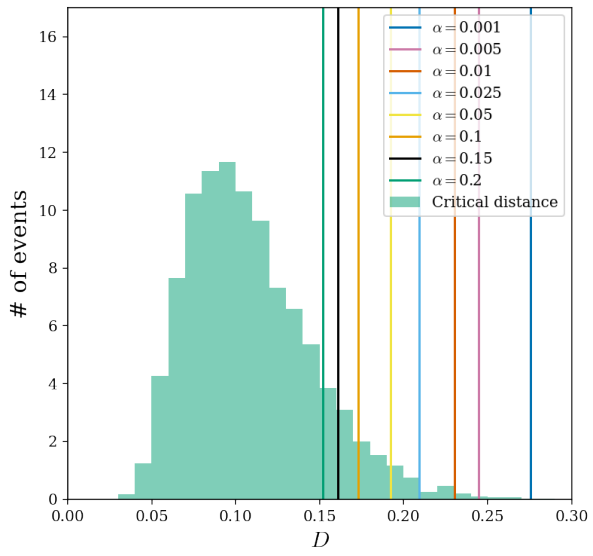
It is important to note that, the outcome of the test will depend on the threshold we set. For example, if we set a high threshold (e.g., 0.1), we will be more likely to reject the null hypothesis and less likely to conclude that the two datasets come from the same distribution, even if the test statistic is relatively large. Conversely, if we set a low threshold (e.g., 0.01), we will be more likely to reject the null hypothesis and conclude that the two datasets do not come from the same distribution, even if the test statistic is only slightly larger than the threshold.



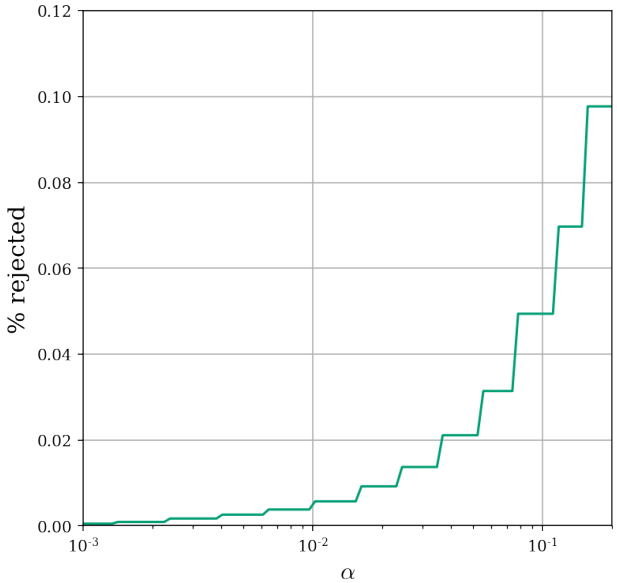
(a) Cumulatives for two datasets of 100 randoms variable (green and blue), and the cumulative function given by Eq. (28) (pink).



(b) Distribution of the maximum distance between two dataset of 100 random variables for 100 experiences. The rejected level is also displayed (see legend for the colors information).



(c) Distribution of the maximum distance between two dataset of 100 random variables for 10000 experiences. The rejected level is also displayed (see legend for the colors information).



(d) percentage of rejected Hypothesis in terms of α.

Figure 14: Results of the Kolmogorov hypothesis test.

In our case, we want to verify if the two samples come from the same PDF. Then we need to modify Eq. 29 as

$$D = \sup\left(|\mathrm{cum}_{1,n} - \mathrm{cum}_{2,m}|\right) , \tag{30}$$

where $\mathrm{cum}_{1,n}$ ($\mathrm{cum}_{2,m}$ is the cumulative function of the first with n data (second respectively with m data). The coefficient must fulfill the condition

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n\cdot m}} \overset{n=m}{=} c(\alpha)\sqrt{\frac{2}{n}} , \tag{31}$$

$$c(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right)\cdot\frac{1}{2}} , \tag{32}$$

where alpha is the threshold we set up. Some values of $c(\alpha)$ are resume in table 2

α	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.138	1.224	1.358	1.48	1.628	1.731	1.949

Table 2: Different values of $c(\alpha)$ for the Kolmogorov test.

Fig. 14a shows the three cumulative from the first two sets of 100 energies and from the Eq. (28). We see that around $E_\omega \sim 50$[K] the two generated cumulative deviate the most from each other. The maximal distance is $D = 0.08$. We can repeat this calculation N_{exp} times and fill the results in a histogram. Figs. 14b and 14c shows the repetition of 100 and 10000 respectively. In the first histogram, there is not enough data to find a tendency in the distribution of the maximal distance D, but we can already see for α sufficiently high some D are already rejected for $\alpha \gtrsim 0.025$. We increased the number of experiences to 10000 and we see that the shape of the histogram is similar to the expected Kolmogorov distribution. Fig. 14d shows the percentage of rejected experiments in terms of α for the $N_{\mathrm{exp}} = 10000$ sample. In Fig. 14d. For example, if we consider $\alpha = 0.1$, there are $\sim 95\%$ of D that do not reject the hypothesis: which means that in 95% of the cases, we cannot reject the null hypothesis that the two datasets originate from the same distribution, at significance level 10%. In fact, α represents the percentage of rejected hypotheses.

We also increased the number of generated energies for each experiment. This has the effect of globally reducing D. Indeed, when $N \to \infty$ the cumulative cum_k will converge to C_f. In consequence, the maximum difference between two cum_k will reduce. For $N_1, N_2 = 200$ we have only $\sim 1\%$ of rejected hypothesis for $\alpha = 0.01$.

7 Task-6

The Fisher information which is given by

$$I_{E_\omega}(T) = \mathbb{E}\left[-\left.\frac{\partial^2 \ln L(E_\omega \mid T)}{\partial^2 T}\right|_{T=T_0}\right] , \tag{33}$$

gives an estimation to the variance $V(\hat{T})$ of an estimated parameter as

$$V(\hat{T}) = \frac{1}{I_{E_\omega}(T)} . \tag{34}$$

Note this is equality because our PDF is Gaussian which is a special case. However, for a generic PDF, this should be an inequality where the Fisher information is a lower limit to the variance, which means that the precision of an estimated parameter is fundamentally limited by the Fisher information of the likelihood function. If we take the results of the moment estimation method we have the variance $V_{\text{moment}} = 0.0367[\text{K}^2]$. But if we calculate it with Eq. (33) we find $V_I = 0.0360[\text{K}^2]$. Which confirms the statement.

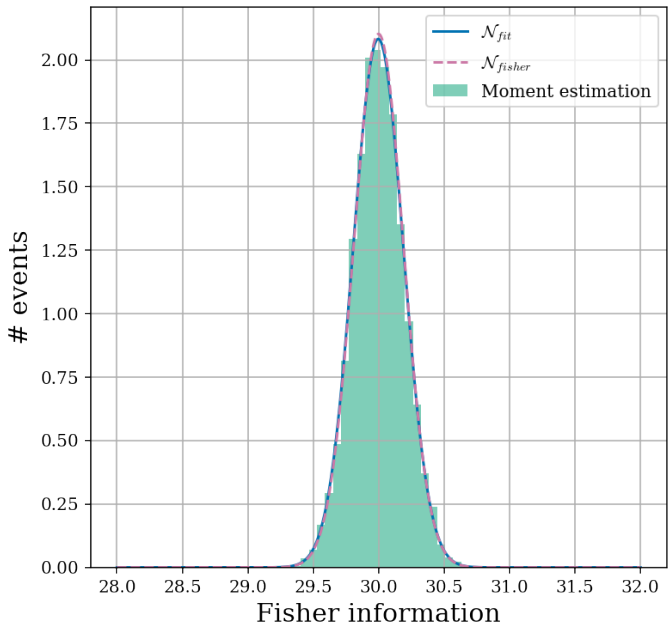


Figure 15: Plot of the moment histogram with the fitted envelope and the fisher information Gaussian envelope

Fig. 15 shows the comparison between the two Gaussian envelopes. With the two very close variances, the two curves are almost indistinguishable.