Statistics: Final Project 1

Gilles Koziol

Contents

1 **Basic statistical properties** 2

2 **Convergence** 6

3 **χ^2 distribution** 8

4 **Parameter estimation** 9

5 **Hypothesis testing** 14

6 ***Bonus:* Measurement with systematic errors** 16

Introduction

In this project, we'll consider a box that contains ^{40}K atoms that can be described as fermions. The mass of each atom is simply given by $m_{^{40}K} = 40\, m_0$, where m_0 is the atomic mass : $m_0 = 1.66 \cdot 10^{-27}$ kg.

Thanks to statistical mechanics, we know that the energy density ϵ of fermions at low temperature can be described by :

$$f(\epsilon, T) = A\frac{\sqrt{\epsilon}}{e^{\beta(\epsilon-\mu(T))}+1}$$

The coefficient A is a constant and here, this constant is given : $A = 3.5 \cdot 10^{31}$. T is obviously the temperature. We'll set it at $T = 3K$. μ is the chemical potential of the system. Here it is given that $\mu(T) = 30\, k_B T$.[1]

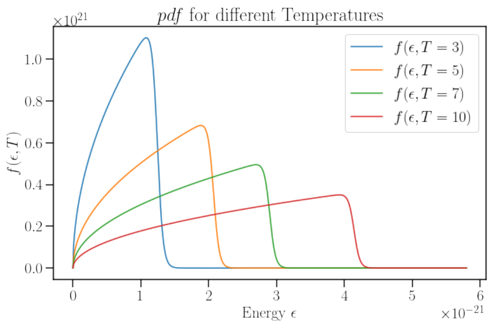

Figure 1: Graph of the *pdf* for different temperatures

It is also useful for the rest to look now at figure 1 and 2 ,which are the plots of the pdf for different temperatures. We can see that the function has roughly the shape of $\sqrt{\epsilon}$ until $x = \mu$. At this point, the functions decreases

[1] We see with this relation that μ has the same units of $k_B T$, which is an energy in J.

rapidly. The size of the interval in which the function decreases is $\mathcal{O}(k_B T)$. This behaviour is because in statistical physics, any density function are related by the density quantity we want to compute (the energy density $\sim \sqrt{\epsilon}$ in our case) times the Fermi-Dirac function. At low temperature[2] the Fermi-Dirac function behaves almost like a Heaviside-function, centered at $x = \mu$, which "kills" all the higher energy values.

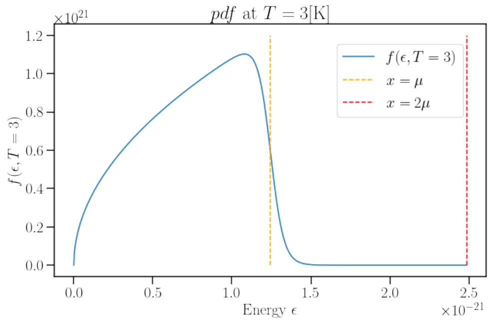

Figure 2: Graph of the pdf at $T = 3$[K]

[2] $T = 3K$ is a good approximation of "low" temperature

1 Basic statistical properties

a) Show that the energy density distribution is actually a Probability Density Function (pdf)

In order to be a pdf, a distribution needs to fulfill two conditions :

1) The pdf has to be greater or equal to zero for any given value of energy ϵ:

$$f(\epsilon) \geq 0 \,\forall \epsilon \,\in [0, \infty[$$

Proof. Let's analyse both the numerator and the denominator of f:

The numerator contains the constant A which is positive and also the square root of the energy, which is also positive since the energy is positive.

The denominator contains the exponential, which is always positive, even if the energy ϵ is greater than μ. In this case we have a negative exponential that converges to 0 asymptotically but is never smaller. We also have the $" + 1"$ that is obviously positive.

Overall, both the numerator and the denominator are positive, which proves the 2^{nd} property. □

2) The integral on the whole domain has to be 1. In our case, the energy has to be positive, so f must verify:

$$\int_0^\infty f(\epsilon)\, d\epsilon = 1$$

Proof. We can see in the literature [3] that this function has no analytical primitive and needs to be solved numerically. There are two ways to proceed.

[3] See : https://arxiv.org/ftp/arxiv/papers/0811/0811.0116.pdf

The first one is the lazy one, that is to use the built in function `NIntegrate` from Mathematica and see that the result is close to 1. The code is the following:

$$\texttt{NIntegrate}\left[\,\frac{A\sqrt{\epsilon}}{e^{\beta(\epsilon-\mu)}+1},\{\epsilon,0,2\mu\}\right.$$
$$,\texttt{Method}\rightarrow\texttt{"TrapezoidalRule"}$$
$$\left.,\texttt{MaxPoints}\rightarrow 5000\,\right]$$

And the output is `1.02272`, which is close to 1. The upper limit of the integral has been set to 2μ because of the behaviour of the pdf that we explained in the introduction.

The second way to do it is to build a small program that approximates the area under the function with trapezes. The idea of it is to take a given number n of subdivision of the interval $[0, 2\mu]$ and compute the area of the $n-1$ trapezes below the curve. The total area is the result of the integral. We can see on figure 3 that the result of the integral converges asymptotically to 1, which is exactly what we have to show. Note that for $n = 5000$ points, the numerical approximation gives 1.02271. The result is very close to the one provided by Mathematica, which is a good sign, because we also used a trapezoidal integration method.[4]

□

Since the function has the temperature as a parameter, if we change it, the function changes and is not normalised anymore. To correct it, we wanted to change the A parameter so that the function is normalised for all T. In order to do this, we computed numerically the value of the new coefficient, namely $A(T) = 1/\int_0^{2\cdot 30 k_B T} f(\epsilon, T)d\epsilon$ for a lot of temperature values. After that, we used the `scipy.interpolate` library to interpolate the given result with a third degree polynomial function. As we can see on figure 4, the coefficient changes a lot, especially for low temperature values. With this correction factor, we can express the function as :

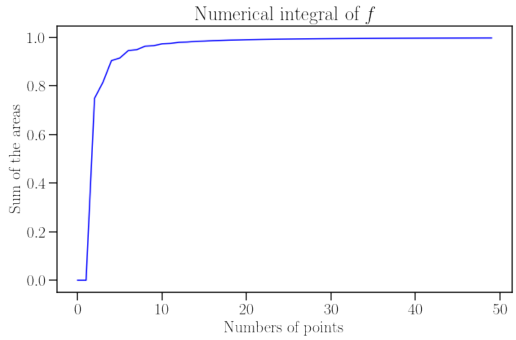

Figure 3: Numerical result of the integral performed for various numbers of trapeze under the curve. We clearly see that the result converges to 1.

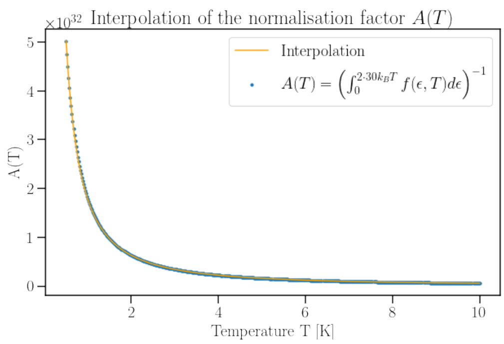

Figure 4: Normalisation factor as a function of the temperature

[4] In order to have a better approximation, one could use more advanced methods, like the Simpson's method but it is clearly not relevant here.

$$f(\epsilon) = A(T)\frac{\sqrt{\epsilon}}{e^{\beta(\epsilon - 30k_BT)}+1}$$

After that, we also computed numerically the value of the expectation value and the variance. We found that there was a difference between the given one and the computed one. The values we found are the following :

$$\mathbb{E}_{num} = 7.493 \cdot 10^{-22}$$
$$var = 1.09 \cdot 10^{-43}$$

The numerical value of the given expectation value is $\mathbb{E}_{th} = 7.455 \cdot 10^{-22}$, which is different from the obtained one by a factor 0.9949. This is really small, but it can solve some issues as we will see later.

b) Build a MC code to generate $N \gg 1$ data points,and verify their correctness by comparing them with the curve of the analytical form

To perform this task, we simply used the accept-reject method described in the lecture. We used a uniform envelope with a height of 1.05 the maximum of the "y" array created by python. We used this method because to solve the equation $f'(\epsilon) = 0$ is not a simple task. The obtain result can be seen on figure 5:

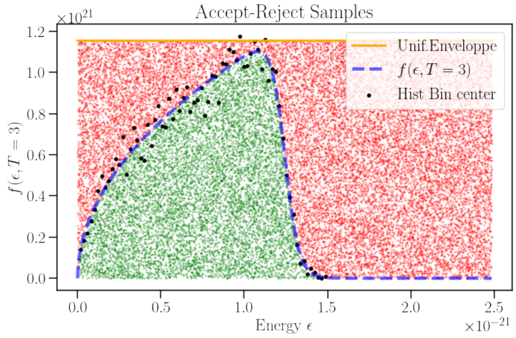

Figure 5: Graph of the accepted and rejected samples for $N = 10^4$ events. We see that the histogram's bin centers in black fits nicely the pdf shown in blue. The option `density = True` has been put in order that the area under the histogram is normalised to 1.

c) Compute the sample mean $\langle\epsilon\rangle$ and compare with the expectation value : $\mathbb{E}[\epsilon] \approx \frac{3}{5}n\mu$

For this part, one can use the facilities of the built-in `numpy mean` to explicitly compute the sample mean. We ran a MC simulation that generated $N = 10^4$ events, and we find a sample mean of $\langle\epsilon\rangle = 7.477 \cdot 10^{-22}$.

If we want compute the mean with a small code that we build ourselves, it is a good idea to compute the mean by using the bins, each of them having a height of h_i, a constant width of w and centered at x_i. With those parameters, the mean of the bin content is :

$$\langle\epsilon\rangle = \sum_{i \in N_{bins}} h_i \cdot w \cdot x_i$$

with this method, we find $\langle\epsilon\rangle = 7.477 \cdot 10^{-22}$, which is as expected the same value of the `numpy` mean.

There is a difference between the sample mean and the expectation value, because we used a finite number of events. We'll see later that the mean is actually converging to the expectation value.

Note: Remember that in a physical solid material like ^{40}K, there are $\mathcal{O}(10^{23})$ fermions, which provides a much better statistic, which will make the mean converge to the expectation value to a huge precision. We'll see it later. For a more visual result, we can take a look at figure 6 that shows visually the result of the mean and the expectation value.

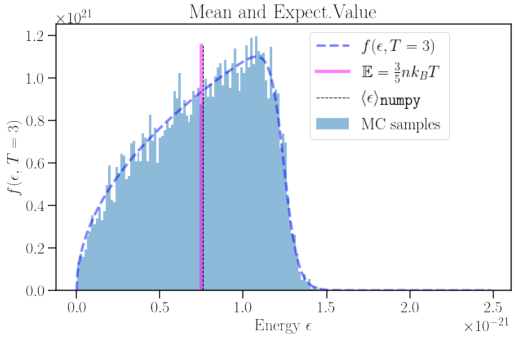

Figure 6: Histogram of the MC generated samples for $N = 10^4$ and 100 bins. We see that the expectation value, and the mean are really close to each other.

d) Calculate $\mathbb{E}/k_BT$.

This task is a direct consequence of the previous task, where we simply inspected "by eye" that the results were approximately the same. Here, we divide the expectation value by k_BT. If we inject the value of μ and n in the expectation value, we have (with the correction we added, it is the same computation but with an additional 0.9949 factor):

$$\mathbb{E} = \frac{3}{5}n\mu = \frac{3}{5}n \cdot 30\, k_BT = 18k_BT$$

,which gives 18 if we divide by k_BT. That means that the ratio $r = \langle\epsilon\rangle/k_BT$ should roughly give the same result. With the mean from `numpy`, we find a ratio of $r = 17.96$ and for the mean form the self-built code, we also find $r = 17.96$, which are close to 18 as expected. Again, we only have 10^4 events so the difference is also expected.

If we perform the same computations with $N = 10^7$ events, the ratio are : $r = 18.01$ for the two methods. We see that those result are closer to 18.

e) Compute the variance, skewness and kurtosis of the distribution and comment.

This time, we won't compute the desired quantities by hand, because it is already implemented in the packages `numpy` and `scipy.stats`.

For the variance, the result is : $\sigma^2 = 1.09 \cdot 10^{-43}$, which implies a standard deviation of $\sigma = 3.31 \cdot 10^{-22}$. This result seems coherent because it is in fact the same that we computed numerically.

For the skewness, the result is $skew = -0.28$, which means that the distribution is asymmetrical (as we could already tell with the histogram). Moreover, it is negatively skewed, and that means that the distribution is slightly shifted on the right. This was an expected result since the envelope of the pdf is $\sqrt{\epsilon}$, which is an increasing function and thus, the function is right skewed.

For the kurtosis, the result is $k = -0.912$. The kurtosis gives us information about the tails of the distribution : If the distribution has long tails, the kurtosis is big (*e.g.* the exponential pdf). If the distribution has small or no tails (*e.g.* a uniform distribution), the kurtosis will be negative. This is our case since the kurtosis is smaller than zero. The fact that the tails are small is because of the behaviour of the $\sqrt{\epsilon}$ that quickly grows near 0. On the end of the distribution, the temperature of $T = 3K$ implies that the Fermi-Dirac distribution acts almost like a Heaviside function, so the tail is very small.

2 Convergence

a) Take $N_{exp} = 1$ and show the law of large numbers based on the data points you generated

We will here show the strong law of large numbers. This law tells that if the condition[5] :

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

(where σ_i^2 is the variance of each event), then the mean of the sample $\langle\epsilon\rangle$ converges to the expectation value $\mathbb{E}_{new}$. Here, we are going to show that the mean converges to the expectation value by taking the mean of an increasing number of events, up to 10^8 events. The result is shown on figure 7. We clearly see that the mean converges to the expectation value.

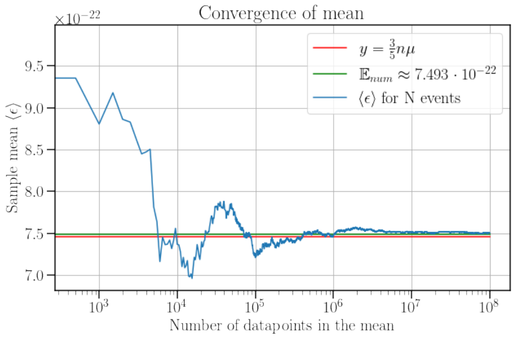

Figure 7: Graphic of the mean for an increasing number of sample, given a list of $N = 10^8$ MC generated sample.

We see here that the mean converges to the green line, whose height is the expectation value that we computed numerically. We see that the convergence is evident. This precisely show that with $N \sim \mathcal{O}(10^{23})$ events like in a physical solid material, the mean converges to the expectation value with a high accuracy.

[5]Here, this condition is satisfied because σ_i are constants and the sum of the inverse of squares converges.

An other way to visualise the convergence is to see that the ratio r that we defined before also converges. As we can see on figure 8, where we divided the ratio by 18, so that it converges to 1, the convergence is also evident.

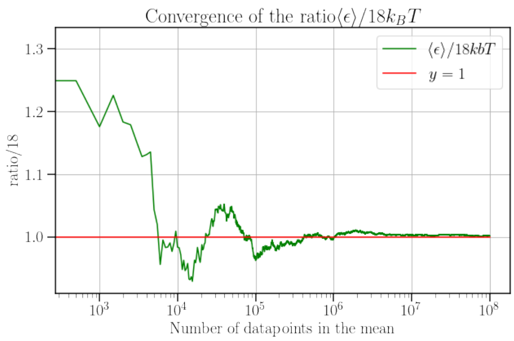

Figure 8: Graphic of the ratio for an increasing number of event. The red line represents the expected value : 1

b) Take $N_{exp} \gg 1$ and show the validity of the central limit theorem for $\langle\epsilon\rangle$

For this task, we did 10^4 experiment, each of them represents $N = 10^4$ events. For each experiment, we computed the mean and append it to a list that we put in a histogram. According to the central limit theorem, the sample mean should be distributed normally around the expectation value of the pdf (since all experiment comes from the same pdf). The standard deviation should be given by the standard deviation of the pdf divided by $\sqrt{N}$, the number of samples per experiment. In our case, since the pdf has no analytic primitive, we used the numerical value computed before. In order to see if the mean $\langle\epsilon\rangle$ is normally distributed, we also plot a Gaussian with mean $\mu = \mathbb{E}_{num}$ and the standard deviation given from the numerical integral divided by $\sqrt{N}$. The result is shown on figure 9.

We can see that the sample mean is effectively normally distributed, which is exactly

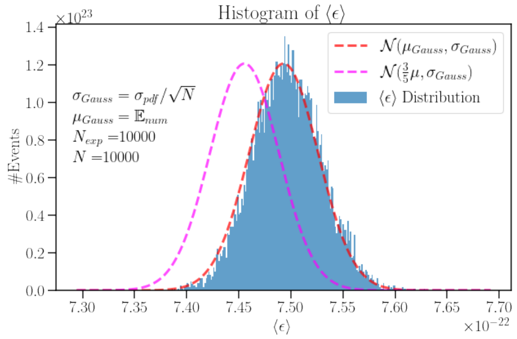

Figure 9: Distribution of the means for $N = N_{exp} = 10^4$.

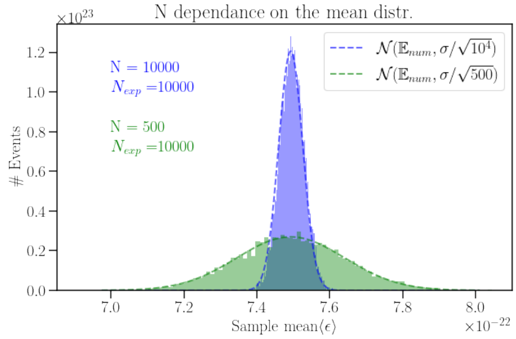

Figure 10: Histogram of the $\langle\epsilon\rangle$ distribution for two different sample sizes.

what we expected. We also added a magenta Gaussian centered at the expectation value $\mathbb{E} = \frac{3}{5}n\mu$ and we clearly see that there is an offset. This has been corrected by redefining the amplitude coefficient and thus the expectation value.

An important factor in the central limit theorem is the number of element N in the sample to compute a single mean, since the distribution should have a standard deviation $\sim \sigma/\sqrt{N}$. Indeed, if we compute the mean of two data samples of $N = 10^4$ and $N_2 = 500$ events, compute their mean and repeat the procedure $N_{exp} = 10^4$ times, we can put the obtained $\langle\epsilon\rangle$ distribution. This is exactly what figure 10 shows. We clearly see that for $N = 500$, the distribution is much flatter.

The number of experiment does not influence the shape of the distribution, it simply implies that we have more means, which implies that the shape of the distribution is less noisy.

c) Compute the variance of $\langle\epsilon\rangle$ and compare it with the value you expect.

This step is to make sure that the variance of the means $\langle\epsilon\rangle$ corresponds to the mean of the Gaussian. If we take the ratio of the two, the result is 0.998 (for $N = 10^4$), which is obviously close to 1. This shows that the means from the samples are effectively normally distributed. Here, we only used $N = 10^4$ events and $N_{exp} = 10^4$. Remember also that according to the central limit theorem, the variance of the means is $\sigma/\sqrt{N}$. Therefore, in a physical solid, those numbers are of order $\mathcal{O}(10^{23})$ and the variance of the means tends to be very small (indeed, of order $\mathcal{O}(\sigma \cdot 10^{-23/2} \approx 10^{-11})$).

3 χ^2 distribution

a) What is the expected p.d.f. for the number of entries per bin?

The number entries per bin follows a Poisson distribution. However, since we generated a lot of samples, the number of events per bins is large. In this case, we have seen that the Normal distribution is a good approximation for the Poisson distribution as we can see on figure 11.

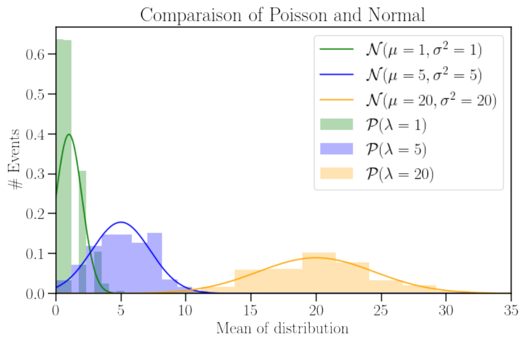

Figure 11: Histogram of Poisson distribution and Gaussian function for different values of distribution mean.

As a result, since the bin content is larger than ~ 5, we can consider that the bin content is normally distributed around a mean $\mu(\epsilon)$ that is given by the pdf itself. In order to try this hypothesis, we'll compute the χ^2 and compare it with the theoretical distribution.

b) Show that the χ^2 of the obtained entries per bin follows a χ^2 distribution.

In order to do this, we will generate N_{exp}, each one of them containing $N \gg 1$ events. For each bin, we are going to compute the quantity :

$$X^2 = \frac{(\text{CONTENT} - \text{EXPECTATION})^2}{\text{EXPECTATION}}$$

,where "CONTENT" is the number of entries in a specific bin (of width w) and EXPECTATION is the number of entries that one would expect in theory. The expectation is given by $f_i = N \cdot \int_{\epsilon_i - w/2}^{\epsilon_i + w/2} f(\epsilon) d\epsilon$. This can be computed easily with `quad`. It is also possible to use the approximation that the function vary sufficiently slowly to consider the function as a constant on the interval of the bin width w. Therefore, the expectation value for each bin is :

$$f_i = N \cdot f(\epsilon_i) \cdot w$$

We will use both, and compare the result to see if the approximation is valid. If we compute this for, say $N_{exp} = 2000$, $N = 2000$ and $N_{bins} = 10$, we get the figure 12. We have set the height of the bins to be normalised, so that the theoretical χ^2 distribution with $10 - 1 = 9$ degrees of freedom (because there's only one parameter[6]) has the same height.

We see that the distribution that we computed with the integral fits very nicely to the theoretical χ^2 distribution, whereas the distribution computed with the approximation has a nice shape but it is completely offset. The reason is probably that for a low bin number the approximation is not that accurate because of the sudden decrease of f around $\epsilon = \mu$.

That means that our assumption to consider the number entries per bin as a Gaussian distribution centered at $\mathbb{E}_{num}$ was indeed good, but we can not assume that the function vary sufficiently slowly to use the above approximation (at least for a low bin number).

c) How does the χ^2 distribution change with the number of bins?

If we now do the same process for different number of bins, we expect the peak of the dis-

[6]The total number of events is fixed, so the content of one bin can be deduced from the others

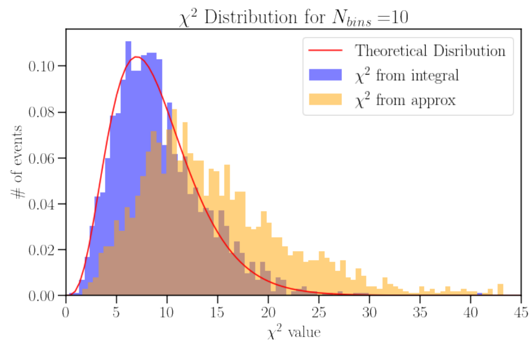

Figure 12: Histogram of the χ^2 values for 2000 separate experiment. The thick line represent the χ^2 distribution with $10-1=9$ degrees of freedom.

tribution to move to the right, since the degree of freedom is $N_{bins}-1$. To probe this, we did exactly the same, with $N_{bins}=\{5,10,15,25\}$. On top of that, we plot the expected χ^2 distributions with $\{4,9,14,24\}$ degrees of freedom. The result is shown on figure 13.

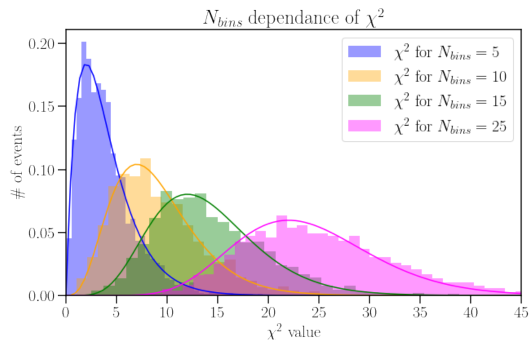

Figure 13: Histograms of the χ^2 obtained with different numbers of bin. The straight lines are the χ^2 distributions with $N_{bins}-1$ degrees of freedom.

The theoretical expectations fits very well to the samples. It is however hard to say if the peak is located at $N_{bins}-1$ or N_{bins} visually, because as N_{bins} gets bigger and bigger, we can approximate $N_{bins}-1 \approx N_{bins}$.

4 Parameter estimation

We now generate a list of $N=5000$ energies and we will try to recover the T parameter (it is equivalent to recover the μ parameter since $\mu=30k_BT$). We'll use different methods and compare them.

a) Compute the log likelihood $\ln L(x_1,x_2,...,x_N|\mu)$ at a given μ. Then, use the maximum likelihood method to estimate and its variance.

A good way to recover the parameter is the log likelihood. Given a parameter, it returns the likelihood that the given parameter is the right one. It can be computed with :

$$-\ln\left(L(\epsilon_1\dots\epsilon_N|T)\right)=-\sum_{i=1}^{N}f(\epsilon_i,T_{test})$$

This can be easily computed with a `for` loop. For instance, if we compute it for a sample of 5000 random energies with $T=3K$, we get: $-\ln(L)=-240638.7$. If we want to estimate the T parameter, we need to compute the log likelihood for a lot of test parameters and select the parameter for which the log likelihood is the minimum[7]. Once it is done, we can vertically shift all the likelihood so that the best parameter has a likelihood of zero. We tried 500 parameters between $T=0.5$K and $T=7$K. The result is shown on figure 14. We can see that the minimum is clearly around $T=3K$. To estimate the error, we use the same method as in exercise 7.5 from the tutorials. Indeed, the 1σ limit is given by the half of the difference of the points T_1 and T_2, for which the likelihood is 0.5 (since we normalised it).

If we now want the estimation of the μ parameter, we simply apply the classical formula

[7] We see here why it is important that the function is normalised for all T values.

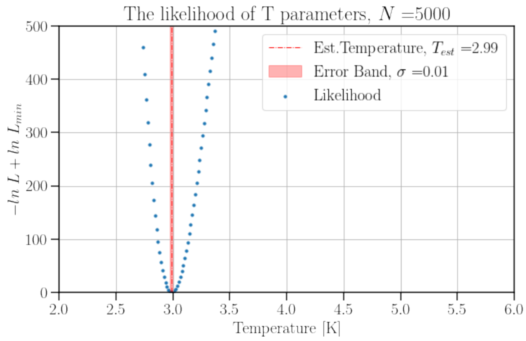

Figure 14: Graph of $-\ln(L)$ for various initial parameters. The parameter that minimises the log likelihood is in this case $T = 2.99 \pm 0.01K$.

of error propagation ($\mu(T) = 30k_BT$):

$$\delta\mu = \sqrt{\left(\frac{\partial\mu}{\partial T}\delta T\right)^2} \approx 0.54 \cdot 10^{-23}$$

We have found the μ parameter and its standard deviation : $\mu = (123.8 \pm 0.5) \cdot 10^{-23}[J]$. The value we should obtain is $\mu = 124.2 \cdot 10^{-23}[J]$, which is included in the error band of the estimated parameter,

We can perform this analysis for a large number of experiment ($N_{exp} = 10^4$), each of them containing $N = 5000$ samples and put the obtained data in a histogram. If we compare the obtained result with a Gaussian distribution centered at $T = 3$[K] and with a standard deviation defined as the mean of the standard deviations of the estimated parameter: $\sigma = \langle\sigma\rangle$, we obtain the histogram shown on figure 15. This shows that our temperature estimation is coherent. There is apparently no bias because the distribution we try to fit has $T = 3$[K] as parameter and the parameter distribution has a mean of $T = 2.99$[K]. We also tried to see if the bias is dependant of N by performing the simulation for different sample sizes, but it seems that it doesn't change a lot, since the mean for $N_2 = 500$ is $T = 2.99$[K] as well. The errors that we derived seems to be consistent because the standard error of the distribution matches the mean of the errors on the parameter.

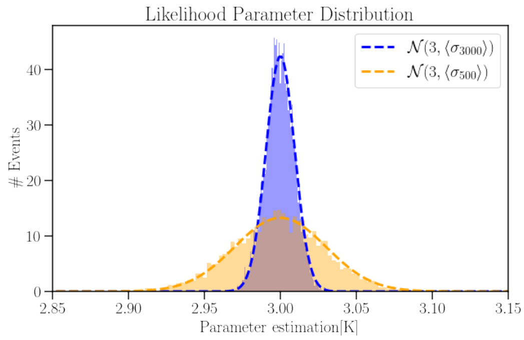

Figure 15: Distribution of the estimated temperature via the Log-likelihood method.

In order to reduce the computation time, it is also possible to use the `minimize_scalar` form the `scipy.optimize` library. It allows us to find the parameter that minimizes the given function. As we can see on figure 16, the shape of the parameter estimation distribution is the same but there are less bins.

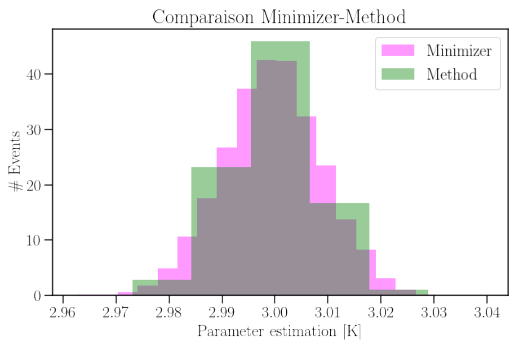

Figure 16: Comparison of the results obtained with the two different techniques.

The reason is because with the "hand made" technique, we use a bunch of points for which we test the likelihood. It implies that the obtained values are discrete, whereas for the minimizer, the obtained parameters are continuously distributed in a given interval. Since the results are the same, and it saves some time, we will use this method for the next computations.

b) Compute the goodness of fit at various μ and use the least squares method to estimate μ and its variance

Now, we will do a computation that is similar to the one of the χ^2. Indeed, we will define an array containing the N parameter to test and for each one of them, we will compute the following quantity:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - h(\epsilon_i, T_{estimate})}{\sigma_i} \right)^2$$

,where y_i is the i^{th} bin content. $h(\epsilon_i, T_{estimate})$ is the value of the function at $x = \epsilon_i$ with parameter $T_{estimate}$. Since in the chapter on the χ^2 distribution we concluded that the approximation $h(\epsilon_i, T_{estimate}) \approx N \cdot f(\epsilon, T_{estimate}) \cdot w$ was not really good, we will here again compute the numerical integral between two bin edges, so that $h_i = N \cdot \int_{\epsilon_i - w/2}^{\epsilon_i + w/2} f(\epsilon, T) d\epsilon$. Since the bin content is greater than ~ 5, we assume that the bin content is normally distributed and thus the standard deviation of each bin content is $\sigma_i \sim \sqrt{y_i}$.

We can again use the minimizer to search the T parameter that minimizes the χ^2. This parameter is going to be our temperature estimate. Now, to compute the error on this estimation, the process is going to be similar to the one of the likelihood. The only difference is that now, the "1σ" limit is defined as the points for which the chi square is one unit above the one of the estimated parameter[8].

$$\chi^2_{1\sigma} = \chi^2_{min} + 1$$

The obtained result is shown on figure 17.

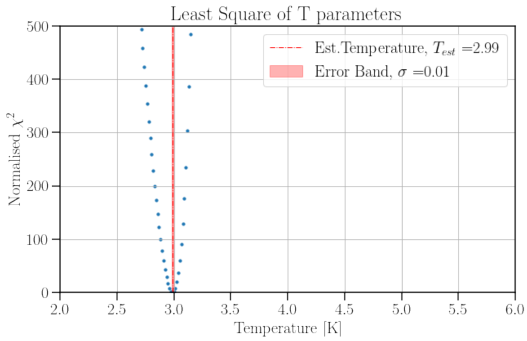

Figure 17: Graphic of the normalised least square. The parameter that minimises the least square is in this case T= 2.99K

To obtain the μ parameter, we can again apply the error propagation formula to find: $\mu = (123.8 \pm 0.3) \cdot 10^{-23} [J]$. Here again, the estimated parameter is satisfying because the theoretical value is included in the error bands.

We can here too, perform a set of two times 10^4 experiment, each one of them containing $N = 5000$ and $N_2 = 500$ samples. The estimated parameter distribution is shown on figure 18. As we can see, the case with 5000 samples per experiment is behaving as one could expect: peaked around $T = 3$[K] (the distribution mean is at $T = 2.99$[K]) and with a Gaussian shape, whose standard deviation is given by the mean of the parameter's standard deviation.

On the other hand, the case with $N_2 = 500$ has some issues. A lot of parameters were estimated at $T \approx 7$[K] as we can see on figure 19. There are however some parameters

[8] The reason is (lecture chap 10.1) that the log likelihood of the "pdf associated to one of the points (x_i, y_i) is given by" : $-2\ln(L) = \sum_i \left(\frac{y_i - h(x_i, a)}{\sigma_i}\right)^2$. The factor of "2" implies $2 \times 0.5 = 1$.

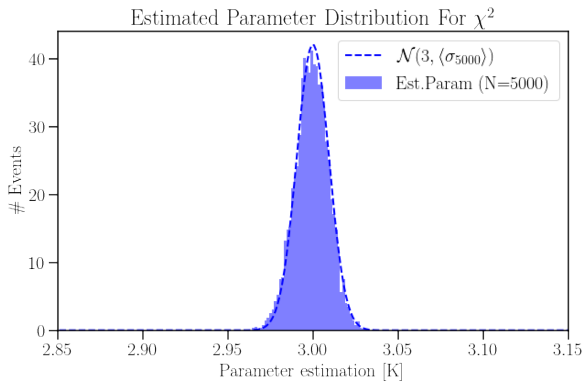

Figure 18: Distribution of the estimated T parameter for $N_{exp} = 10^4$ and $N = 5000$.

around $T = 3$[K] (their mean is 2.99[K]), but they represent only 38% of the cases, which is not a lot.

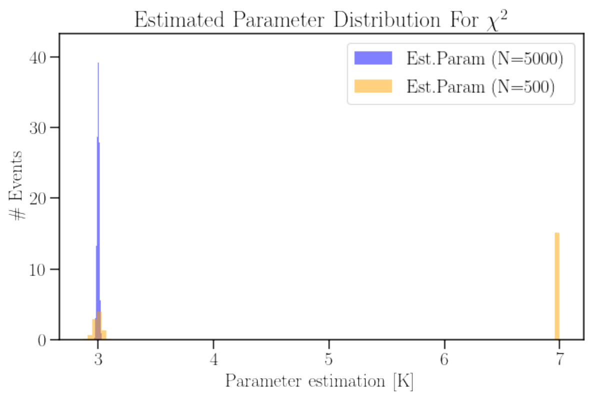

Figure 19: Estimated parameter distribution with two different sample size

The reason of this instability seems to be an other minimal region of the χ^2 for higher temperature values (see figure 20), and that it requires a much bigger sample to find the correct parameter.

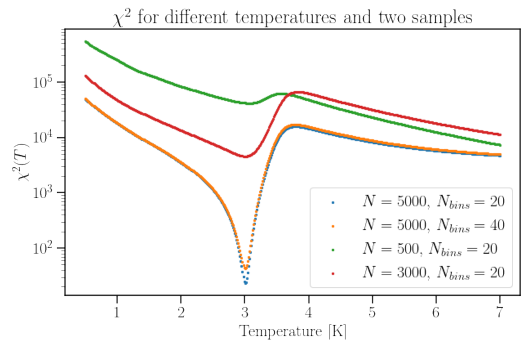

Figure 20: Example of the χ^2 for different bin numbers and sample sizes. We see that the bin number is not as determinant as the sample size

Notes on experiments Because we used the numerical integral to estimate the value of the function at the bin center, it requires a lot of computation time. It is reduced with the use of the minimizer, but the computation time could still be of the order of a couple of hours for $N = 5000$ and $N_{exp} = 5000$. We wanted to reduce the computation time so that the whole jupyter notebook could run entirely during the night. We noticed that for the χ^2 method, the computation time of a single experiment increases linearly with the bin number as we can see on figure 21.

We see also that for the generated sample, the result was the same no matter what bin number has been chosen. Therefore, we put a low bin number : 20.

c) The parameter μ can also be estimated by computing the sample mean $\langle\epsilon\rangle$ of Eq. (1) with the MC integral. Perform this estimation of μ and its corresponding variance.

In order to do this, we are going to use the method of moments. Indeed, we know the expectation value for $\langle\epsilon\rangle$ as a function of the μ parameter (or in function of T, it is equivalent).

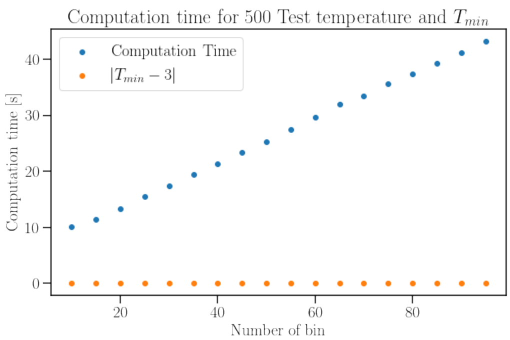

Figure 21: Graphic of the computation time as a function of the bin number. We also added to the graph the result of the least square minus 3 for each case.

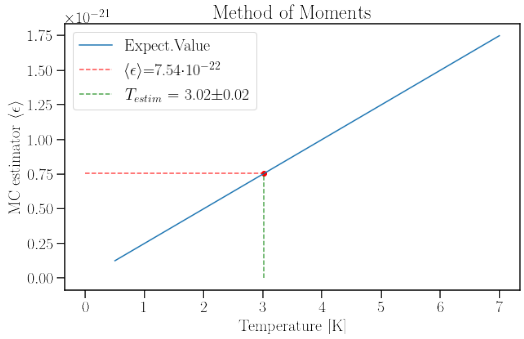

Figure 22: Graph of the expectation value as a function of the temperature, with the sample mean and the corresponding estimated parameter, here 3.02[K].

Analytically, the mean is of the form (0.9949 is the correction between the theoretical mean and the computed one):

$$\begin{aligned}\mathbb{E}_{new}(\epsilon) = \int \epsilon f(\epsilon, T) d\epsilon &= \frac{3}{5 \cdot 0.9949} n\mu \\ &= \frac{3}{5 \cdot 0.9949} n 30 k_B T\end{aligned}$$

What we are going to do is to compute the mean of the data sample we have using the estimator of the j^{th} momenta :

$$\langle \epsilon^j \rangle = \frac{1}{N} \sum_i^N \epsilon_i^j \approx \mathbb{E}_{new}(\epsilon^j)$$

We can then invert this relation for the mean (*i.e:* j=1) to find :

$$\mu_{est} = \frac{5 \cdot 0.9949}{3} \langle \epsilon \rangle \quad , \quad T_{est} = \frac{5 \cdot 0.9949}{3 \cdot 30 k_B} \langle \epsilon \rangle$$

If we plot the relation between the temperature and the expected value with a line representing the mean of the generated sample and the corresponding temperature estimation, we get the figure 22. The corresponding μ value is : $\mu = (125.0 \pm 0.8) \cdot 10^{-23} [J]$.

We see that the error on the estimation with the method of moments is also relatively small and leads to a small uncertainty. If we perform the same analysis $N_{exp} = 10^4$ times, we obtain the figure 23.

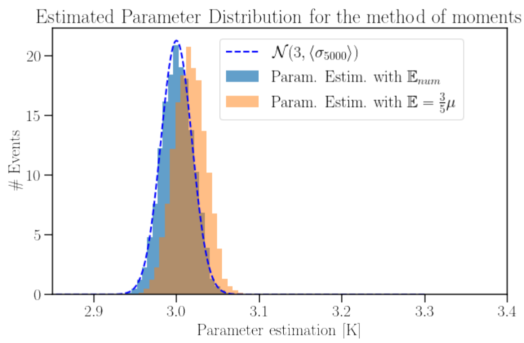

Figure 23: Distribution of the estimated temperature via the method of moments.

We see that the distribution is normally distributed with a mean of $T = 3.00$[K], which indicates that the estimator is unbiased as written in the lecture notes. Additionally, since this method uses the expectation value to deduce the parameter, we also did the same computation with the mean that hasn't been corrected (in orange). We can see that the deviation is obvious, which means that the correction on the expectation value was necessary in order to have a correct parameter estimation. Moreover, it also means that the correction was precise.

d) Compare the results from the three questions above and discuss them.

As we could see, the three methods we've used gave some good results. However, we can see on figure 24 that the three methods are all centered around $T = 3$[K], but the width is not the same. We clearly see that the χ^2 and the likelihood give roughly the same results in terms of estimate parameters and errors. For the method of moments, the distribution is wider.

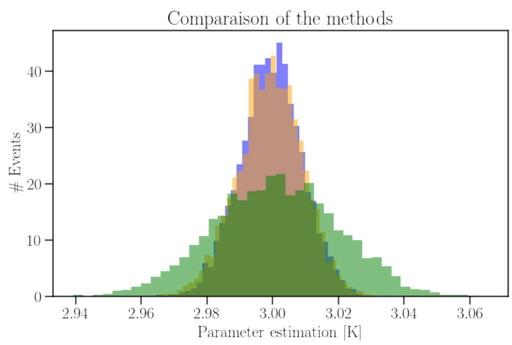

Figure 24: Comparison of the three methods we've used to estimate the parameter

From a more practical point of view, the chi square method requires a lot more computation time, whereas the likelihood and the method of moments are very fast, because they don't need as much computation. Moreover, the likelihood is not as sensitive as the χ^2 method. We saw that for a smaller sample the χ^2 method was unstable because there are other minimal points.

5 Hypothesis testing

Now, we are going to test the hypothesis that two experiment are obtained from the same pdf. This will be our "null hypothesis": H_0, and the hypothesis H_1 is that the two samples follow different pdf. To do this, we are going to set a threshold : 5% seems to be a standard, so we'll stick to it. The Kolmogorov test procedure is the following :

1) Get the cumulative function of the generated sample. We use the facilities of `scipy.stats.percentileofscore` to do this. This function takes two parameters: an array and a number. It returns the percentage of the array elements that are smaller than the number. If we divide this number by 100 (to get a percentage) and perform a `for` loop for each x in a `np.linspace` we get the cumulative distribution for the generated sample. An example of the obtained result is shown on figure 25.

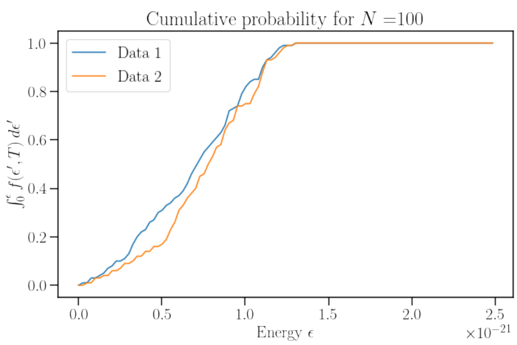

Figure 25: Example of the cumulative function for two experiment, each one of them containing $N = 100$ events.

We see that the behaviour is as one could expect, because the end of the cumulative is really flat, which is coherent with the "Kroenecker-Delta" behaviour of the energy density.

2) Take the maximal absolute distance between the two cumulative functions.

3) Compare the maximal distance with the quantity that we will call $\mathcal{D}_{crit}$ [9]:

$$\mathcal{D}_{crit} = c(\alpha) \cdot \sqrt{\frac{m+n}{m \cdot n}}$$

, where m and n are the sample size, and since they are equal, the square root simplifies as $\sqrt{2/n}$. The $c(\alpha)$ is the critical value, which is a function of the statistical significance α. For $\alpha = 5\%$, we have $c(\alpha) = 1.358$. If the max distance is smaller than $\mathcal{D}_{crit}$, then we can't reject the null hypothesis, but if it is bigger, then the null hypothesis can be rejected at level α.

If we take a pair of two experiment containing $N = 1000$ events, then $\mathcal{D}_{crit} = 0.061$. For the two generated samples, the maximal distance is 0.033. In this particular case, the maximal distance is smaller than $\mathcal{D}_{crit}$, which implies that we can not reject the null hypothesis (*i.e.* that the two samples have the same pdf).

After some research, we found that there exists a function from the library `scipy.stats` that, given two sets of data, compute the maximal distance. The result for the data set is the same, which indicates that we did it properly.

We now repeat this $N_{exp} = 100$ times, each one of them, we compute the maximal distance, compare it to the critical value $\mathcal{D}_{crit}$, and check if we can reject the null hypothesis. It appears that in 97% of the cases, it is not possible to reject the null hypothesis. This is obviously the expected result, since the two data samples are effectively distributed with the same pdf. If we do the same with $N_{exp} = 10^5$, the null hypothesis can not be rejected for about 95.5%, which is coherent because the α value has been precisely set to 5%.

If we try to change the α value to 0.01, then it is harder to reject the null hypothesis and the result is coherent, because 99.89% of the times, the null hypothesis was not rejected.

As a last test, we tried to lower the sample size to see if there was a difference. We set $N = 200$ and $\alpha = 5\%$. On the 10^5 pairs of experiment generated, we could not reject the null hypothesis in only 20% of the cases. It shows that having a bigger sample size implies better results, because in this case, the pdf was also the same for both pairs of data.

[9] `https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test`

6 *Bonus:* Measurement with systematic errors

For this task, we will first generate $N = 5000$ data points and we will assume that the energy value can not be determined precisely, which means that we will assume that the uncertainty of measurement follows a normal distribution centered at zero and with a standard deviation $\sigma = 2k_BT$. That means that to each data point, we will add a random value $x_{err} \sim \mathcal{N}(0, 2k_BT)$. To visualise the effect of the uncertainty of measurement, one can take a look at figure 26.

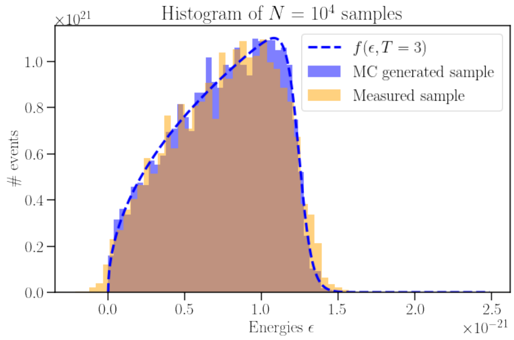

Figure 26: Histograms of the generated sample and the measured energies.

We see that there are some values of energies that are smaller than zero. For mathematical and physical reasons, those values are not acceptable[10], so we will only consider positive (or null) values.

Then, we again try to estimate the temperature and the μ parameter via the three methods we've seen before. The first one is the likelihood.

We see that the log-likelihood has its minimum at $T = 3.05 \pm 0.01$[K], which corresponds to $\mu = (126.5 + -0.3) \cdot 10^{-23}[J]$, which is above the expected result. Indeed, we performed this

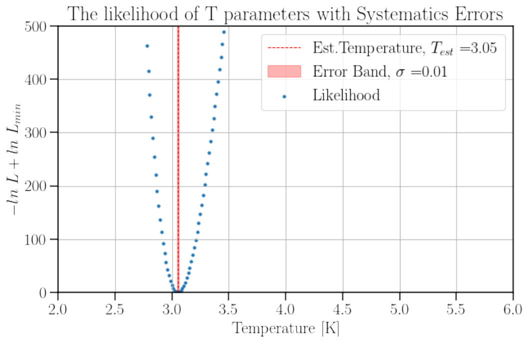

Figure 27: Graph of the likelihood for a sample of size $N = 5000$.

experiment $N_{exp} = 5000$ times and as shown on figure 28, we observe that the estimated parameter is biased.

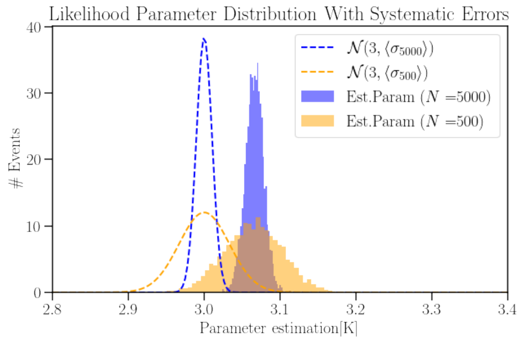

Figure 28: Distribution of the estimated parameter obtained with the log-likelihood method and with systematic error measurements.

To be complete, we did this experiment for $N = 500$ to be sure that there is no obvious N dependence on the bias. The two distributions are centered at $T = 3.06$[K] for both sample sizes. The reason of this shift is probably due to the shape of the distribution. Indeed, when the temperature increase, the energy distribution is flatter, which is also the effect of adding a small measurement error as we can see on figure 26.

[10]It is in fact impossible to get negative energies and we do not consider complex values here.

The second estimation method is the χ^2. We took the same data sample that we analysed with the likelihood. The result is shown on figure 29.

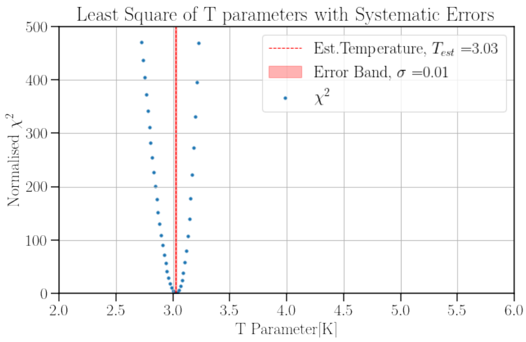

Figure 29: χ^2 of the data sample with systematic errors.

We also see that the estimated temperature $T_{estim} = 3.03 \pm 0.01$[K], which is higher than the one we expect. Again, we performed this $N_{exp} = 5000$ times, which leads to the histogram shown on figure 30. The Gaussian represents the result we would expect if the distribution wasn't biased.

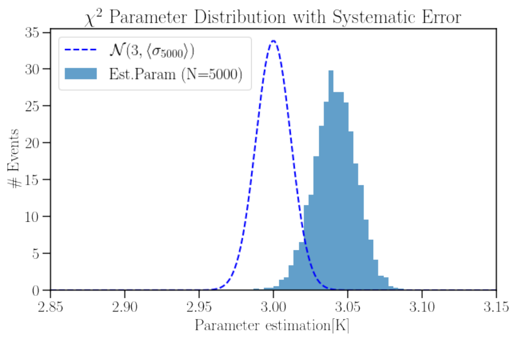

Figure 30: Distribution of the parameter obtained via the least square method

The distribution is also biased, because the mean of the distribution is at approximately $T = 3.04$[K]. The reason of this bias is probably the same that the one of the likelihood. Indeed, those two methods are based on the values of the bin contents, which are modified because of the systematic error measurements. It is however a good sign that the two methods give approximately the same result, even if it is slightly biased.

On the other hand, the method of moments is based on the mean, and since the errors are normally distributed, their mean is zero. If we call $\tilde{\epsilon}_i$ the measured energies and ϵ_i the MC generated samples (without errors), we have:

$$\begin{aligned}
\langle\tilde{\epsilon}\rangle &= \frac{1}{N}\sum_{i=1}^{N}\tilde{\epsilon}_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\epsilon_i + x_{err} \\
&= \frac{1}{N}\sum_{i=1}^{N}\epsilon_i + \underbrace{\frac{1}{N}\sum_{i=1}^{N}x_{err}}_{=0} \\
&= \langle\epsilon\rangle
\end{aligned}$$

This means that we expect that the systematic errors do not influence the method of moments[11].

For an experiment with $N = 5000$ samples, we obtain the figure 31. As we can see the result is $T = 3.00$[K]. The corresponding μ parameter is : $\mu = (124.4 \pm 0.8) \cdot 10^{-23}[J]$.

If we perform $N_{exp} = 5000$ experiments (see figure 32), we clearly see that the systematic errors does not affect the estimated parameter, since the mean of the distribution has a value of $T = 3.00$[K].

To be complete, we can compare the three parameter distributions shown on figure 33. It is now obvious that the method of moments is the most accurate one and less sensitive to the systematic errors. The likelihood is this time a bit more biased than the χ^2 method, but the errors are roughly the same.

[11]This is true because the systematic errors are distributed normally and it wouldn't be true in general

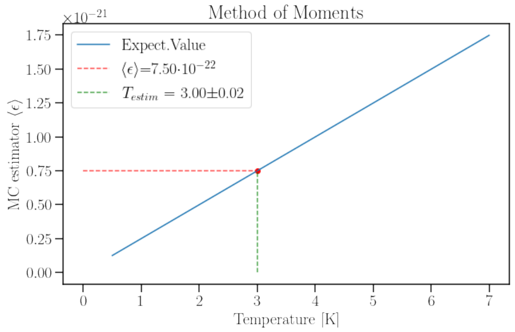

Figure 31: Obtained result with the method of moments, with systematic errors.

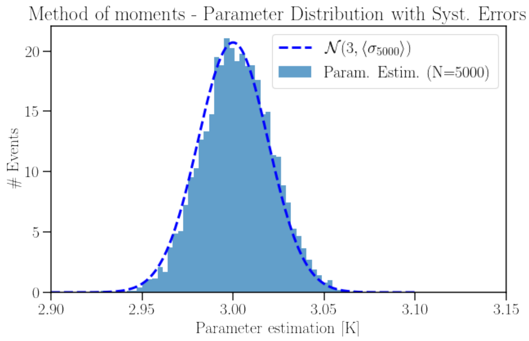

Figure 32: Distribution of the estimated parameter via the method of moments

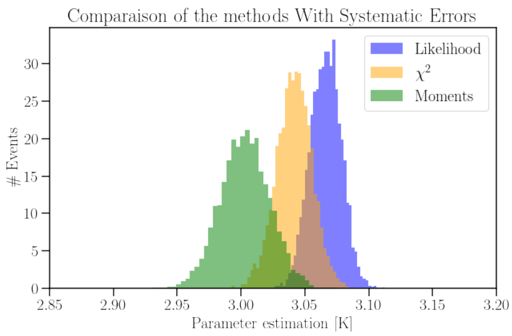

Figure 33: Comparison of the estimated parameters obtained with the three methods, and with systematic errors.