Simulation of Random Variables, Hamiltonian Dynamics, and Sampling via Markov Chains

Author:

Alfonso Mateos Vicente

Tutor:

Noé Blassel

École des Ponts
ParisTech

Ingénierie Mathématique et Informatique
École des Ponts ParisTech
France
September 01, 2023

Contents

1 Simulation of Random Variables

The aim of this section is to introduce the simulation of random variables from a probability distribution alredy given. We will focus on two methods: Inverse CDF method (also known as Inversion Method) and Rejection Sampling method (also known as Acceptance-Rejection method).

1.1 Inverse CDF method

Inverse CDF method is a basic method to generate pseudo-random numbers from any probability distribution given its cumulative distribution function. First of all, we can make the following supposition: Let X be a random variable with cumulative distribution function F and probability density function f. Then, the cumulative distribution function of $Y = F^{-1}(U)$ behaves like F, so the probability density function of Y is f. Knowing this, the idea is to generate a uniform random variable U and then apply the inverse of the cumulative distribution function F^{-1} to obtain a random variable with the desired distribution. The Algorithm 1 is as follows:

Algorithm 1: Inverse CDF mehtod

1. Generate a set of random numbers $U \sim \mathcal{U}(0,1)$;
2. Find the inverse of the cumulative distribution function F^{-1} ;
3. Apply the inverse to the set of random numbers $X = F^{-1}(U)$;

Let's see this with some examples. Using the exponential distribution, we know its probability density function is $f(x) = \lambda e^{-\lambda x}$. Also, we alredy know its cumulative distribution function which is $F(x) = 1 - e^{-\lambda x}$. Then, the inverse of the cumulative distribution function is:

$$F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x) \tag{1}$$

So, we can generate a set of random numbers $U \sim \mathcal{U}(0,1)$ and apply the inverse to obtain a set of random numbers $X = F^{-1}(U) \sim \mathcal{E}(\lambda)$. The Figure 1 shows the histogram of the generated random numbers and the probability density function of the exponential distribution.

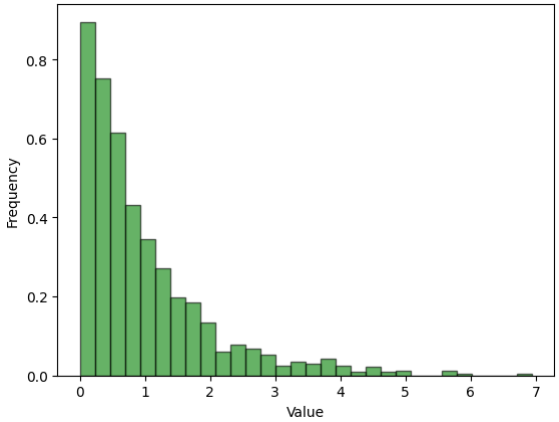

Figure 1: Histogram of Equation (1) with $\lambda = 1$ applied to $\mathcal{U}$.

With this picture, we can see that the histogram of the generated random numbers aligns remarkably well with the target distribution, however, we need to make more trials to be sure. We alredy know the theoreticall mean and variance of the exponential distribution, so we can compare them with the mean of the samples while the number of trials increases.

We define the error as:

$$error = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|$$

where x_i is the i-th random number generated and μ is the theoretical mean of the distribution.

Recall that the uniform and exponential means are:

$$\mu_{\mathcal{U}} = \frac{a+b}{2}$$

$$\mu_{\mathcal{E}} = \frac{1}{\lambda}$$

The Figure 2 and Figure 3 shows the error to the theoretical mean as the number of trials increases for the uniform and exponential distribution.

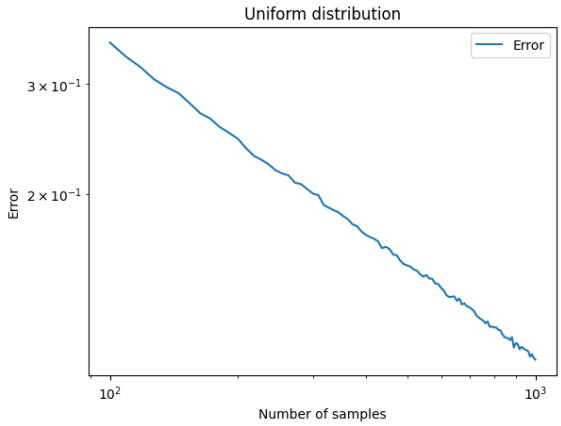

Figure 2: Error to the theoretical mean as the number of trials increases for the uniform distribution. Average of 10000 trials per point.

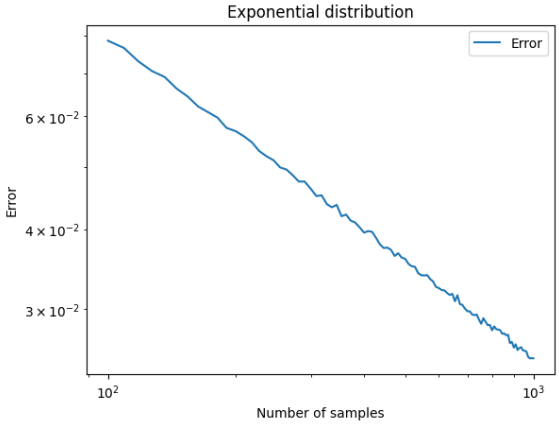

Figure 3: Error to the theoretical mean as the number of trials increases for the exponential distribution. Average of 10000 trials per point.

In drawing conclusions from the observed alignment of the histograms of generated random numbers with the target distribution, there appears to be a noteworthy correlation, as evidenced by Figures 2 and 3. These figures tentatively suggest a coherence between the theoretical and empirical means, hinting at a plausible reliability of the inverse CDF methods used. The error plots seem to indicate a convergence, providing a preliminary yet cautious optimism about the accuracy of the generated numbers in adhering to the desired distributions.

1.2 Rejection Sampling method

The Rejection Sampling method is a method to generate pseudo-random numbers for any probability distribution given its probability density function. The idea is to generate a set of random numbers from a probability distribution that is easy to sample from and then reject the numbers that are not in the desired distribution. The Algorithm 2 is as follows:

Algorithm 2: Rejection Sampling mehtod

1. Generate a set of random numbers $X \sim g(x)$;
2. Generate a set of random numbers $U \sim \mathcal{U}(0, 1)$;
3. If $U \leq \frac{f(X)}{Mg(X)}$ then accept X, otherwise reject X ;

Being $g(x)$ the probability density function with which we alredy know how to generate random numbers, denoted as the "enveloping" distribution; $f(x)$ the target probability density function; and M a factor we can choose manually and can be optimized.

Let's see this with one example. We want to generate a set of random numbers from the following probability density function:

$$f(x) = 0.3e^{-0.2x^2} + 0.7e^{-0.2(x-5)^2} \tag{2}$$

1.2.1 Selection of the "enveloping" distribution

In this subsection we will explain the process of selecting an optimised "enveloping" distribution. In fact, we could use any distribution that envelops the desired one, but this will probably be very

inefficient because there will be many samples that will not be under the objective function, so our goal is to obtain a distribution that envelops the desired one but with the smallest possible space between them. In this paper we have chosen the normal distribution as the "enveloping" distribution because it is easy to sample and is a good candidate for enveloping the desired distribution. Knowing this, we have three parameters to optimise: the mean, μ; the variance, σ; and the scale parameter, M. Note that for this problem, we have to define the bounds as a, b.

First of all, we can define the scale parameter as:

$$M = \max_{x \in \mathbb{R}} \frac{f(x)}{g(x)} \tag{3}$$

In this way, we ensure that whatever the function is, we will envelope it. Knowing this parameter, we only have to choose μ and σ in the order of minimizing M. So we have a problem of optimization with two variables. The problem is the following:

$$\begin{aligned} \min_{\mu,\sigma} \quad & M(\mu, \sigma) \\ \text{s.t.} \quad & M(\mu, \sigma) = \max_{x \in \mathbb{R}} \frac{f(x)}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \\ & \sigma \geq 0, a \leq \mu \leq b \end{aligned} \tag{4}$$

We can solve this problem with multiple methods. In our case, we have chosen the Nelder-Mead's method (also called downhill simplex method).

1.2.2 Results of the Rejection Sampling method

Once we have selected the "enveloping" distribution, we can apply the Rejection Sampling method. In our case, the enveloping function is the N($\mu \approx 3.644$,$\sigma \approx 3.041$) and the scalator parameter $M \approx 1.545$. The Figure 4 shows the histogram of the generated random numbers and the probability density function of the desired and the easy distribution.

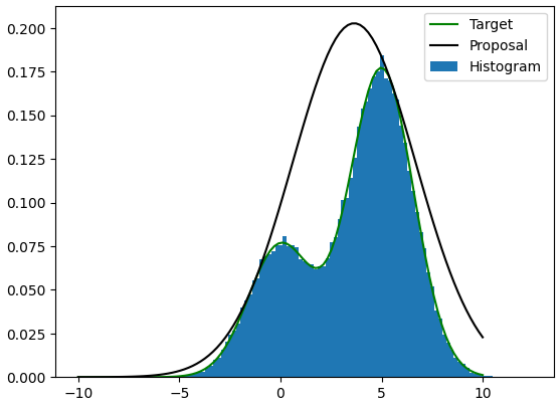

Figure 4: Histogram of the generated random numbers, Equation (2) and PDF of $\mathcal{N}(3.644, 3.041)$ scaled by $M \approx 1.545$.

First of all, note that the distribution chosen with the Equation 4 fits perfectly with the target distribution. Moreover, we can see that the histogram of the generated random numbers aligns

remarkably well with the target distribution, however, we need to make more trials to be sure. As we did in the previous section, we can compare the mean of the samples with the theoretical mean as the number of trials increases. The Figure 5 shows the error to the theoretical mean. As disclaimer, we will use the following approximation for the mean: $\frac{1}{n}\sum_{i=1}^{n} x_i$ and the following approximation for the variance: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

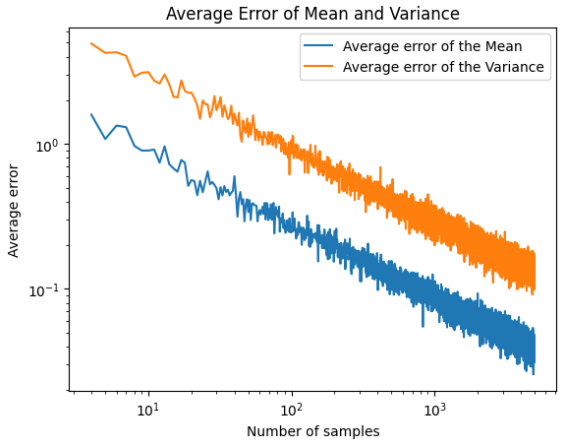

Figure 5: Error to the theoretical mean as the number of trials increases

In drawing conclusions from the observed alignment of the histograms of generated random numbers with the target distribution, there appears to be a noteworthy correlation, as evidenced by Figure 5. This figure tentatively suggests a coherence between the theoretical and empirical means and variances, hinting at a plausible reliability of the rejection sampling methods used. The error plot seems to indicate a convergence, providing a preliminary yet cautious optimism about the accuracy of the generated numbers in adhering to the desired distributions.

1.3 Empirical validation of the Law of Large Numbers

The Law of Large Numbers posits that, as the number of trials in an experiment increases, the average of the results obtained should converge towards the expected value of the probability distribution in question. Formally, the law can be expressed as follows:

Theorem 1.1 (Law of Large Numbers)**.** Let $\{X_1, X_2, ...\}$ be a sequence of independent and identically distributed random variables drawn from a distribution of mean μ. And let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the average of the first n elements in the sequence. Then, as n approaches infinity, the random variables $\bar{X}_n$ converge in probability to μ:

$$\bar{X}_n \to \mu \;\; as \; n \to \infty \tag{5}$$

Given our established ability to generate random numbers reflecting specific probability distributions, we can select a distribution, generate a set of numbers corresponding to it, and then contrast the calculated averages and means of the target distribution while escalating the number of trials. To illustrate, employing the normal distribution with a mean of 0 provides the results depicted in Figure 6.

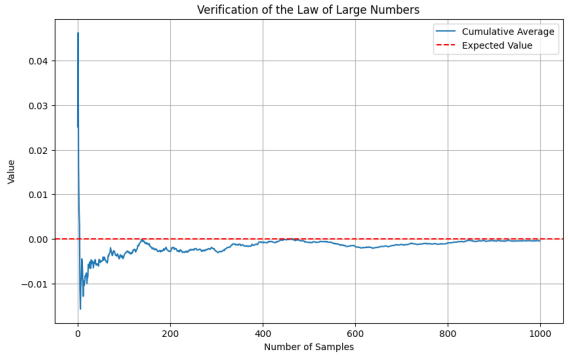

Figure 6: Comparison of the method's average with the mean of the normal distribution (0)

Refining Figure 6 to display the absolute error in relation to the mean, and plotting it on a logarithmic scale, yields Figure 7. So the error is

$$error = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

where x_i is the i-th random number generated and μ is the theoretical mean of the distribution.

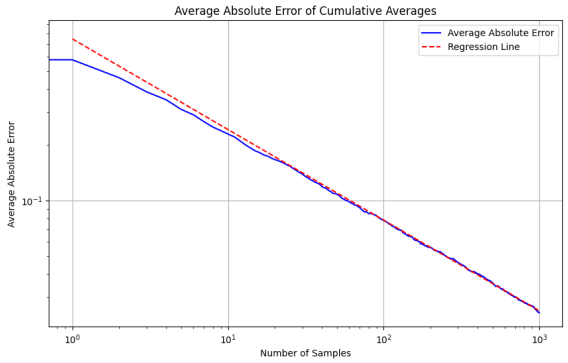

Figure 7: Logarithmic depiction of the absolute error relative to the mean. Average of 1000 trials per point.

The visual representation in Figure 7 substantiates that the error indeed diminishes as the number of trials augments, validating the assertion of the Law of Large Numbers that the experimental mean approaches the theoretical mean with an increasing number of observations. Note that the slope of the regression line is approximately -0.5.

1.4 Empirical Validation of the Central Limit Theorem

The Central Limit Theorem (CLT) posits a pivotal foundational theorem in probability theory, signifying that, irrespective of the shape of the original distribution, the distribution of sample means will approximate a normal distribution as the sample size burgeons. To empirically validate this theorem, we can simulate a series of random numbers from any given probability distribution and systematically calculate their mean. By perpetuating this process, we can construct a histogram of the means and scrutinize whether it converges to a normal distribution, aligning with the theorem's

prediction. Formally, the theorem can be expressed as follows:

Theorem 1.2 (Lindeberg–Lévy CLT)**.** Suppose $\{X_1, X_2, ..., X_n\}$ is a sequence of independent and identically distributed random variables drawn from a distribution of mean $\mathbb{E}(X_i) = \mu$ and finite variance $\mathbb{V}(X_i) = \sigma^2$. Then, as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{6}$$

For illustrative purposes, consider the uniform distribution on $[0, 1]$, which yields a histogram as delineated in Figure 8.

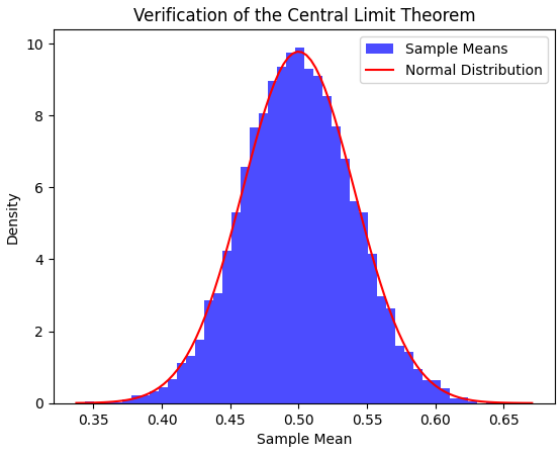

Figure 8: Histogram illustrating the convergence of the means of the uniform distribution with $n_{samples} = 10000$

In congruence with the methodology espoused in previous sections, we have juxtaposed the empirical mean of the samples with the theoretical mean, progressively augmenting the number of trials. Figure 9 delineates the deviation from the theoretical mean, and Figure 10 represents the deviation from the theoretical variance, both as functions of the number of trials. Therefore the formula of the deviation is:

$$error_{mean} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|$$

where x_i is the i-th random number generated and μ is the theoretical mean of the distribution.

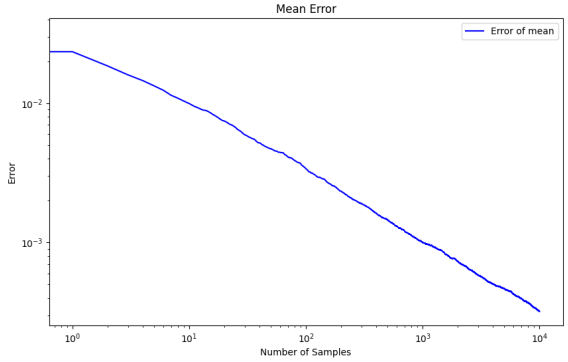

Figure 9: Error from the theoretical mean as a function of the number of trials. Average of 10000 trials per point.

And for the case of the variance:

$$error_{variance} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \sigma^2|$$

where x_i is the i-th random number generated and σ^2 is the theoretical variance of the distribution.

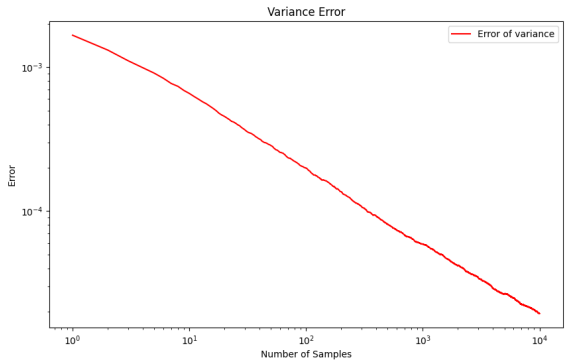

Figure 10: Error from the theoretical variance as a function of the number of trials. Average of 10000 trials per point.

A discernible insight gleaned from Figure 9 and Figure 10 is the palpable decrement in deviations from the theoretical values as the trials proliferate.

Finally, another way to validate the CLT is to plot the variance as the number of variables increases and calculate the slope of the regression line. If the theory is correct, the slope should be around -1 since the variance of the normal distribution is $\mathbb{V}(\bar{X}) = \frac{sigma^2}{n}$, therefore, the variance is inversely proportional to the number of variables. The Figure 11 shows the variance as the number of variables increases and the regression line.

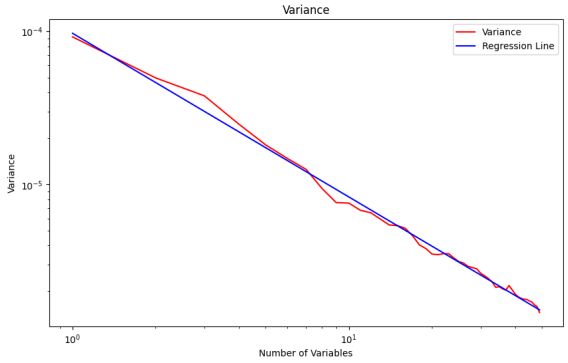

Figure 11: Deviation from the theoretical variance as a function of the number of trials. Average of 1000 trials per point.

Computing the slope of the linear regression we obtain that it is -1. This phenomenological observation substantiates the assertions of the Central Limit Theorem, illuminating the convergence of the distribution of sample means to a normal distribution as the sample size escalates, validating the theoretical underpinnings of the theorem through empirical exploration.

1.5 Variance reduction techniques

In the theory of Monte Carlo methods, variance reduction techniques are a pivotal tool to increase the precision of the estimates of the expected value of a random variable. In this section, we will focus on three techniques: Control variates, Importance sampling and Antithetic variates. Also, we will introduce an example problem in which we will apply these techniques to compare them.

First of all, let introduce the variance reduction formally. Let $\{X_1, X_2, ..., X_n\}$ be a sequence of independent and identically distributed random variables drawn from a distribution of mean $\mathbb{E}(X_i) = \mu$ and finite variance $\mathbb{V}(X_i) = \sigma^2$. Then, the expected value of a random variable is defined as follows:

$$\hat{\mathbb{E}}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

So we want to make zero the variance of our estimation. Since the variance is defined as follows:

$$\mathrm{Var}(\hat{\mathbb{E}}(X)) = \frac{\mathrm{Var}(X)}{n} \tag{8}$$

Consequently, we are presented with two avenues for optimization: increasing the value of n, or diminishing the variance of X. Assuming that n is predetermined and unalterable, our focus would then shift to minimizing the variance of X.

Since we are going to introduce and compare the three techniques, first of all we need to introduce the problem we are going to solve. The problem is that we want to estimate the following integral:

$$I = \int_0^1 x^2 \, dx \tag{9}$$

It should be noted that the selection of the integral for this demonstration was intentional; a

readily solvable integral was chosen for its ease of analytical computation, allowing for a straightforward comparison with theoretical values. Nonetheless, the methods illustrated herein are equally applicable and potent for evaluating integrals that pose substantial challenges to analytical computation.

1.5.1 Preliminaries

First of all we have to compute which is the estimation we alredy can have without applying any variance reduction technique. We can compute the integral analytically. We know that the mean of the variable in a probability space is defined as follows:

$$\mathbb{E}(g(X)) = \int_a^b g(x)f(x)\,dx \tag{10}$$

Since $f(x) = \frac{1}{b-a}$ for the uniform distribution on $[a, b]$, we can apply Equation (10) to Equation (9) and we obtain:

$$I = \int_0^1 x^2\,dx = \mathbb{E}(g(X)) = \int_0^1 x^2 \frac{1}{1-0}\,dx = \frac{1}{3} \tag{11}$$

So, we only have to generate a set of random numbers $X \sim \mathcal{U}(0, 1)$, apply $f(x) = x^2$ and get the mean of the sample which is the estimate of the integral. The Figure 12 shows the error to the theoretical value as the number of trials increases. The error to the theoretical value is defined as:

$$error = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu|$$

where x_i is the i-th random number generated and μ is the theoretical mean of the distribution. Note that this formula will be used in all the following sections.

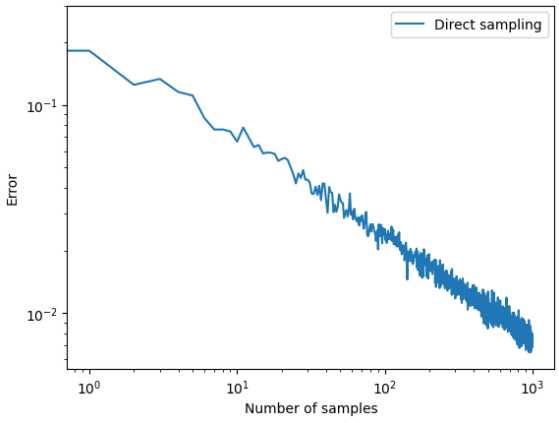

Figure 12: Error to the theoretical value as the number of trials increases using the direct method. Average of 100 trials per point.

To conclude, we can compute the variance of the estimation as follows:

$$\mathrm{Var}(X) = \mathrm{Var}(\hat{\mathbb{E}}(X))\,n \tag{12}$$

In this case, we have computed that $\mathrm{Var}(X) \approx 8.977 \cdot 10^{-2}$.

1.5.2 Control variates

Control variates is a variance reduction technique with the following idea: Let μ the parameter we want to estimate, and assume we have a statistic Y such that $\mathbb{E}(Y) = \tau$. Then, we can estimate μ by estimating $\mathbb{E}(Y)$ as $\hat{\mathbb{E}}(Y)$ and correcting the bias with the following formula:

$$\hat{\mu} = \mu + c(\tau - \hat{\mathbb{E}}(Y)) \tag{13}$$

Being c a constant which minimize the variance of the estimation. It is computed as follows:

$$c = -\frac{Cov(\mu, \tau)}{\mathrm{Var}(\tau)} \tag{14}$$

Proof. Using the Equation (13) we can compute the variance of the estimation as follows:

$$\begin{gathered}\mathrm{Var}(\hat{\mu}) = \\ \mathbb{E}[(\mu + c(\tau - \hat{\mathbb{E}}(Y)))^2] - \mathbb{E}[\mu + c(\tau - \hat{\mathbb{E}}(Y))]^2 = \\ \mathbb{E}[\mu^2 + c^2(\tau - \hat{\mathbb{E}}(Y))^2 + 2c\sigma(\tau - \hat{\mathbb{E}}(Y))] - \mathbb{E}[\mu]^2 - c^2\mathbb{E}[\tau - \hat{\mathbb{E}}(Y))]^2 - 2c\mathbb{E}[\sigma(\tau - \hat{\mathbb{E}}(Y))] = \\ \mathrm{Var}(\mu) + c^2\mathrm{Var}(\tau) + 2cCov(\mu, \tau)\end{gathered}$$

Therefor, since we want to minimize the variance of the estimation, we can differentiate with respect to c and equal to zero:

$$\frac{\partial \mathrm{Var}(\hat{\mu})}{\partial c} = 2c\mathrm{Var}(\tau) + 2Cov(\mu, \tau) = 0$$

And we obtain the following expression for c:

$$c = -\frac{Cov(\mu,\tau)}{\mathrm{Var}(\tau)}$$

□

In our case, we can use the knowledge of the mean of the uniform distribution to correct the bias, so in each iteration we can compute the mean of the sample of random numbers of $\mathcal{U}(0, 1)$, and knowing that the mean of the uniform distirbution is 0.5 we can apply Equation (13) to obtain the estimate of the integral. The Figure 13 shows the error to the theoretical value as the number of trials increases.

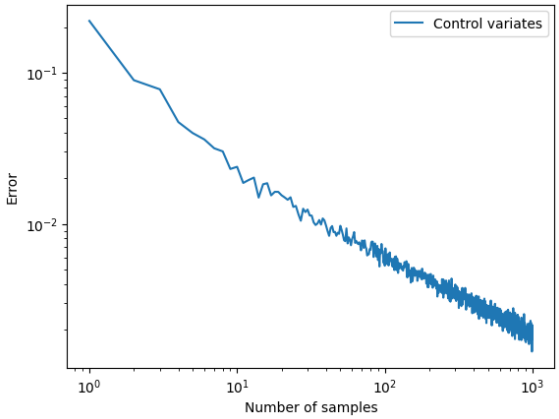

Figure 13: Error to the theoretical value as the number of trials increases using the control variates method. Average of 100 trials per point.

In this case, we have computed that $\mathrm{Var}(X) \approx 5.168 \cdot 10^{-3}$.

1.5.3 Importance sampling

Importance sampling is another variance reduction technique. Instead of using the knowledge of another estimator to reduce the bias of our sample, as done in the control variates method, in this case the idea is using the knowledge of the actual function we want to integrate, so instead of using a uniform distribution, we can use another distribution that is more similar to the function we want to integrate, in order to try more samples in the areas where the function is more important.

In our example, we know that the function is a parabola, so instead of using a sample which follows an uniform distribution, maybe we can use a distribution that fits better with the shape of the function. In this case, we have chosen $Beta(2.9, 1)$. In order to illustrate this, in the Figure 14 we can see the function we want to integrate, the PDF of $Beta(2.9, 1)$ and the PDF of $\mathcal{U}(0, 1)$.

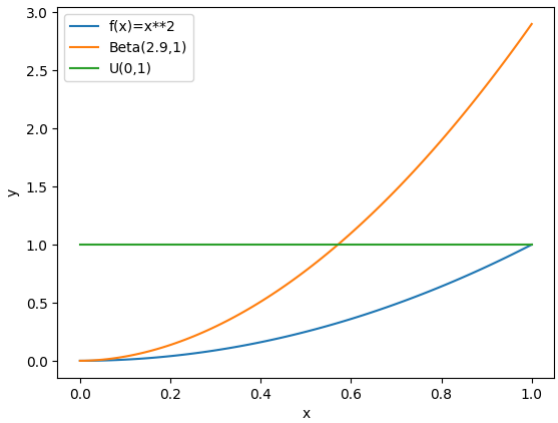

Figure 14: Function we want to integrate, PDF of $Beta(2.9, 1)$ and PDF of $\mathcal{U}(0, 1)$

As we can see, the PDF of $Beta(2.9, 1)$ fits better with the shape of the function we want to integrate, so we can expect that the error will decrease faster than the direct method and the control variates method.

So the idea is to generate a sample of random numbers $X \sim Beta(2.9, 1)$, and take $Y = \frac{f(X)}{g(X)}$ as the estimator of the integral, being $f(x) = x^2$ and $g(x)$ the PDF of $Beta(2.9, 1)$. The Figure 15 shows the error to the theoretical value as the number of trials increases.

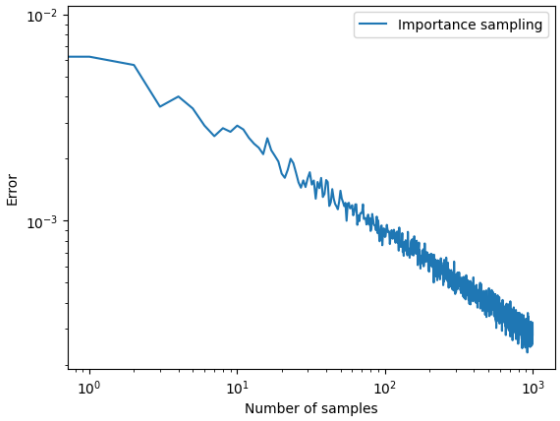

Figure 15: Error to the theoretical value as the number of trials increases using the importance sampling method. Average of 100 trials per point.

In this case, we have computed that $\mathrm{Var}(X) \approx 1.274 \cdot 10^{-4}$.

1.5.4 Stratified sampling

The idea is to divide the interval of the stratified sampling is pretty simple. We divide the interval in n subintervals, and we generate a sample of random numbers for each subinterval. Doing this, we are trying to reduce the variance of the mean of the distribution we are considering, so the sample is more representative of the distribution. The Figure 16 shows the error to the theoretical value as the number of trials increases.

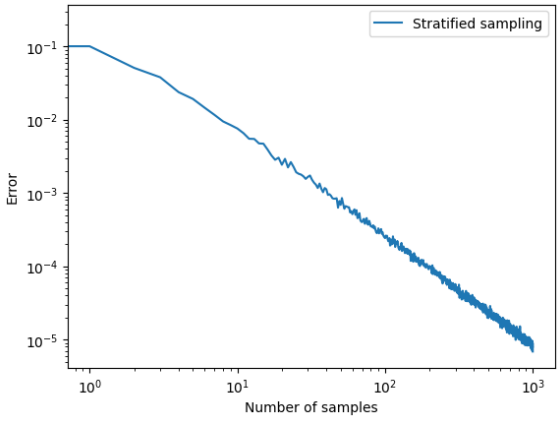

Figure 16: Error to the theoretical value as the number of trials increases using the stratified sampling method. Average of 100 trials per point.

In this case, we have computed that $\mathrm{Var}(X) \approx 1.108 \cdot 10^{-7}$.

1.5.5 Antithetic variates

The Antithetic variates method consists on taking for each random sample, its antithetic, i.e. the symmetric with respect to the mean of the distribution. The idea is that the variance of the mean of the distribution is reduced, since the mean of the antithetic is the same as the mean of the distribution. In our case, we have generated a sample $\mathcal{U} \sim \mathcal{U}(0, 0.5)$ and its antithetic $\mathcal{U}' = \{1 - x : x \in \mathcal{U}\}$. Taking $X = \mathcal{U} \bigcup \mathcal{U}'$, we can apply Equation (7) to $f(X)$ to obtain the estimate of the integral. The Figure 17 shows the error to the theoretical value as the number of trials increases.

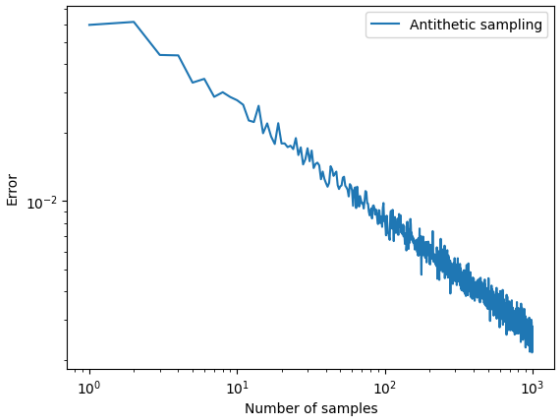

Figure 17: Error to the theoretical value as the number of trials increases using the antithetic variates method. Average of 100 trials per point.

In this case, we have computed that $\mathrm{Var}(X) \approx 1.178 \cdot 10^{-2}$.

1.5.6 Comparative Review

After unraveled the intricacies of each method and confirming the convergence of each approach described in this study, we now turn to a more focused comparison of these techniques. The goal is simple: find out which method works best for our specific problem. See Figure 18 for a visual representation of the error in relation to the theoretical value as we increase the number of trials for each method.

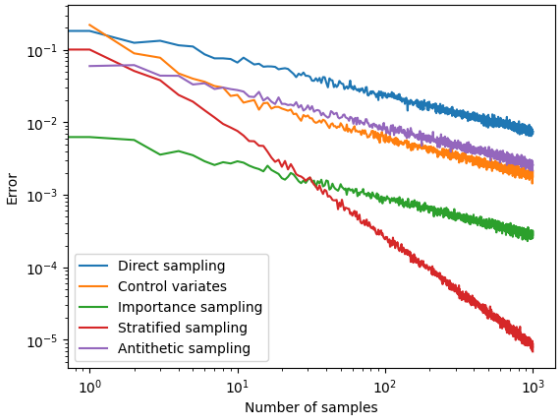

Figure 18: Error variation with increased trials for each method.

Looking at the data, it's clear that stratified sampling takes the lead with its quicker convergence compared to the other methods. However, claiming it as the undisputed champion would be premature, especially when we've explored just one problem. To firm up this initial finding, we need to dig deeper and explore a variety of problems.

But before moving on to look at more problems, let's look at the comparative table of variances for each method:

Method	Variance
Direct sampling	$8.977194 \cdot 10^{-2}$
Control variates sampling	$5.168373 \cdot 10^{-3}$
Importance sampling	$1.274508 \cdot 10^{-4}$
Stratified sampling	$1.107865 \cdot 10^{-7}$
Antithetic sampling	$1.178326 \cdot 10^{-2}$

Analyzing the table, we discern significant disparities in variance between the different sampling methods, accentuating the prominence of stratified sampling, which registers the minimal variance, $1.107865 \cdot 10^{-7}$. This numerical inferiority in variance corroborates the preliminary observation about its superior convergence rate, offering more stable and reliable estimates. Stratified sampling outperformed other methods in our study primarily due to its targeted approach of dividing the population into homogeneous strata. This division ensures proportional representation of all subgroups, reducing sampling bias and variance within each stratum. Consequently, it yields more accurate and representative results. Furthermore, by adjusting the sample size within each stratum based on variability, stratified sampling enhances overall efficiency and precision. These methodological strengths contribute to its superior performance in terms of reliability and convergence rates, especially in diverse and complex datasets. Contrastingly, direct sampling exhibits the maximum variance, revealing its comparative inefficiency and instability in procuring estimates for this specific problem. The remaining methods, while overshadowed by the efficacy of stratified sampling, still exhibit markedly lower variances than direct sampling, with importance sampling making a notable contribution with a variance of $1.274508 \cdot 10^{-4}$. These numerical insights underscore the necessity to meticulously select the appropriate sampling technique based on the inherent characteristics of the problem at hand, and they hint at the potential benefits of exploring hybrid approaches or enhancements to existing methods to optimize variance reduction.

Next, we'll broaden our investigation to a more complex problem, applying the methods we've discussed to estimate the following integrals:

$$Parabola = \int_0^1 x^2 \, dx \tag{15}$$

$$Gaussian = \int_0^1 e^{-x^2} \, dx \tag{16}$$

$$Sine = \int_0^1 \sin(x) \, dx \tag{17}$$

$$Polynomial = \int_0^1 x^3 - 2x^2 + x \, dx \tag{18}$$

$$Exponential = \int_0^1 e^x \, dx \tag{19}$$

Using these examples, we can now draw comparisons between the methods in a broader context. Figure 19 illustrates the error of each method applied to each function.

This error is defined as:

$$error = \mid method(f, a, b, iters) - \int_a^b f dx \mid$$

where $method$ is the method we are using, f is the function we are integrating, a and b are the limits of the integral and $iters$ is the number of iterations we are using, in this case $iters = 1000$.

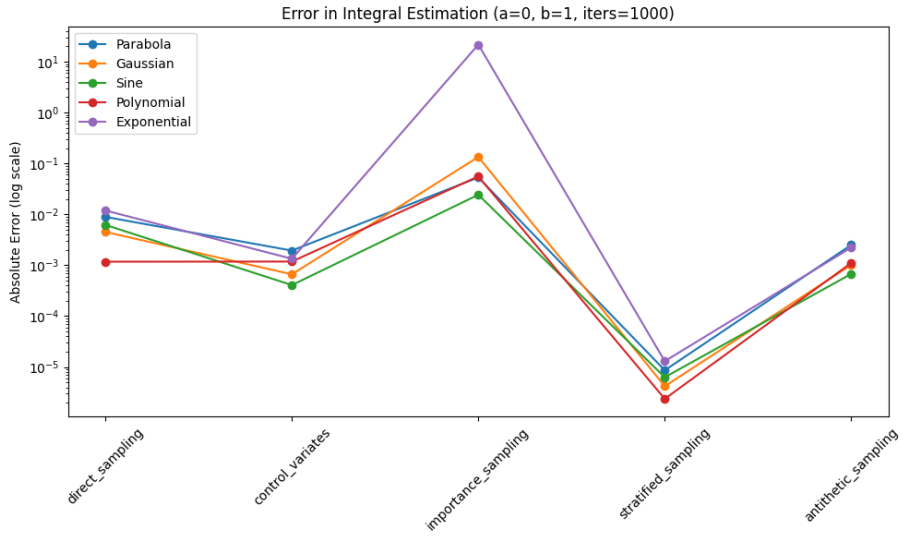

Figure 19: Method-wise error in each function.

Notably, for the importance sampling method, we opted for a normal distribution with the mean and variance of each function to maintain fairness in comparison, avoiding manual distribution selection for every function.

Reviewing the data, stratified sampling again stands out for its rapid convergence. However, the importance sampling lags, a likely outcome of applying a general normal distribution for all functions, each requiring a unique distribution. But to outrightly conclude that stratified sampling is the go-to method would be a rush to judgment, given our exploration is based on limited scenarios. Thorough exploration involving varied problems is essential to cement our initial conclusions while maintaining the professional and scientific rigor of our exploration.

2 Hamiltonian dynamics

Hamiltonian dynamics, also called ”Hamiltonian mechanics”, is a reformulation of the clasical mechanics which describes the temporal evolution of a physical system in terms of pairs of variables: the generalized coordinates q_i and their momenta p_i.

$$p(t) = \begin{bmatrix} p_1(t) \\ p_2(t) \\ \vdots \\ p_n(t) \end{bmatrix} \quad q(t) = \begin{bmatrix} q_1(t) \\ q_2(t) \\ \vdots \\ q_n(t) \end{bmatrix}$$

In this section we consider the time evolution of a isolated system described at a microscopic level, i.e. a system of particles. The state of the system is described by the position of the particles q_i and their momenta p_i. We denote D the dimension of the positions and momenta variables. Therefore $D = 3N$ when the system is composed of N particles in a 3-dimensional physical space. Also, we assume that for the system there is a function $H(q, p, t)$ which describes the energy of the system.

The Hamiltonian dynamics is defined by the following equations:

$$\begin{cases} \frac{dq(t)}{dt} = \nabla_p H(q(t), p(t)) \\ \frac{dp(t)}{dt} = -\nabla_q H(q(t), p(t)) \end{cases} \tag{20}$$

With initial condition $p(0) = p^0,\ q(0) = q^0$ that should be provided. Now, introducing the following matrix:

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$$

And denoting $y = (q, p)$, we can rewrite the Hamiltonian dynamics equations as follows:

$$\frac{dy}{dt} = J\nabla H(y) = J \begin{bmatrix} \nabla_q H(y) \\ \nabla_p H(y) \end{bmatrix} \tag{21}$$

A Hamiltonian function is the sum of the kinetic energy and the potential energy of the system:

$$H(q, p) = K(p) + V(q)$$

A very common physical interpretation of the Hamiltonian function is the following:

$$H(q, p) = V(q) + \frac{1}{2} p^T M^{-1} p$$

Where M is the mass matrix of the system, i.e. $M = diag(m_1, m_2, ..., m_n)$ with m_i the mass of the i-th particle, note that we are supposing that the mass is stable over time. In this case, we can reformulate the Hamiltonian dynamics equations as follows:

$$\begin{cases} \frac{dq(t)}{dt} = M^{-1} p(t) \\ \frac{dp(t)}{dt} = -\nabla_q V(q(t)) \end{cases} \tag{22}$$

Therefore, if we consider this equations in terms of positions we get:

$$M \frac{d^2 q(t)}{dt^2} = -\nabla_q V(q(t))$$

Which is the Newton's second law of motion. So, we can see that the Hamiltonian dynamics is a generalization of the Newton's second law of motion.

One of the most important property of the Hamiltonian dynamics is the following:

Theorem 2.1 (Conservation of energy)**.** Let $H(q,p)$ be the Hamiltonian function of a system. Then, the energy of the system is conserved over time, i.e. $H(q(t),p(t)) = H(q^0,p^0)$ for all t.

Proof. Deriving the Hamiltonian function with respect to time we get:

$$\begin{aligned}\frac{dH}{dt} = \frac{\partial H}{\partial q}\frac{dq}{dt} + \frac{\partial H}{\partial p}\frac{dp}{dt} = \\ = \nabla_q H \frac{dq}{dt} + \nabla_p H \frac{dp}{dt} = \\ = \nabla_q H J \nabla_p H - \nabla_p H J \nabla_q H = 0\end{aligned}$$

Since the hamiltonian H is defined as the total energy of the system, we can conclude that the energy of the system is conserved over time. □

2.1 Symplectic schemes

In this section we will introduce the symplectic schemes, which are a family of numerical methods to solve the Hamiltonian dynamics equations. The idea is to discretize the Hamiltonian dynamics equations (21) in order to obtain a numerical approximation of the solution. The symplectic schemes are a family of numerical methods which preserve the symplectic structure of the Hamiltonian dynamics equations, i.e. the energy of the system is conserved over time.

Definition 2.1. For an open set $U \subset \chi$, a mapping $g : U \to \mathbb{R}^{2D}$ of class C^1 is symplectic if $\nabla g(q,p)$ satisfies

$$(\nabla g)^T J \nabla g = J, \quad \forall (q,p) \in U$$

2.1.1 Analytical solution

Let us introduce a simple example problem to illustrate the symplectic schemes: the harmonic oscillator. The harmonic oscillator is a system composed of a particle of mass m attached to a spring with spring constant k. The position of the particle is denoted by $q(t)$. The potential energy of the system is defined as follows:

$$V(q) = \frac{1}{2}kq^2$$

Therefore, the Hamiltonian function of the system is defined as follows:

$$H(q,p) = \frac{1}{2m}p^2 + \frac{1}{2}kq^2 \tag{23}$$

Applying this Hamiltonian function to the Hamiltonian dynamics equations (21) we obtain the following equations:

$$\begin{cases} \frac{dq(t)}{dt} = \frac{1}{m}p(t) \\ \frac{dp(t)}{dt} = -kq(t) \end{cases}$$

Let's proceed step by step to solve the given system of differential equations for the harmonic oscillator. We are given the Hamiltonian:

$$H(q,p) = \frac{1}{2m}p^2 + \frac{1}{2}kq^2$$

And the Hamilton's equations:

$$\begin{cases} \frac{dq(t)}{dt} = \frac{1}{m}p(t) \\ \frac{dp(t)}{dt} = -kq(t) \end{cases}$$

We can combine the two first-order differential equations into a single second-order differential equation by substituting the expression for $\dot{x}(t)$ into the derivative $\dot{p}(t)$. Substituting $\dot{x}(t) = \frac{1}{m}p(t)$ into the derivative of the second equation gives us:

$$\frac{d^2q(t)}{dt^2} = -\frac{k}{m}q(t)$$

This is a second-order homogeneous linear differential equation.

To solve this second-order differential equation, we can use the characteristic equation method. Assume a solution of the form:

$$q(t) = e^{rt}$$

where r is a constant to be determined. Substituting this into the second-order equation gives:

$$r^2e^{rt} + \frac{k}{m}e^{rt} = 0$$

Since e^{rt} is never zero, we can divide through by it to get the characteristic equation:

$$r^2 + \frac{k}{m} = 0$$

Solving for r gives us:

$$r = \pm i\sqrt{\frac{k}{m}}$$

where i is the imaginary unit.

Given that the roots are complex, the general solution of the equation is:

$$\begin{cases} q(t) = A\cos(\omega t) + B\sin(\omega t) \\ p(t) = -m\omega A\sin(\omega t) + m\omega B\cos(\omega t) \end{cases}$$

where A and B are arbitrary constants determined by initial conditions and $\omega = \sqrt{\frac{k}{m}}$ is the angular frequency of the oscillator. In order to determine the constants A and B, we need initial conditions. Specifically, we need the initial position $x(0)$ and initial velocity $\dot{x}(0)$. Let's say, for example:

$$q(0) = q_0 \quad \text{and} \quad \dot{q}(0) = v_0$$

Substituting these into the general solution and its derivative gives:

$$q_0 = A \quad \text{and} \quad v_0 = B\omega$$

Thus, if initial conditions are provided, A and B can be determined to give the particular solution for the system. For example, if we take $x_0 = 1$ and $v_0 = 0$, we get that $A = 1$ and $B = 0$. Therefore, the solution of the harmonic oscillator is:

$$\begin{cases} q(t) = \cos(\omega t) = \cos(\sqrt{\frac{k}{m}}t) \\ p(t) = -m\omega \sin(\omega t) = -\sqrt{km}\sin(\sqrt{\frac{k}{m}}t) \end{cases}$$

The Figure 20 shows the phase space of the harmonic oscillator with $k = 1$ and $m = 1$.

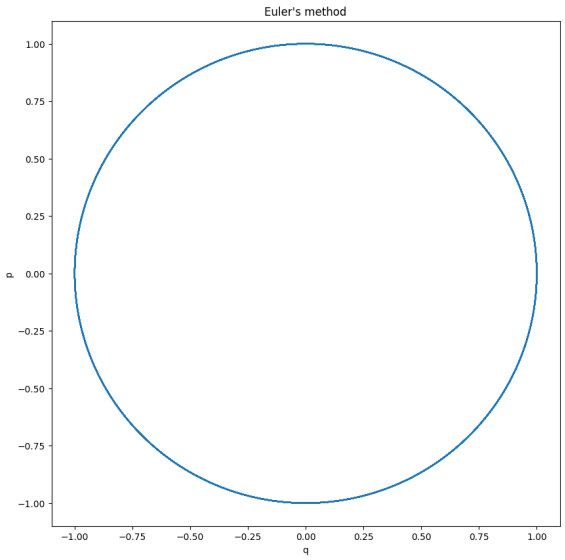

Figure 20: Phase space of the harmonic oscillator with $k = 1$ and $m = 1$ and initial conditions $q_0 = 1$ and $v_0 = 0$.

2.1.2 Euler method

The Euler method is a numerical method to solve ordinary differential equations. The idea is to discretize the differential equation in order to obtain a numerical approximation of the solution. The Euler method discretizes the differential equation as follows:

$$\begin{cases} p_{t+\Delta t} &= p_t - \Delta t \frac{\partial H}{\partial q}(q_t, p_t) = p_t - \Delta t \nabla_q V(q_t) \\ q_{t+\Delta t} &= q_t + \Delta t \frac{\partial H}{\partial p}(q_t, p_t) = q_t + \Delta t M^{-1} p_t \end{cases} \tag{24}$$

Being Δt the time step of the discretization. Applying this sequence to the harmonic oscillator, which equations are described in (23), we obtain the following phase space:

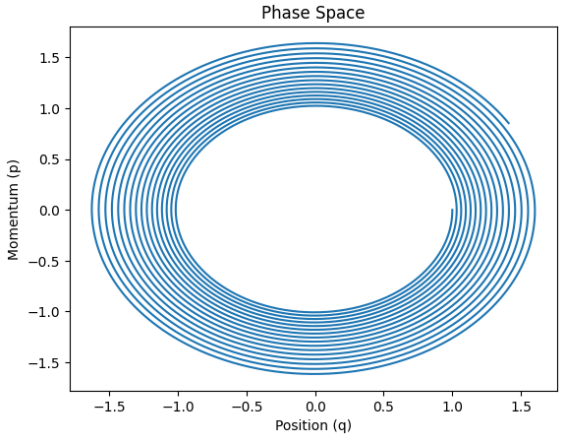

Figure 21: Phase space of the harmonic oscillator using the Euler method with $\Delta t = 0.01$.

As we can discern, this phace space is not even near to the real phase space of the harmonic oscillator. The Figure 21 shows the phase space of the harmonic oscillator using the Euler method with $\Delta t = 0.1$. We can see that the phase space is not a closed curve, which is the correct behaviour of the harmonic oscillator. Also, the energy is not conserved over time, which is another property of the harmonic oscillator. The Figure 22 shows the energy of the harmonic oscillator using the Euler method with $\Delta t = 0.1$.

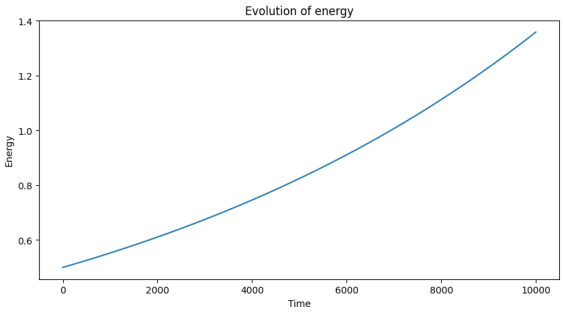

Figure 22: Energy of the harmonic oscillator using the Euler method with $\Delta t = 0.01$.

As we can see, the energy is not conserved over time, which is another property of the harmonic oscillator. The energy is increasing over time, which is not the correct behaviour of the harmonic oscillator. With this results we can conclude that the Euler method is not a good method to solve the Hamiltonian dynamics equations. This is a consequence of the fact that the Euler method does not preserve the symplectic structure of the Hamiltonian dynamics equations. Therefore, in the next section we will introduce a new method that preserves the symplectic structure of the Hamiltonian dynamics equations which is a slight modification of the Euler method.

2.1.3 Failure of standard methods

As we alredy know, the Hamiltonian dynamics is a standard ordinal differential equation (ODE), so it can be approximated by any standard integration scheme. However, as we saw using the Euler scheme, the energy increases over the time so it doesn't work. To provide a further view of this, in

this section, we want to introduce the mathematical approach to see why this methods are failing. Considering again the problem of the harmonic oscillator, whose Hamiltonian is provided in 23. And taking the euler scheme given in 24 we can rewrite it to:

$$\begin{pmatrix} p^{n+1} \\ q^{n+1} \end{pmatrix} = A \begin{pmatrix} p^n \\ q^n \end{pmatrix}$$

Being $A = \begin{pmatrix} 1 & -k\Delta t \\ \frac{1}{m}\Delta t & 1 \end{pmatrix}$. Now we diagonalize this matrix, so we obtain that its eigenvalues are:

$$\lambda_1 = 1 + i\Delta t\sqrt{\frac{k}{m}} = i + i\Delta t\omega, \quad \lambda_2 = 1 - i\Delta t\sqrt{\frac{k}{m}} = i - i\Delta t\omega$$

Since $\lambda_1\lambda_2 = 1 + (\Delta t\omega)^2 > 1$ the energy exponentially increases in the time. Following this way of reasoning, it can be easily checked that the Runge-Kutta scheme and the implicit Euler scheme also fail. So the next step of the project is to construct schemes that can conserve the energy.

2.1.4 Constructing Symplectic Schemes

To develop an integrator, we split the Hamiltonian $H(q,p)$ into simpler components:

$$H(q,p) = H_1(q,p) + H_2(q,p)$$

With:

$$H_1(q,p) = \frac{1}{2}p^T M^{-1} p \quad \text{(kinetic energy)}$$
$$H_2(q,p) = V(q) \quad \text{(potential energy)}$$

After the Hamiltonian is decomposed, the associated dynamics become:
For $H1$:

$$\dot{q} = M^{-1}p$$
$$\dot{p} = 0$$

For $H2$:

$$\dot{q} = 0$$
$$\dot{p} = -\nabla V(q)$$

Now, let's talk about flows. Think of a flow as a way to evolve a point in phase space over time based on our differential equations.

For the dynamics associated with H_1, we can integrate with respect to time, t:

$$\int \dot{q}dt = \int M^{-1}pdt$$

This gives:

$$q(t) = q(0) + tM^{-1}p$$

Similarly, for momentum, $p(t) = p(0)$ because $\dot{p} = 0$.

Thus, the flow for H_1, denoted as ϕ_1^t, evolves as:

$$\phi_1^t(q, p) = (q + tM^{-1}p, p)$$

For H_2, given $\dot{q} = 0$, $q(t) = q(0)$. But for momentum, we get:

$$\int \dot{p}dt = \int -\nabla V(q)dt$$

$$p(t) = p(0) - t\nabla V(q)$$

Thus, the flow for H_2, ϕ_2^t, is:

$$\phi_2^t(q, p) = (q, p - t\nabla V(q))$$

Combining these flows yields our symplectic schemes. The sequence of combining matters.

1. Kinetic (ϕ_1) followed by Potential (ϕ_2):

$$\begin{cases} q_{n+1} &= q_n + \Delta t M^{-1} p_n \\ p_{n+1} &= p_n - \Delta t \nabla V(q_{n+1}) \end{cases}$$

2. Potential (ϕ_2) followed by Kinetic (ϕ_1):

$$\begin{cases} p_{n+1} &= p_n - \Delta t \nabla V(q_n) \\ q_{n+1} &= q_n + \Delta t M^{-1} p_{n+1} \end{cases}$$

Theorem 2.2 (Symplecity of the Hamiltonian flow)**.** Let $H(q, p)$ be a $C^2(U)$ function, where U is an open set of $\mathbb{R}^{2D}$. Then, for any fixed $t \in \mathbb{R}$ such that the flow ϕ^t is defined, the mapping ϕ^t is symplectic.

Proof. Proof in [4, Chapter 2.1.2]. □

2.1.5 Symplectic Euler method

The symplectic Euler method is a slight modification of the Euler method that preserves the symplectic structure of Hamiltonian systems. This new method discretizes the Hamiltonian dynamics equations (21) as follows:

$$\begin{cases} p_{t+\Delta t} &= p_t - \Delta t \frac{\partial H}{\partial q}(q_t, p_t) = p_t - \Delta t \nabla_q V(q_t) \\ q_{t+\Delta t} &= q_t + \Delta t \frac{\partial H}{\partial p}(q_{t+1}, p_{t+1}) = q_t + \Delta t M^{-1} p_{t+1} \end{cases} \tag{25}$$

Being Δt the time step of the discretization. Applying this sequence to the harmonic oscillator, which equations are described in (23), we obtain the following phase space:

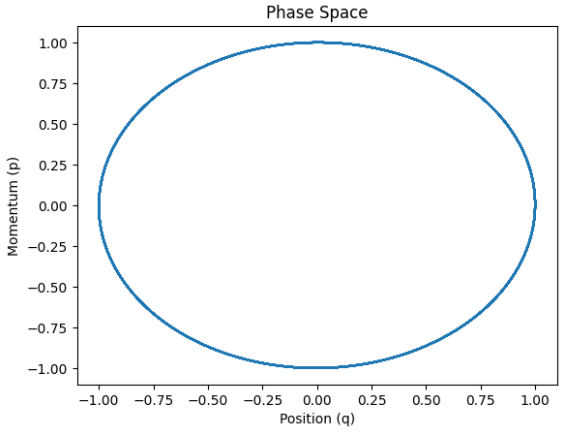

Figure 23: Phase space of the harmonic oscillator using the symplectic Euler method with $\Delta t = 0.01$.

The Figure 23 shows the phase space of the harmonic oscillator using the symplectic Euler method. We can see that the phase space is a closed curve, which is the correct behaviour of the harmonic oscillator.

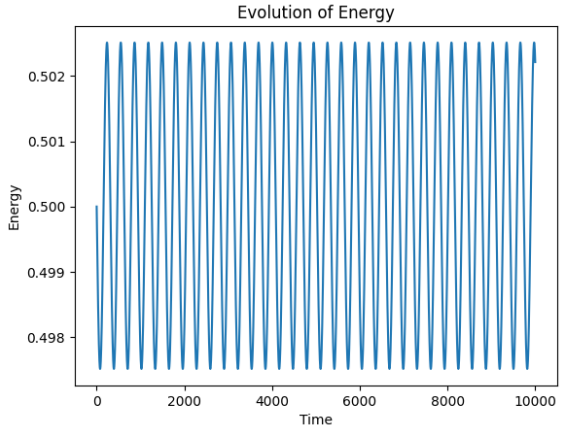

Figure 24: Energy of the harmonic oscillator using the symplectic Euler method with $\Delta t = 0.01$.

The Figure 24 shows the energy of the harmonic oscillator using the symplectic Euler method. We can see that the energy is conserved over time, as we wanted for the Hamiltonian dynamics. Note that is not a straight line but the average of the energy is the same, so we can conclude that the energy is conserved over time.

Now, if we define the next map:

$$\Gamma_{\Delta t}^{Euler} = \max_{n \in \mathbb{N}} \{|H_{\Delta t}(p^n, q^n) - H_{\Delta t}(p^0, q^0)|\}$$

The Figure 25 shows the function $\Gamma_{\Delta t}^{Euler}$ over $\frac{1}{\Delta t}$.

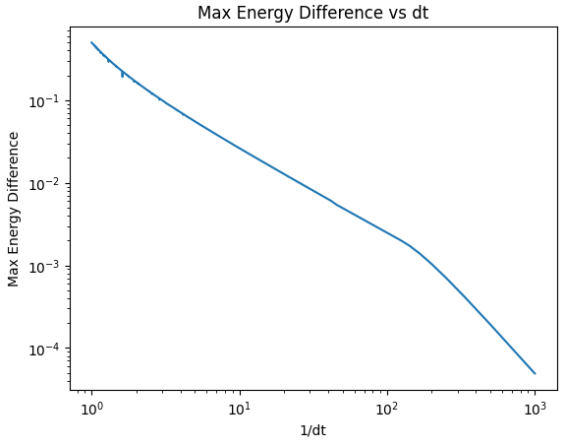

Figure 25: Function $\Gamma_{\Delta t}^{Euler}$ over $\frac{1}{\Delta t}$

To conclude the section, we can see in Figure 25 that the difference decreases as the time step decreases, which is the expected behaviour of any scheme that preserves the symplectic structure of the Hamiltonian dynamics equations. Therefore, we can conclude that the symplectic Euler method is a good method to solve the Hamiltonian dynamics equations.

2.1.5.1 Linear Stability Analysis of the Sympletic Euler Scheme

Using the equation 25 and replacing all the termusing the harmonic oscillator problem, one can rewrite the scheme as:

$$\begin{pmatrix} p^{n+1} \\ q^{n+1} \end{pmatrix} = A \begin{pmatrix} p^{n} \\ q^{n} \end{pmatrix}$$

With

$$A = \begin{pmatrix} 1 & -\Delta t k \\ \frac{\Delta t}{m} & 1 \end{pmatrix}$$

Noting $\xi = \Delta t \sqrt{\frac{k}{m}}$, the eigenvalues are:

$$\lambda_1 = 1 + i\xi, \;\; \lambda_2 = 1 - i\xi$$

In linear stability analysis, a scheme is said to be stable if the eigenvalues of the amplification matrix lie inside the unit circle in the complex plane. This means that their absolute values should be less than or equal to 1 for all possible values of ξ. In this case, their magnitudes are:

$$|\lambda_1| = |\lambda_2| = 1 + \xi^2$$

For stability we require $|\lambda| \leq 1$. This transalates to the condition:

$$1 + \xi^2 \leq 1$$

Therefore, the sympletic euler scheme isn't stable for any ξ.

2.1.6 Stormer-Verlet method

The Störmer-Verlet method, often simply referred to as the Verlet method, is a numerical technique used to integrate ordinary differential equations of the form $\dot{y} = f(y)$. It is particularly popular in molecular dynamics simulations and other problems modeled by Hamiltonian systems. The method can be derived directly from the Taylor series expansion of the solution. The central idea behind the method is to use information from both the current and previous time steps to predict the value at the next time step.

The Störmer-Verlet method applied to the Hamiltonian equations is as follows (using dHdq and dHdp to denote the partial derivatives of H with respect to q and p, respectively):

$$\begin{cases} p^{n+\frac{1}{2}} = p^n - \frac{\Delta t}{2} \nabla_q V(q^n) \\ q^{n+1} = q^n + \Delta t M^{-1} p^{t+1/2} \\ p^{n+1} = p^{n+\frac{1}{2}} - \frac{\Delta t}{2} \nabla_q V(q^{n+1}) \end{cases} \tag{26}$$

The numerical flow of this scheme is noted as ϕ_t^{Verlet}. In matter of fact, note that:

$$\Phi_{\Delta t}^{Verlet} = \phi_{\Delta t/2}^2 \circ \phi_{\Delta t}^1 \circ \phi_{\Delta t}^2$$

Therefore, it is easy to prove that, by construction, is a sympletic scheme. So given the problem of the harmonic oscillator using the Störmer-Verlet method, the update equations become:

$$\begin{cases} p^{n+\frac{1}{2}} = p^n - \frac{\Delta t}{2} k q^n \\ q^{n+1} = q^n + \Delta t M^{-1} p^{t+1/2} \\ p^{n+1} = p^{n+\frac{1}{2}} - \frac{\Delta t}{2} k q^{n+1} \end{cases} \tag{27}$$

The Figure 26 shows the phase space of the harmonic oscillator using the Störmer-Verlet method.

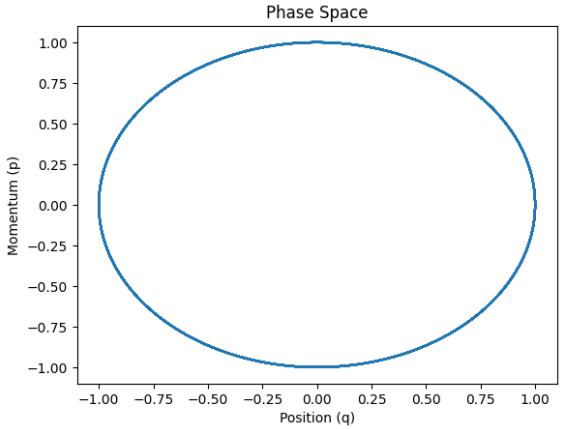

Figure 26: Phase space of the harmonic oscillator using the Störmer-Verlet method with $\Delta t = 0.01$.

The Figure 26 shows the phase space of the harmonic oscillator using the Störmer-Verlet method. We can see that the phase space is a closed curve, which is the correct behaviour of the harmonic oscillator.

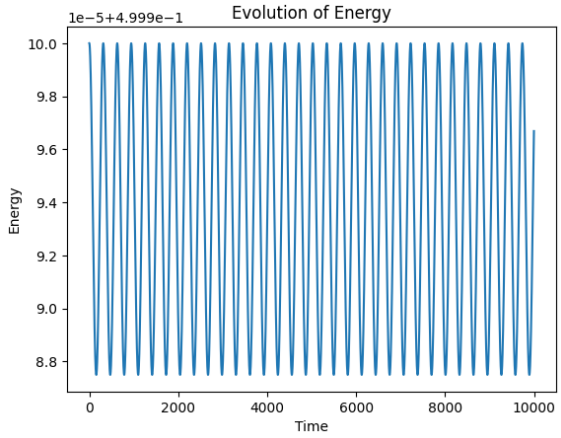

Figure 27: Energy of the harmonic oscillator using the Störmer-Verlet method with $\Delta t = 0.01$.

The Figure 27 shows the energy of the harmonic oscillator using the Störmer-Verlet method. We can see that the energy is conserved over time, as we wanted for the Hamiltonian dynamics. Like the behaviour of the Euler Sympletic method, we can see that in this case, the energy continues to oscillate around the average value, however the scale of this oscillation is much smaller than in the Euler Sympletic method. To see this, we define the following function:

$$\Gamma_{\Delta t}^{Verlet} = \max_{n \in \mathbb{N}} \{|H_{\Delta t}(p^n, q^n) - H_{\Delta t}(p^0, q^0)|\}$$

The Figure 28 shows the function $\Gamma_{\Delta t}^{Verlet}$ over $\frac{1}{\Delta t}$.

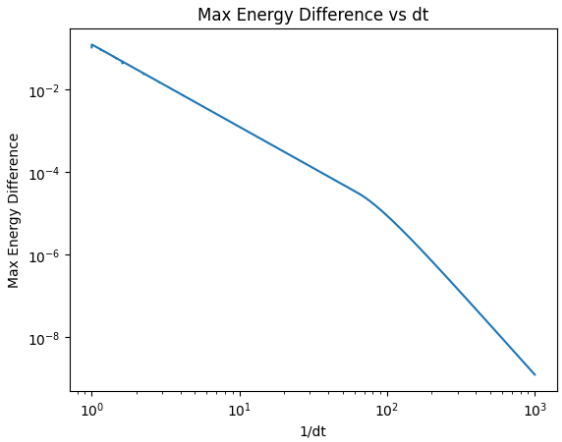

Figure 28: Function $\Gamma_{\Delta t}^{Verlet}$ over $\frac{1}{\Delta t}$.

So now we are prepared to overview the difference between both methods. The Figure 29 shows the function $\Gamma_{\Delta t}^{Euler}$ and $\Gamma_{\Delta t}^{Verlet}$ over $\frac{1}{\Delta t}$.

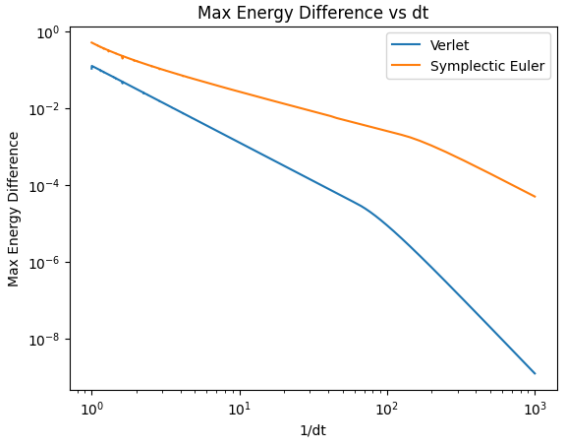

Figure 29: Function $\Gamma_{\Delta t}^{Euler}$ and $\Gamma_{\Delta t}^{Verlet}$ over $\frac{1}{\Delta t}$

As we can see, the Störmer-Verlet method is much better than the Euler Sympletic method. Probably, this is a consequence of the fact that the Störmer-Verlet method is a second order method, while the Euler Sympletic method is a first order method. Therefore, we can conclude that the Störmer-Verlet method is a better method to solve the Hamiltonian dynamics equations than the Euler Sympletic method.

2.1.6.1 Linear Stability Analysis of the Verlet Scheme

Using the equation 27 and replacing all the terms by its definition, and considering that $\omega = \sqrt{\frac{k}{m}}$, we can rewrite the scheme as:

$$\begin{pmatrix} p^{n+1} \\ q^{n+1} \end{pmatrix} = A \begin{pmatrix} p^n \\ q^n \end{pmatrix}$$

With

$$A = \begin{pmatrix} 1 - \frac{(\omega\Delta t)^2}{2} & \Delta t \\ -\omega^2\Delta t(1 - \frac{(\omega\Delta t)^2}{4}) & 1 - \frac{(\omega\Delta t)^2}{2} \end{pmatrix}$$

Noting $\xi = \frac{(\omega\Delta t)^2}{2}$, the eigenvalues are the solutions of λ in:

$$(1 - \xi - \lambda)^2 + \xi(2 - \xi) = 0$$

Therefore, the eigenvalues are:

$$\begin{cases} \lambda_1 = 1 - \xi + i\sqrt{\xi(2-\xi)} & \lambda_2 = 1 - \xi - i\sqrt{\xi(2-\xi)} \;\; if\ \xi(2-\xi) \geq 0 \\ \lambda_1 = 1 - \xi + \sqrt{\xi(\xi-2)} & \lambda_2 = 1 - \xi - \sqrt{\xi(\xi-2)} \;\; if\ \xi(2-\xi) \leq 0 \end{cases}$$

In linear stability analysis, a scheme is said to be stable if the eigenvalues of the amplification matrix lie inside the unit circle in the complex plane. This means that their absolute values should be less than or equal to 1 for all possible values of ξ. This ensures that errors do not grow unboundedly as we march forward in time.

Given the eigenvalues derived above:

$$\lambda_1 = \begin{cases} 1 - \xi + i\sqrt{\xi(2-\xi)} \text{ if } \xi(2-\xi) \geq 0 \\ 1 - \xi + \sqrt{\xi(\xi-2)} \text{ if } \xi(2-\xi) \leq 0 \end{cases}$$

$$\lambda_2 = \begin{cases} 1 - \xi - i\sqrt{\xi(2-\xi)} \text{ if } \xi(2-\xi) \geq 0 \\ 1 - \xi - \sqrt{\xi(\xi-2)} \text{ if } \xi(2-\xi) \leq 0 \end{cases}$$

The eigenvalues are complex conjugates. Their magnitudes are:

$$|\lambda_1| = |\lambda_2| = \sqrt{(1-\xi)^2 + \xi(2-\xi)}$$

For stability, we require $|\lambda_1| \leq 1$. This translates to the condition:

$$(1-\xi)^2 + \xi(2-\xi) \leq 1$$

This inequality leads to a tautology so it is always stable. This is consistent with the observed property of the Störmer-Verlet scheme which conserves the energy over time and doesn't let errors gro unboundedly. This, in combination with its second-order accuracy, makes it preferable over the Euler sympletic method for problems modeled by Hamiltonian systems.

2.1.7 Backward Error Analysis

In this section, we will introduce the backward error analysis. The backward error analysis is a technique to analyze the error of a numerical method. We can define the idea of this technique as: For a given problem $P(x)$, an exact input x and an approximate solution $\hat{x}$, we can define the backward error as the smallest perturbation Δx such that $P(x + \Delta x) = \hat{x}$. Therefore the formal definitions, extracted with little modifications from [2], is:

Definition 2.2 (Backward error)**.** Let $\hat{x}$ be an approximate solution to the equation $P(d) = x$. Then the backward error of $\hat{x}$ is defined by:

$$\eta = \min\{\epsilon : P(x + \Delta x) = \hat{x}, ||\Delta x|| \leq \epsilon\}$$

The following Figure 30 extracted from [3] illustrated the idea of the backward error analysis.

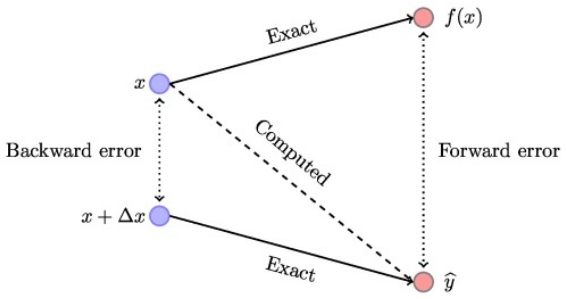

Figure 30: Backward error analysis.

2.2 Solar System Simulation

In this section, we'll harness the power of Hamiltonian dynamics to craft a simulation of the solar system. We'll adopt the Stormer-Verlet scheme, renowned for its energy-conserving properties, as our primary numerical method.

First of all, let's do a brief introduction to the problem we want to solve. We can see the solar system as a set of N bodies, each of them with a mass m_i a position q_i and a velocity v_i. The position and velocity are vectors in $\mathbb{R}^2$. The gravitational force between two bodies is given by:

2.2.1 Modeling the Solar System: Gravitational Dynamics

The motion of celestial bodies in the solar system, primarily planets, moons, and the Sun, can be described by Newton's law of universal gravitation. For N bodies, the gravitational force exerted on the i^{th} body due to the j^{th} body is given by:

$$\mathbf{F}_{ij} = -\frac{Gm_i m_j(\mathbf{q}_i - \mathbf{q}_j)}{|\mathbf{q}_j - \mathbf{q}_i|^3}$$

where G is the gravitational constant, approximately 6.67430×10^{-11} Nm^2/kg^2, $\mathbf{q}_i$ and $\mathbf{q}_j$ are the position vectors of the i^{th} and j^{th} bodies respectively and m_i and m_j are the masses of the i^{th} and j^{th} bodies respectively. Therefore, the total gravitational force acting on the i^{th} body due to all other bodies is:

$$\mathbf{F}_i = \sum_{j=1, j\neq i}^{N} \mathbf{F}_{ij}$$

To use Hamiltonian mechanics and the Störmer-Verlet scheme, we can represent the system in terms of its Hamiltonian, which is the sum of its kinetic and potential energies. For the i^{th} body:

1. Kinetic Energy, T_i:

$$T_i = \frac{1}{2} m_i \mathbf{v}_i \cdot \mathbf{v}_i$$

where $\mathbf{v}_i$ is the velocity of the i^{th} body.

2. Potential Energy, U_{ij}, due to the interaction between the i^{th} and j^{th} bodies:

$$U_{ij} = -\frac{Gm_i m_j}{|\mathbf{q}_j - \mathbf{q}_i|}$$

The total Hamiltonian H for the system is the sum of the kinetic and potential energies for all body pairs:

$$H = \sum_{i=1}^{N} T_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} U_{ij}$$

This Hamiltonian can be used to derive the equations of motion using Hamilton's equations, which can then be solved using the Störmer-Verlet scheme.

In the next section, we'll dive deeper into the implementation details, but this sets up the mathematical foundation for our solar system simulation based on gravitational interactions.

2.2.2 Störmer-Verlet Scheme for Gravitational Dynamics

Now that we have the Hamiltonian for the system, we can derive use it inside the Störmer-Verlet scheme to derive the equations of motion. The Störmer-Verlet scheme for the i^{th} body is:

$$\begin{cases} \mathbf{p}_i^{n+\frac{1}{2}} = \mathbf{p}_i^n - \frac{\Delta t}{2} \nabla_{\mathbf{q}_i} U(\mathbf{q}_i^n) \\ \mathbf{q}_i^{n+1} = \mathbf{q}_i^n + \Delta t M^{-1} \mathbf{p}_i^{n+1/2} \\ \mathbf{p}_i^{n+1} = \mathbf{p}_i^{n+\frac{1}{2}} - \frac{\Delta t}{2} \nabla_{\mathbf{q}_i} U(\mathbf{q}_i^{n+1}) \end{cases}$$

where $U(\mathbf{q}_i)$ is the potential energy of the i^{th} body due to all other bodies. The total potential energy of the system is the sum of the potential energies of all body pairs:

$$U = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} U_{ij}$$

where U_{ij} is the potential energy of the i^{th} body due to the j^{th} body. Therefore, the gradient of the potential energy of the i^{th} body is:

$$\nabla_{\mathbf{q}_i} U(\mathbf{q}_i) = \nabla_{\mathbf{q}_i} U(\mathbf{q}_i) = \sum_{j=1, j \neq i}^{N} \nabla_{\mathbf{q}_i} U_{ij}(\mathbf{q}_i, \mathbf{q}_j)$$

Given that:

$$U_{ij} = -\frac{G m_i m_j}{|\mathbf{q}_j - \mathbf{q}_i|}$$

We can compute the gradient of U_{ij} with respect to $\mathbf{q}_i$. Firstly, we can express the distance $|\mathbf{q}_j - \mathbf{q}_i|$ as r_{ij}:

$$r_{ij} = |\mathbf{q}_j - \mathbf{q}_i|$$

Thus, the potential energy becomes:

$$U_{ij} = -\frac{G m_i m_j}{r_{ij}}$$

To find the gradient of U_{ij} with respect to $\mathbf{q}_i$, we need to differentiate with respect to $\mathbf{q}_i$:

$$\nabla_{\mathbf{q}_i} U_{ij} = \frac{\partial U_{ij}}{\partial r_{ij}} \frac{\partial r_{ij}}{\partial \mathbf{q}_i}$$

First, differentiate U_{ij} with respect to r_{ij}:

$$\frac{\partial U_{ij}}{\partial r_{ij}} = G m_i m_j \frac{1}{{r_{ij}}^2}$$

Next, differentiate r_{ij} with respect to $\mathbf{q}_i$:

$$r_{ij} = \sqrt{(\mathbf{q}_j - \mathbf{q}_i) \cdot (\mathbf{q}_j - \mathbf{q}_i)}$$

$$\frac{\partial r_{ij}}{\partial \mathbf{q}_i} = \frac{\mathbf{q}_j - \mathbf{q}_i}{r_{ij}}$$

Putting it all together:

$$\nabla_{\mathbf{q}_i} U_{ij} = G m_i m_j \frac{1}{{r_{ij}}^2} \frac{\mathbf{q}_i - \mathbf{q}_j}{r_{ij}} = G m_i m_j \frac{\mathbf{q}_i - \mathbf{q}_j}{{r_{ij}}^3}$$

Thus, the gradient of the potential energy U_{ij} with respect to $\mathbf{q}_i$ is:

$$\nabla_{\mathbf{q}_i} U_{ij} = -G m_i m_j \frac{\mathbf{q}_i - \mathbf{q}_j}{|\mathbf{q}_j - \mathbf{q}_i|^3}$$

Now, plugging this back into the equation for the gradient of the total potential energy with respect to $\mathbf{q}_i$:

$$\nabla_{\mathbf{q}_i} U(\mathbf{q}_i) = \sum_{j=1, j \neq i}^{N} -G m_i m_j \frac{\mathbf{q}_i - \mathbf{q}_j}{|\mathbf{q}_j - \mathbf{q}_i|^3}$$

This is the gradient of the potential energy with respect to the position of the i^{th} body, taking into account all the interactions with other bodies in the system. Before executing the Störmer-Verlet scheme, let's show the potential energy of the system. The Figure 31 shows the potential energy of the system.

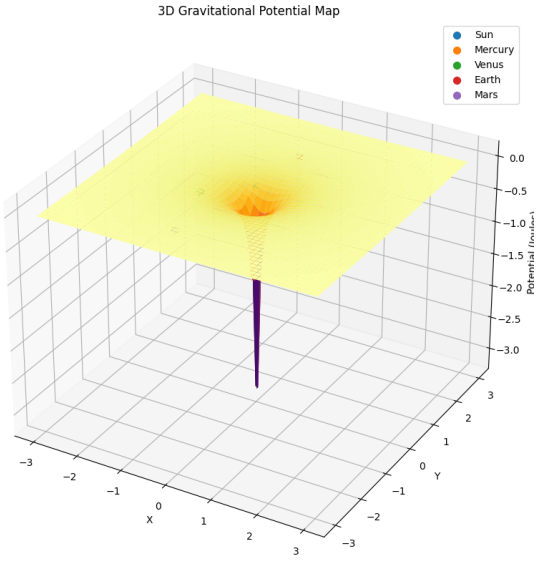

Figure 31: Potential energy of the part of the solar system.

As we could have expected, the Sun has, by far, the most negative potential energy, since it is the body with the highest mass. Also we can observe that the potential energy is negative, which is the expected behaviour of the gravitational potential energy.

2.2.3 Implementation

First of all, we need to define the parameters of the simulation. The Table 1 shows the parameters of the simulation.

Celestial Body	Mass (Relative to Sun)	Position (AU)	Mean velocity (AU/Year)
Sun	1.0000	0.000	0.00
Mercury	1.6505e-7	0.390	9.99
Venus	2.4335e-6	0.720	7.38
Earth	2.9860e-6	1.000	6.28
Mars	3.2085e-7	1.520	5.08
Jupiter	9.4950e-4	5.187	2.76

Table 1: Data of Celestial Bodies in the Solar System

Therefore, the units that we have used are:

- Mass: $M_\odot$ (Solar Mass)
- Distance: AU (Astronomical Unit)
- Time: Earth year

This is because we want to do a simulation of the solar system inside a machine that uses discrete numbers, therefore if we use the SI units, we could get some troubles with the precision of the machine, while if we use the units that we have used, we get numbers that are not too big or too small, so we can get a good precision. Also, using this numbers we get that the gravitational constant is:

$$G = 4\pi^2 \, \frac{\text{AU}^3}{\text{year}^2 \, M_\odot}$$

Note that we have used the real masses of the planets and the Sun, and the real distance between them. Therefore, we can see that the simulation is very similar to the real solar system. All of these parameters can be found on the NASA website [1]. Now we are prepared to implement the simulation.

For illustrative purposes, we'll simulate the orbits of the Sun, Mercury, Venus, Earth and Mars. The Figure 32 shows the orbits of the planets of the solar system during 365 days and with a time step of 1 day.

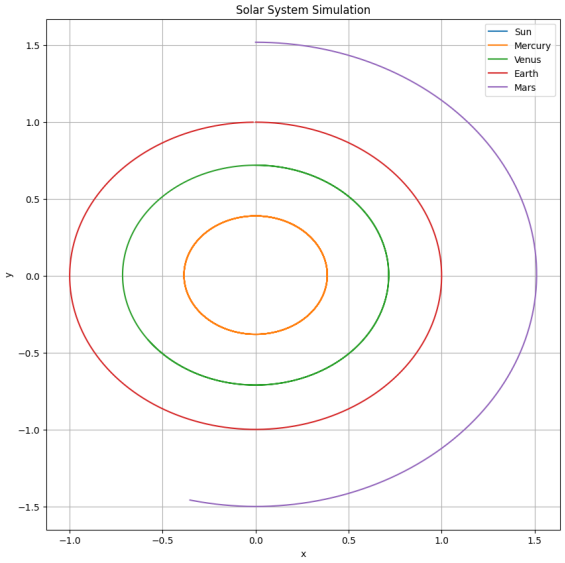

Figure 32: Simulation of the orbits of part of the planets of the solar system using the Störmer-Verlet scheme for $\Delta t = 1$ day and $T = 365$ days.

We can see that the orbits are closed curves, which is the expected behaviour of the planets of the solar system. Also, we can see that the orbits are not perfect circles, which is also the expected behaviour of the planets of the solar system.

Now let's see the energy of the system. The Figure **??** shows the energy of the system for the same simulation.

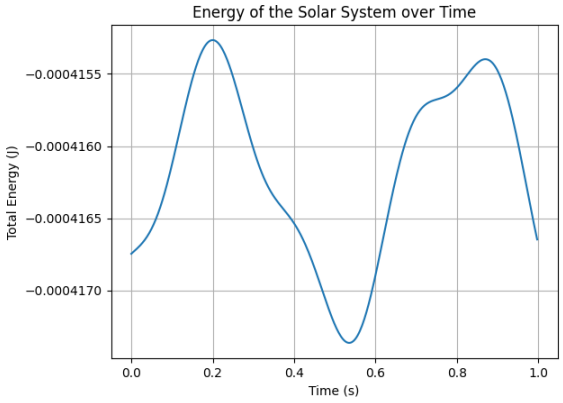

Figure 33: Energy of the simulation of part of the planets of the solar system using the Störmer-Verlet scheme for $\Delta t = 1$ day and $T = 365$ days.

We can see that the energy is conserved over time, as we wanted for the Hamiltonian dynamics. Note that is not a straight line but the average of the energy is the same and we can see a slight of oscillations that may be repeated in the future if the simulation is done for more time. So, let's do the simulation for 20 years. The Figure 34 shows the energy of the system for the same simulation.

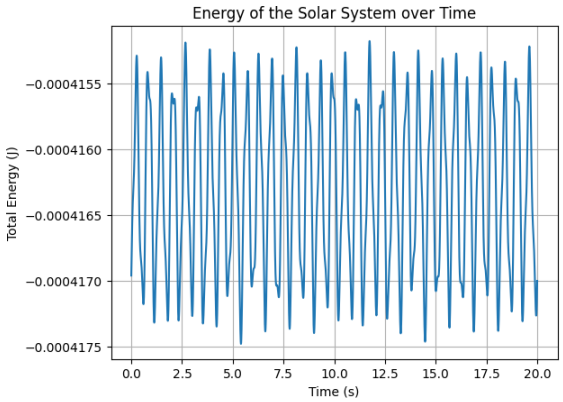

Figure 34: Energy of the simulation of part of the planets of the solar system using the Störmer-Verlet scheme for $\Delta t = 1$ day and $T = 20$ years.

As predicted, we can observe that the energy is preserved.

Now that we have the system created and we have seen that the energy is preserved, we can do some experiments.

2.2.3.1 Experiment 1: Expanding to the N-Body Problem

In our initial simulation, as detailed in the referenced table, we included data for Jupiter but did not incorporate it into the actual simulation. This was primarily due to Jupiter's significant distance from the inner planets, which could potentially complicate the visual clarity of our model. However, given Jupiter's substantial mass – the largest among the planets – its inclusion is crucial for a more comprehensive understanding of the solar system's dynamics. This necessitates an expansion of our model to an N-body problem, allowing for the gravitational influence of not just the Sun but also the other planets to be accounted for.

To explore this, we have extended our simulation to include Jupiter, aiming to observe its impact on the orbital paths of the other planets. The updated Figure 35 illustrates the orbits of the solar system's planets over a period of 20 years with a daily time step, now incorporating Jupiter. This addition provides a more realistic and complex representation of the gravitational interplay within our solar system.

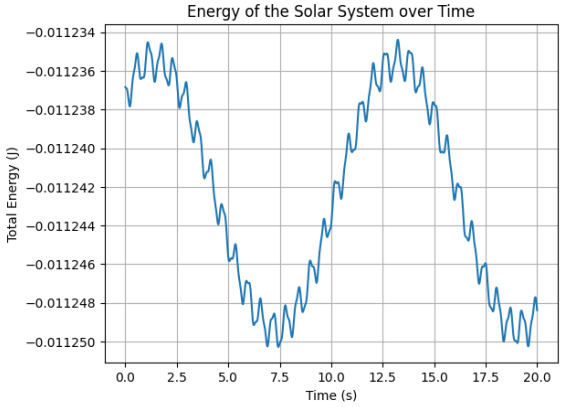

Figure 35: Simulation of the orbits of the planets of the solar system using the Störmer-Verlet scheme for $\Delta t = 1$ day and $T = 20$ years, now including Jupiter.

2.2.3.2 Experiment 2: Using Symplectic Euler Method

We alredy know that the Störmer-Verlet method performs better than the Euler Symplectic method, but let's see how the solar system behaves using the Euler Sympletic method. The Figure 36 shows the orbits of the planets of the solar system using the Euler Symplectic method for $\Delta t = 1$ day and $T = 20$ years.

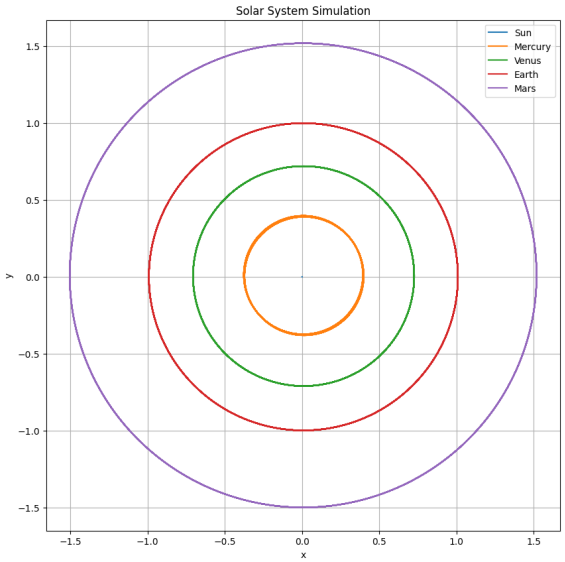

Figure 36: Simulation of the orbits of the planets of the solar system using the Euler Sympletic method for $\Delta t = 1$ day and $T = 20$ years.

Also we can plot the energy of the system. The Figure 37 shows the energy of the system for the same simulation.

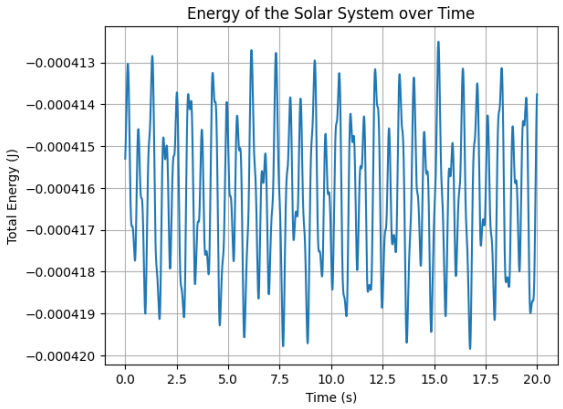

Figure 37: Energy of the simulation of the orbits of the planets of the solar system using the Euler Sympletic method for $\Delta t = 1$ day and $T = 20$ years.

Comparing it to the results obtained using the Störmer-Verlet method, again we can see that the first one is better in the following table:

Method	$\max H(x) - \min H(x)$	$\max \lvert H(X) - \mathbb{E}(H(X)) \rvert$
Symplectic Euler	7.29e-6	3.64e-6
Störmer-Verlet	2.32e-6	1.16e-6

Table 2: Data of Celestial Bodies in the Solar System

Therefore, in this problem we can conclude that the Störmer-Verlet method is 3.14 times better in terms of conservation of the Energy than the Euler Sympletic method.

References

[1] Planetary fact sheet, 2023. Accessed: 2023-11-02.

[2] V. Frayssé. Hdr thesis, 2018. Accessed: 2023-11-02.

[3] Nicholas J. Higham. What is backward error?, 3 2020. Accessed: 2023-11-15.

[4] Gabriel Stoltz. *An Introduction to Computational Statistical Physics*. Master de Mathématiques et Applications Spécialité Mathématiques de la Modélisation, 1 2023.