

# FYS4411 - Computational Physics II

## Project 2

Dorthea Gjestvang  
Even Marius Nordhagen

May 10, 2018

- Github repository containing programs and results:  
<https://github.com/evenmn/FYS4411/tree/master/Project%202>

### **Abstract**

This project aims to simulate a quantum dot by using a restricted Boltzmann machine (RBM) in combination with Variational Monte Carlo (VMC) with either Metropolis or Gibbs sampling. Using a form of the wave function given by the RBM, we find the optimal ground state wave function with the gradient decent method, and calculate an upper bound limit of the ground state energy and onebody density for one electron, as well as for two electrons with and without interaction. Our results are benchmarked with analytical expectations.

Using one electrons without interaction, the Metropolis algorithm reproduces the benchmarks. When turning interaction, the Metropolis algorithm yields a lower upper bound estimate of the ground state energy, compared to Gibbs. This is due to the wave function used in Gibbs sampling is not the exact ground state wave function. The onebody density changes when introducing interaction, which is as expected.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Theory</b>	<b>4</b>
2.1	Presentation of potential . . . . .	4
2.2	Solving this with machine learning . . . . .	5
2.2.1	Machine learning . . . . .	5
2.2.2	Neural network . . . . .	6
2.2.3	Restricted Boltzmann Machines . . . . .	7
2.3	Energy calculation . . . . .	9
2.3.1	Exact energy for non-interacting case . . . . .	9
2.4	Onebody density . . . . .	10
2.5	Virial theorem . . . . .	11
2.6	Error estimation . . . . .	11
<b>3</b>	<b>Method</b>	<b>12</b>
3.1	Variational Monte Carlo . . . . .	12
3.2	Metropolis algorithm . . . . .	13
3.2.1	Brute force . . . . .	13
3.2.2	Metropolis-Hastings . . . . .	14
3.3	Gibbs sampling . . . . .	15
3.4	Gradient descent . . . . .	16
3.5	Blocking method . . . . .	17
<b>4</b>	<b>Code</b>	<b>18</b>
4.1	Structure . . . . .	18
4.2	Implementation . . . . .	19
4.2.1	General implementation . . . . .	19
4.2.2	VMC . . . . .	19
4.2.3	Gradient Decent . . . . .	19
<b>5</b>	<b>Results</b>	<b>20</b>
5.1	Energy calculations, without interaction . . . . .	20
5.1.1	Playing with learning rate and number of hidden nodes	22
5.2	Energy calculations, with interaction . . . . .	23
5.2.1	Playing with learning rate and number of hidden nodes	25
5.3	Onebody density . . . . .	25
5.4	Average energies and distance . . . . .	26

<b>6</b>	<b>Discussion</b>	<b>26</b>
<b>7</b>	<b>Conclusion</b>	<b>29</b>
<b>8</b>	<b>Appendix A - Local energy and quantum force calculations for Metropolis</b>	<b>30</b>
8.1	Local energy . . . . .	30
8.2	Quantum force . . . . .	31
<b>9</b>	<b>Appendix B - Gibbs local energy and gradient calculation</b>	<b>32</b>
9.1	Local energy . . . . .	32
9.2	Gradient for gradient decent method . . . . .	32
<b>10</b>	<b>Appendix C - Derivation of the wave function</b>	<b>32</b>
<b>11</b>	<b>References</b>	<b>34</b>

# 1 Introduction

In this project, we study a two-electron system in a two-dimensional harmonic oscillator. This system is often referred to as a quantum dot. Quantum dots are a highly relevant topic in modern physics, and the technology has a wide range of applications, ranging from solar panel development to medical imaging [7].

The goal is to find an upper bound limit of the ground state energy and the onebody density of the system, using a restricted Boltzmann machine (RBM). We represent the wave function of the system as a probability distribution for the RBM, as first done by [6], who named this kind of network a *neural network quantum state* (NQS). We calculate estimates of the ground state energy. The system is evaluated for one electron, along with two electrons with and without interaction. The theory is presented in section 2.

The RBM is used to find the optimal wave function for the state, while we use Variational Monte Carlo (VMC) with either the Metropolis algorithm or Gibbs sampling to move the system towards the most likely state. The NQS wave function is optimized with the gradient decent method. The methods used are presented in section 3, and the code structure and implementation are presented in section 4. The upper bound estimate of the ground state energy and the onebody density are given in section 5, and our results are discussed in section 6.

## 2 Theory

### 2.1 Presentation of potential

In this project, we simulate a system of  $P$  electrons trapped in a harmonic oscillator potential, with a Hamiltonian given by

$$\hat{H} = \sum_{i=1}^P \left( -\frac{1}{2} \nabla_i^2 + \frac{1}{2} \omega^2 r_i^2 \right) + \sum_{i < j} \frac{1}{r_{ij}} \quad (1)$$

where  $\omega$  is the harmonic oscillator potential and  $\mathbf{r}_i = (x_i, y_i)$  is the position of electron  $i$ . The term  $\frac{1}{r_{ij}}$  is the interacting term, where  $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$  is the distance between a given pair of interacting electrons. Natural units have been used, such that  $\hbar = c = m_e = e = 1$ .

Since electrons are fermions, we need an antisymmetric wave function under exchange of two coordinates, and we need to take the Pauli principle

into account. A Slater determinant is therefore needed for multiple fermions to ensure that the total wave function is antisymmetric. In this project we will study particles in the ground state only, and according to the Pauli principle we can in this case study a maximum of two particles with spin  $s = \pm 1/2$ . The Slater determinant for two particles reads

$$\Psi_T = \begin{vmatrix} \Phi_1(\mathbf{r}_1) & \Phi_2(\mathbf{r}_1) \\ \Phi_1(\mathbf{r}_2) & \Phi_2(\mathbf{r}_2) \end{vmatrix} = \Phi_1(\mathbf{r}_1)\Phi_2(\mathbf{r}_2) - \Phi_2(\mathbf{r}_1)\Phi_1(\mathbf{r}_2) \quad (2)$$

where  $\Phi_i(\mathbf{r})$  is the single particle wave function (SPF) of state  $i$ . This contains a spatial part and a spin part, and we assume that it can be splitted up such that the spin part takes the antisymmetry property and does not affect the energy. Therefore we only need a symmetric spatial part to calculate the energies.

## 2.2 Solving this with machine learning

When solving a system of particles as the one described in the previous system with Variational Monte Carlo, as described in [3], we would need an ansatz for the wave function, where we use our physical intuition to create the form of a wave function with different variational parameters, and then let it be up to the computer to find the optimal parameters through a minimization method. However, this method is only as good as the physical intuition; if the form of the wave function is unrealistic, the results will be the same

This challenge can be mended by using machine learning. There are several different types of machine learning systems, and the one we will present and utilize in this project has the ability to learn and sample from a probability distribution. This is perfect for quantum mechanical problems, as we know from quantum mechanics the wave function  $\Psi$  is nothing more than a probability density, giving that  $\Psi^2$  is a probability distribution that says something about where a given particle most probably can be found. As we are solely interested in the energy of the two-fermion system, and not the exact wave function, the fact that the machine learning program does not explicitly give the wave function is therefore of no consequence. We still have to give some guidelines for the form of the probability distribution, but the machine learning program can sample from a larger variety of probability distributions compared to the form used in VMC.

### 2.2.1 Machine learning

With the goal of solving the quantum mechanical system presented in section 2.1 in mind, we should start by explaining what machine learning is. Machine

learning is the idea that a computer can be trained to learn to yield certain outputs, without directly being told exactly what to give. Examples on this is pattern recognition, where the computer first is shown for example pictures of wolves and huskies. After training the computer on pictures where the computer sees huskies and wolves and is told the correct answer, it should after a sufficiently long training period, be able to recognize huskies and wolves by itself.

The example described above is what we call supervised learning, where the correct output answer is known during the training program. A machine learning program could also be unsupervised, where the correct answer is unknown, or based on reinforcement learning, where the the program learns by conducting trial-and-error experiments.

### 2.2.2 Neural network

This sound amazing, and maybe even impossible. Therefore the question now is: how to program computers to learn, just like humans? The answer is, fittingly, that we should make the program run like the the human brain by implementing what is called a neural network. Inspired by neurons in the human brain, a neural network is a programmed network of variables, called nodes, that communicate in a given manner. Each node preforms a simple process: based on the input it receives, and how that input is weighted, it decides wheter or not to fire. The mathematical model of an example of a neural node was presented by McCulloch and Pitts in 1943 [2], shown in figure 1, where the input is marked  $x_i$ , the weights deciding how much the input should count is  $W_i$ , and the output from the node based on  $x_i$  and  $W_i$  is called  $h$ .

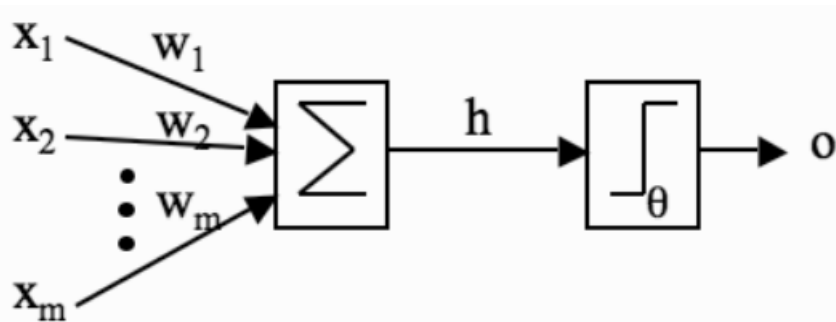


Figure 1: McCullochs and Pitts' model of neutrons visualized. Image reproduced from [2]

The neurons are arranged in layers in the neural network, one visible layer that receives the input, and up to several hidden layers. The layers are arranged such that the output values from the visible nodes is the input values of the visible nodes. The nodes can also have bias values, that shift the output value  $h$  with a certain number, and the nodes in different layers can be connected in different ways. An example of a neural network with two layers is shown in figure 2.

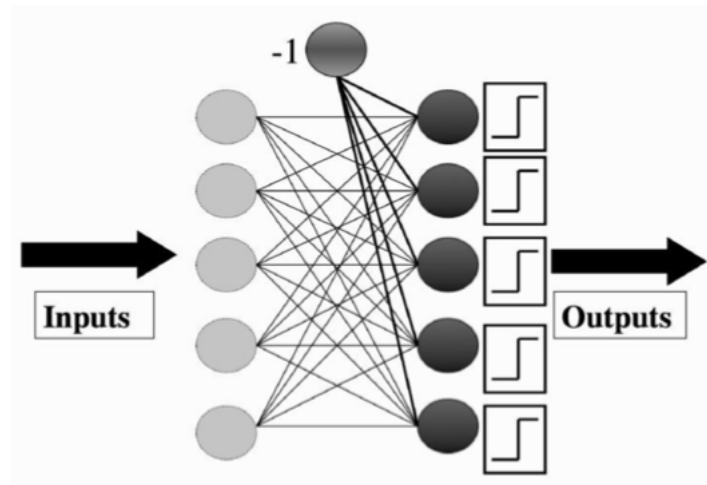


Figure 2: An example of a neural network with one visible and one output layer of nodes, and with bias  $-1$  on the hidden nodes. Image reproduced from [2]

The idea behind machine learning is that the weights  $W_i$ , that decides how much a node puts emphasis on a given input, can be changed, and thus change the system's response to the same input. We will explain with the example from earlier, with huskies and the wolves: first, the program are shown pictures of wolves and huskies, and it is told the correct answer. The weights  $W_i$  are then updated such that when a husky is shown, the output "husky" is generated, and the same for wolves. After a sufficiently long training period, the program's weights are optimized for recognizing wolves and huskies. When shown a picture, the program should then by itself be able to determine wheter it is a wolf or a husky that it sees.

### 2.2.3 Restricted Boltzmann Machines

There are plenty of ways to put together a neural network, as one can modify the number of nodes and layers, the bias, and also which nodes that are allowed to communicated. In this project, we will be using the so-called

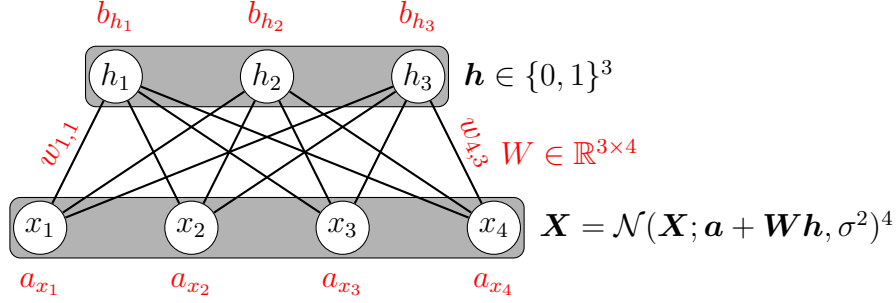


Figure 3: An example of a Gaussian-Binary RBM with four visible nodes ( $\mathbf{X}$ ) and three hidden nodes ( $\mathbf{h}$ ), where  $\mathbf{W}$  are the weights, and  $\mathbf{a}$  and  $\mathbf{b}$  are the bias weights.

Restricted Boltzmann Machine (RBM). It is a two-layer network. The reason why it is named "restrictive" is that there are no connections between nodes in the same layer, but every node in the previous layer is connected to all the nodes in the next layer. The RBM can learn to draw samples from a probability distribution, which is just what we want in our project. In addition, we want to use a Gaussian-Binary RBM (see figure 3), where the hidden nodes have binary values, while the positions of the particles can take continuous values, as they are, in fact, positions.

The joint probability distribution is known from statistical mechanics,

$$F(\mathbf{X}, \mathbf{H}) = \frac{1}{Z} e^{-\beta E(\mathbf{X}, \mathbf{H})} \quad (3)$$

where we set  $\beta = 1/kT = 1$  and  $Z$  is the partition function, which can be ignored since it will vanish anyway.

The system energy of a Gaussian-Binary RBM is given by

$$E(\mathbf{X}, \mathbf{H}) = \sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^N b_j H_j - \sum_{i,j=1}^{M,N} \frac{X_i W_{ij} H_j}{\sigma_i^2} \quad (4)$$

[8]. We then define the NQS wave function as  $\Psi = F(\mathbf{X}, \mathbf{H})$ , which has a general form of the probability distribution. In equation 5 the form of the wave function is shown, when we have used that we have  $M$  visible nodes and  $N$  hidden nodes, and that the hidden nodes should take binary values.

$$\Psi = \frac{1}{Z} e^{\sum_i \frac{(X_i - a_i)^2}{2\sigma_i^2}} \prod_j^N (1 + e^{b_j + \sum_i \frac{X_i W_{ij}}{\sigma_i^2}}) \quad (5)$$

See appendix C for derivation.



## 2.3 Energy calculation

We want to calculate the energy of the two-fermion system, given the wave function described in 5. As explained in [3], the energy of the system is the expectation value of the Hamiltonian, but this is hard to compute directly.

$$E_L(\mathbf{r}) = \frac{1}{\Psi_T(\mathbf{r})} \hat{H} \Psi_T(\mathbf{r}). \quad (6)$$

By defining the local energy given in equation 6, the energy can be expressed as equation 7, and this can be solved with a Monte Carlo loop.

$$E = \int |\Psi_T^2| E_L d\mathbf{r} \quad (7)$$

### 2.3.1 Exact energy for non-interacting case

The wave function of one electron in a two-dimensional harmonic oscillator is given by

$$\phi_{n_x, n_y} = A H_{n_x} \sqrt{\omega} x H_{n_y} \sqrt{\omega} y e^{-\frac{\omega(x^2 + y^2)}{2}} \quad (8)$$

where  $H_{n_x} \sqrt{\omega} x$  is the Hermite polynomials and A is the normalization constant, and  $n_x$  and  $n_y$  are the principal quantum numbers. If we denote  $\phi_1$  the spatial wave function for the first electron, and  $\phi_2$  the spatial wave function for the second electron, then the total spatial wave function for the two-particle system is  $\phi_1 \phi_2$ . As explained in section 2.1, the spin wave functions are anti-symmetric, and therefore it is fine that the spatial wave function is symmetric.

We know that the energy of a harmonic oscillator in one dimension is given by

$$E_n = \omega \left( n + \frac{1}{2} \right) \quad (9)$$

where  $n$  is the principal quantum number.

For a one-electron system in two dimensions the energy can be shown to take the value

$$H |\phi_1\rangle = E_1 |\phi_1\rangle = \omega(n_{x,1} + n_{y,1} + 1) |\phi_1\rangle \rightarrow E_1 = \omega(n_x + n_y + 1) \quad (10)$$

where  $n_x$  and  $n_y$  are the principal quantum number for the  $x$  and  $y$  direction respectively, and we have used natural units, and thus all the energies are given in atomic units a.u.

When calculating the energy of the two-electron system we then get the following

$$H |\phi_1 \phi_2\rangle = |H\phi_1\rangle |\phi_2\rangle + |\phi_1\rangle |H\phi_2\rangle = E_1 |\phi_1 \phi_2\rangle + E_2 |\phi_1 \phi_2\rangle \quad (11)$$

For the ground state  $n_x = n_y = 0$ ,  $E_1 = E_2 = \omega$  and therefore the ground state energy is simply

$$E_1 + E_2 = 2\omega. \quad (12)$$

For the non-interacting case with  $\omega = 1$  it is known the ground state energy is 3 a.u. [5].

## 2.4 Onebody density

A quantity that often is handy to calculate is the onebody density. The onebody density is the density of particles in the potential, and may give a better understanding of the placement of the particles than their positions. The onebody density  $\rho_i$  with respect to a given particle  $i$  is given by equation 13, which can be solved by Monte Carlo integration, as described in [3].

$$\rho_i(\mathbf{r}) = \int_{-\infty}^{\infty} d\mathbf{r}_1 \dots d\mathbf{r}_{i-1} d\mathbf{r}_{i+1} \dots d\mathbf{r}_N |\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N)|^2. \quad (13)$$

Apart from the fact that the one body density is a good visual representation of the positions of the particles, it is also a quantity that is often calculated experimentally. Therefore, it is simpler to compare computational simulations to experimental results when having calculated the onebody density.

Without interaction, the onebody density of particle  $i$  is simply its wave function squared, which is known to be

$$\rho_i(\mathbf{r}) = A^2 e^{-r_i^2} \quad (14)$$

for a particle in the groundstate of a harmonic oscillator potential with  $A$  as the normalization constant. Numerically we solve this by dividing the area around a particle into bins. Thereafter we count the number of particles in each bin, and divide by the area to get the density. The area of bin  $k$  is

$$A_k = \pi r_k^2 - \pi r_{k-1}^2 = 2\pi r_k dr - \pi dr^2 \quad (15)$$

where  $r_k$  is the outer radius of bin  $k$  and  $dr$  is the bin width.

## 2.5 Virial theorem

The virial theorem is an equation in mechanics that gives a relation between the time-averaged kinetic energy  $\langle T \rangle$  and the potential forces acting on  $N$  particles confined in a stable system. The theorem takes the following form:

$$\langle T \rangle = -\frac{1}{2} \sum_{i=0}^N \langle \mathbf{F}_i \cdot \mathbf{r}_i \rangle \quad (16)$$

where  $\mathbf{F}_i$  is the sum of forces acting on particle  $i$ .

In a two-particle system where the potential is proportional to  $r^n$ , then the virial theorem simplifies to

$$2\langle T \rangle = n\langle V_{tot} \rangle \quad (17)$$

## 2.6 Error estimation

As a physicist one should always be sceptical to measurements presented without an estimation of the uncertainties. When someone says they have measured their height to be 1.90 cm, is the uncertainty  $\pm 1\text{cm}$  or  $\pm 10\text{cm}$ ? If they used  $\pm 1\text{cm}$ , we have a fairly good idea of the height of the person, but if it is  $\pm 10\text{cm}$ , we do not really know anything about the real height of the person.

The error in a measurement is the difference between the true value of the measurand,  $x$ , and the measured value  $x_i$ :

$$error = x - x_i \quad (18)$$

However, the true value  $x$  is rarely known. When  $x$  is not known, it is thus impossible to give the error in the measurement. We must therefore be content with providing estimates of the error. It is common to give this uncertainty as the standard deviation  $\sigma$ ; when performing a measurement  $n$  times, such that we have  $n$  measured points  $x_1, \dots, x_n$  where the average value is  $\langle x \rangle$ , then the standard deviation is defined as

$$\sigma_\theta = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \langle x \rangle)^2} \quad (19)$$

When giving the uncertainty, one writes  $x_i \pm \sigma$ , where  $\sigma$  is so that  $\approx 68\%$  of the measured values lies within the interval  $\pm \sigma$ .

There are different kinds of sources of error when conducting computer experiments. Systematic errors originate from faults in the theory used in the experiment, and this might give consistently wrong results. Systematic errors are hard to handle, as one might get seemingly good looking results. One way to avoid systematic results are to benchmark the code with either experimental results, or with other’s computer simulation results.

Another source of error is the statistical error, which is based on how many measurements you have done. In the example with the height measurement in the beginning of this section, if the experimentalist only measured the height one time, we can be fairly certain that this is not the true value of the person. If they, however, repeated the same measurement several times and then maybe used the average as the estimate of the measurand, we can be more confident that they have a better estimate, compared to when they only used one point. A common way to estimate the statistical uncertainty is with equation 20.

$$\sigma^2 \approx \langle x^2 \rangle - \langle x \rangle^2 \quad (20)$$

However, as explained in [3], this does not account for the covariance between measurements. This leads to equation 20 being an underestimation of  $\sigma^2$ , and is thus more a guideline for the size of the uncertainty, more than an actual estimate of it. Luckily, there are computational methods that can calculate  $\sigma^2$  with the covariance contribution included. The method used in this project is the blocking method, presented in section 3.5.

## 3 Method

### 3.1 Variational Monte Carlo

The Variational Monte Carlo method, often referred to as VMC, is a method in computational physics. The goal of VMC is to approximate the ground state energy of a quantum mechanical system by using the variational principle. The variational principle says that for any given trial wave function  $\Psi_T$ , the energy  $E$  given this  $\Psi_T$  is always more than or equal to the ground state energy  $E_0$  of the system:

$$E = \frac{\langle \Psi_T | H | \Psi_T \rangle}{\langle \Psi_T | \Psi_T \rangle} \geq E_0 \quad (21)$$

In VMC, the user inputs a trial wave function for the quantum mechanical system. Over a Monte Carlo loop, the particles in the system are moved

randomly, and for each proposed move, the new position of the particle is accepted or rejected depending on whether the new position is more probable according to  $\Psi_T$  or not, compared to the previous position. For the ground state wave function, the position which yields a minimum in the energy of the quantum mechanical system is the most probable position. The result is therefore that the particles in the system are moved towards lower energies. When the method has converged, the energy  $E$  is an upper limit estimate of the true ground state energy  $E_0$ . If  $\Psi_T$  in addition is the true ground state wave function, then  $E = E_0$ .

### 3.2 Metropolis algorithm

Above, we described that the VMC method had to reject or accept proposed moves of the particles in the system. The task of handling this is often left up to the Metropolis algorithm. The Metropolis algorithm uses the theory that the probability of a system to transition from a state  $i$  to state  $j$  is given by

$$W_{i \rightarrow j} = T_{i \rightarrow j} \cdot A_{i \rightarrow j} \quad (22)$$

where  $T_{i \rightarrow j}$  is the transition probability and  $A_{i \rightarrow j}$  is the acceptance probability. If we denote the probability for being in a given state for  $p$ , and then look at the transition from a state  $i$  to a state  $j$  and back again, then it can be shown that we obtain the following relation between the transition and acceptance probabilities [3]:

$$\frac{A_{j \rightarrow i}}{A_{i \rightarrow j}} = \frac{p_i T_{i \rightarrow j}}{p_j T_{j \rightarrow i}} = w \quad (23)$$

As the Metropolis algorithm's job is to evaluate probable moves between states, the expression in equation 23 is used as the acceptance criteria. Thus the ratio  $\frac{A_{j \rightarrow i}}{A_{i \rightarrow j}}$  describes if the system is more likely to move one way or the other. There are several methods that can be used for accepting and rejecting the moves, and these use different expressions for  $T$  and  $p$  in equation 23. In this project, we will study the brute force and the Hastings sampling algorithms.

#### 3.2.1 Brute force

The brute force got its name because when proposing and accepting moves, the algorithm does so without "thinking" which way the system will most probably move. A random particle with position  $\mathbf{R}$  is chosen, and its new position  $\mathbf{R}_{new}$  is proposed at random as follows:

$$\mathbf{R}_{new} = \mathbf{R} + \mathbf{r} \cdot \text{step}. \quad (24)$$

where  $\mathbf{r}$  is random numbers and  $\text{step}$  is the steplength.

The brute force algorithm does not consider the transitions probabilities, and simply puts  $T_{i \rightarrow j} = T_{j \rightarrow i}$ . To check the probability of a system being found in a given state, it uses  $p(\mathbf{R}) = |\Psi_T(\mathbf{R})|^2$  for both position, such that the probability  $w$  for the transition is given by

$$w = \frac{P(\mathbf{R}_{new})}{P(\mathbf{R})} = \frac{|\Psi_T(\mathbf{R}_{new})|^2}{|\Psi_T(\mathbf{R})|^2}. \quad (25)$$

A random number  $r \in [0, 1]$  is then drawn, and compared to  $w$ . If  $w$  is larger than  $r$ , then the new position is accepted, and if not, the move is rejected and the particle is moved back to its original position:

$$\text{New position: } \begin{cases} \text{accept} & \text{if } w > r \\ \text{reject} & \text{if } w \leq r. \end{cases} \quad (26)$$

This way, the algorithm will over many iterations move the particles towards the most probable state.

### 3.2.2 Metropolis-Hastings

The Metropolis-Hastings algorithm, also known as importance sampling, have a more refined approach to choosing new positions and accepting them, compared to the brute force method. Instead of proposing the new position at random, the Metropolis-Hastings algorithm considers the quantum force  $\mathbf{F}$ , given by equation 8.2, which says something about which direction the particle is pushed in.

$$\mathbf{F}(\mathbf{R}) = \frac{2\nabla\psi_T(\mathbf{R})}{\psi_T(\mathbf{R})} \quad (27)$$

The equation that describes the motion of a particle pushed around by the drift force  $\mathbf{F}$ , is called the Langevin equation. The Langevin equation can be derived from the Fokker-Planck equation, which describes the time-evolution of the probability density function. The Langevin equation can be solved using the Euler method. The result, presented in equation 28, describes how the Metropolis-Hastings algorithm chooses a new position for the random particle:

$$\mathbf{R}_{new} = \mathbf{R} + D\mathbf{F}(\mathbf{R})\Delta t + \xi\sqrt{\Delta t} \quad (28)$$

where  $D = \frac{1}{2}$  is the diffusion constant and  $\Delta t$  is the timestep, that decides how long steps the method should take for each move. This is a much more educated guess compared to the brute force algorithm.

When regarding the transition probabilities, the Metropolis-Hastings algorithm does not go the easy way with  $T_{i \rightarrow j} = T_{j \rightarrow i}$ . Instead it uses that the transition probabilities are given as the Green's function, also found using the Fokker-Planch equation:

$$\begin{aligned} G(R_{new}, R, \Delta t) &= \frac{1}{(4\pi D \Delta t)^{3N/2}} \exp\left[-(R_{new} - R - D\Delta t F(R))^2 / 4D\Delta t\right] \\ &= T_{R \rightarrow R_{new}} \end{aligned} \quad (29)$$

such that the probability parameter  $w$  becomes

$$w = \frac{G(R, R_{new}, \Delta t) |\Psi_T(R_{new})|^2}{G(R_{new}, R, \Delta t) |\Psi_T(R)|^2}. \quad (30)$$

Thereafter a random number  $r$  is drawn and compared to  $w$ , just like in the brute force method. This more complicated way of proposing and accepting positions makes the Metropolis-Hastings slightly slower than the brute force method, but also quicker to converge towards the ground state. When using the Metropolis-Hastings algorithm, we need the expression for the quantum force for the NQS wave function. The calculations and the final expression of  $\mathbf{F}$  are shown in appendix 8.

One thing that is worth to note with the brute force and the Metropolis-Hastings algorithms when doing machine learning projects, is that both of them in this project are implemented so that only the positions, i.e., the visible nodes, are updated. This makes it so that we lose the opportunity to modify half of the system, which maybe yield a worse convergence than if we varied the whole system.

### 3.3 Gibbs sampling

As an alternative to the Metropolis algorithm, one can use the Gibbs sampling method. In both the brute force method and Metropolis-Hastings, one has to account for that proposed particle moves can be rejected, and this "wastes" one Monte Carlo cycle with doing several flops, when no change is inflicted on the system.

The Gibbs sampling algorithm differs from the Metropolis in several ways. In section 2.2.3 we wrote that we chose the trial wave function  $\Psi = F(\mathbf{X}, \mathbf{H})$ ,

where  $F$  is defined in equation 3. In Gibbs sampling, we instead use the trial wave function  $\Psi_T = \sqrt{F}$ .

Furthermore, instead of updating only the visible nodes, as is done in Metropolis, the Gibbs sampling algorithm is a two-step sampling process where first a new position is chosen, and then one of the hidden nodes are updated. The positions are sampled from a Gaussian probability distribution, shown in equation 31, while the probability for sampling  $h = 1$  is shown in equation 32.

$$P(\mathbf{X}|\mathbf{h}) = \mathcal{N}(\mathbf{X}; \mathbf{a} + \mathbf{W}\mathbf{h}, \sigma^2) \quad (31)$$

$$P(H_j = 1|\mathbf{X}) = \frac{1}{1 + e^{-b_j - \frac{\mathbf{X}^T \mathbf{W}_{cols,j}}{\sigma^2}}} \quad (32)$$

When we say that the visible and hidden nodes here are "sampled", we mean that they are drawn from the probability distribution, and then automatically accepted. This is different from Metropolis, and should mean that the Gibbs sampling method converges faster than both brute force and Metropolis-Hastings, as no Monte Carlo cycles are lost on the move not being accepted.

As we use a new trial wave function in Gibbs sampling, the expression for the local energy changes. The expression for the local energy using the Gibbs trial wave function is calculated in appendix 9.

### 3.4 Gradient descent

We use stochastic gradient descent to update the weights in order to minimize the local energy. The process itself is quite similar to that one typically used in VMC where the weights are our variational parameters, but compared to a standard VMC we have a lot more parameters to vary. The updating algorithm for updating the variational parameter  $\alpha_i$  goes as

$$\alpha_i^+ = \alpha_i - \eta d\alpha_i \quad (33)$$

where the gradient of the local energy with respect to parameter  $\alpha_i$  is

$$d\alpha_i = \frac{\partial \langle E_L \rangle}{\partial \alpha_i} = 2 \left( \left\langle E_L \frac{1}{\Psi} \frac{\partial \Psi}{\partial \alpha_i} \right\rangle - \langle E_L \rangle \left\langle \frac{1}{\Psi} \frac{\partial \Psi}{\partial \alpha_i} \right\rangle \right) \quad (34)$$

where  $\eta$  is the learning rate. We need to do this for all the parameters, and for our purposes it is convenient to vectorize the gradient such that  $d\mathbf{a}$  and



$d\mathbf{b}$  are vectors and  $d\mathbf{W}$  is a matrix. The gradients to implement are

$$d\mathbf{a} = \frac{1}{\Psi} \frac{\partial \Psi}{\partial \mathbf{a}} = \frac{\mathbf{X} - \mathbf{a}}{\sigma^2} \quad (35)$$

$$d\mathbf{b} = \frac{1}{\Psi} \frac{\partial \Psi}{\partial \mathbf{b}} = \frac{1}{1 + e^{-v}} \quad (36)$$

$$d\mathbf{W} = \frac{1}{\Psi} \frac{\partial \Psi}{\partial \mathbf{W}} = \frac{\mathbf{X}}{\sigma^2} \frac{1}{1 + e^{-v}} \quad (37)$$

which means that we need to calculate  $d\mathbf{b}$  and  $d\mathbf{W}$  elementwise, i.e,  $db_k = (1 + e^{-v_k})^{-1}$ .  $\mathbf{v}$  is defined as

$$\mathbf{v} \equiv \mathbf{b} + \frac{\mathbf{X}^T \cdot \mathbf{W}}{\sigma^2}. \quad (38)$$

### 3.5 Blocking method

In section 2.6, we described the need of a proper estimation of the uncertainty in computational simulations, where the covariance was included in the calculation of  $\sigma$ . A quick and easy way to get a proper estimate of the uncertainty, is by using the blocking method. If we have a data set  $x_1, x_2, \dots, x_n$  with  $n$  data points, where the mean value is  $\langle x \rangle$ . We can calculate the standard deviation using equation 19, however, as explained before, this does not include the covariance, as the data points  $x_1, x_2, \dots, x_n$  may be correlated. A way to get around this, is by transforming the data set such that

$$x'_i = \frac{1}{2}(x_{2i} + x_{2i+1}) \quad (39)$$

The number of points in the transformed data set is now  $n'$ , and every  $x'_i$  consists of two of the original data points. The transformed  $x'_i$ 's are referred to as blocks, and the standard deviation in each block  $\sigma_i$  can now be calculated. An estimate of the total standard deviation  $\hat{\sigma}$  when we have  $n'$  transformed data points is

$$\hat{\sigma} = \frac{\sigma_1 + \dots + \sigma_{n'}}{n'} \quad (40)$$

We now increase the block size, such that there are even more of the original data point in each block, and calculate the total standard deviation from equation 40 again. If we continue doing this and plot the resulting estimates of  $\hat{\sigma}$ , we will observe that it first increases, until it flats out. When we have reached this plateau, we know that the data blocks  $x'_i$  are no longer correlated. The covariance is then 0, and we have thus obtained a proper estimation of the uncertainty.

## 4 Code

In this section, we give an overview of the structure and implementation of our code, to make it easier for potential users to familiarize themselves with the code.

### 4.1 Structure

In order to keep the code structured, we have split the running of the program up in several files. The user only needs to interact with `main.cpp`, as all the parameters for running the machine learning project are set here. Examples are that the user can choose between brute force, Metropolis-Hastings or Gibbs sampling, and set the number of visible and hidden nodes, along with the number of iterations. `main.cpp` calls on the function `GradientDecent` in `gradient_decent.cpp`, where the machine learning structure is implemented. The machine learning structure consists of the VMC-loop, where the nodes are updated, and the gradient decent method, where the weights are updated. `GradientDecent` uses the class `WaveFunction` in `wavefunction.cpp` to calculate the value of the wave function, along with the energy of the system. Depending on whether the user chose brute force, Metropolis-Hastings or Gibbs sampling, `GradientDecent` calls functions in either `wavefunction.cpp`, `hastings_tools.cpp` or `gibbs_tools.cpp` to update the positions and calculate the transition probabilities. The test functions are found in `test.cpp`. The structure can be seen in figure 4.

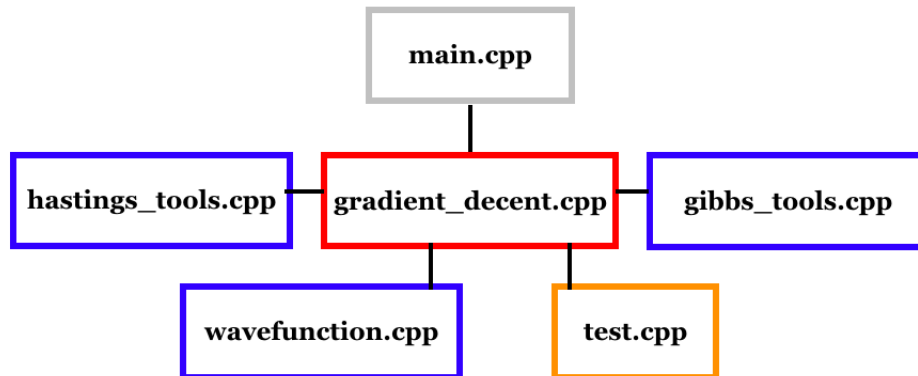


Figure 4: The structure of the machine learning program, showing the communication between the different files

## 4.2 Implementation

### 4.2.1 General implementation

```
Set up vectors X and h for visible and hidden nodes, fill with
    ↪ random numbers, X between -1 and 1, h with 0 and 1

Set opp vectors a, b, W for the weights, fill with random numbers
    ↪ between -1 and 1

Initialize class WaveFunction

for iter iterations
    for MC Monte Carlo cycles
        #do VMC

        #do GradientDecent

    end when stopping criterias == true, when method has
        ↪ converged
    print final energy
```

### 4.2.2 VMC

```
for M Monte Carlo iterations

    if Metropolis:
        move random particle
        accept or reject move based on transition probability

    if Gibbs:
        sample new position and hidden node

    calculate energy of new position
    update sum of E.L

print estimated E
```

### 4.2.3 Gradient Decent

```
for iter iterations:

    for MC Monte Carlo cycles:
        #do VMC

    Check if stopping criteria == true:
        if true, break and print final values

    Calculate gradient for the weights a, b and W
    Update a, b and W
```

## 5 Results

Firstly, we calculate the energy of one particle in one dimension. Secondly, we extend the system to two interacting electrons in a two-dimensional harmonic oscillator. Furthermore the onebody densities with and without interaction are presented and finally we study how the total energy is distributed between kinetic energy and potential energy for various harmonic oscillator frequencies.

### 5.1 Energy calculations, without interaction

We first study the energy of a non-interacting electron in one dimension, and in figure 5 we compare the energies produced by the three methods as a function of number of iterations.

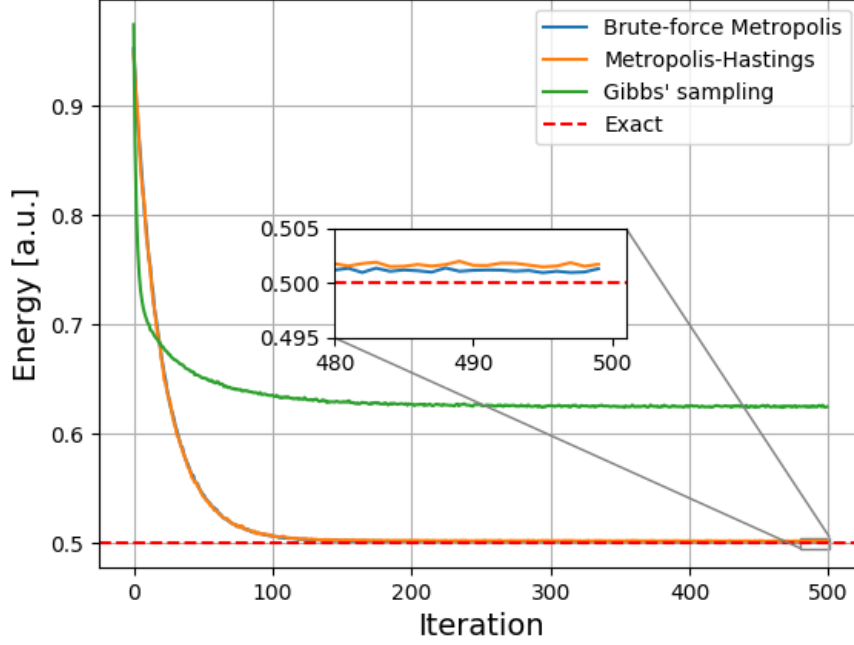


Figure 5: Energy as a function of number of iterations up to 500 iterations for one particle in one dimension. Brute-Force (BF) Metropolis (blue line), Metropolis-Hastings (MH) (orange line), Gibbs' (green line) and analytical value (dashed red line). All the samplings were conducted with  $MC=2^{20} = 1048576$ ,  $\eta = 0.01$ , two hidden nodes,  $\omega = 1.0$  and  $\sigma = 1.0$ . For BF we used steplength  $dx = 1.0$ , while for MH the timestep was  $dt = 0.01$ . See text for further description.

We observe that the energies produced by brute-force and Metropolis-Hastings are very similar, and in fact it is hard to distinguish them. They also give lower energies to the analytical energy compared to Gibbs' sampling. In table 1 some chosen energies from figure 5 are listed, and we observe that brute-force has actually the smallest absolute error for all number of iterations.

Table 1: Energy as a function of number of iterations up to 500 iterations for one particle in one dimension.

Iterations	BF	MH	Gibbs	Exact
100	0.50633(23)	0.50734(13)	0.63524(89)	0.5
200	0.50149(08)	0.50232(36)	0.62690(54)	0.5
300	0.50136(03)	0.50198(11)	0.62562(52)	0.5
400	0.50121(02)	0.50198(07)	0.62453(52)	0.5
500	0.50132(02)	0.50173(06)	0.62471(52)	0.5

### 5.1.1 Playing with learning rate and number of hidden nodes

To find the lowest energy, we experiment with different learning rates and a various number of hidden nodes. We use brute-force Metropolis and in figure 6 one can find energy plotted as functions of number of iterations. Initially the energy is overall higher for more hidden nodes, but we get a closer energy to the exact energy for lower number of hidden nodes.

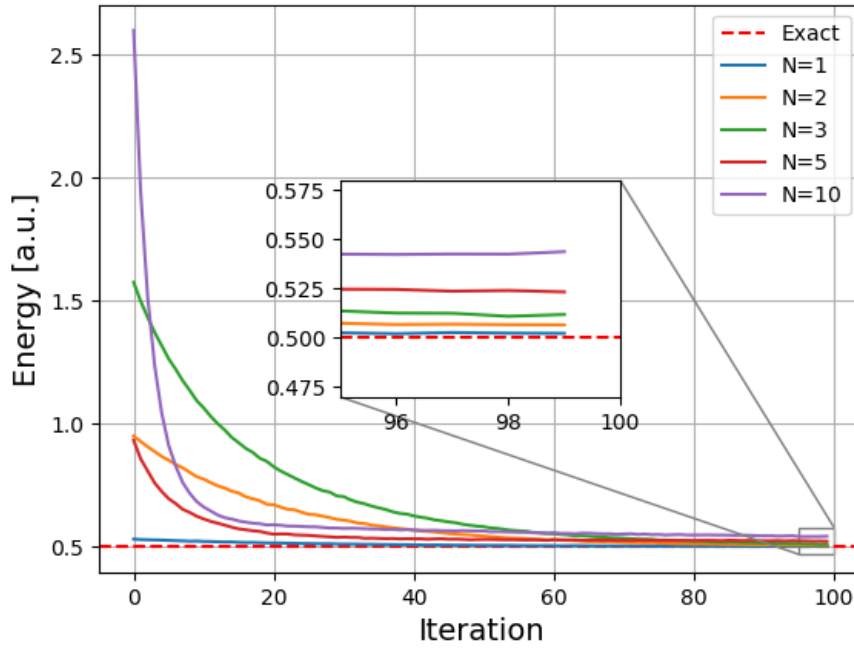


Figure 6: Energy as functions of iterations for  $N=1,2,3,5$  and 10. Brute-Force Metropolis is used,  $MC=5000000$ ,  $\eta = 0.01$  and  $dx = 1.0$ .

Since the optimal number of hidden nodes is one in this case, we will use  $N=1$  while optimizing the learning rate. In figure 7 we compare the energy convergence for various  $\eta$ 's. Initially set all the weights and positions to 0.2, such that we they all start at the same point, and again we use brute-force Metropolis with  $dx = 1.0$ .

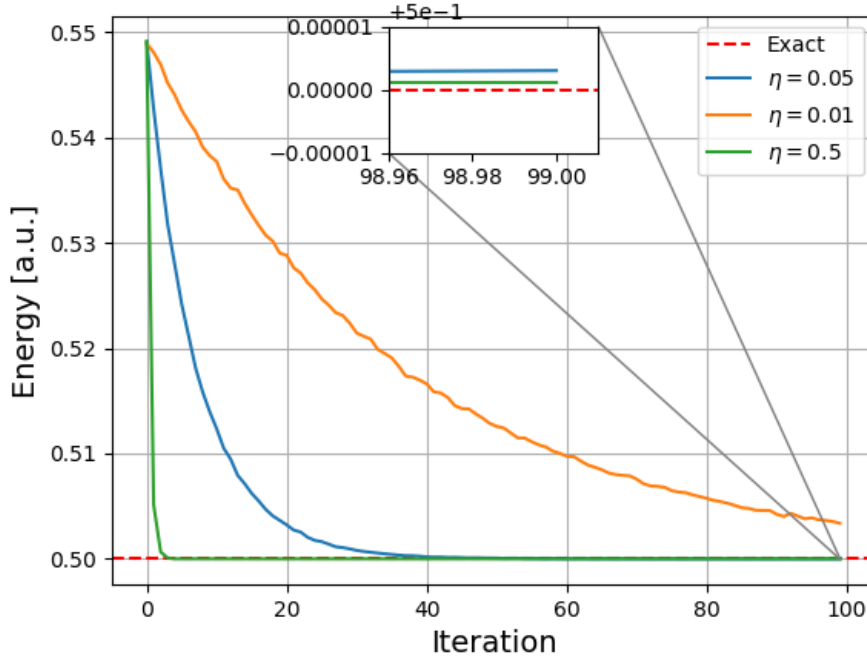


Figure 7: Energy as functions of iterations for  $\eta=1,2,3,5$  and 10. Brute-Force Metropolis is used,  $MC=5e6$ ,  $N=1$  and  $dx = 1.0$ .

The smaller learning rate the slower it converges, and after 100 iterations both  $\eta=0.5$  and  $\eta=0.05$  are really close to the exact value.

## 5.2 Energy calculations, with interaction

We now go further to study the results from two interacting fermions.

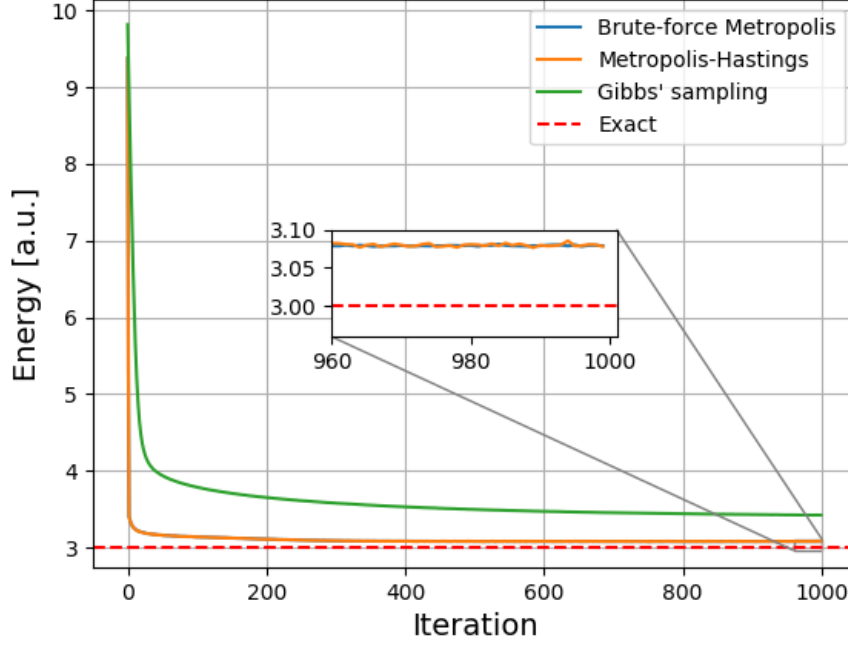


Figure 8: Energy as a function of number of iterations up to 1000 iterations for two particles in two dimensions. Brute-Force Metropolis (blue line), Metropolis-Hastings (orange line), Gibbs' (green line) and analytical value (dashed red line). All the samplings were conducted with  $MC=2^{25} = 33554432$ ,  $\eta = 0.01$ , two hidden nodes,  $\omega = 1.0$  and  $\sigma = 1.0$ . For BF we used steplength  $dx = 1.0$ , while for MH the timestep was  $dt = 0.01$ . See text for further description.

Table 2: Energy as a function of number of iterations up to 500 iterations for one particle in one dimension.

Iterations	BF	MH	Gibbs	Exact
200	3.10897	3.10712	3.65337	3.0
400	3.08195	3.08410	3.52992	3.0
600	3.07979	3.07890	3.4714	3.0
800	3.07895	3.07970	3.44103	3.0
1000	3.07881	3.07791	3.42241	3.0



### 5.2.1 Playing with learning rate and number of hidden nodes

Again we play with the number of nodes and the learning rate

## 5.3 Onebody density

In figure 9 the onebody densities for two particles in two dimensions with and without interaction is plotted. From the energy calculations, we found Metropolis-Hastings method to be the better one for two particles, and henceforth we will therefore stick to it.

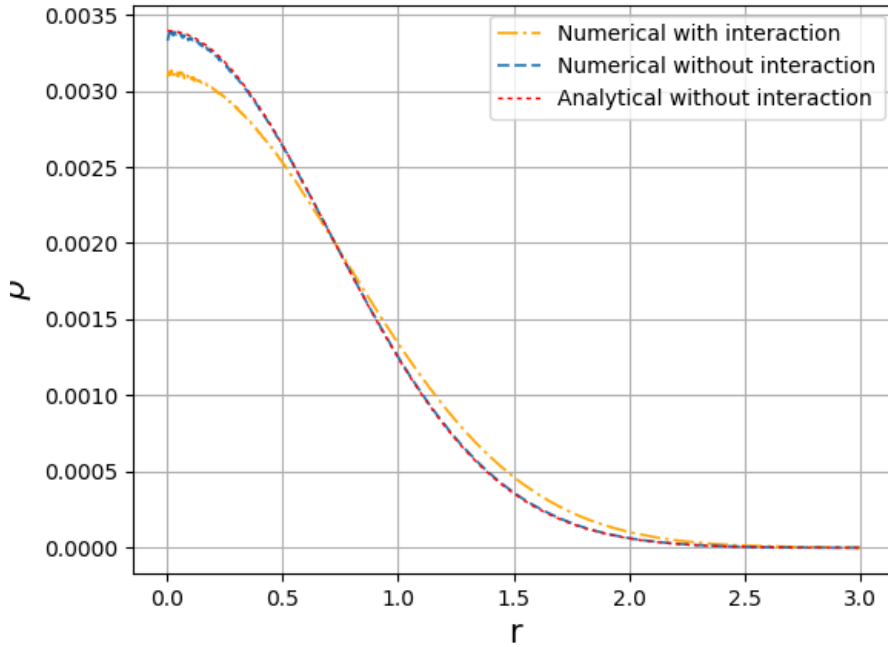


Figure 9: Onebody densities plotted as a function of the distance from origin. The orange (dashed and dotted) represents the numerical calculation of two interacting particles, the blue (dotted) line represents the numerical calculation of two non-interacting particles, while the red (dotted) line represents the analytical calculation of two non-interacting particles. The graphs are normalized to 1 in the interval  $r \in [0, 3]$ , and they are conducted in the way discussed in section 2.4. We used 1000 bins, MC=1e9 Monte-Carlo cycles and the Metropolis-Hastings method.

For the non-interacting case we see that the numerical calculation reproduces the analytical onebody density, but it's slightly noisy for small  $r$ 's.

The onebody density for two interacting fermions is broader and has a lower maximum.

## 5.4 Average energies and distance

In table 3 average kinetic, interaction and potential energy is presented along with the mean distance for both the interacting and non-interacting case and for various frequencies. We observe that both the interaction energy and the potential energy increase as the frequency increase, while the mean distance decreases.

Table 3: In this table we present average kinetic ( $\langle T \rangle$ ), interaction ( $\langle V_{int} \rangle$ ) and external potential energies ( $\langle V_{ext} \rangle$ ) for various  $\omega$ 's, along with the average distance between particles ( $\langle |\mathbf{r}_i - \mathbf{r}_j| \rangle$ ) in the rightmost column. We used the brute-force Metropolis method with  $dx = 1.0$  and found the averages after convergence.

$\omega$	$E_{tot}$ [a.u.]	$\langle T \rangle$ [a.u.]	$\langle V_{ext} \rangle$ [a.u.]	$\langle V_{int} \rangle$ [a.u.]	$\langle  \mathbf{r}_i - \mathbf{r}_j  \rangle$
<b>Without interaction</b>					
0.01	0.86104	0.86104	1.6997e-4	-	1.7483
0.05	0.86911	0.86410	0.0041111	-	1.6904
0.1	0.87778	0.86231	0.015474	-	1.6492
0.5	1.1947	0.91850	0.27619	-	1.3217
1.0	2.0004	0.98093	1.0195	-	1.2685
<b>With interaction</b>					
0.01	1.1663	0.98781	9.4224e-4	0.17756	5.8220
0.05	1.2075	0.99888	0.021559	0.18710	5.5416
0.1	1.2736	0.98968	0.069086	0.21483	4.8911
0.5	2.0287	0.90810	0.56948	0.55025	2.4452
1.0	3.1638	0.92120	1.1673	1.0753	1.4412

## 6 Discussion

Initially we calculated the energy for one particle in one dimension to ensure that the method worked properly. Using equation 9 to calculate the ground state energy, we get that the energy for one particle in one dimension with  $\omega = 1$  should be  $\frac{1}{2}$ . The result of the calculation with the machine learning program with different sampling algorithms can be seen in figure

5. We see that both the brute-force Metropolis and Metropolis-Hasting converges towards the analytical energy. Additionally, from table 1, we see that the variance decreases when the two methods converges towards  $\frac{1}{2}$ , as we increases the number of gradient decent cycles used in the minimalization. This is a good indication that our methods are approaching the correct value, even if we did not have an analytical answer to compare them to.

However, from figure 5, one can see that Gibbs' sampling converges to a higher energy than the analytic value. The reason for that is probably related to how we define the wave function. For brute force and Metropolis-Hastings, we use the probability distribution defined in 3 to represent the wave function, while the wave function used in Gibbs' sampling is the square root of the probability distribution. It seems this representation of the wave function is further away from the true wave function, and thus Gibbs converges towards a higher energy. This is in accordance with the variational principle; a worse guess on trial wave function will yield a upper bound estimate of the ground state energy  $E$  that is larger than the true ground state energy  $E_0$ .

For the case where the Coulomb repulsion is turned on, the results of the convergence of the Metropolis algorithms and the Gibbs sampling method are shown in figure ?? and table 2. Again, as described above, we see that the brute force and Metropolis-Hastings algorithms converges towards the exact value of 3 a.u, while Gibbs sampling approaches a higher energy. The fraction of deviation between the calculated and analytical value is roughly the same for the interacting and non-interacting case for all three methods, which can be seen by comparing table 1 and 2.

We varied the number of hidden nodes  $N$  used in the machine learning program, as seen in figure 6. We see that varying the number of hidden nodes changes the rapidness of the method's convergence. For few hidden nodes, the convergence of the energy seems to happen slower, while for many hidden nodes, it decends steeply towards the analytical energy. **This can maybe be explained by when there are more free parameters that can be varied, the optimalization occurs faster. → Discuss with Vilde.**

We also see that we get better estimates of the ground state energy when using fewer hidden nodes. This is as expected, since more hidden nodes means more variables that has to be updated for each cycle, and thus a less precise answer. **→ Discuss with Vilde**

Furthermore, we observed the effects of changing the learning rate  $\eta$  used in the gradient decent method. The results are shown in figure 7. We see that the convergence rapidness was heavily dependent on the learning rate, where

a low learning rate gave a slow convergence. This is what we expected, as the learning rate defines how much the weights  $W$ ,  $a$  and  $b$  are updated in each gradient decent iteration, that is, how large steps the method takes for each iteration. Larger steps will lead to the method approaching the convergence more rapidly. As seen in this figure, if we limit the number of gradient decent iterations, then the choice of  $\eta = 0.01$  has not converged by the end of 100 minimalization iterations. A good choice of  $\eta$  is thus necessary for the method to work optimally.

The onebody density calculated for two particles in two dimensions with the Metropolis-Hastings method is plotted in figure 9. We are able to reproduce the analytical onebody density without interaction. When interaction is introduced, the distribution of particles becomes broader and lower. This is explained by the two electrons now being subject to a repulsive force which pushed them apart, so that the average distance between the electrons increases. This creates a broader onebody density distribution.

In table 3, we see that the estimates of the average distance between the two electrons  $\langle |\mathbf{r}_i - \mathbf{r}_j| \rangle$  increases with increasing harmonic oscillator frequency  $\omega$ . This is as expected, as the higher  $\omega$ , the steeper the sides of the potential well, and thus the electrons are forced together. We also observe that the average distance for the same  $\omega$  increases when the interaction is turned on, which as explained above, corresponds to the Coulomb repulsion forcing the electrons apart.

As explained in section 2.5, the virial theorem holds for particles in stable systems, and the relation between the time-averaged kinetic and potential energy for potentials that are proportional to  $r^n$  is shown in equation 17. In the non-interacting case, the only potential affecting the particles is the harmonic oscillator potential, which is proportional to  $r^2$ . Thus equation 17 gives that the time-averaged kinetic and potential energies should be equal. From table 3, we see that we are able to reproduce this for two particles without interaction and  $\omega = 1.0$ . However, for different values of  $\omega$ , it seems that the method becomes unstable, and we are no longer able to reproduce the virial theorem. In the interaction case, both the harmonic oscillator potential and the Coulomb interaction contributes to the total potential, and the total potential is thus no longer on the form  $r^n$ . Thus the simple relation between the time-averaged kinetic and potential energies in equation 17 no longer applies.

Another concern regarding table 3 is that the energies in the non-interacting case are not on the form  $2\omega$ , as they should have been, according to equation

12. This is likely due to the same instability described above, which makes us unable to reproduce the virial theorem.

## **7 Conclusion**

## 8 Appendix A - Local energy and quantum force calculations for Metropolis

### 8.1 Local energy

The energy can be calculated by introducing a local energy, where the average local energy goes to the true energy with a sufficient amount of samples, as explained in 2. This local energy can be splitted up in a kinetic part, a part from the harmonic oscillator potential and a interacting part,

$$E_L = \sum_{k=1}^M (E_{\text{KIN},k} + E_{\text{EXT},k}) + E_{\text{POT}}. \quad (41)$$

As discussed in [3], the kinetic part can be expressed as

$$E_{\text{KIN},k} = \frac{1}{\Psi_T} \nabla_k^2 \Psi_T \quad (42)$$

$$= (\nabla_k \ln \Psi_T)^2 + \nabla_k^2 \ln \Psi_T \quad (43)$$

where we have used that

$$\frac{1}{\Psi_T} \nabla_k \Psi_T = \nabla_k \ln \Psi_T. \quad (44)$$

$$\frac{\partial}{\partial X_k} \ln \Psi_T = -\frac{X_k - a_k}{\sigma^2} + \sum_{j=1}^N \frac{W_{kj}}{\sigma^2} \text{Logistic}(v(j)) \quad (45)$$

$$\frac{\partial^2}{\partial X_k^2} \ln \Psi_T = -\frac{1}{\sigma^2} + \sum_{j=1}^N \frac{W_{kj}^2}{\sigma^4} \text{Logistic}^2(v(j)) e^{v(j)} \quad (46)$$

where

$$\text{Logistic}(x) = \frac{1}{1 + e^x} \quad (47)$$

and

$$v(j) = b_j + \sum_{i=1}^M \frac{X_i W_{ij}}{\sigma^2}. \quad (48)$$

Thus the local energy can be expressed as

$$E_L = \sum_{i=1}^N \frac{\mathbf{W}_{*i}^T \mathbf{W}_{*i}}{\sigma^4} \text{Logistic}^2(v(i)) \quad (49)$$

$$+ \sum_{i,j=1}^{N,N} \frac{\mathbf{W}_{*i}^T \mathbf{W}_{*j}}{\sigma^4} \text{Logistic}(v(i)) \text{Logistic}(v(j)) \quad (50)$$

$$- 2 \sum_{i=1}^N \frac{\mathbf{W}_{*i}^T (\mathbf{X} - \mathbf{a})}{\sigma^4} \text{Logistic}(v(i)) \quad (51)$$

$$+ \frac{(\mathbf{X} - \mathbf{a})^T \cdot (\mathbf{X} - \mathbf{a})}{\sigma^4} - \frac{M}{\sigma^2} + \mathbf{X}^T \mathbf{X} + E_{\text{INT}} \quad (52)$$

## 8.2 Quantum force

In equation 8.2, the equation for the quantum force is given, but we repeat it here:

$$\mathbf{F}(\mathbf{R}) = \frac{2\nabla\psi_T(\mathbf{R})}{\psi_T(\mathbf{R})}$$

When implementing importance sampling, we need the expression for the quantum force for the NQS wave function. As shown above,

$$\frac{1}{\Psi_T} \nabla_k \Psi_T = \nabla_k \ln \Psi_T. \quad (53)$$

and

$$\frac{\partial}{\partial X_k} \ln \Psi_T = -\frac{X_k - a_k}{\sigma^2} + \sum_{j=1}^N \frac{W_{kj}}{\sigma^2} \text{Logistic}(v(j)) \quad (54)$$

The expression for the quantum force is thus

$$\mathbf{F}(\mathbf{R}) = -2\frac{X_k - a_k}{\sigma^2} + 2 \sum_{j=1}^N \frac{W_{kj}}{\sigma^2} \text{Logistic}(v(j)) \quad (55)$$

## 9 Appendix B - Gibbs local energy and gradient calculation

### 9.1 Local energy

We need the expression for the local energy with the new trial wave function  $\Psi_T = \sqrt{F(\mathbf{X}, \mathbf{H})}$ . We denote this new wave function  $\Psi^* = \sqrt{\Psi}$ . Earlier, we saw that we had the following relation:

$$\frac{1}{\Psi_T} \nabla_k^2 \Psi_T = (\nabla_k \ln \Psi_T)^2 + \nabla_k^2 \ln \Psi_T \quad (56)$$

Using  $\Psi^* = \sqrt{\Psi}$ , we now get

$$\frac{1}{\Psi^*} \nabla_k^2 \Psi^* = (\nabla_k \ln \Psi^*)^2 + \nabla_k^2 \ln \Psi^* \quad (57)$$

$$= (\nabla_k \ln \Psi^{\frac{1}{2}})^2 + \nabla_k^2 \ln \Psi^{\frac{1}{2}} \quad (58)$$

$$= \frac{1}{4} (\nabla_k \ln \Psi_T)^2 + \frac{1}{2} \nabla_k^2 \ln \Psi_T \quad (59)$$

We see that there are two constants which differ the local energy caculated with  $\Psi$  and  $\Psi^*$ .

### 9.2 Gradient for gradient decent method

When calculating the gradients for  $\mathbf{W}$ ,  $\mathbf{a}$  and  $\mathbf{b}$  when using the wave function  $\Psi^* = \sqrt{\Psi}$ , we must remember that:

$$\nabla \ln \Psi^{\frac{1}{2}} = \frac{1}{2} \nabla \ln \Psi \quad (60)$$

We thus get a factor 2 difference in the gradients that are used to uptade the weights  $\mathbf{W}$ ,  $\mathbf{a}$  and  $\mathbf{b}$ , which we must remember to add when running Gibbs sampling.

## 10 Appendix C - Derivation of the wave function

We have seen that the probability of having a set of positions  $\mathbf{X}$  with a set of hidden nodes  $\mathbf{h}$  is given by

$$F(\mathbf{X}, \mathbf{H}) = \frac{1}{Z} e^{-\beta E(\mathbf{X}, \mathbf{H})} \quad (61)$$



where we set  $\beta = 1/kT = 1$ ,  $Z$  is the partition function and  $E(\mathbf{X}, \mathbf{h})$  is the system energy

$$E(\mathbf{X}, \mathbf{H}) = \sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^N b_j H_j - \sum_{i,j=1}^{M,N} \frac{X_i W_{ij} H_j}{\sigma_i^2} \quad (62)$$

such that

$$F_{\text{RBM}}(\mathbf{X}, \mathbf{h}) = e^{\sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2}} e^{\sum_{j=1}^N \left( b_j h_j + \sum_{i=1}^M \frac{X_i W_{ij} h_j}{\sigma_i^2} \right)}. \quad (63)$$

We skip the partition function because it will not affect the results (it is just a normalization constant). The probability of a set of positions only is therefore the sum over all sets of  $\mathbf{h}$ ,  $\{\mathbf{h}\}$ :

$$\begin{aligned} F_{\text{RBM}}(\mathbf{X}) &= \sum_{\{\mathbf{h}\}} e^{\sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2}} \prod_{j=1}^N e^{b_j h_j + \sum_{i=1}^M \frac{X_i W_{ij} h_j}{\sigma_i^2}} \\ &= \sum_{h_1} \sum_{h_2} \dots \sum_{h_N} e^{\sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2}} \times \\ &\quad e^{b_1 h_1 + \sum_{i=1}^M \frac{X_i W_{i1} h_1}{\sigma_i^2}} e^{b_2 h_2 + \sum_{i=1}^M \frac{X_i W_{i2} h_2}{\sigma_i^2}} \dots e^{b_N h_N + \sum_{i=1}^M \frac{X_i W_{iN} h_N}{\sigma_i^2}} \\ &= e^{\sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2}} \prod_{j=1}^N \sum_{h_j=0}^1 e^{b_j h_j + \sum_{i=1}^M \frac{X_i W_{ij} h_j}{\sigma_i^2}} \\ &= e^{\sum_{i=1}^M \frac{(X_i - a_i)^2}{2\sigma_i^2}} \prod_j^N (1 + e^{b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2}}) \end{aligned} \quad (64)$$

## 11 References

- [1] Morten Hjorth-Jensen. Computational Physics 2: Variational Monte Carlo methods, Lecture Notes Spring 2018. Department of Physics, University of Oslo, (2018).
- [2] S. Marsland, *Machine Learning: An algorithmic Perspective, Second edition* (2015)
- [3] D. Gjestvang, E. M. Nordhagen *Computational Physics II: Project 1* (2018)
- [4] V. Flugsrud and M. Hjort-Jensen Project 2, The restricted Boltzmann machine applied to the quantum many body problem. Deadline June 1 Department of Physics, University of Oslo, (2018).
- [5] M. Taut *Two electrons in an external oscillator potential: Particular analytic solutions of a Coulomb correlation problem*. Phys. Rev. **A 48**, 3561 (1993).
- [6] G.Carleo and M. Troyer *Solving the quantum many-body problem with artificial neural networks* Science 355, Issue 6325, pp. 602-606 (2017)
- [7] A. Chilton *The Properties and Applications of Quantum Dots* (2014) <https://www.azoquantum.com/Article.aspx?ArticleID=31> downloaded May 6th 2018
- [8] G.E. Hinton *A Practical Guide to Training Restricted Boltzmann Machines*, Lecture Notes in Computer Science, vol **7700**. Springer, Berlin, Heidelberg (2012)