Thesis notes

Vilde Flusgrud

January 30, 2018

1 Restricted Boltzmann Machine (RBM)

The joint probability distribution is defined as [5]

$$F_{rbm}(\mathbf{V},\mathbf{H}) = \frac{1}{Z} e^{-\frac{1}{T_0}E(\mathbf{V},\mathbf{H})} \tag{1}$$

where

$$Z = \int\int \frac{1}{Z} e^{-\frac{1}{T_0}E(\mathbf{v},\mathbf{h})}\, \mathrm{d}\mathbf{v}\, \mathrm{d}\mathbf{h} \tag{2}$$

It is common to ignore T_0 by setting it to one.

2 Gaussian-Binary RBM

Here we have [5]

$$E(\mathbf{V},\mathbf{H}) = \sum_i^M \frac{(V_i - a_i)^2}{2\sigma_i^2} - \sum_j^N b_j H_j - \sum_{i,j}^{M,N} \frac{V_i w_{ij} H_j}{\sigma_i^2} \tag{3}$$

If $\sigma_i = \sigma$ then

$$E(\mathbf{V},\mathbf{H}) = \frac{||\mathbf{V}-\mathbf{a}||^2}{2\sigma^2} - \mathbf{b}^T\mathbf{H} - \frac{\mathbf{V}^T\mathbf{W}\mathbf{H}}{\sigma^2} \tag{4}$$

OBS: in the eq above, check the factor two's in the denominator with σ.

2.1 $\psi = \sqrt{F_{rbm}}$ (the mistake)

Fill in more computations in these [5]:

$$P(V_i|\mathbf{h}) = \mathcal{N}(V_i; a_i + \mathbf{w}_{i*}\mathbf{h}, \sigma^2) \tag{5}$$

$$P(\mathbf{V}|\mathbf{h}) = \prod_i^M \mathcal{N}(V_i; a_i + \mathbf{w}_{i*}\mathbf{h}, \sigma^2) \tag{6}$$

$$= \mathcal{N}(\mathbf{V}; \mathbf{a} + \mathbf{W}\mathbf{h}, \sigma^2) \tag{7}$$

$$P(H_j|\mathbf{v}) = \frac{e^{(b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{\sum_{h_j} e^{(b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2})h_j}} \tag{8}$$

$$= \frac{e^{(b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{1 + e^{b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2}}} \tag{9}$$

$$P(\mathbf{H}|\mathbf{v}) = \prod_j^N \frac{e^{(b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{1 + e^{b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2}}} \tag{10}$$

$$\tag{11}$$

Meaning

$$P(H_j = 1|\mathbf{v}) = \frac{e^{b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2}}}{1 + e^{b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2}}} \tag{12}$$

$$= \frac{1}{1 + e^{-b_j - \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2}}} \tag{13}$$

$$P(H_j = 0|\mathbf{v}) = \frac{1}{1 + e^{b_j + \frac{\mathbf{v}^T \mathbf{w}_{*j}}{\sigma^2}}} \tag{14}$$

Some observations (v being coordinate of a particle): Something that will increase the probability of $H_j = 1$ is any of the particles being far away from the origin in positive direction if w_{ij} is positive, or in negative direction (if w_{ij} is negative). If the particle is far in a direction opposite to the sign of w_{ij} it decreases the probability of $H_j = 1$.

(Is this meaningful..? Shouldn't the placing of the coordinate system be arbitrary? Or does it have to do with the potential, what is the relationship between HO potential and location? Well yes it's centered at the origin for ground state. So what is the meaning of the hidden nodes having higher likelihood of being activated for the above mentioned cases..?)

Then, when it's decided which h's are activated and not, we move to $P(V_i|\mathbf{h})$. Here, an active h means that its corresponding weight gets to contribute to the mean of the normal distribution of v_i. The weights w_{*j} gets to contribute to the decision of h_j, while the weights w_{i*} gets to contribute to the decision of v_i. So while the size of the sum $v^T w_{*j}$ contributed to the likelihood of h_j being activated, h_j only brings with it the weight w_{ij} to skew the mean of the distribution of v_i.

What is the meaning of this? What do we want given h_j has been activated? It can be a sign that a lot of coordinates are far from origin in the direction of the weights w_{*j}. Thus, we want to skew the mean of $\mathbf{v}$ by

the weights $\mathbf{w}_{*\mathbf{j}}$ respectively. This will add to or subtract from the bias $\mathbf{a}$ to make up the mean of the distribution of the new positions.

We expect from the HO potential ground state that there should be most positions around the origin.

Also, given this partly documented behaviour, how should we expect the weights w_{i*} to be tuned (and the biases)?

2.2 $\psi = F_{rbm}$ (correct)

Here $P(\mathbf{X})$ refers to the probability distribution we wish to sample from according to quantum mechanics given $\psi(\mathbf{X})$. We have $P(\mathbf{X}, \mathbf{H}) = |F_{rbm}(\mathbf{X}, \mathbf{H})|^2$. And $P(\mathbf{X}) = |\psi(\mathbf{X})|^2 = |\sum_{\mathbf{h}} F_{rbm}(\mathbf{X}, \mathbf{h})|^2$. Question: If $F_{rbm}(\mathbf{X}, \mathbf{H})$ is full on complex etc. Is it guaranteed that finding $P(\mathbf{X}, \mathbf{H})$ and $P(\mathbf{X})$ 'independently', by way of F_{rbm} as expressed above, gives the same result as if I were to find $P(\mathbf{X})$ from $\sum_{\mathbf{h}} P(\mathbf{X}, \mathbf{h})$? Should this be guaranteed? In the first case you first add the complex number, then take the modulus. In the second you first take the modulus, then add real numbers. The latter should possibly miss some information (the inference term). Is this OK? That $P(\mathbf{X}, \mathbf{H})$ and $P(\mathbf{X})$ will not have the usual relation between them?

How to sample from $P(X)$ given $P(X, H)$? When doing Gibbs sampling, according to Wikipedia (should find better source): "The marginal distribution of any subset of variables can be approximated by simply considering the samples for that subset of variables, ignoring the rest." Thus, when we use our samples of $\mathbf{x}$'s to calculate quantum mechanical variables, these samples are approximating the probability distribution $P(X)$ when they have been Gibbs sampled from $P(X, H)$. Let's find P(X, H):

$$P(\mathbf{X}, \mathbf{H}) = |F_{rbm}|^2 \tag{15}$$

$$= \frac{1}{Z^2} e^{-2E(\mathbf{X}, \mathbf{H})} \tag{16}$$

$$= \frac{1}{Z^2} e^{-\frac{|\mathbf{X}-\mathbf{a}|^2}{\sigma^2} + 2\mathbf{b}^T\mathbf{H} + 2\frac{\mathbf{X}^T\mathbf{W}\mathbf{H}}{\sigma^2}} \tag{17}$$

$$\tag{18}$$

Then

2.2.1 Derive $P(\mathbf{X}|\mathbf{h})$

$$P(\mathbf{X}|\mathbf{h}) = \frac{P(\mathbf{X},\mathbf{h})}{\int P(\mathbf{x},\mathbf{h})\,\mathrm{d}\mathbf{x}} \tag{19}$$

$$= \frac{e^{2\mathbf{b}^T\mathbf{h}} \prod_i e^{-\frac{(X_i-a_i)^2}{\sigma^2}+2\frac{X_i\mathbf{w}_{i*}\mathbf{h}}{\sigma^2}}}{e^{2\mathbf{b}^T\mathbf{h}} \int \prod_i e^{-\frac{(x_i-a_i)^2}{\sigma^2}+2\frac{x_i\mathbf{w}_{i*}\mathbf{h}}{\sigma^2}}\,\mathrm{d}\mathbf{x}} \tag{20}$$

$$= \prod_i \frac{e^{-\frac{(X_i-a_i)^2}{\sigma^2}+2\frac{X_i\mathbf{w}_{i*}\mathbf{h}}{\sigma^2}}}{\int e^{-\frac{(x_i-a_i)^2}{\sigma^2}+2\frac{x_i\mathbf{w}_{i*}\mathbf{h}}{\sigma^2}}\,\mathrm{d}x_i} \tag{21}$$

$$\tag{22}$$

Remember that if two random events A and B are independent, their joint probability equals the product of their probabilitites: $P(A,B) = P(A)P(B)$. Since the X_i are independent variables in the restricted Boltzmann machine, it makes sense that we see the following form above: $P(\mathbf{X}|\mathbf{h}) = \prod_i P(X_i|\mathbf{h})$. (correct?)

We now want to rewrite the exponent to get the expression for $P(X_i|\mathbf{h})$ into the form of a Gaussian distribution, which generally has the form

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{23}$$

We write

$$-\frac{(X_i-a_i)^2}{\sigma^2} + 2\frac{X_i\mathbf{w}_{i*}\mathbf{h}}{\sigma^2} \tag{24}$$

$$= \frac{1}{\sigma^2}(-X_i^2 + 2X_ia_i - a_i^2 + 2X_i\mathbf{w}_{i*}\mathbf{h}) \tag{25}$$

$$= \frac{1}{\sigma^2}(-X_i^2 + 2X_ia_i - a_i^2 + 2X_i\mathbf{w}_{i*}\mathbf{h} \tag{26}$$

$$+ (\mathbf{w}_{i*}\mathbf{h})^2 - (\mathbf{w}_{i*}\mathbf{h})^2 + 2(\mathbf{w}_{i*}\mathbf{h})a_i - 2(\mathbf{w}_{i*}\mathbf{h})a_i) \tag{27}$$

$$= \frac{-(X_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2 + (\mathbf{w}_{i*}\mathbf{h})^2 + 2(\mathbf{w}_{i*}\mathbf{h})a_i}{\sigma^2} \tag{28}$$

Inserting this back in, we find

$$P(\mathbf{X}|\mathbf{h}) = \prod_i \frac{e^{\frac{-(X_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2 + (\mathbf{w}_{i*}\mathbf{h})^2 + 2(\mathbf{w}_{i*}\mathbf{h})a_i}{\sigma^2}}}{\int e^{\frac{-(x_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2 + (\mathbf{w}_{i*}\mathbf{h})^2 + 2(\mathbf{w}_{i*}\mathbf{h})a_i}{\sigma^2}} \mathrm{d}x_i} \tag{29}$$

$$= \prod_i \frac{e^{\frac{(\mathbf{w}_{i*}\mathbf{h})^2 + 2(\mathbf{w}_{i*}\mathbf{h})a_i}{\sigma^2}} e^{\frac{-(X_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2}{\sigma^2}}}{e^{\frac{(\mathbf{w}_{i*}\mathbf{h})^2 + 2(\mathbf{w}_{i*}\mathbf{h})a_i}{\sigma^2}} \int e^{\frac{-(x_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2}{\sigma^2}} \mathrm{d}x_i} \tag{30}$$

$$= \prod_i \frac{e^{\frac{-(X_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2}{\sigma^2}}}{\int e^{\frac{-(x_i - a_i - \mathbf{w}_{i*}\mathbf{h})^2}{\sigma^2}} \mathrm{d}x_i} \tag{31}$$

$$= \prod_i \mathcal{N}(X_i; a_i + \mathbf{w}_{i*}\mathbf{h}, \frac{\sigma^2}{2}) \tag{32}$$

$$= \mathcal{N}(\mathbf{X}; \mathbf{a} + \mathbf{W}\mathbf{h}, \frac{\sigma^2}{2}) \tag{33}$$

$$\text{and} \quad P(X_i|\mathbf{h}) = \mathcal{N}(X_i; a_i + \mathbf{w}_{i*}\mathbf{h}, \frac{\sigma^2}{2}) \tag{34}$$

$$\tag{35}$$

2.2.2 Derive $P(\mathbf{H}|\mathbf{x})$

$$P(\mathbf{H}|\mathbf{x}) = \frac{P(\mathbf{x}, \mathbf{H})}{\sum_{\mathbf{h}} P(\mathbf{x}, \mathbf{h})} \tag{36}$$

$$= \frac{e^{-\frac{|\mathbf{X} - \mathbf{a}|^2}{\sigma^2}} \prod_j e^{(2b_j + 2\frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{\sum_{\mathbf{h}} e^{-\frac{|\mathbf{X} - \mathbf{a}|^2}{\sigma^2}} \prod_j e^{(2b_j + 2\frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})h_j}} \tag{37}$$

$$= \prod_j \frac{e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{\sum_{h_j} e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})h_j}} \tag{38}$$

$$= \prod_j \frac{e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{1 + e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})}} \tag{39}$$

$$\text{and} \quad P(H_j|\mathbf{x}) = \frac{e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})H_j}}{1 + e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})}} \tag{40}$$

$$\tag{41}$$

From this we see that the probability for $H_j = 1$ and $H_j = 0$ respectively

is

$$P(H_j = 1|\mathbf{x}) = \frac{1}{1 + e^{-2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})}} \tag{42}$$

$$P(H_j = 0|\mathbf{x}) = \frac{1}{1 + e^{2(b_j + \frac{\mathbf{x}^T \mathbf{w}_{*j}}{\sigma^2})}} \tag{43}$$

3 Quantum Mechanics (How to go from RBM to Ψ)

Find the marginal probability amplitude $F_{rbm}(X)$:

$$F_{rbm}(\mathbf{X}) = \sum_{\mathbf{h}} F_{rbm}(\mathbf{X}, \mathbf{h}) \tag{44}$$

$$= \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{X}, \mathbf{h})} \tag{45}$$

Now, in this case where for the ground state the wave function is positive definite, we set

3.1 $\psi = \sqrt{F_{rbm}}$ (the mistake)

$$|\Psi(\mathbf{X})|^2 = F_{rbm}(\mathbf{X}) \tag{46}$$

$$\Rightarrow \Psi(\mathbf{X}) = \sqrt{F_{rbm}(\mathbf{X})} \tag{47}$$

$$= \frac{1}{\sqrt{Z}} \sqrt{\sum_{\{h_j\}} e^{-E(\mathbf{X},\mathbf{h})}} \tag{48}$$

$$= \frac{1}{\sqrt{Z}} \sqrt{\sum_{\{h_j\}} e^{-\sum_i^M \frac{(X_i-a_i)^2}{2\sigma^2} + \sum_j^N b_j h_j + \sum_{i,j}^{M,N} \frac{X_i w_{ij} h_j}{\sigma^2}}} \tag{49}$$

$$= \frac{1}{\sqrt{Z}} e^{-\sum_i^M \frac{(X_i-a_i)^2}{4\sigma^2}} \sqrt{\sum_{\{h_j\}} \prod_j^N e^{b_j h_j + \sum_i^M \frac{X_i w_{ij} h_j}{\sigma^2}}} \tag{50}$$

$$= \frac{1}{\sqrt{Z}} e^{-\sum_i^M \frac{(X_i-a_i)^2}{4\sigma^2}} \sqrt{\prod_j^N \sum_{h_j} e^{b_j h_j + \sum_i^M \frac{X_i w_{ij} h_j}{\sigma^2}}} \tag{51}$$

$$= \frac{1}{\sqrt{Z}} e^{-\sum_i^M \frac{(X_i-a_i)^2}{4\sigma^2}} \prod_j^N \sqrt{e^0 + e^{b_j + \sum_i^M \frac{X_i w_{ij}}{\sigma^2}}} \tag{52}$$

$$= \frac{1}{\sqrt{Z}} e^{-\sum_i^M \frac{(X_i-a_i)^2}{4\sigma^2}} \prod_j^N \sqrt{1 + e^{b_j + \sum_i^M \frac{X_i w_{ij}}{\sigma^2}}} \tag{53}$$

$$\tag{54}$$

3.2 $\psi = F_{rbm}$ (correct)

$$\Psi(\mathbf{X}) = F_{rbm}(\mathbf{X}) \tag{55}$$

$$= \frac{1}{Z} \sum_{\{h_j\}} e^{-E(\mathbf{X},\mathbf{h})} \tag{56}$$

$$= \frac{1}{Z} \sum_{\{h_j\}} e^{-\sum_i^M \frac{(X_i-a_i)^2}{2\sigma^2} + \sum_j^N b_j h_j + \sum_{i,j}^{M,N} \frac{X_i w_{ij} h_j}{\sigma^2}} \tag{57}$$

$$= \frac{1}{Z} e^{-\sum_i^M \frac{(X_i-a_i)^2}{2\sigma^2}} \prod_j^N (1 + e^{b_j + \sum_i^M \frac{X_i w_{ij}}{\sigma^2}}) \tag{58}$$

$$\tag{59}$$

3.3 Alternative, if $h_j \in \{-1, 1\}$, not $\{0, 1\}$

$$\Psi(\mathbf{X}) = F_{rbm}(\mathbf{X}) \tag{60}$$

$$= \frac{1}{Z} \sum_{\{h_j\}} e^{-\sum_i^M \frac{(X_i - a_i)^2}{2\sigma^2} + \sum_j^N b_j h_j + \sum_{i,j}^{M,N} \frac{X_i w_{ij} h_j}{\sigma^2}} \tag{61}$$

$$= \frac{1}{Z} e^{-\sum_i^M \frac{(X_i - a_i)^2}{2\sigma^2}} \prod_j^N \sum_{h_j} e^{b_j h_j + \sum_i^M \frac{X_i w_{ij} h_j}{\sigma^2}} \tag{62}$$

$$= \frac{1}{Z} e^{-\sum_i^M \frac{(X_i - a_i)^2}{2\sigma^2}} \prod_j^N (e^{-b_j - \sum_i^M \frac{X_i w_{ij}}{\sigma^2}} + e^{b_j + \sum_i^M \frac{X_i w_{ij}}{\sigma^2}}) \tag{63}$$

$$= \frac{1}{Z} e^{-\sum_i^M \frac{(X_i - a_i)^2}{2\sigma^2}} \prod_j^N 2\cosh(b_j + \sum_i^M \frac{X_i w_{ij}}{\sigma^2}) \tag{64}$$

$$= \frac{1}{Z} e^{-\sum_i^M \frac{(X_i - a_i)^2}{2\sigma^2}} \prod_j^N 2\cosh(b_j + \sum_i^M \frac{X_i w_{ij}}{\sigma^2}) \tag{65}$$

$$\tag{66}$$

4 Expressions needed for QMC

We have that the gradient is [3]

$$G_i = \frac{\partial \langle E \rangle}{\partial \alpha_i} = 2(\langle E \frac{1}{\Psi} \frac{\partial \Psi}{\partial \alpha_i} \rangle - \langle E \rangle \langle \frac{1}{\Psi} \frac{\partial \Psi}{\partial \alpha_i} \rangle) \tag{67}$$

where $\alpha_i = a_1, ..., a_M, b_1, ..., b_N, w_{11}, ..., w_{MN}$.

We use that $\frac{1}{\Psi} \frac{\partial \Psi}{\partial \alpha_i} = \frac{\partial \ln \Psi}{\partial \alpha_i}$ and first find

4.1 $\psi = \sqrt{F_{rbm}}$ (the mistake)

$$\ln \Psi(\mathbf{X}) = -\frac{1}{2} \ln Z - \sum_m^M \frac{(X_m - a_m)^2}{4\sigma^2} + \frac{1}{2} \sum_n^N \ln(1 + e^{b_n + \sum_i^M \frac{X_i w_{in}}{\sigma^2}}) \tag{68}$$

Giving

$$\frac{\partial}{\partial a_m} \ln \Psi = \frac{1}{2\sigma^2}(X_m - a_m) \tag{69}$$

$$\frac{\partial}{\partial b_n} \ln \Psi = \frac{1}{2(e^{-b_n - \frac{1}{\sigma^2} \sum_i^M X_i w_{in}} + 1)} \tag{70}$$

$$\frac{\partial}{\partial w_{mn}} \ln \Psi = \frac{X_m}{2\sigma^2(e^{-b_n - \frac{1}{\sigma^2} \sum_i^M X_i w_{in}} + 1)} \tag{71}$$

4.2 $\psi = F_{rbm}$ (correct)

$$\ln \Psi(\mathbf{X}) = -\ln Z - \sum_m^M \frac{(X_m - a_m)^2}{2\sigma^2} + \sum_n^N \ln(1 + e^{b_n + \sum_i^M \frac{X_i w_{in}}{\sigma^2}}) \tag{72}$$

Giving

$$\frac{\partial}{\partial a_m} \ln \Psi = \frac{1}{\sigma^2}(X_m - a_m) \tag{73}$$

$$\frac{\partial}{\partial b_n} \ln \Psi = \frac{1}{e^{-b_n - \frac{1}{\sigma^2}\sum_i^M X_i w_{in}} + 1} \tag{74}$$

$$\frac{\partial}{\partial w_{mn}} \ln \Psi = \frac{X_m}{\sigma^2(e^{-b_n - \frac{1}{\sigma^2}\sum_i^M X_i w_{in}} + 1)} \tag{75}$$

4.3 Local energy

We also need an expression for E, which here represents the local energy, $E = E_L$. It is given

$$E = \frac{1}{\Psi}\hat{\mathbf{H}}\Psi \tag{76}$$

5 Harmonic Oscillator

The Hamiltonian for the Harmonic Oscillator system is [1]

$$\hat{\mathbf{H}} = \sum_p^P (-\frac{1}{2}\nabla_p^2 + \frac{1}{2}\omega^2 r_p^2) + \sum_{p<q} \frac{1}{r_{pq}} \tag{77}$$

where the first summation term represents the standard harmonic oscillator part and the latter the repulsive interaction between two electrons. Natural units ($\hbar = c = e = m_e = 1$) are used, and P is the number of particles.

Thus we get (D being the number of dimensions)

$$E = \frac{1}{\Psi}\mathbf{H}\Psi \tag{78}$$

$$= \frac{1}{\Psi}(\sum_{p}^{P}(-\frac{1}{2}\nabla_p^2 + \frac{1}{2}\omega^2 r_p^2) + \sum_{p<q}\frac{1}{r_{pq}})\Psi \tag{79}$$

$$= -\frac{1}{2}\frac{1}{\Psi}\sum_{p}^{P}\nabla_p^2\Psi + \frac{1}{2}\omega^2\sum_{p}^{P}r_p^2 + \sum_{p<q}\frac{1}{r_{pq}} \tag{80}$$

$$= -\frac{1}{2}\frac{1}{\Psi}\sum_{p}^{P}\sum_{d}^{D}\frac{\partial^2\Psi}{\partial x_{pd}^2} + \frac{1}{2}\omega^2\sum_{p}^{P}r_p^2 + \sum_{p<q}\frac{1}{r_{pq}} \tag{81}$$

$$= \frac{1}{2}\sum_{p}^{P}\sum_{d}^{D}(-(\frac{\partial}{\partial x_{pd}}\ln\Psi)^2 - \frac{\partial^2}{\partial x_{pd}^2}\ln\Psi + \omega^2 x_{pd}^2) + \sum_{p<q}\frac{1}{r_{pq}} \tag{82}$$

$$\tag{83}$$

Now if each visible node in the Boltzmann machine represents one coordinate of one particle, this can be written as

$$E = \frac{1}{2}\sum_{m}^{M}(-(\frac{\partial}{\partial v_m}\ln\Psi)^2 - \frac{\partial^2}{\partial v_m^2}\ln\Psi + \omega^2 v_m^2) + \sum_{p<q}\frac{1}{r_{pq}} \tag{84}$$

Where we have that

5.1 $\psi = \sqrt{F_{rbm}}$ (the mistake)

$$\frac{\partial}{\partial x_m}\ln\Psi = -\frac{1}{2\sigma^2}(x_m - a_m) + \frac{1}{2\sigma^2}\sum_{n}^{N}\frac{w_{mn}}{e^{-b_n - \frac{1}{\sigma^2}\sum_i^M x_i w_{in}} + 1} \tag{85}$$

$$\frac{\partial^2}{\partial x_m^2}\ln\Psi = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n}^{N}\omega_{mn}^2\frac{e^{b_n + \frac{1}{\sigma^2}\sum_i^M x_i w_{in}}}{(e^{b_n + \frac{1}{\sigma^2}\sum_i^M x_i w_{in}} + 1)^2} \tag{86}$$

5.2 $\psi = F_{rbm}$ (correct)

$$\frac{\partial}{\partial x_m}\ln\Psi = -\frac{1}{\sigma^2}(x_m - a_m) + \frac{1}{\sigma^2}\sum_{n}^{N}\frac{w_{mn}}{e^{-b_n - \frac{1}{\sigma^2}\sum_i^M x_i w_{in}} + 1} \tag{87}$$

$$\frac{\partial^2}{\partial x_m^2}\ln\Psi = -\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\sum_{n}^{N}\omega_{mn}^2\frac{e^{b_n + \frac{1}{\sigma^2}\sum_i^M x_i w_{in}}}{(e^{b_n + \frac{1}{\sigma^2}\sum_i^M x_i w_{in}} + 1)^2} \tag{88}$$

5.3 Monte Carlo

Having now the expressions for E and $\frac{1}{\Psi}\frac{\partial \Psi}{\partial \alpha_i}$ needed for the gradient we can use the following formula to obtain their expectation values from a Monte Carlo simulation.

$$\langle U \rangle = \frac{1}{N} \sum_{n}^{N} U_n \tag{89}$$

where N is the number of Monte Carlo samples and U_n is the value calculated at each sampling.

6 Simplest case analytical

In the simplest case with zero hidden nodes ($N = 0$) and one visible node ($M = 1$) (meaning one particle in one dimension, no interaction) we get ($\sigma = 1$ and $\omega = 1$):

$$\Psi = \frac{1}{Z} e^{-\frac{1}{2}(x-a)^2} \tag{90}$$

with normalizing giving

$$1 = \int_{-\infty}^{\infty} |\Psi|^2 \, \mathrm{d}x \tag{91}$$

$$= \frac{1}{Z^2} \int_{-\infty}^{\infty} e^{-(x-a)^2} \, \mathrm{d}x \tag{92}$$

$$= \frac{1}{Z^2} \sqrt{\pi} \tag{93}$$

$$\Rightarrow Z = \sqrt{\sqrt{\pi}} \tag{94}$$

$$\Rightarrow \Psi = \frac{1}{\sqrt{\sqrt{\pi}}} e^{-(x-a)^2} \tag{95}$$

Our goal is to solve $\frac{\partial \langle E_L \rangle}{\partial a} = 0$ for a. We have that

$$\langle E_L \rangle = \int_{-\infty}^{\infty} \Psi^* E_L \Psi \, \mathrm{d}x \tag{96}$$

so we first find

$$E_L = \frac{1}{\Psi} \hat{\mathbf{H}} \Psi \tag{97}$$

$$= \frac{1}{2} \sum_{m}^{M} (-(\frac{\partial}{\partial v_m} \ln \Psi)^2 - \frac{\partial^2}{\partial v_m^2} \ln \Psi + \omega^2 v_m^2) \tag{98}$$

$$= \frac{1}{2} (-(a-x)^2 - (-1) + x^2) \tag{99}$$

$$= \frac{1}{2} (1 + x^2 - (a-x)^2) \tag{100}$$

Giving us the integral

$$\langle E_L \rangle = \int_{-\infty}^{\infty} \Psi^* E_L \Psi \, \mathrm{d}x \tag{101}$$

$$= \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} (1 + x^2 - (a-x)^2) e^{-(a-x)^2} \, \mathrm{d}x \tag{102}$$

$$= \frac{1}{2\sqrt{\pi}} (\sqrt{(}\pi) + \frac{\sqrt{\pi}}{2}(2a^2 + 1) - \frac{\sqrt{\pi}}{2}) \tag{103}$$

$$= \frac{1}{2}(1 + a^2) \tag{104}$$

We thus solve the initial equation to find

$$0 = \frac{\partial \langle E_L \rangle}{\partial \alpha_i} \tag{105}$$

$$= a \tag{106}$$

$$\Rightarrow a = 0 \tag{107}$$

This gives the lowest energy as $E_L = \frac{1}{2}$ as expected.

7 Notes on Bayes' theory

Monte Carlo methods is an approach to Bayesian inference where one obtains a sample from the posterior. Bayesian inference is the process of inductive learning via Baye's rule.

Bayesian learning - some definitions [2]:
Statistical induction = the process of learning about the general characteristics of a population from a subset ofmembers of that population.
θ = parameter which expresses numerical values of population characteristics
y = dataset made up by numerical descriptions of a the subset of the population.
$\mathcal{Y}$ = The sample space = the set of all possible datasets, from which a single y will result.
Θ = The parameter space = the set of possible parameter values
1.) For each numerical value $\theta \in \Theta$, our **prior distribution** $p(\theta)$ describes our belief that θ represents the true population characteristics.
2.) For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, our **sampling model** $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true.
3.) Once we obtain the data y, the last step is to update our beliefs about θ: For each numerical value of $\theta \in \Theta$, our **posterior distribution** $p(\theta|y)$ describes our belief that θ is the true value, having observed dataset y.

Obtain the **posterior distribution** from the **prior distribution** and **sam-**

pling model via **Bayes' rule**:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \tag{108}$$

8 Relating this to RBM/some general probability notes

Probability of intersection of two events is the same as the joint probability: $P(A \cap B) = P(A, B)$. Probability of a continious variable means it's a probability distribution function (PDF).

The product rule gives (where we used that joint probability is symmetrical, $P(A, B) = P(B, A)$):

$$P(A, B) = P(B|A)P(A) \tag{109}$$
$$P(B, A) = P(A|B)P(B) \tag{110}$$
$$\Rightarrow P(B|A)P(A) = P(A|B)P(B) \tag{111}$$
$$\Rightarrow P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{112}$$

Thus we see that we recovered Baye's rule by using the product rule and the symmetry of the joint probability distribution.

Does these rules apply to multivariate distributions like $P(\mathbf{V}, \mathbf{H})$? If so it tells us that when sampling $\mathbf{V}$ using

$$P(\mathbf{V}|\mathbf{H}) = \frac{P(\mathbf{H}|\mathbf{V})P(\mathbf{V})}{P(\mathbf{H})} \tag{113}$$

then, interpreted in terms of Bayes, $P(\mathbf{V})$ is our prior distribution which describes our belief that $\mathbf{V}$ represents the true population charcteristics, $P(\mathbf{H}|\mathbf{V})$ is our sampling model which describes our belief that $\mathbf{H}$ would be the outcome of our study of a population subset if we knew $\mathbf{V}$ to be true, and $P(\mathbf{V}|\mathbf{H})$ is our posterior distribution - our belief that $\mathbf{V}$ is the true value, having observed the data $\mathbf{H}$.

Similarly, when we sample $\mathbf{H}$ from $P(\mathbf{H}|\mathbf{V})$ the same would apply to the formula

$$P(\mathbf{H}|\mathbf{V}) = \frac{P(\mathbf{V}|\mathbf{H})P(\mathbf{H})}{P(\mathbf{V})} \tag{114}$$

Using [4] could be relevant in this section. Maybe a lot more is Stanford lecture notes on Generative Learning algorithms: [6] (in your masters folder).

9 Sampling

As it is, our desired stationary distribution we wish to sample according to is $P(X, H)$. Our proposal distribution is $P(X|H)$. Comparing Metropolis Hastings (left) to our two-step Gibbs sampling (right) of the x variable. Sampling of h will be same just swithing the variables. First Metropolis definitions:
$P(X)$ is our desired stationary distribution from which we want to sample states. For us: $P(x, h)$.
$g(x'|x)$ is the proposal distribution. For us: $P(x|h)$
$A(x'|x)$ is our acceptance distribution, which we get from $P(x'|x) = g(x'|x)A(x'|x)$.

10 Code flow/testing rbm HO

Components that need to work (testing aspects in square brackets):

- Initializations
  - Set parameters
  - Random setting of α_i and $\mathbf{X}$ [get to know cpp random generator system better. then check set ok. maybe implement unit test. seperate the operation into a function to make testing clearer? it can take seed/no seed as argument, and the test case can be with a seed?]
- Minimization cycle
  - Sampling loop
    - Sample new $\mathbf{H}$
    - Sample new $\mathbf{X}$ [For this one and previous: get to know cpp random generator system better. then check set ok. maybe implement unit test. maybe seperate into functions, to make testing with specific parameter/seed etc easier. Check that calc of logisitc function and Gaussian mean works.]
    - If thermalized:
      - Sample expectation values for the gradient. Means computing derivatives of $\ln \Psi$ wrt $\mathbf{X}$ and α_i. [CHECK ALL ASPECTS OF FORMULAS WORK. Means moving into functions, to test simple cases.]
    - Compute expectation values and gradient [Small, but could be moved into a function that can be tested]
    - Update α_i according to minimization algorithm [HOW to test this??? but MUST be tested well. SGD, ASGD, the other one.. resource "Unit tests for stochastic optimization": `https://arxiv.org/pdf/1312.6055.pdf`]

11 Look into

- **Issues with the current system.**
  - The square root problem
    - Should have $\Psi = \sqrt{P(x)}$ (when Ψ positive definite) but $\Psi = P(x)$ gave better result?
    - When testing I corrected the square root as well as fixing a sign issue in the boltzmann energy at once. Test fixing the sign seperately, to see if this at least improves.
    - What is the implication of deriving Ψ from $P(X)$, but sampling from $P(X|H)$?
      - Compare to standard Metropolis Hastings. (without Importance Sampling - only uses Ψ in acceptance. With, maybe uses it in Proposal distribution also?)
    - The weird sum factor. Is this ok? Not 'just' exponentials, like the common trial wf in lecture notes etc?
      - Check fex `https://arxiv.org/pdf/1506.03752.pdf` article which introduces similar term to Jastrow for bosons to account for mid range correlations
    - How to make the parameters complex like [**?**] suggests, while keeping the probability real..?? And how to go from less information to more?
    - Looking in the article [**?**] it seems inconsistent whether they are taking the square root or not? I have left to write out the derivation to see if I'm correct that they haven't taken the sqrt in the final expression.
  - Follow the derivation of the sampling method step by step, compare to traditional Gibbs and Metropolis Hastings. Some questions that have arisen from a brief look is
    - Are they saying $P(h) = \sum_h P(\sigma, h) = 1$? Is this a property of the binary RBM..? (ICPT notes, equation 2.15).
    - In the Wiki Metropolis-Hastings article, they have the transition probability $P(x'|x)$, related to our desired stationary distributin $P(x)$ and the proposal distribution $g(x'|x)$. But in ICPT Notes, they refer to T as the transition probability, but when writing out the expression for A, they have put T in the place of g. Did they mix up the two?
  - Whitening. I don't think I need to do additional standardization of the data (the positions) since I'm sampling it from the assumed distribution. However, to do: Change the implementation so you can run for different values of σ other than 1. With $\sigma = 1$ the

variance in what we sample X from is far greater than the changes in μ offered by the rbm parameters (often at the order 10^{-2}, so maybe overshadowing their effect?

- Could do: check effect of having x and y as variables Ψ, not just x^2 and y^2 as in the normal trial wf. But as seen above there are many more differences at play here so could be hard to seperate out this effect.
- Implement the TESTS as explained earlier in the document
- Analytical case to benchmark simplest number of parameters? What to do, three unknowns, does not make sense to set number of hidden units to zero (especially not in the code, what to sample from then)

- **Results to extract and document from the current program, the HO system.**
  - Document the effect of the number of hidden variables. [**?**] says increasing it should "in principle" improve the results, but this seems uncertain. If it doesn't, it means we have one more parameter that needs to be 'found', which makes the method harder to use.
  - Document effect of different σ values in the RBM.
  - Most obvious: document ability to reach analytical benchmarks. Include plots to show how the algorithm procedes (both the optimization and the sampling at the optimized rbm).
    - possible to interpret what the hidden variables are modeling?
  - All the different options in minimizing.
    - SGD - document effect of η
    - ASGD - document effect of 5 different parameters
    - Stochastic reconfiguration (Sorella - Weak binding be- tween two aromatic rings: Feeling the van der Waals at- traction by quantum Monte Carlo methods.)

- **Changes in the RBM**
  - Use symmetries exhibited by the Hamilitonian to reduce numbers of variational parameters, by formulating the RBM in symmetry conserving fashion? ([**?**] did this with translational invariance)
  - Something to account for relative distance, if the results for this cannot be improved by elements mentioned earlier.
    - Think more about this.
    - Test mean-covariance RBM which lets two visibles connect

        * Multiply with a Jastrow like you would in normal vmc? Does this even make sense..?
    - Rectified linear hidden units (make the hidden units also continuous)
    - Several RBM layers - Deep belief network

- **Other systems**
    - Hydrogen - vanlig å bruke Gaussisk basis
    - Boson system (see FYS4411, 2016, project 1)
    - Important: What should be the benchmarks here? Should I implement code for more common methods to use as benchmark?

- **Other models**
    - Autoencoder
    - CNN

- **Writing**
    - Formulate a table of contents/list of chapters. Fex:
        * Introduction
        * Theory
            · Quantum many body problem, modeled systems, vmc?
            · Machine learning - neural networks - RBM, sampling, learning/minimization
        * Results (alternatively, John Anders: Advanced theory, implementation, results). Display effects documented above. What's important to include here, for a good thesis?

- **Goals for the code?**

- **Other**
    - What are the goals here. Dimension reduction, feature extraction. Detecting interesting patterns that may serve as an input on how to guess wave functions better (how to extract, and translate/use this information?). Use as a prestep in minimizing? Similar to what Alocais did, what was that? Or was this reference about me splitting my minimization into several steps/methods?

References

[1] Morten Hjorh-Jensen. *Computational Physics Lecture Notes*. University of Oslo Department of Physics, 2015.

[2] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer-Science+BusinessMedia, 2009.

[3] Jørgen Høgberget. Quantum monte-carlo studies of generalized many-body systems. Master's thesis, University of Oslo, 2013.

[4] Mark Liberman and Stephen Isard. Joint probability, conditional probability and bayes' theorem. `http://www.ling.upenn.edu/courses/cogs501/Bayes1.html`, 2011. [Online; accessed 07-December-2017].

[5] Jan Melchior, Nan Wang, and Laurenz Wiskott. Gaussian-binary restricted boltzmann machines on modeling natural image statistics. *PLOS ONE*, 2017.

[6] Andrew Yan-Tak Ng. *CS229 Lecture Notes Generative Learning Algorithms*. Stanford University Department of Computer Science, 2017.