1 1-D part

Definition 1.1 (Potts chain)**.** *A Potts chain is $\mathcal{C}$ chain of spins s_i with $i \in \mathbb{Z}$ and each s_i can assume any integer value from 1 to the vocabulary size V*

Definition 1.2 (Windowed Hamiltonian)**.** *A windowed Hamiltonian H_i is an Hamiltonian that acts only on the spins inside his finite window of interaction $\mathcal{W}_i$. Without loss of generality, we are going to assume that the lowest energy state has a value of zero*

Definition 1.3 (Pseudo-Convolutional Hamiltonian)**.** *A pseudo-convolutional Hamiltonian $H = \sum_i H_i$ is an Hamiltonian that can be written as the sum of several windowed Hamiltonians H_i all with the same window width $\mathcal{W}$. For sake of simplicity we are going to assume that there exists an upper bound to the highest energy of every windowed Hamiltonian $E_i^{max} < E^{max}$*

The pseudo-convolutional Hamiltonians, in matrix form, are band matrices, meaning that after a certain distance from the diagonal, all of their elements are equal to zero. An example of a band matrix is the matrix B in equation 1

$$B = \begin{bmatrix} B_{11} & B_{12} & 0 & \cdots & \cdots & 0 \\ B_{21} & B_{22} & B_{23} & \ddots & \ddots & \vdots \\ 0 & B_{32} & B_{33} & B_{34} & \ddots & \vdots \\ \vdots & \ddots & B_{43} & B_{44} & B_{45} & 0 \\ \vdots & \ddots & \ddots & B_{54} & B_{55} & B_{56} \\ 0 & \cdots & \cdots & 0 & B_{65} & B_{66} \end{bmatrix} \tag{1}$$

Definition 1.4 (Stored Pattern)**.** *A stored pattern $\mathcal{P}$ is a particular sequence of spins $(\dots, s_{-1}, s_0, s_1, \dots)$ such that the energy of the pseudo-convolutional Hamiltonian H associated to this pattern is equal to zero. If more than one stored pattern is present, they can be numbered as $\mathcal{P}^n = (\dots, s_{-1}^n, s_0^n, s_1^n, \dots)$.*

Theorem 1.1. *Let H be a pseudo-convolutional Hamiltonian with $N > 1$ stored patterns. At non-zero temperature the system will be unable to converge to single saved pattern*

Proof. Suppose that our Potts chain starts out equal to our first stored pattern $\mathcal{C} = \mathcal{P}^1$. Now we want to know if the formation of a single domain barrier is thermodynamically favorable.

$$\Delta F = \Delta E - T\Delta S < 0 \tag{2}$$

For that to be true, the Free energy of the system must decrease upon the formation of a domain barrier.

Upon the formation of a domain barrier, The windowed Hamiltonians that intersect it will have a non zero, positive energy. Therefore $\Delta E > 0$, however, we know that the energy of each window Hamiltonian is smaller than $E^{\max}$ and no more that $\mathcal{W}-1$ windows can be affected by a domain wall, therefore

$$0 \leq \Delta E \leq (\mathcal{W}-1)E^{\max} \tag{3}$$

At the same time we know that in a sequence long L there can be $L-1$ possible places where a domain wall can appear, and for each of this possible places it can lead to any of the $N-1$ other patterns saved, therefore there are $(L-1)(N-1)$ possible configurations where the system has a single domain wall. This means that the change of the entropy of the system is

$$\Delta S = \log[(N-1)(L-1)] \tag{4}$$

Putting equations 3 and 4 all together we have that

$$\Delta F \leq (\mathcal{W}-1)E^{\max} - T\log[(N-1)(L-1)] \tag{5}$$

In the thermodynamic limit ($L \to \infty$) we have that the right hand side of the equation becomes eventually negative, therefore the domain barriers are inevitable □

Definition 1.5 (Auto-Regressive Model)**.** *During inference time, given some input tokens $\{s_i \,|\, i_{first} \leq i \leq i_{last}\}$ an auto-regressive model M return a V-dimensional $(p_1, \dots, p_V)$ vector with an estimate of the probability for the next token in the sequence to predict $i_{pred} = i_{last}+1$.*

$$M(s_{i_{first}}, \dots, s_{i_{last}}) = (p_1, \dots, p_V) \tag{6}$$

Theorem 1.2. *It is possible to associate pseudo-convolutional Hamiltonian to any auto-regressive model*

Proof. Through Botzmann's equation it's possible to turn a probability distribution of equation 6 to some energy levels

$$p_c = \frac{1}{Z}e^{-\frac{E_c}{T}} \quad \text{with} \quad c = 1 \dots V \tag{7}$$

Without loss of generality, we can assume $T=1$ and set the energy associated with every prediction turns out to be

$$E_c = -\log p_c + \text{const} \quad \text{with} \quad c = 1 \dots V \tag{8}$$

Where we can set the constant in such a way that the lowest energy state has a energy equal to zero.
We can now define a windowed Hamiltonian

$$H_{i_{\text{pred}}}(s_{i_{\text{first}}}, \dots, s_{i_{\text{last}}}, s_{i_{\text{pred}}}) = -\log\left[M(s_{i_{\text{first}}}, \dots, s_{i_{\text{last}}}) \cdot s_{i_{\text{pred}}}\right] + \text{const} \tag{9}$$

And the full pseudo-convolutional Hamiltoninan can now be seen as the sum of all the $H = \sum H_{i_{\text{pred}}}$ of the sequence.

The generation process can now be seen as sampling from the Boltzmann distribution given from

$$p_{\text{sequence}} = \frac{1}{Z} e^{-\frac{1}{T} H(\text{sequence})} \tag{10}$$

□

Corollary 1.2.1. *Autoregressive models with fixed window size are incapable of generating infinite length, coherent output*

Proof. From theorem 1.2 we know that autoregressive models can be modelled by pseudo-convolutional Hamilonians, which we know that from Theorem 1.1 are not able to converge to any single pattern □