1 1D Case

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.

Later on, to highlight links with Natural Language Processing (NLP) we will use interchangeably the word spins and tokens.

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 $E_i^{\max} < E^{\max}$

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 stored 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 $\{s_i \,|\, i_{\text{first}} \leq i \leq i_{\text{last}}\}$ 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 $i_{\text{pred}} = i_{\text{last}} + 1$.

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

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 □

2 2D Case

We have seen before that in 1D case a Windowed Hamiltonian cannot lead us to a coherent phase and how this applies to auto-regressive models. Now we are going to see what happens in 2D, but first we are going to generalize all the definitions given in section 1 in 2D.

Definition 2.1 (Potts grid)**.** A Potts chain is $\mathcal{C}$ chain of spins $s_{i,j}$ with $(i, j) \in \mathbb{Z}^2$ and each $s_{i,j}$ can assume any integer value from 1 to the vocabulary size V.

Definition 2.2 (Windowed Hamiltonian)**.** A windowed Hamiltonian $H_{i,j}$ is an Hamiltonian that acts only on the spins inside his finite window of interaction $\mathcal{W}_{i,j}$. Without loss of generality, we are going to assume that the lowest energy state has a value of zero. Any window shape is allowed as long as it can fit inside a finite size square.

Definition 2.3 (Pseudo-Convolutional Hamiltonian)**.** A pseudo-convolutional Hamiltonian $H = \sum_{i,j} H_{i,j}$ is an Hamiltonian that can be written as the sum of several windowed Hamiltonians H_i all with the same window with size $\mathcal{W}$. For sake of simplicity we are going to assume that there exists an upper bound to the highest energy of every windowed Hamiltonian $E_i^{\max} < E^{\max}$

Theorem 2.1. *Let H be a pseudo-convolutional Hamiltonian acting on a Potts grid with $N > 1$ stored patters. At thermal equilibrium, there exists a critical temperature T_c below which the system will converge to a single stored pattern*

Proof. The following proof will be a generalization of the Peierls argument.
We now start with a $L \times L$ grid of V-dimensional Potts spins with $N > 1$ saved patterns. 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 like in figure 1 is thermodynamically favorable.

We now imagine starting in a state of a large 2D system with the spins on

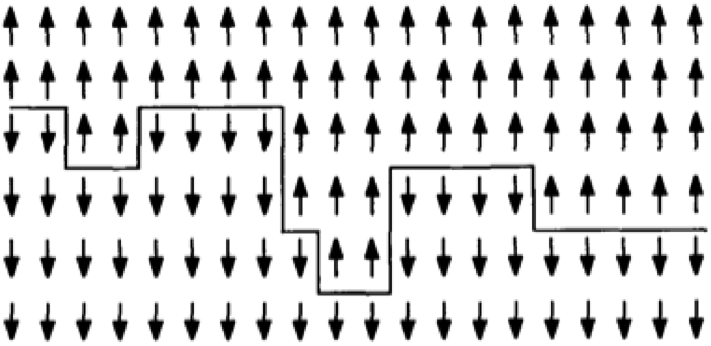

Figure 1: A Domain wall in 2D

the boundary frozen in the pattern $\mathcal{P}^1$ configuration. We again wish to compute the free energy difference of inserting a domain wall at the origin that has a different sign. Now our domain wall boundary consists not just of a pair of points, but of some perimeter of length P. Each spin with its window intersecting the boundary creates an energy penalty of at least $E^{\min}$. The number of such spins is linearly proportional to the perimeter length P and the area of the window of at least equal to 1 $\mathcal{W} \geq 1$ (and at least just one element) so the total change in energy is

$$\Delta E \geq PE^{\min} \tag{11}$$

We can give an upper bound on the number of domain barrier is $(N-1)P^2 3^P$. This is because the domain is a connected component, so the boundary is a self-avoiding curve. The curve has perimeter P, so it must fit inside a box of side-length P centered at the origin. Each step along the perimeter of this curve has at most 3 choices for where to step next (or else it would backtrack and self-intersect). Since the total length is P and there are at most P^2 starting points at which to begin the curve, there are at most $P^2 3^P$ such domain walls. Furthermore any domain wall can appear between the starting pattern $\mathcal{P}^1$ and any other stored pattern, therefore the number of configuration is multiplied by $(N-1)$

$$\Delta S \leq \log(N-1) + 2\log P + P\log 3 \tag{12}$$

Therefore for $P \to \infty$

$$\Delta F \geq PE^{\min} - TP\log 3 \tag{13}$$

This means that for

$$T \leq E^{\min}/\log 3 \tag{14}$$

We have an ordered phase that converges to one of the stored patterns. □

2.1 Some problems with this proof

The problem with this proof is that we assume that the free energy is actually minimized. This is true at thermal equilibrium, otherwise you could extract energy from the system. However some energy landscapes are so rugged that the free energy cannot be completely minimized.
This means that even though the lowest free energy configuration is ordered, the system might never be able to reach it because it can take a very long, or infinite amount of time to reach it. We will tackle this problem in sections 4 and 5

3 More complex Topologies

As you have seen from the two examples, determining whether or not an ordered phase can exists boils down to a counting problem.

1. Start with the system being equal to one of the patterns stored
2. Create a domain wall
3. Estimate the energy gained by the system
4. Count the number of such domain walls
5. See if the free energy increases or decreases as the size of the domain walls goes to infinity

This can be applied to systems with very different topologies, we are now going to explore that

Definition 3.1 (Graph Hamiltonian)**.** Let G be the adjacency matrix of a graph, then a Graph Hamiltoninan H is a Hamiltonian that can be written as

$$H = H * G \tag{15}$$

where the (*) operator represents the element-wise multiplication

Definition 3.2 (Entropy Scaling)**.** Let H be a Graph Hamiltonian, and P be the perimeter length, or surface area of a domain wall, as the perimeter length increases, the number of possible configurations of domain barrier increases, thus increasing the entropy of the system ΔS. We say that the Entropy gained scales as f_S if

$$\Delta S = O(f_S(P))$$

Definition 3.3 (Energy Scaling)**.** Let H be a Graph Hamiltonian, and P be the perimeter length, or surface area of a domain wall, as the perimeter length increases, the the Higher and Lower bound of the energy gained ΔE scale as respectively $O(f_E^{\text{high}}(P))$ and $O(f_E^{\text{low}}(P))$. If $f_E^{\text{high}} = f_E^{\text{low}} \equiv f_E$ we say that the energy gained scales as f_E

$$\Delta E = O(f_E(P))$$

Theorem 3.1. *If $O(f_S) = O(f_E) = O(f)$ there exists a ordered phase*

Proof.

$$\Delta F = \Delta E - T\Delta S = \lim_{P\to\infty} O(f(P)) - TO(f(P)) \tag{16}$$

If we now do $\lim_{T\to 0}$ the term on the right disappears, therefore the creation of a domain wall increases the free energy and therefore a coherent phase is favored □

Theorem 3.2. *Let H be a Graph Hamiltonian with $N > 1$ stored patterns. At thermal equilibrium, the ability to converge to a ordered phase doesn't depend on N*

Proof. The change in entropy due to the creation of a domain barrier can always be written as

$$\Delta S = \log\left[(N-1)N_{\text{barriers}}\right] = \log N_{\text{barriers}} + \log(N-1) \tag{17}$$

Where N_{barriers} is the number of barriers of a certain shape. In the thermodynamic limit, the term proportional to the number of barriers increases, while the one proportional to the number of patterns stored stays constant, therefore can be ignored as it doesn't change the entropy scaling □

Remark. The importance of this last theorem is that, since the number and the shape of stored pattern doesn't affect the thermodynamics of the problem we might as well stick with a system with just 2 ground state equal to all spin ups and all spin downs

Theorem 3.3. *Let $H = \sum_i H_i$, if there exists two energies E_{max}, E_{min} which are the biggest and smallest non-zero energy level of all the windowed Hamiltonians H_i. At thermal equilibrium, the ability to converge to a ordered phase doesn't depend from the energy levels and the window sizes*

Proof. The proof will be similar to the steps done to reach equation 11.
Let $\mathcal{W}$ be the biggest window size, and 1 the smallest window size of any H_i, and let P be the perimeter length of our domain wall. The energy gain by creating such a domain wall is bounded by

$$PE^{\text{min}} \leq \Delta E \leq \mathcal{W}PE^{\text{max}} \tag{18}$$

In both cases we have that

$$E = O(P) \tag{19}$$

□

Remark. The importance of this last theorem is that, since the strength and window size of the interaction don't matter, we might as well consider next-neighbor and constant strength interactions

Theorem 3.4. *It only depends on the topology WRITE PROOF*

As an example on determining whether or not, a coherent phase can exist we focus on the Connected Tree of Spins

Definition 3.4 (Connected Tree of Spins)**.** ADD IMAGES
A connected tree of spins is a tree structure where each node has C children. Every spin interacts with his parent node, his child node and his next-neighbors

Remark. This type of data structure lends itself well to the task of text generation. We can imagine that in the last row is where the characters of the text are located. The nodes in the row above host information related to the collection of characters, effectively acting as word-level embedding. The nodes in the row above host information related to the collection of works, effectively acting as sentence-level embedding, and so on until the tree ends.

Theorem 3.5. *The Ising model on a connected tree does have a condensed phase*

Proof. REWRITE THIS PROOF BETTER AND WITH IMAGES
The Hamiltonian is

$$H = -J \sum_{\langle i,j \rangle} s_i s_j \tag{20}$$

Where $\langle i, j\rangle$ means that is summed over all the couple of nodes connected in the tree.
The energy Δ required to create a perimeter of length P is equal to

$$\Delta E = 2JP$$

Similar to Peierls argument, the number of starting positions is $\approx L^{\log L}/(\log L)!$. For each starting position the number of turn a perimeter can take, in this geometry is either 2 or 3, and thus the number of perimeters starting from one starting position is less than 3^P. Thus the change in entropy is

$$\Delta S \leq P \log 3 + \dots \tag{21}$$

By the end we have that

$$\Delta F \geq 2JP - TP \log 3 \tag{22}$$

This means that for low enough temperature we have an ordered phase □

4 The role of disorder

The problem with the theorems stated so far is that some of this systems can exhibit spin-glass like behavior, this is due to the intrinsic quenched disorder present in machine learning algorithms.

For example lets say our system searches trough a family of Hamiltonians dependent from a set of parameters θ, then the Free energy will depend on theta as such

$$F(\theta) = -T \log \left[\int e^{-H(\{s_i\}|\theta)/T} D s_i \right] \tag{23}$$

However different parameters will yield give different free energies, and since they are learned they follow a distribution $p(\theta|\mathcal{D})$ dependent on the dataset $\mathcal{D}$. A more meaningful variable will be the expected free energy $\langle F\rangle$

$$\langle F\rangle = \int F(\theta)p(\theta|\mathcal{D})d\theta \tag{24}$$

Since the loss function is the negative log likelyhood of the parameters $l(\theta|\mathcal{D}) = -\log p(\theta|\mathcal{D})$ we can combine this with equations 23 and 24 to get

$$F = -T\int e^{-l(\theta|\mathcal{D})}\log\left[\int e^{-H(\{s_i\}|\theta)/T}Ds_i\right]D\theta \tag{25}$$

This systems, more often than not, exhibit glassy behaviors, and as such, must be treated with extra care.

5 Local Hopfield Networks

We are now going to focus on Hopfield networks to study the stability of this systems as a function of the topology and the number of stored patterns.

Definition 5.1 (Hopfield Network)**.** An Hopfield network is a system described by the Hamiltonian

$$H = -\sum_{\mu}^{N} F\left(\sum_{i}^{L} X_i^{\mu}\sigma_i\right) \tag{26}$$

where N is the number of patterns stored and L is the sequence length

Definition 5.2 (Local Hopfield Network)**.** The Hamiltonian of a windowed Hopfield networks is a sum over many Hopfield networks, each of which interacts inside its own window

$$H = -\sum_{j}^{L}\sum_{\mu}^{N} F\left(\sum_{\langle i,j\rangle} X_i^{\mu}\sigma_i\right) \tag{27}$$

Remark. A nice way to imagine local Hopfield network is as a patchwork of several overlapping Hopfield networks ADD IMAGE

Theorem 5.1. *A Local Hopfield network with an energy function that is the sum of several sub-Hopfield networks with window size of W has a storage capacity equal to that of any given sub-network*

Proof.

$$\Delta E = \sum_{\langle j,k\rangle}\sum_{\mu}^{N} F\left(X_k^{\mu}X_k^{\nu} + \sum_{\langle i,j\rangle\neq k} X_i^{\mu}X_i^{\nu}\right) - F\left(-X_k^{\mu}X_k^{\nu} + \sum_{\langle i,j\rangle\neq k} X_i^{\mu}X_i^{\nu}\right) \tag{28}$$

Now we are going to define the average local change in energy.

$$\Delta E_{\text{loc}}(j) \equiv \sum_{\mu}^{N} F\left(X_k^{\mu} X_k^{\nu} + \sum_{\langle i,j\rangle \neq k} X_i^{\mu} X_i^{\nu}\right) - F\left(-X_k^{\mu} X_k^{\nu} + \sum_{\langle i,j\rangle \neq k} X_i^{\mu} X_i^{\nu}\right) \tag{29}$$

for each j we have a sub-Hopfield network, and when averaging the j dependence goes away. this means that

$$\langle \Delta E \rangle = \sum_{\langle j,k\rangle} \langle \Delta E \rangle_{\text{loc}} = W \langle \Delta E \rangle_{\text{loc}} \tag{30}$$

Now we calculate the variances, first the change in energy can be written as

$$\Delta E^2 = \sum_{\langle j_1,k\rangle} \sum_{\langle j_2,k\rangle} \Delta E_{\text{loc}}(j_1) \Delta E_{\text{loc}}(j_2) \tag{31}$$

Now we calculate the average of the term inside the sum.
When we flip a bit in one window, the change in energy in the other window will be close to it.

$$\Delta E_{\text{loc}}(j_2) = \Delta E_{\text{loc}}(j_1) + \delta \tag{32}$$

Where δ is a probability distribution independent from $\Delta E_{\text{loc}}(j_1)$[1]. Since

$$\langle \Delta E_{\text{loc}}(j_1) \rangle = \langle \Delta E_{\text{loc}}(j_2) \rangle$$

we have that $\langle \delta \rangle = 0$ This means that

$$\begin{aligned} \langle \Delta E_{\text{loc}}(j_1) \Delta E_{\text{loc}}(j_2) \rangle &= \left\langle \Delta E_{\text{loc}}^2(j_1) \right\rangle + \langle \Delta E_{\text{loc}}(j_1) \delta \rangle = \\ &= \left\langle \Delta E_{\text{loc}}^2 \right\rangle + \langle \Delta E_{\text{loc}}(j_1) \rangle \langle \delta \rangle = \\ &= \left\langle \Delta E_{\text{loc}}^2 \right\rangle \end{aligned} \tag{33}$$

Where from the first to the second row we have used the fact that δ is independent from $\Delta E_{\text{loc}}(j_1)$, and form the second to the third row we have used the fact that $\langle \delta \rangle = 0$.
Therefore equation 31 becomes

$$\left\langle \Delta E^2 \right\rangle = W^2 \left\langle \Delta E_{\text{loc}}^2 \right\rangle \tag{34}$$

and the variance is

$$\Sigma^2 = W^2 \left(\left\langle \Delta E_{\text{loc}}^2 \right\rangle - \langle \Delta E_{\text{loc}} \rangle^2 \right) = W^2 \Sigma_{\text{loc}} \tag{35}$$

Suppose that the probability distribution is a Gaussian with mean $\langle \Delta E \rangle$ and variance Σ^2. Then, following the line of reasoning done in [1], the probability

[1] How really accurate is this? I'm sure that IF there is a dependence it is going to be VERY small, should we consider it? and how?

of making an error is the probability that after the spin flips, the energy of the system decreases.

$$
\begin{aligned}
P =& \int_{\Delta E}^{\infty} \frac{1}{\sqrt{2\pi\Sigma^2}} e^{-\frac{x^2}{2\Sigma^2}} dx = \\
=& \int_{W\Delta E_{\text{loc}}}^{\infty} \frac{1}{\sqrt{2\pi W^2 \Sigma_{\text{loc}}^2}} e^{-\frac{x^2}{2W^2\Sigma_{\text{loc}}^2}} dx = \\
=& \int_{\Delta E_{\text{loc}}}^{\infty} \frac{1}{\sqrt{2\pi\Sigma_{\text{loc}}^2}} e^{-\frac{z^2}{2\Sigma_{\text{loc}}^2}} dz = \\
=& P_{\text{loc}}
\end{aligned}
\tag{36}
$$

Where in the last passage we defined $z = x/W$.
This means that a Hopfield network with an energy function that is the sum of several overlapping sub-Hopfield networks with window size of W has a storage capacity of any given sub-network □

References

[1] D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," *Advances in neural information processing systems*, vol. 29, 2016.