In this note, we will explore some ways to understand the behavior of some classic examples of Markov chains on the hypercube, namely the lazy random walk and the Glauber dynamics of the Curie–Weiss model (which is a generalization of the former). We will see how to understand these chains in terms of their projections onto characterizing one-dimensional subspaces.

1 Lazy random walk on the hypercube

Consider the following Markov chain on the hypercube $Q = \{-1, +1\}^n$: at each step, a uniformly random coordinate is chosen and resampled from a fair coin flip ($+1$ or -1 with probability $\frac{1}{2}$ each). Let's denote the coordinates of $x \in Q$ by $x(i)$, and the Markov chain by X_t, so that $X_t(i)$ is the ith coordinate of the Markov chain at time t. Last time Zoe showed us how a coupon collector argument shows that the mixing time of this chain is $\leq n \log n + O(n)$. But in fact, the mixing time is equal to $\frac{1}{2} n \log n + O(n)$. Intuitively, this is because we don't need to resample every bit in order to be approximately uniform (which is the stationary distribution of this chain); we can miss $\sqrt{n}$ bits.

However, we cannot use a *Markovian* coupling (like the one considered last time) to prove this. A Markovian coupling of two chains (X_t, Y_t) is a coupling which is also a Markov chain on the product space. It should be relatively straightforward to see that the best Markovian coupling is the one that Zoe presented, and if that coupling misses $\sqrt{n}$ bits then the two chains will be unequal with high probability. In many more complicated situations, it may be the case that we can only think of a Markovian coupling. But in the present setting, we know that the distribution we are aiming for is uniform, and we can also get ahold of the distribution of X_t fairly easily.

The punch line will be the following: the sum of entries of a uniformly random element of Q has fluctuations of order $\sqrt{n}$ around 0. So, if we miss $c\sqrt{n}$ bits, then the distance from uniformity will be high if c is large and low if c is small. The expected amount of time before hitting $n - c\sqrt{n}$ bits is

$$\sum_{i=0}^{n-c\sqrt{n}} \frac{n}{n-i} = n \sum_{j=c\sqrt{n}}^{n} \frac{1}{j} \approx n \int_{c\sqrt{n}}^{n} \frac{1}{x}\, dx = n\big(\log(n) - \log(c\sqrt{n})\big) = \tfrac{1}{2} n \log n - n \log c.$$

Note also that the error term in the approximation of the sum by the integral is $O(n^{-1/2})$, so the error term is accurate. So, at least intuitively, if we wait for time $t = \frac{1}{2} n \log n + sn$, then if s is large and negative the chain will be far from mixed and if s is large and positive, the chain will be close to mixed.

1.1 More careful analysis

Let us examine directly the distribution of the chain after t time steps, started from X_0 being the all $+1$ configuration for example (i.e. $X_0(i) = +1$ for all i). Let τ_i be the first time when bit i is selected to be resampled, and let $I_t = \{i : \tau_i > t\}$. Then, given I_t, the distribution of the state X_t has all $X_t(i)$ for $i \notin I_t$ being independent ± 1 variables, and $X_t(i)$ for $i \in I_t$ being $+1$. By symmetry, given $|I_t|$, the set I_t is a uniformly random subset of $[n] = \{1, \ldots, n\}$ with size $|I_t|$. So another way to think about this is as follows: first sample Y uniformly from Q, then sample a size $|I_t|$ from its distribution and choose a uniformly random subset of size $|I_t|$, and then set $X_t(i) = 1$ for all of those indices and $X_t(j) = Y(j)$ for the other ones.

Therefore, to understand how close this is to the uniform distribution on Q, we need to understand the distribution of $|I_t|$. In the next section, we will calculate moments to see that for any fixed s, with probability at least $1 - \gamma$ we have

$$|I_{\frac{1}{2} n \log n + sn}| = \sqrt{n} \cdot e^{-s} + O_{s,\gamma}(n^{\frac{1}{4}}).$$

Now let's see how this can be used to derive a precise understanding of the mixing of the chain.

Since I_t is uniformly random given its size, the distribution of the set of $+1$ values in X_t is that of a uniformly random subset with size $\text{Binomial}(n - |I_t|, \frac{1}{2}) + |I_t|$. So the sum of entries in X_t has the same

distribution as the following mixture over the possible values of $|I_t|$:

$$S_t = |I_t| + \sum_{i=1}^{n-|I_t|} R_i,$$

where R_i are independent Rademacher ± 1 variables. Again by uniformity of the set of $+1$s, the total variation distance between X_t and a uniform sample from Q is the same as the total variation distance between S_t and $S = \sum_{i=1}^{n} R_i$.

Now we give the idea which can be made rigorous using a local central limit theorem and a more careful handling of the concentration. First, we can replace $|I_{\frac{1}{2}n\log n+sn}|$ by $\sqrt{n}e^{-s}$ since we have concentration. Then we have

$$S_{\frac{1}{2}n\log n+sn} \approx \text{Normal}(\sqrt{n}e^{-s}, n-\sqrt{n}e^{-s}), \qquad S \approx \text{Normal}(0,n).$$

So when s is very positive, the distance between these distributions is small and when s is very negative, the distance is small. It's helpful to consider a picture and recall the definition of total variation distance as related to the L^1 norm between the PDFs/PMFs of the distributions.

In particular, we observe the *cutoff phenomenon* wherein the timescale at which the TV distance between X_t and the stationary distribution decreases from close to 1 to close to 0 is much smaller than the timescale *before* that happens. Here, it takes time $\frac{1}{2}n\log n$ to "start" mixing, and only further time of $O(n)$ to finish mixing after that.

1.2 Moment calculations

First, the expectation:

$$\mathbb{E}[|I_t|] = \sum_{i=1}^{n} \mathbb{P}[\tau_i > t] = n \cdot \mathbb{P}[\tau_1 > t] = n \cdot \left(1-\frac{1}{n}\right)^t.$$

Using the bounds $e^{-x-x^2} \le 1-x \le e^{-x}$ for $|x| \le \frac{1}{2}$, we find that

$$ne^{-\frac{t}{n}-\frac{t}{n^2}} \le \mathbb{E}[|I_t|] \le ne^{-\frac{t}{n}}.$$

Next, let's calculate the second moment:

$$\mathbb{E}\big[|I_t|^2\big] = \sum_{i,j=1}^{n} \mathbb{P}[\tau_i, \tau_j > t] = \mathbb{E}[|I_t|] + n(n-1)\mathbb{P}[\tau_1, \tau_2 > t] \le n\left(1-\frac{1}{n}\right)^t + n^2\left(1-\frac{2}{n}\right)^t.$$

Combining these, we may upper bound the variance as follows:

$$\begin{aligned}
\text{Var}[|I_t|] &\le n\left(1-\frac{1}{n}\right)^t + n^2\left(\left(1-\frac{2}{n}\right)^t - \left(1-\frac{1}{n}\right)^{2t}\right) \\
&\le ne^{-\frac{t}{n}} + n^2\left(e^{-\frac{2t}{n}} - e^{-\frac{2t}{n}-\frac{2t}{n^2}}\right) \\
&\le ne^{-\frac{t}{n}} + n^2 e^{-\frac{2t}{n}}\left(1-e^{-\frac{2t}{n^2}}\right) \\
&\le ne^{-\frac{t}{n}} + n^2 e^{-\frac{2t}{n}}\frac{2t}{n^2} \\
&= ne^{-\frac{t}{n}} + 2te^{-\frac{2t}{n}}.
\end{aligned}$$

Now let's plug in $t = \frac{1}{2}n\log n + sn$ for some fixed s. We find that

$$\mathbb{E}\Big[|I_{\frac{1}{2}n\log n+sn}|\Big] = n\exp\left(-\frac{1}{2}\log n - s\right)\cdot\left(1+O\left(\frac{\log n}{n}\right)\right) = \sqrt{n}\cdot e^{-s}\cdot(1+o(1)),$$

and

$$\text{Var}\Big[|I_{\frac{1}{2}n\log n+sn}|\Big] \le \sqrt{n}\cdot e^{-s} + \log(n)\cdot e^{-2s} = \sqrt{n}\cdot e^{-s}\cdot(1+o(1)).$$

Thus, by Chebyshev's inequality, with probability at least $1-\gamma$ we have

$$|I_{\frac{1}{2}n\log n+sn}| = \sqrt{n}\cdot e^{-s} + O_{s,\gamma}(n^{\frac{1}{4}}).$$

This is what we set out to prove.

1.3 Interpretation in terms of ballistic and diffusive behavior

Another perspective which is helpful to keep in mind is the following: by symmetry, we are essentially waiting for the sum of entries to mix to $S = \sum_{i=1}^{n} R_n$, which is the sum of entries of a uniformly random element of Q. This is like a random walk W_t on $[-n, n]$ with a strong drift towards zero when $|W_t| \gg \sqrt{n}$; the only thing that can prevent it from moving ballistically towards 0 is choosing the same index to update multiple times. So it takes time $\frac{1}{2}n\log n$ to go from n to some (possibly large) constant times $\sqrt{n}$.

However, when $|W_t| \lesssim \sqrt{n}$, the drift is weak and so W_t essentially exhibits diffusive behavior. Thus in order to "fill out" the central hump of the distribution, or just to move the $\Omega(\sqrt{n})$ steps required for this, we need to wait for $\Omega(\sqrt{n}^2 = n)$ time steps. This gives more meaning to the two different parts of the expression

$$t = \tfrac{1}{2}n\log n + sn.$$

2 The Curie–Weiss model

The Curie–Weiss model is a special case of the Ising model, which Cecilia will discuss in a bit more detail next time. It's a probability distribution on the hypercube $Q = \{-1, +1\}^n$ where the probability of x is proportional to

$$\exp\left(\frac{\beta}{n}\sum_{i,j} x(i)x(j)\right).$$

In other words, elements with more agreement among the bits will have higher probability. Later Cecilia will discuss a different version where we restrict the sum over all $\{i, j\}$ to just the edges of a graph on $[n]$, such as a two or three-dimensional square lattice, to recover the standard Ising model (and remove the factor of $\frac{1}{n}$); that way the interactions will essentially only be local, which is more realistic for lattices of atoms in three-dimensional space. But for now, we are summing over everything, which gives the following convenient representation. Define the magnetization as follows:

$$m(x) = \frac{1}{n}\sum_{i=1}^{n} x(i).$$

Then the probability of x is proportional to $\exp\big(\beta\cdot m(x)^2\cdot n\big)$. By symmetry, we will thus also be able to reduce the dynamics of this model to the dynamics of the magnetization.

2.1 Glauber dynamics: high level picture

We'll consider the Glauber dynamics (X_t) where at each step a uniformly random index i is selected to be resampled, and then the distribution of $X_{t+1}(i)$ takes value $+1$ with probability

$$\frac{e^{\beta\cdot(m_i(X_t)+1)^2\cdot n}}{e^{\beta\cdot(m_i(X_t)-1)^2\cdot n} + e^{\beta\cdot(m_i(X_t)+1)^2\cdot n}} = \frac{e^{\beta\cdot 2m_i(X_t)}}{e^{-\beta\cdot 2m_i(X_t)} + e^{\beta\cdot 2m_i(X_t)}} \approx \frac{e^{\beta\cdot 2m(X_t)}}{e^{-\beta\cdot 2m(X_t)} + e^{\beta\cdot 2m(X_t)}},$$

where $m_i(x) = \frac{1}{n}\sum_{j\neq i} x(i) = m(x) + O(n^{-1})$. The reason for the formula on the left-hand side above is so that the Curie–Weiss distribution is the stationary distribution; in fact it is easy to see that this will satisfy the Detailed Balance equation, so the dynamics is reversible with respect to the Curie–Weiss distribution. In any case, by the above formula, the average value of $X_{t+1}(i)$, given that i was chosen, is $\approx \tanh(2\beta m(X_t))$. This means there is a drift of the magnetization towards an attracting fixed point of the map $m \mapsto \tanh(2\beta m)$, and this drift is of order 1 as long as $m(X_t)$ is order-1 separated from the fixed point.

If there are multiple attracting fixed points, then we are in the slow mixing regime since any configuration started near one of the fixed points will stay near that one and not get a chance to go near the other one for an exponential amount of time. Let's consider the case where there is a unique attracting fixed point, which is $m = 0$. This is the case for $\beta \le \frac{1}{2}$, since $\tanh(x)$ has derivative 1 at 0. In this case, like in the lazy random walk on the hypercube, we will again see something like $O(n \log n)$ time before enough indices are resampled that the all $+1$ starting point is transformed to something with $o(1)$ magnetization. The mixing time will thus be of the form

$$cn \log n + f(n),$$

where $f(n)$ is the amount of time it takes to mix after it is $o(1)$.

This function will depend on the parameter β, and will be particularly different for a certain choice of β where the model is at *criticality*. The behavior for $\beta = 0$ is exactly what we saw earlier, and in that case $\tanh(2\beta m) = 0$. one might expect that the behavior is qualitatively the same as long as the graph of $\tanh(2\beta m)$ is not tangent to the graph of the identity m, but it may change otherwise. Indeed, for $\beta < \frac{1}{2}$, we will have $f(n) = \Theta(n)$, and thus observe the cutoff phenomenon. But for $\beta = \frac{1}{2}$ we will get $\Theta(n^{3/2})$ and so in particular *will not* observe cutoff.

2.2 Understanding the distribution near zero magnetization

Let's figure out the distribution of $m(X)$, where X is from the Curie–Weiss distribution. The probability of any particular x with $m(x) = m$ is $e^{\beta m^2 n}$, so let's see how many such states there are. For m at least $\Omega(1)$ away from ± 1, using Stirling's formula, there are

$$\begin{aligned}
\binom{n}{\frac{1+m}{2}n} &= \frac{n!}{(\frac{1+m}{2}n)!(\frac{1-m}{2}n)!} \\
&\sim \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi\frac{1+m}{2}n}\left(\frac{\frac{1+m}{2}n}{e}\right)^{\frac{1+m}{2}n}\sqrt{2\pi\frac{1-m}{2}n}\left(\frac{\frac{1-m}{2}n}{e}\right)^{\frac{1-m}{2}n}} \\
&= \frac{1}{\sqrt{n\pi\frac{1-m^2}{2}}}\exp\left(n\log n - \frac{1+m}{2}n\log\left(\frac{1+m}{2}n\right) - \frac{1-m}{2}n\log\left(\frac{1-m}{2}n\right)\right) \\
&= \frac{1}{\sqrt{n\pi\frac{1-m^2}{2}}}\exp\left(n\left(-\frac{1+m}{2}\log\frac{1+m}{2} - \frac{1-m}{2}\log\frac{1-m}{2}\right)\right).
\end{aligned}$$

We are interested in the behavior near $m = 0$, meaning that the factor outside of the exponential is roughly constant. Let $H(m)$ denote the entropy function inside the exponent, so that the probability of having magnetization $m \approx 0$ is approximately proportional to

$$\exp\big(\big(\beta m^2 + H(m)\big)n\big).$$

Note that the Taylor expansion of $H(m)$ around 0 is

$$H(m) = H(0) - \frac{m^2}{2} - \frac{m^4}{12} - O(m^6).$$

So for $\beta < \frac{1}{2}$ the probability of having magnetization $m \approx 0$ is approximately proportional to

$$\exp\left(-\left(\frac{1}{2} - \beta\right)m^2 n\right),$$

meaning it is approximately Gaussian with variance $\frac{1}{n(\frac{1}{2}-\beta)}$. In other words, the *sum of entries* has variance approximately $\frac{n}{\frac{1}{2}-\beta}$, or fluctuations of order $\sqrt{n}$. This means that similar reasoning to the $\beta = 0$ case goes through, and we find that the distribution of the Glauber dynamics X_t after $t = \frac{1}{2}n\log n + sn$ steps goes from very far to very close the Curie–Weiss model as s goes from very negative to very positive.

On the other hand, if $\beta = 0$ then the first term in the Taylor expansion cancels, and we are left with a distribution of the form

$$\exp\left(-\frac{m^4}{12}n\right).$$

This is a non-Gaussian distribution, but the same general idea still applies: we need to find the width of the section with roughly constant value of the above density, which is where we observe diffusive behavior.

The above density is roughly constant as long as $m^4 \lesssim n^{-1}$, or in other words as long as $m \lesssim n^{-\frac{1}{4}}$. This means that the distribution is mostly spread across configurations with total sum of entries $\lesssim n^{\frac{3}{4}}$. Therefore the dynamics of $m(X)$ will exhibit quasi-ballistic behavior for time $\frac{1}{4}n\log n$ and then diffusive behavior afterwards. It will then need to go a distance of $n^{\frac{3}{4}}$ just via diffusion, which will take $\Theta(n^{\frac{3}{2}})$ time. Since $n^{\frac{3}{2}} \gg n\log n$, the mixing time here is of order $n^{\frac{3}{2}}$. Moreover, we *do not* observe the cutoff phenomenon here.

For a bit of intuition about why the cutoff phenomenon does not hold here, it is useful to consider the random walk on a cycle or a line: here, the chain mixes in the same amount of time as it takes to start mixing, since the distributions are sort of overlapping at the right scale roughly the whole time. Perhaps one reason why cutoff should occur is because of this decomposition into two different time regimes, with the "warm-up" regime being of a larger order than the "diffusion mixing" regime. But I think this is not very well understood in general.