Introduction

The Ising model is a mathematical model of a magnet, where at each site (vertex) of a finite graph $G = (V, E)$ (with n vertices, usually taken to be a lattice), there is a "spin" which is either $+1$ or -1. The energy of a spin configuration $\sigma \in \Omega = \{\pm 1\}^V$ is given by

$$H(\sigma) = -\sum_{x \sim y} \sigma_x \sigma_y.$$

This means that configurations where spins disagree across an edge have higher energy.

The Gibbs measure, which gives the probability of a spin configuration, is given by

$$\mathbb{P}[\sigma] \propto \exp(-\beta H(\sigma)),$$

where $\beta \in [0, \infty]$ is a parameter called the inverse temperature. Under the Gibbs measure, configurations with lower energy (i.e. with more aligned spins) have higher probability. The parameter β controls the importance of the energy. For $\beta = 0$, the distribution is uniform (i.e. energy is irrelevant), while for $\beta = \infty$, the measure is concentrated on just two states, the "all-plus" and "all-minus" states. For other values of β, the Gibbs measure interpolates between these two behaviors.

However, for certain families of graphs such as growing portions of a lattice, this interpolation is not smooth in the large-graph limit. There is a critical inverse temperature β_c where a phase transition occurs; for $\beta < \beta_c$, the Gibbs measure behaves much like the uniform case, with nearly independent spins and no long-range order, while for $\beta > \beta_c$, the Gibbs measure is concentrated near the all-plus and all-minus states, and there is long range order (i.e. distant spins are positively correlated). At the critical point $\beta = \beta_c$, a rich self-similar fractal structure emerges, and there is polynomial decay of correlations (as opposed to exponential decay in high temperature or no decay in low temperature).

One way to (approximately) sample from the Gibbs measure is to use Markov-chain Monte-Carlo. A natural Markov chain on the space of spin configurations is the Glauber dynamics, where each site has an independent Poisson clock. When a site's clock rings, the spin at that site is resampled according to the conditional distribution of the spin at that site, given the spins in the rest of the graph. Since the spins only interact across edges, only the spins of the neighbors can affect this conditional distribution. Specifically, since there is a lot of cancellation, we resample according to the rule

$$\mathbb{P}[\sigma_x = \pm 1] = \frac{\exp(\pm\beta \sum_{y \sim x} \sigma_y)}{\exp(\beta \sum_{y \sim x} \sigma_y) + \exp(-\beta \sum_{y \sim x} \sigma_y)}. \tag{$*$}$$

At low temperatures, this chain mixes exponentially slowly, since it will get stuck near one of the two low-energy states; getting to the other low-energy states requires $\approx n$ low-probability events (flipping to the "wrong" spin) to happen in quick succession, which has exponentially small probability (in $n = |V|$). However, at high temperatures the chain mixes quickly (in time $O(\log n)$). In this regime, as proved by Lubetzky and Sly, the chain also exhibits the "cutoff" phenomenon, which means that there is a specific time $t_{\mathfrak{m}}$ such that before $t_{\mathfrak{m}}$ the chain is far from stationarity, whereas aftter $t_{\mathfrak{m}}$ the chain is close to stationarity. Our objective is to understand this result.

Definitions of Mixing and Cutoff

We will use the total variation distance between two probability measures on Ω:

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{\sigma \in \Omega} |\mu(\sigma) - \nu(\sigma)| = \sup_{A \subset \Omega} |\mu(A) - \nu(A)| = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \mathbb{P}[X \neq Y].$$

For a Markov chain (X_t) on Ω with stationary distribution π, define

$$d(t) = \sup_{\sigma \in \Omega} \|P^t(\sigma, \cdot) - \pi\|_{\mathrm{TV}},$$

where P is the transition matrix of the Markov chain, and

$$t_{\mathrm{mix}}(\varepsilon) = \inf\{t : d(t) \le \varepsilon\}.$$

Typically, as in our case, there is an implicit parameter n (e.g. the size of the graph). As $n \to \infty$, typically $t_{\mathrm{mix}}(\varepsilon) \to \infty$ for each ϵ as well. We say that there is "cutoff" if $t_{\mathrm{mix}}(\varepsilon)$ is not very dependent on ε, i.e. if $t_{\mathrm{mix}}(\varepsilon) = (1 + o(1))t_{\mathrm{mix}}(\varepsilon')$ for any $\varepsilon, \varepsilon' \in (0, 1)$. Intuitively, this means that all the mixing happens "at once" at a certain time. This sounds strange, but in fact the cutoff phenomenon is quite common.

Note that to prove an upper bound on the mixing time, we need a lower bound on $d(t)$, which corresponds to finding an event A for which $P^t(\sigma, A)$ and $\pi(A)$ are very different. To prove a lower bound on the mixing time, we need an upper bound on $d(t)$, which corresponds to finding a coupling $X \sim P^t(\sigma, \cdot)$ and $Y \sim \pi$ for which the probability $\mathbb{P}[X \neq Y]$ is small.

Case Study: $\beta = 0$

If $\beta = 0$ (infinite temperature), then $(*)$ is $\frac{1}{2}$, i.e. the spin is resampled as an independent coin flip. The stationary distribution of this chain is uniform. Let's quickly analyze the mixing of this chain. For now, let's think about time as discrete instead of continuous, so that at each time step one bit is chosen to be resampled. This can be thought of as a $\frac{1}{2}$-lazy random walk on the Boolean hypercube $\{\pm 1\}^V$ (forgetting about the edges of G).

First, notice that once all bits are selected and resampled, the state of the chain is exactly uniform, and so the chain has mixed (and $d(t) = 0$). By coupon-collecting, this happens at time $t \approx n \log n$. However, this is more than what we need. To understand why, recall that under the uniform distribution on the hypercube, which can be thought of as flipping n independent coins, there is a natural spread of $\sqrt{n}$ (the variance of the number of heads is n). So, intuitively, we only need to resample $n - \sqrt{n}$ of the bits in order to be close to the uniform distribution (as long as the bits that we miss are random!). By the same calculation that gave us coupon-collecting, this takes time

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

Indeed, this chain has cutoff at time $t = \frac{1}{2} n \log n$, with cutoff window $O(n)$. If we translate this back to continuous time, since each of the n sites acts in parallel, we get cutoff at time $t = \frac{1}{2} \log n$, with cutoff window $O(1)$. Let's see how to actually prove this cutoff (still working in discrete time).

First, notice that for any set $J \subseteq V$, the function

$$f_J(\sigma) = \prod_{x \in J} \sigma_x$$

is an eigenvector for the transition matrix of this chain. Indeed, we have

$$P(\sigma, \tau) = \begin{cases} \frac{1}{2} & \text{if } \sigma = \tau, \\ \frac{1}{2n} & \text{if } \sigma \text{ differs from } \tau \text{ at exactly one site,} \\ 0 & \text{otherwise.} \end{cases}$$

In the second case, if the location where σ and τ differ is in J, then $f_J(\tau) = -f_J(\sigma)$, and if it is not in J then $f_J(\tau) = f_J(\sigma)$. Thus

$$\begin{aligned} \sum_{\tau} P(\sigma, \tau) f_J(\tau) &= \frac{1}{2} f_J(\sigma) - \frac{|J|}{2n} f_J(\sigma) + \frac{n - |J|}{2n} f_J(\sigma) \\ &= \frac{n - |J|}{n} f_J(\sigma), \end{aligned}$$

so the eigenvalue is $1 - \frac{|J|}{n}$. These eigenvectors are orthonormal: if $x \in J_1 \setminus J_2$, then

$$\mathbb{E}_\pi[f_{J_1}(\sigma) f_{J_2}(\sigma)] = \mathbb{E}_\pi[\sigma_x] \cdot \mathbb{E}_\pi[f_{J_1 \setminus \{x\}}(\sigma) f_{J_2}(\sigma)] = 0,$$

and $\mathbb{E}_\pi[f_J(\sigma)^2] = \mathbb{E}_\pi[1] = 1$.

Now notice that

$$\|P^t(\sigma, \cdot) - \pi\|_{\mathrm{TV}} = \frac{1}{2} \left\| \frac{P^t(\sigma, \cdot)}{\pi(\cdot)} - 1 \right\|_{L^1(\pi)} \leq \frac{1}{2} \left\| \frac{P^t(\sigma, \cdot)}{\pi(\cdot)} - 1 \right\|_{L^2(\pi)},$$

so we can use this spectral information to bound $d(t)$ from above. Specifically, since

$$\frac{P(\sigma, \tau)}{\pi(\tau)} = \sum_{J \subseteq V} f_J(\sigma) f_J(\tau) \left(1 - \frac{|J|}{n}\right),$$

we have

$$\begin{aligned}
\left\| \frac{P^t(\sigma, \cdot)}{\pi(\cdot)} - 1 \right\|_{L^2(\pi)}^2 &= \left\| \sum_{\varnothing \neq J \subseteq V} f_J(\sigma) f_J(\cdot) \left(1 - \frac{|J|}{n}\right)^t \right\|_{L^2(\pi)}^2 \\
&= \sum_{\varnothing \neq J \subseteq V} f_J(\sigma)^2 \left(1 - \frac{|J|}{n}\right)^{2t} \\
&= \sum_{j=1}^{n} \left(1 - \frac{j}{n}\right)^{2t} \binom{n}{j} \\
&\leq \sum_{k=1}^{n} e^{-2jt/n} \binom{n}{j} \\
&= (1 + e^{-2t/n})^n - 1.
\end{aligned}$$

In other words,

$$4d(t) \leq (1 + e^{-2t/n})^n - 1.$$

Plugging in $t = \frac{1}{2} n \log n + cn$, we get

$$\begin{aligned}
4d(t) &\leq \left(1 + \frac{1}{n} e^{-2c}\right)^n - 1 \\
&\leq e^{e^{-2c}} - 1 \\
&\leq 2e^{-2c},
\end{aligned}$$

with the last inequality holding as long as $c > 1$. Thus, as $c \to \infty$, we have $d(\frac{1}{2} n \log n + cn) \to 0$.

Now for the lower bound on $d(t)$. Notice that the Hamming weight $W(\sigma)$ (number of +1s, say) of a spin configuration $\sigma \sim \pi$ is distribued as a Binomial$(n, \frac{1}{2})$ random variable, so it has mean $\frac{n}{2}$ and variance $O(n)$. On the other hand, suppose σ_t^+ is the state of the Glauber dynamics at time t, started from the all-plus configuration, and let R_t denote the number of untouched spins. Then

$$\mathbb{E}[W(\sigma_t^+) | R_t] = R_t + \frac{n - R_t}{2} = \frac{1}{2}(n + R_t).$$

Now notice that by writing $R_t = \sum_{x \in V} \mathbf{1}_{\{x \text{ is untouched at time } t\}}$, we get

$$\mathbb{E}[R_t] = n \left(1 - \frac{1}{n}\right)^t,$$

and so by the tower property of expectation we obtain

$$\mathbb{E}[W(\sigma_t^+)] = \frac{n}{2} \left[1 + \left(1 - \frac{1}{n}\right)^t\right].$$

We also have $\mathrm{Var}(W(\sigma_t^+)) = O(n)$, and so Chebyshev's theorem implies that $d(t)$ is close to 1 as long as $\left(1-\frac{1}{n}\right)^t \geq \frac{C}{\sqrt{n}}$ for some large constant C (i.e. $d(t) \to 1$ as $C \to \infty$).

Now notice that $\left(1-\frac{1}{n}\right)^t \geq \frac{1}{2}e^{-t/n}$ for $t \leq n^2$ and $n > 1$, so if $t = \frac{1}{2}n\log n - cn$, we get

$$\left(1-\frac{1}{n}\right)^t \geq \frac{1}{2}e^{-\frac{1}{2}\log n + c} = \frac{e^c}{2}\frac{1}{\sqrt{n}}.$$

So, as $c \to \infty$ we have $d(\frac{1}{2}n\log n - cn) \to 1$.

Main Result and Proof Idea

As we have seen, there is cutoff in the infinite-temperature case, at the time when the number of unresampled spins is about $\sqrt{n}$. The main theorem is a manifestation of the intuition that the high-temperature regime is similar to the infinite temperature case.

Instead of considering the number of untouched spins/Hamming weight, we will consider the expected magnetization at a vertex,

$$\mathfrak{m}_t = \mathbb{E}[\sigma_t^+(v)],$$

where σ_t^+ is the Glauber dynamics started from the all-plus state. In the $\beta = 0$ case, the number of untouched spins being roughly $\sqrt{n}$ corresponds to having a magnetization of roughly $\frac{1}{2\sqrt{n}}$. The constant factor 2 is not too important, so we will define

$$t_{\mathfrak{m}} = \inf\left\{t : \mathfrak{m}_t < \frac{1}{\sqrt{n}}\right\}.$$

This is the location where the high-temperature Glauber dynamics experiences cutoff. By the way, here and in the rest of the talk we switch back to continuous time.

Theorem 1. *For any $d > 1$, there exists $\beta_0 = \beta_0(d)$ such that whenever G is a d-regular transitive graph and $0 \leq \beta < \beta_0$, the continuous-time Glauber dynamics for the Ising model on G satisfies $t_{mix}(\epsilon) = t_{\mathfrak{m}} + O_\epsilon(1)$.*

The lower bound on the mixing time follows from a very similar argument to the one for the $\beta = 0$ case. Using the magnetization as the distinguishing statistic instead of the Hamming weight, we can apply Chebyshev's inequality to see that at time $t = t_{\mathfrak{m}} - s$ the chain is far from mixed: $d(t_{\mathfrak{m}} - s) \to 1$ as $s \to \infty$.

The fact which roughly corresponds to the calculation that we did above with the Hamming weights and untouched spins is the submultiplicativity of the expected magnetization. Specifically, we have

$$e^{-s}\mathfrak{m}_t \leq \mathfrak{m}_{t+s} \leq e^{-(1-\beta d)s}\mathfrak{m}_t. \tag{$**$}$$

This is used to show that $\mathfrak{m}_{t_{\mathfrak{m}}-s}$ is larger than $\frac{e^s}{\sqrt{n}}$, which allows Chebyshev's inequality to be used when s is large, since the magnetization at the stationary distribution has mean 0 and variance of order $\frac{1}{n}$, and the magnetization of σ_t^+ also has variance of order $\frac{1}{n}$. These variance bounds probably follow from a comparison with the $\beta = 0$ case, but I'm not sure of the details.

I won't prove the submultiplicativity of the expected magnetization, but we will use it again in the proof of the upper bound on the mixing time, which will be our focus for the rest of the talk. That is to say, we want to show that $d(t_{\mathfrak{m}} + s) \to 0$ as $s \to \infty$. This is where information percolation enters.

The idea of information percolation is as follows. We can couple the Glauber dynamics starting at all initial spin configurations together by using the same Poisson clocks on the vertices, and by coupling the distributions of the resampled spins together in a monotone way (perhaps by using the same uniform random variables to make the decisions at each step). Then, given the randomness used to run the Glauber dynamics, we can look backwards from time $t_{\mathfrak{m}} + s$ to determine how much dependence there is on the initial configuration.

Each vertex in G will look backwards to the last time its Poisson clock rang, and use the uniform random variable to decide what the spin should be. This may require examining the neighboring spins at that time in order to make the decision, but the key insight is that when the temperature is high enough, $(*)$ is close to $\frac{1}{2}$, and so we can couple the dynamics in such a way that there is a good chance that a vertex will not

have to even examine its neighbors to decide what spin to take; it will just be a uniform random spin. This is called an "oblivious" update.

We will construct a network of dependence in the space-time slab in this way. The point is that when temperature is high, the top layer of this network may not be connected to the bottom layer since there are many oblivious updates which essentially sever the connection between the top and the bottom. If the entire top of the network is disconnected from the entire bottom (the initial configuration), then the top is independent of the bottom and hence it must be in the stationary distribution, since it is coupled perfectly with a sample started from the stationary distribution.

As in the $\beta = 0$ case, however, this is more than we need for mixing. As long as the connections between the top and bottom are sparse enough, the defects from stationarity will be small enough as to be indistinguishable from standard fluctuations of the stationary distribution. This is the idea that we will pursue in the rest of the talk.

Information Percolation Proof Sketch

Fix a time $t_{\mathrm{m}} + s$ at which we will analyze the clusters described above. The randomness of the coupled Glauber dynamics can be captured by a sequence of pairs (T_x^i, U_x^i) at each vertex x, where T_x^i is the time at which x's Poisson clock rings for the ith time, and U_x^i is a uniform random variable. The result of the resample at x at time T_x^i will be -1 exactly when

$$U_x^i < \frac{\exp(-\beta\Sigma)}{\exp(\beta\Sigma) + \exp(\beta\Sigma)} = \frac{1}{2}(1 + \tanh(\beta\Sigma)),$$

and it will be $+1$ otherwise, where Σ denotes the sums of spins at neighbors of x. Notice that $|\Sigma| \le d$, so

$$\frac{1}{2}(1 + \tanh(b\Sigma)) \ge \frac{1}{2}(1 - \tanh(\beta d)).$$

So, if $U_x^i < \frac{1}{2}(1-\tanh(\beta d))$, then we will always get a -1 upon resampling, no matter what the configuration at the neighbors of x. Similarly, if $U_x^i > \frac{1}{2}(1 + \tanh(\beta d))$, then we will always get a $+1$. Such updates are called oblivious, and the above calculation shows that we have probability $1 - \tanh(\beta d)$ for any particular update to be an oblivious update.

This gives a branching process in the space-time slab $V \times [0, t_{\mathrm{m}} + s]$. To get $\sigma_x^{t_{\mathrm{m}}+s}$, we look backwards in time, starting from $(x, t_{\mathrm{m}} + s)$, until an update occurs at time T_x^i. If the update is not oblivious, we need to use information about the neighbors (and not just U_x^i) to determine the spin $\sigma_x^{T_x^i}$. In this case, the process branches and we add a spatial edge $e \times \{T_x^i\}$ for each edge e incident to x, and then continue the process backwards in time from each of the neighboring vertices. If the update is oblivious however, we terminate the process then and there. The only way there could be further temporal edges down the line is if another vertex branches back and uses information about x.

This produces a graph in the space-time slab, which we denote by $\mathscr{H}_x$ ($\mathscr{H}$ for the "history" of x). The slice at time t is $\mathscr{H}_x(t) = \mathscr{H}_x \cap V \times \{t\}$. We let $\mathscr{H}_A = \bigcup_{x \in A} \mathscr{H}_x$ for any $A \subseteq V$, and similarly for $\mathscr{H}_A(t)$. The information percolation clusters are the connected components of $\mathscr{H}_V$. Note that each cluster can be written as $\mathscr{H}_A$ for some $A \subseteq V$, but not all $\mathscr{H}_A$ form a complete cluster.

If $\mathscr{H}_V(0) = \varnothing$, then the configuration at time $t_{\mathrm{m}} + s$ is independent of the configuration at time 0. This means that, given the randomness encoded by the T_x^is and U_x^is, the final configuration will be the same, regardless of the initial configuration. In particular, we could have started with a sample from the stationary distribution and ended up with the same final state. This means that if we happen to have $\mathscr{H}_V(0) = \varnothing$, we will have ended up with a truly exact sample from the stationary distribution. This is similar to the "coupling from the past" technique.

However, as was the case with $\beta = 0$, we do not need to go so far. As long as $\mathscr{H}_V(0)$ is sparse enough, the fluctuations in the stationary distribution can absorb the imperfections, and we will have a good approximation of the stationary distribution. To analyze the clusters more precisely, we split them into three types, denoted by the colors RED, BLUE, and GREEN.

- A cluster $\mathscr{H}_A$ is RED if $\mathscr{H}_A(0) \neq \varnothing$.
- A cluster $\mathscr{H}_A$ is BLUE if $\mathscr{H}_A(0) = \varnothing$ and A is a singleton.
- A cluster $\mathscr{H}_A$ is GREEN otherwise (i.e. if $\mathscr{H}_A(0) = \varnothing$, but A is not a singleton).

If $\mathscr{H}_A$ is a RED cluster, then the spins in A depend on the initial configuration. If $\mathscr{H}_A$ is a BLUE cluster, then $A = \{v\}$ and the spin at v is a coin flip, independent of everything else. The GREEN clusters represent the potentially complex behavior of interactions between sites under the stationary distribution.

Since the GREEN clusters are difficult to analyze, we will simply condition on them. Once this is done, the distance to stationarity comes down to a fight between the RED and BLUE clusters. Under true stationarity (with $\mathscr{H}_V(0) = \varnothing$), there would only be BLUE and GREEN clusters, and so conditioning on the GREEN clusters would mean that the rest of the spin configuration consists of i.i.d. coin flips. Thus, one way to get a handle on the distance to stationarity is to determine how far the configuration is from being i.i.d., conditioned on the GREEN clusters. We will invoke a lemma of Miller and Peres to do this:

Lemma 2. *Let W be a finite set, and let ν be the uniform measure on $\{\pm 1\}^W$. For each $R \subseteq W$, let φ_R be a measure on $\{\pm 1\}^R$, and let μ be the measure on $\{\pm 1\}^W$ obtained by sampling a subset $R \subseteq W$ via some measure $\tilde{\mu}$, generating the spins of R via φ_R, and finally sampling $W \setminus R$ uniformly. Then*

$$4\|\mu - \nu\|_{\mathrm{TV}}^2 \leq \mathbb{E}2^{|R \cap R'|} - 1,$$

where R, R' are i.i.d. with law $\tilde{\mu}$.

Proof. We have

$$\begin{aligned} 4\|\mu - \nu\|_{\mathrm{TV}}^2 &= \left\| \frac{\mu}{\nu} - 1 \right\|_{L^1(\nu)}^2 \\ &\leq \left\| \frac{\mu}{\nu} - 1 \right\|_{L^2(\nu)}^2 \\ &= \sum_{\sigma \in \{\pm 1\}^W} \left(\frac{\mu(\sigma)}{\nu(\sigma)} - 1 \right)^2 \nu(\sigma) \\ &= 2^{|W|} \sum_{\sigma \in \{\pm 1\}^W} \mu(\sigma)^2 - 1, \end{aligned}$$

where we have used the fact that $\nu(\sigma) = 2^{-|W|}$, and μ is a probability measure. Now notice that for any σ,

$$\mu(\sigma) = \sum_{R \subseteq W} \tilde{\mu}(R) \varphi_R(\sigma|_R) 2^{-(|W| - |R|)},$$

since to sample from μ, we sample a subset R, then sample the spins in R via φ_R, and finally sample the rest of the spins uniformly. Therefore we have

$$\begin{aligned} 2^{|W|} \sum_{\sigma \in \{\pm 1\}^W} \mu(\sigma)^2 &= 2^{|W|} \sum_{\sigma \in \{\pm 1\}^W} \sum_{R, R' \subseteq W} \tilde{\mu}(R) \tilde{\mu}(R') \varphi_R(\sigma|_R) \varphi_{R'}(\sigma|_{R'}) 2^{-2|W| + |R| + |R'|} \\ &= 2^{-|W|} \sum_{R, R' \subseteq W} \tilde{\mu}(R) \tilde{\mu}(R') 2^{|R| + |R'|} \sum_{\sigma \in \{\pm 1\}^W} \varphi_R(\sigma|_R) \varphi_{R'}(\sigma|_{R'}). \end{aligned}$$

Now, for any fixed R and R', since $\varphi_R(\sigma|_R)\varphi_{R'}(\sigma|_{R'})$ only depends on $\sigma|_{R \cup R'}$, we have

$$\sum_{\sigma \in \{\pm 1\}^W} \varphi_R(\sigma|_R) \varphi_{R'}(\sigma|_{R'}) = 2^{|W| - |R \cup R'|} \sum_{\sigma \in \{\pm 1\}^{R \cup R'}} \varphi_R(\sigma|_R) \varphi_{R'}(\sigma|_{R'}).$$

The sum on the right-hand side is the probability that configurations $\sigma \sim \varphi_R$ and $\sigma' \sim \varphi_{R'}$ agree in $R \cap R'$. Since this is ≤ 1,

$$2^{|W|} \sum_{\sigma \in \{\pm 1\}^W} \mu(\sigma)^2 = \sum_{R, R' \subseteq W} \tilde{\mu}(R) \tilde{\mu}(R') 2^{|R| + |R'| - |R \cup R'|},$$

which finishes the proof via the inclusion-exclusion formula. ∎

We will apply this lemma with R representing V_{RED}, the set of RED vertices (i.e. $\mathscr{H}_{\text{RED}}(t_{\mathfrak{m}}+s)$), conditional on $\mathscr{H}_{\text{GREEN}}$ the set of GREEN clusters. In particular, we will take $W = V \setminus V_{\text{GREEN}}$, so that the measure we end up with is indeed uniform outside of V_{RED}.

To see why this is useful, first notice that by Jensen's inequality, we have

$$\begin{aligned} d(t_{\mathfrak{m}}+s) &\le \max_{\sigma_0\in\Omega} \mathbb{E}\left[\|\mathbb{P}_{\sigma_0}[\sigma_{t_{\mathfrak{m}}+s}\in\cdot|\mathscr{H}_{\text{GREEN}}] - \mathbb{P}_{\pi}[\sigma_{t_{\mathfrak{m}}+s}\in\cdot|\mathscr{H}_{\text{GREEN}}]\|_{\text{TV}}\right] \\ &= \sup_{\mathscr{H}_{\text{GREEN}}} \max_{\sigma_0\in\Omega} \|\mathbb{P}_{\sigma_0}[\sigma_{t_{\mathfrak{m}}+s}\in\cdot|\mathscr{H}_{\text{GREEN}}] - \mathbb{P}_{\pi}[\sigma_{t_{\mathfrak{m}}+s}\in\cdot|\mathscr{H}_{\text{GREEN}}]\|_{\text{TV}}\,. \end{aligned}$$

Now, since $\sigma_{t_{\mathfrak{m}}+s}$ can be coupled exactly with the stationary distribution on V_{GREEN}, projecting onto the complement $V\setminus V_{\text{GREEN}}$ cannot decrease the total variation distance. In other words,

$$\begin{aligned} d(t_{\mathfrak{m}}+s) &\le \sup_{\mathscr{H}_{\text{GREEN}}} \max_{\sigma_0\in\Omega} \left\|\mathbb{P}_{\sigma_0}[\sigma_{t_{\mathfrak{m}}+s}|_{V\setminus V_{\text{GREEN}}}\in\cdot|\mathscr{H}_{\text{GREEN}}] - \mathbb{P}_{\pi}[\sigma_{t_{\mathfrak{m}}+s}|_{V\setminus V_{\text{GREEN}}}\in\cdot|\mathscr{H}_{\text{GREEN}}]\right\|_{\text{TV}} \\ &\le 2 \sup_{\mathscr{H}_{\text{GREEN}}} \max_{\sigma_0\in\Omega} \left\|\mathbb{P}_{\sigma_0}[\sigma_{t_{\mathfrak{m}}+s}|_{V\setminus V_{\text{GREEN}}}\in\cdot|\mathscr{H}_{\text{GREEN}}] - \nu_{V\setminus V_{\text{GREEN}}}\right\|_{\text{TV}}, \end{aligned}$$

where ν_W is the uniform measure on $\{\pm 1\}^W$. Thus, the lemma implies that

$$d(t_{\mathfrak{m}}+s) \le \left(\sup_{\mathscr{H}_{\text{GREEN}}} \mathbb{E}\left[2^{|V_{\text{RED}}\cap V'_{\text{RED}}|}\middle|\mathscr{H}_{\text{GREEN}}\right] - 1\right)^{1/2}.$$

It remains to analyze this exponential moment.

To do this, we need a few lemmas. For $A\subseteq V$, we say that $A\in\text{RED}$ if A is the intersection of a complete RED cluster with $V\times\{t_{\mathfrak{m}}+s\}$. In particular, we don't allow any strict subsets of such sets, or unions of tops of disjoint clusters. For any $A\subseteq V$, define

$$\Psi_A = \sup_{\mathscr{H}_{V\setminus A}} \mathbb{P}\left[A\in\text{RED}\middle|\mathscr{H}_{V\setminus A}, A\in\text{RED or } A\subseteq V_{\text{BLUE}}\right].$$

Then we have the following exponential upper bound:

Lemma 3. *For any $d\ge 2$ and $\lambda>0$ there exist $\beta_0, C_0>0$ such that if $\beta<\beta_0$ then for any $A\subseteq V$ and large enough n, we have*

$$\Psi_A \le C_0 \mathfrak{m}_{t_{\mathfrak{m}}+s} e^{-\lambda\mathfrak{W}(A)},$$

where $\mathfrak{W}(A)$ is the size of the smallest connected subgraph containing A.

For intuition, notice that $A\in\text{RED}$ if the histories $\mathscr{H}_v$ for $v\in A$ are all connected, and $\mathscr{H}_A(0)\neq\varnothing$. The requirement that the histories of A spatially connect is the source of the $e^{-\lambda\mathfrak{W}(A)}$ factor, since the projection of the cluster of A on V must be a connected subgraph containing A. The factor of

$$\mathfrak{m}_{t_{\mathfrak{m}}+s} = \mathbb{E}[\sigma^+_{t_{\mathfrak{m}}+s}(v)] = \mathbb{P}[\sigma^+_{t_{\mathfrak{m}}+s}(v)\neq\sigma^-_{t_{\mathfrak{m}}+s}(v)]$$

occurs because this is the probability that the history of any single vertex will reach the bottom, as it is exactly in this scenario that the extremal configurations do not couple.

Then we have a lemma which relates the worst-case probability Ψ_A that A is the top of a RED cluster to the actual sets of RED vertices V_{RED}:

Lemma 4. *Let $\{Y_{A,A'} : A, A'\subseteq V\}$ be a family of independent indicator random variables satisfying*

$$\mathbb{P}[Y_{A,A'}=1] = \Psi_A\Psi_{A'}.$$

The conditional distribution of two independent samples of the set of red vertices, V_{RED} and V'_{RED} can be coupled to the $Y_{A,A'}$s such that

$$|V_{RED}\cap V'_{RED}| \le \sum_{A\cap A'\neq\varnothing} |A\cup A'| Y_{A,A'}.$$

Intuitively, for a vertex v to be in both V_{RED} and V'_{RED}, there must be sets A and A' which are both the top slice of a full RED cluster, and for which $v \in A$ and A'. The reason that the union appears instead of the intersection has to do with the fact that we need to ensure that $Y_{A,A'}$ are independent (I think).

Anyway, using these two lemmas we can finish the proof. Since $|A \cup A'| \leq |A| + |A'|$, we obtain

$$\begin{aligned}
\sup_{\mathscr{H}_{\text{GREEN}}} \mathbb{E}\left[2^{|V_{\text{RED}} \cap V'_{\text{RED}}|} \middle| \mathscr{H}_{\text{GREEN}}\right] &\leq \mathbb{E}\left[2^{\sum_{A \cap A' \neq \varnothing} (|A|+|A'|) Y_{A,A'}}\right] \\
&= \prod_{A \cap A' \neq \varnothing} \mathbb{E}\left[2^{(|A|+|A'|)Y_{A,A'}}\right], && \text{(using independence)} \\
&\leq \prod_{v} \prod_{\substack{A,A' \\ v \in A \cap A'}} \left((2^{|A|+|A'|} - 1)\Psi_A \Psi_{A'} + 1\right) && \text{(double-counting)} \\
&\leq \prod_{v} \prod_{\substack{A,A' \\ v \in A \cap A'}} \left(2^{|A|+|A'|}\Psi_A \Psi_{A'} + 1\right) \\
&\leq \exp\left(n \left(\sum_{A \ni v} 2^{|A|} \Psi_A\right)^2\right).
\end{aligned}$$

Now, using the fact that $e^x - 1 \leq 2x$ for $x \in [0,1]$, we obtain that

$$\begin{aligned}
d(t_{\mathfrak{m}} + s) &\leq \left(\exp\left(n \left(\sum_{A \ni v} 2^{|A|} \Psi_A\right)^2\right) - 1\right)^{1/2} \\
&\leq \sqrt{2n} \sum_{A \ni v} 2^{|A|} \Psi_A.
\end{aligned}$$

Using Lemma 3 with $\lambda = \log(8d)$, we obtain

$$\begin{aligned}
\sum_{A \ni v} 2^{|A|} \Psi_A &\leq C_0 \mathfrak{m}_{t_{\mathfrak{m}}+s} \sum_{k} \sum_{\substack{A \ni v \\ \mathfrak{W}(A) = k}} 2^k e^{-\lambda k} \\
&\leq C_0 \mathfrak{m}_{t_{\mathfrak{m}}+s} \sum_{k \geq 1} 2^k 2^k d^k e^{-\lambda k} \\
&= C_0 \mathfrak{m}_{t_{\mathfrak{m}}+s} \sum_{k \geq 1} (4de^{-\lambda})^k \\
&= C_0 \mathfrak{m}_{t_{\mathfrak{m}}+s}.
\end{aligned}$$

This shows that $d(t_{\mathfrak{m}} + s) \leq C_1 \mathfrak{m}_{t_{\mathfrak{m}}+s} \sqrt{n}$. Finally, we conclude by using the submultiplicativity of the expected magnetization $(**)$ from above, that

$$d(t_{\mathfrak{m}} + s) \leq C_1 e^{-(1-\beta d)s} \mathfrak{m}_{t_{\mathfrak{m}}} \sqrt{n} = C_1 e^{-(1-\beta d)s},$$

using the fact that $\mathfrak{m}_{t_{\mathfrak{m}}} = \frac{1}{\sqrt{n}}$ by definition of $t_{\mathfrak{m}}$. Thus, if β is small enough (as a function of d), we have $d(t_{\mathfrak{m}} + s) \to 0$ as $s \to \infty$.