6.7720/18.619/15.070 Lecture 3
The First & Second Moment Methods: Statistical Inference for the Broadcast Process

Kuikui Liu

February 10, 2025

Acknowledgements & Disclaimers *Several of the topic choices and lecture notes were partly inspired by materials created by Guy Bresler and David Gamarnik, who taught previous iterations of this course. We are grateful for the materials they shared and for the discussions we had with them. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Broadcasting on Trees

In this lecture, we consider a new type of stochastic process on trees called the *broadcast process*.

Definition 1 ((Binary Symmetric) Broadcast Process on $\widehat{\mathbb{T}}_d$)**.** *Fix $d \in \mathbb{N}$ and an error parameter $0 \leq \epsilon \leq 1/2$. The associated* (binary symmetric) broadcast process *on the infinite d-ary tree $\widehat{\mathbb{T}}_d$ rooted at r is a random assignment $\sigma_v \in \{\pm 1\}$ to each vertex v of $\widehat{\mathbb{T}}_d$ generated as follows:*

- *We initialize the process by sampling $\sigma_r \sim \mathsf{Unif}\{\pm 1\}$ for the root vertex.*
- *Independently for each child v of r, we set $\sigma_v = \sigma_r$ with probability $1 - \epsilon$ and $\sigma_v = -\sigma_r$ with probability ϵ.*
- *This process continues recursively in the subtrees rooted at each of the children v of r.*

Remark 1. This process can be vastly generalized to allow for larger state spaces, as well as arbitrary Markov chains for the channels on the edges.

Informally, the vertex r is "broadcasting" its random assignment to its descendants through channels corrupted by an ϵ amount of noise. The basic question is whether or not we can *infer* the assignment of the root vertex given only information about far away vertices. More precisely, if we let $L(n)$ denote the set of vertices at distance exactly n from the root, then our goal is to design an estimator $\widehat{\sigma}_{r,n}$ for σ_r based on only the assignments $\sigma_{L(n)}$ of the vertices in $L(n)$ for n large. For instance, $\widehat{\sigma}_{r,n}$ could be a deterministic function mapping $\{\pm 1\}^{L(n)}$ to $\{\pm 1\}$ which tells us what to guess upon seeing some specific $\sigma_{L(n)}$. At the highest level of generality, our goal is to design a function $\widehat{p} : \{\pm 1\}^{L(n)} \to [0, 1]$ which outputs the probability that we should set our estimator $\widehat{\sigma}_{r,n} \in \{\pm 1\}$ to, say, $+1$.

Definition 2 (Reconstruction Problem)**.** *Fix $d \in \mathbb{N}$ and $0 \leq \epsilon \leq 1/2$. For an estimator $\widehat{\sigma}_{r,n}$ given access to only $\sigma_{L(n)}$, write $\frac{1+b(\widehat{\sigma}_{r,n})}{2} = \Pr[\widehat{\sigma}_{r,n} = \sigma_r]$, where the probability is calculated with respect to a random draw of the broadcast process σ and the randomness of the estimator, and $b(\widehat{\sigma}_{r,n})$ denotes the "advantage" of the estimator over random guessing.*[1] *We also write $\frac{1+b^*}{2}$ for the optimal success probability, where $b^* = b^*\left(n, \widehat{\mathbb{T}}_d, \epsilon\right) \geq 0$ is given by maximizing $b(\widehat{\sigma}_{r,n})$ over all possible estimators $\widehat{\sigma}_{r,n}$.*

[1] A trivial estimator is to just output a uniformly random element of $\{\pm 1\}$. This estimator clearly guesses the correct answer with probability $\frac{1}{2}$.

Theorem 1.1 (Kesten–Stigum'66 [**???**], Bleher–Ruiz–Zagrebnov'95 [**???**])**.** *Fix $d \in \mathbb{N}$, and let $\theta = 1 - 2\epsilon$ for $0 \leq \epsilon \leq 1/2$ (so that $\epsilon = \frac{1-\theta}{2}$). Let θ_{c} satisfy $d \cdot \theta_{\mathsf{c}}^2 = 1$, and correspondingly define $\epsilon_{\mathsf{c}} = \frac{1}{2} - \frac{1}{2\sqrt{d}}$. Then we have the following phase transition for the reconstruction problem:*

$$\lim_{n \to \infty} b^* \left(n, \widehat{\mathbb{T}}_d, \epsilon\right) \begin{cases} = 0, & \text{if } \epsilon > \epsilon_{\mathsf{c}} \text{ (or equivalently } d \cdot \theta^2 < 1\text{)} \\ > 0, & \text{if } \epsilon < \epsilon_{\mathsf{c}} \text{ (or equivalently } d \cdot \theta^2 > 1\text{)} \end{cases}.$$

Remark 2. It turns out that in the regime $\epsilon > \epsilon_{\mathsf{c}}$, we have that there is some $\delta = \delta(\epsilon) > 0$ such that we have an exponential decay rate

$$b^* \left(n, \widehat{\mathbb{T}}_d, \epsilon\right) \lesssim (1 - \delta)^n.$$

At criticality, when $\epsilon = \epsilon_{\mathsf{c}}$, a closer inspection of the proof reveals that $b^* \left(n, \widehat{\mathbb{T}}_d, \epsilon\right) \leq O(1/n)$. This decay rate is much slower than the decay in the regime $\epsilon > \epsilon_{\mathsf{c}}$, but it still yields $b^* \left(n, \widehat{\mathbb{T}}_d, \epsilon\right) \to 0$.

Very roughly speaking, this phase transition occurs due to two competing effects: There is the exponential-in-n growth of the number of vertices which could receive the "signal" σ_r, but there's also the exponential decay of information due to the addition of noise in each step as we increase the distance n.

A Brief Word on Notation Throughout the remainder of the course, for a random variable X, we write $\mathsf{Law}(X)$ for the associated probability measure X according to which X is distributed. We also write $\operatorname{supp}(X)$ for the *support* of the distribution $\mathsf{Law}(X)$, i.e. the set of values which occur with positive probability.

2 The Branching Process Perspective

Theorem 1.1 establishes a phase transition phenomenon for the reconstruction problem on $\widehat{\mathbb{T}}_d$. Rather than diving head-first into bare-handed calculations, let us first try to relate this to the only phase transition phenomenon we've already studied, namely percolation on $\widehat{\mathbb{T}}_d$ (or more generally, Galton–Watson branching processes). Because the broadcast process is about noisy transmission of information, it is natural to track the number of vertices of $\widehat{\mathbb{T}}_d$ which "receive uncorrupted information from the root". One clean way to formalize this is to reframe the broadcast process as follows:

- We initialize the process by sampling $\sigma_r \sim \mathsf{Unif}\{\pm 1\}$ for the root vertex.
- Independently for each child v of r, we set $\sigma_v = \sigma_r$ with probability $\theta = 1 - 2\epsilon$, and otherwise sample $\sigma_v \sim \mathsf{Unif}\{\pm 1\}$.
- This process continues recursively in the subtrees rooted at each of the children v of r.

In this way of looking at the broadcast process, we see that independently to each child v, the parent u "perfectly transmits" its assignment with probability $\theta = 1 - 2\epsilon$. Otherwise, with probability 2ϵ, the child completely ignores its parent's assignment and samples a fresh bit in $\{\pm 1\}$. In the former case, let us mark the corresponding edge $\{u, v\}$ with a 1, indicating the edge is "open"; otherwise, we mark the edge $\{u, v\}$ with a 0, indicating the edge is "closed". The open edges in $\widehat{\mathbb{T}}_d$, i.e. the edges which receive a 1, are collectively distributed as bond percolation on $\widehat{\mathbb{T}}_d$ with edge probability $\theta = 1 - 2\epsilon$.

For each $n \in \mathbb{N}$, let Z_n denote the number of vertices in $L(n)$ which are connected to the root r by a path of open edges (i.e. edges marked by a 1); for each of these vertices, its assignment, as well as the assignments of its ancestors, all perfectly copy the assignment of the root. A key observation is that the collection of random variables $\mathcal{Z} = \{Z_n\}_{n \in \mathbb{N}}$ is a Galton–Watson branching process with binomial offspring distribution $\mathsf{Bin}(d, \theta)$, where recall that $\theta = 1 - 2\epsilon$. This branching process is critical if and only if $d\theta = 1$ and so we expect a phase transition at this point.

In a future lecture, we will discuss the phase transition at $d\theta = 1$ in the context of the famous *(ferromagnetic) Ising model* from statistical physics, which is intimately related to the broadcast process we're studying here. However, unfortunately, this transition point doesn't match the one in Theorem 1.1.

The reason is that even if $\theta > \frac{1+\epsilon}{d}$, it could be that the number of vertices Z_n copying the root is too small relative to the total number of vertices d^n in $L(n)$. In other words, even if there is a "signal" (whose magnitude is Z_n) received by $L(n)$, it is "drowned out" by the sea of noise from the remaining $d^n - Z_n$ vertices who receive random bits independent of the root. To get a better sense of "how large" the signal needs to be, imagine an alternative world where all vertices of $L(n)$ receive independent random bits drawn from $\mathsf{Unif}\{\pm 1\}$, and there is simply no way to estimate the root σ_r better than a coin flip. In this pure-noise world, the expected number of vertices which have same assignment as the root is $\frac{1}{2}d^n$ while the standard deviation is $\frac{1}{2}d^{n/2}$. In order to "rule out" this alternative universe, we would need $\mathbb{E}[Z_n] > \frac{1}{2}d^{n/2}$ for large n. Since $\mathbb{E}[Z_n] = d^n\theta^n$, this inequality is equivalent to $d\theta^2 > 1$, which is the threshold described by Theorem 1.1.

3 Reinterpreting b^*

The problem of recovering σ_r becomes easier if $\epsilon < \epsilon_{\mathsf{c}}$ is small because the value of σ_r has a noticeable effect on the distribution of $\sigma_{L(n)}$. If $\epsilon < \epsilon_{\mathsf{c}}$ is small and $\sigma_r = +1$ (resp. $\sigma_r = -1$), then we expect a disproportionate fraction of the vertices in $L(n)$ to be assigned $+1$ (resp. -1). It turns out we can actually interpret b^* as measuring the size of this effect. More formally, we can express it as the *total variation distance* between the distribution of $\sigma_{L(n)}$ conditioned on $\sigma_r = +1$ and the distribution of $\sigma_{L(n)}$ conditioned on $\sigma_r = -1$. To see this, observe that

$$\begin{aligned}
b^* &= 2 \cdot \sup_{\widehat{p}} \Pr\left[\widehat{\sigma}_{r,n} = \sigma_r\right] - 1 \\
&= 2 \cdot \sup_{\widehat{p}} \left\{ \frac{1}{2} \Pr\left[\widehat{\sigma}_{r,n} = \sigma_r \mid \sigma_r = +1\right] + \frac{1}{2} \Pr\left[\widehat{\sigma}_{r,n} = \sigma_r \mid \sigma_r = -1\right] \right\} - 1 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Law of Total Probability)} \\
&= \sup_{\widehat{p}} \left\{ \Pr\left[\widehat{\sigma}_{r,n} = +1 \mid \sigma_r = +1\right] - \Pr\left[\widehat{\sigma}_{r,n} = +1 \mid \sigma_r = -1\right] \right\} \\
&= \sup_{\widehat{p}} \left\{ \mathbb{E}_\sigma\left[\widehat{p}\left(\sigma_{L(n)}\right) \mid \sigma_r = +1\right] - \mathbb{E}_\sigma\left[\widehat{p}\left(\sigma_{L(n)}\right) \mid \sigma_r = -1\right] \right\} \\
&= \mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = +1\right), \mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = -1\right)\right), \qquad \text{(Lemma 3.1)}
\end{aligned}$$

where the supremum is over all possible functions $\widehat{p} : \{\pm 1\}^{L(n)} \to [0, 1]$ and the probability is with respect to σ drawn from the broadcast process and $\widehat{\sigma}_{r,n}$ drawn independently from $\mathsf{Ber}\left(\widehat{p}\left(\sigma_{L(n)}\right)\right)$ given σ. We justify the final step using the following lemma.

Lemma 3.1. *Let μ, ν be two probability measures over a common state space Ω. Then the* total variation distance $\mathscr{D}_{\mathsf{TV}}(\mu, \nu) \overset{\mathsf{def}}{=} \frac{1}{2}\sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)|$ *may be equivalently written as*

$$\mathscr{D}_{\mathsf{TV}}(\mu, \nu) = \sup_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \sup_f \left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right|,$$

where the supremum in the right-most expression is over all 1*-bounded functions* $f : \Omega \to [0, 1]$.

Proof Sketch. Since Ω is finite, we can view $f : \Omega \to [0, 1]$ and μ, ν as big vectors in $[0, 1]^\Omega$. Then $|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| = |\langle \mu - \nu, f\rangle|$, which is convex in f. This convexity immediately implies the second equality via the natural correspondence between $A \subseteq \Omega$ and its $\{0, 1\}$-indicator function.

For the first equality, observe that since $|\mu(A) - \nu(A)| = |\mu(\Omega \setminus A) - \nu(\Omega \setminus A)|$,

$$|\mu(A) - \nu(A)| = \frac{1}{2}\left|\sum_{\omega \in A}(\mu(\omega) - \nu(\omega))\right| + \frac{1}{2}\left|\sum_{\omega \in \Omega \setminus A}(\mu(\omega) - \nu(\omega))\right| \le \|\mu - \nu\|_{\mathsf{TV}} \qquad \forall A \subseteq \Omega$$

by the Triangle Inequality. On the other hand, this inequality is saturated by the set $A = \{\omega \in \Omega : \mu(\omega) > \nu(\omega)\}$, and so we must have equality. $\square$

3.1 Designing Estimators

Given the interpretation of b^* as the total variation distance between $\mu_n^+ \stackrel{\mathsf{def}}{=} \mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = +1\right)$ and $\mu_n^- \stackrel{\mathsf{def}}{=} \mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = -1\right)$, Lemma 3.1 suggests an estimator $\widehat{\sigma}_{r,n}^{\mathsf{MLE}}$ known as the *maximum likelihood estimator*: If $\sigma_{L(n)} = \tau$ for some $\tau \in \{\pm 1\}^{L(n)}$, then

$$\widehat{\sigma}_{r,n}^{\mathsf{MLE}} = \begin{cases} +1, & \text{if } \mu_n^+(\tau) > \mu_n^-(\tau) \\ -1, & \text{if } \mu_n^+(\tau) < \mu_n^-(\tau) \end{cases}.$$

This is the optimal estimator whose success probability is precisely $\frac{1+b^*}{2}$.

Remark 3. Given $\tau \in \{\pm 1\}^{L(n)}$, one can of course compute the probabilities $\mu_n^+(\tau), \mu_n^-(\tau)$ through brute force enumeration of all possible assignments for the vertices in $L(1), \ldots, L(n-1)$. However, if the goal is to just evaluate whether or not $\mu_n^+(\tau) > \mu_n^-(\tau)$, then we can actually do this in time polynomial in the size of the input τ. The idea is that by Bayes' Rule, $\mu_n^+(\tau) > \mu_n^-(\tau)$ holds if and only if

$$\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right] > \Pr\left[\sigma_r = -1 \mid \sigma_{L(n)} = \tau\right].$$

These two quantities can be computed via recursion. The specific algorithm is known in the literature as *belief propagation*, the *tree recursion*, or the *sum-product algorithm*, and is extremely well-studied. We will discuss this in greater depth in Section 5; see also Appendix A for a derivation of this recursion.

Another natural estimator to consider is simply the one given by majority vote:

$$\widehat{\sigma}_{r,n}^{\mathsf{MAJ}} \stackrel{\mathsf{def}}{=} \mathsf{sign}\left(S_n\right) \qquad \text{where} \qquad S_n \stackrel{\mathsf{def}}{=} \sum_{v \in L(n)} \sigma_v.$$

We note that the maximum likelihood and majority estimators are not equal. We leave this as an exercise.

Exercise 1. *Prove that* $\widehat{\sigma}_{r,n}^{\mathsf{MLE}} \neq \widehat{\sigma}_{r,n}^{\mathsf{MAJ}}$.

While these two estimators have different advantages, i.e. $b\left(\widehat{\sigma}_{r,n}^{\mathsf{MLE}}\right) \neq b\left(\widehat{\sigma}_{r,n}^{\mathsf{MAJ}}\right)$, fortunately for us, it will turn out that $\lim_{n\to\infty} b\left(\widehat{\sigma}_{r,n}^{\mathsf{MAJ}}\right) > 0$ if and only if $\lim_{n\to\infty} b\left(\widehat{\sigma}_{r,n}^{\mathsf{MLE}}\right) > 0$. This will be a consequence of the proof of Theorem 1.1. We note that there are many other natural broadcast processes for which these two regimes *do not coincide* [**???**].

4 Proof of Theorem 1.1 in the Case $\epsilon < \epsilon_{\mathsf{c}}$

In this section, we work in the regime $\epsilon < \epsilon_{\mathsf{c}}$, or equivalently, $d\theta^2 > 1$. From our previous analysis, we have that

$$\begin{aligned} b^* &\geq \mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}\left(\widehat{\sigma}_{r,n}^{\mathsf{MAJ}} \mid \sigma_r = +1\right), \mathsf{Law}\left(\widehat{\sigma}_{r,n}^{\mathsf{MAJ}} \mid \sigma_r = -1\right)\right) \\ &= \mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}\left(S_n \mid \sigma_r = +1\right), \mathsf{Law}\left(S_n \mid \sigma_r = -1\right)\right). \end{aligned}$$

Note the final equality holds due to symmetry between ± 1. More concretely, the random variables $S_n \mid \sigma_r = -1$ and $-S_n \mid \sigma_r = +1$ have the same law, and so

$$\{s \in \mathbb{Z} : \Pr[S_n = s \mid \sigma_r = +1] > \Pr[S_n = s \mid \sigma_r = -1]\} = \{s \in \mathbb{Z} : s > 0\}$$

and we may apply Lemma 3.1.

We will show that when $\epsilon < \epsilon_{\mathsf{c}}$, the right-hand side is lower bounded by a universal constant even in the large n limit. We will achieve this by combining the second moment method with the following lemma.

Lemma 4.1 (Second Moment Lower Bound for Total Variation)**.** *Let X, Y be two real-valued random variables, and let W have law $\frac{1}{2}\mathsf{Law}(X) + \frac{1}{2}\mathsf{Law}(Y)$. Then we have the lower bound*

$$\mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}(X), \mathsf{Law}(Y)\right) \geq \frac{1}{4} \cdot \frac{\left(\mathbb{E}[X] - \mathbb{E}[Y]\right)^2}{\mathrm{Var}\left(W\right)}.$$

We prove this lemma at the end of this section. To complete the proof of Theorem 1.1 in the case $\epsilon < \epsilon_{\mathsf{c}}$, we now compute the quantities involved in the above lower bound when $X_n = S_n \mid \sigma_r = +1$ and $Y_n = S_n \mid \sigma_r = -1$; note that $W_n = S_n$ without any conditioning. If we consider again the Galton–Watson branching process $\mathcal{Z} = \{Z_n\}_{n\in\mathbb{N}}$ we discussed previously, which has offspring distribution $\mathsf{Bin}(d,\theta)$ with mean $d\theta$, then observe that X_n is equal to Z_n plus a sum of $d^n - Z_n$ many (dependent) $\mathsf{Unif}\{\pm 1\}$ random variables. Hence,

$$\mathbb{E}[X_n] = \mathbb{E}[Z_n] = d^n\theta^n \qquad \text{and} \qquad \mathbb{E}[Y_n] = -\mathbb{E}[Z_n] = -d^n\theta^n.$$

We now calculate the variance of S_n.

$$\begin{aligned}
\mathrm{Var}\left(S_n\right) &= \mathbb{E}\left[S_n^2\right] && \text{(Using } \mathbb{E}\left[S_n\right] = 0\text{)}\\
&= \sum_{u,v\in L(n)} \mathbb{E}[\sigma_u\sigma_v]\\
&= \sum_{u\in L(n)} \mathbb{E}\left[\sigma_u^2\right] + \sum_{u\in L(n)}\sum_{\ell=0}^{n-1}\sum_{\substack{v\neq u\\ \mathsf{LCA}(u,v)\in L(\ell)}} \mathbb{E}[\sigma_u\sigma_v].
\end{aligned}$$

Fix a pair of distinct vertices $u, v \in L(n)$ such that their least common ancestor $w = \mathsf{LCA}(u,v)$ lies in $L(\ell)$. Then by the Law of Total Expectation and the fact that an independently sampled $\mathsf{Unif}\{\pm 1\}$ is assigned when a child fails to copy its parent's assignment, we have

$$\begin{aligned}
\mathbb{E}\left[\sigma_u\sigma_v\right] &= \Pr\left[\sigma_u, \sigma_v \text{ both copy } \sigma_w\right]\\
&= \Pr\left[\sigma_u \text{ copies } \sigma_w\right]\cdot\Pr\left[\sigma_v \text{ copies } \sigma_w\right]\\
&= \theta^{2(n-\ell)}.
\end{aligned}$$

Combining with the fact that $\#\left\{v \neq u : \mathsf{LCA}(u,v) \in L(\ell)\right\} = d^{n-\ell-1}(d-1)$ for every $u \in L(n)$, the above simplifies to

$$\mathrm{Var}\left(S_n\right) = d^n + d^{n-1}(d-1)\sum_{k=1}^{n}\left(d\theta^2\right)^k = d^n + d^n\left(\left(d\theta^2\right)^n - 1\right)\cdot\frac{(d-1)\theta^2}{d\theta^2-1}.$$

With this in hand, we find that

$$\begin{aligned}
b^* &\geq \mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}\left(S_n \mid \sigma_r = +1\right), \mathsf{Law}\left(S_n \mid \sigma_r = -1\right)\right) && \text{(Previous display)}\\
&\geq \frac{1}{4}\cdot\frac{\left(\mathbb{E}[X_n]-\mathbb{E}[Y_n]\right)^2}{\mathrm{Var}\left(S_n\right)} && \text{(Lemma 4.1)}\\
&= \frac{d^{2n}\theta^{2n}}{d^n + d^n\left(\left(d\theta^2\right)^n - 1\right)\cdot\frac{(d-1)\theta^2}{d\theta^2-1}}\\
&= \frac{1}{\left(d\theta^2\right)^{-n} + \left(1-\left(d\theta^2\right)^{-n}\right)\cdot\frac{(d-1)\theta^2}{d\theta^2-1}}\\
&\to \frac{d\theta^2-1}{(d-1)\theta^2}. && \text{(In the } n\to\infty \text{ limit, using } d\theta^2 > 1\text{)}
\end{aligned}$$

As a sanity check, this is at most 1 since $\theta \leq 1$. It is positive since $d\theta^2 > 1$ by assumption, which completes the proof of Theorem 1.1.

Proof of Lemma 4.1. Without loss of generality, we may shift X, Y, W by the same constant to ensure they have mean zero, while preserving the total variation distance between $\mathsf{Law}(X), \mathsf{Law}(Y)$. Hence, we assume $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[W] = 0$ in the remainder of the proof.

Observe that $\mathrm{supp}(W) = \mathrm{supp}(X) \cup \mathrm{supp}(Y)$, and so

$$\begin{aligned}
\mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}(X), \mathsf{Law}(Y)\right) &= \frac{1}{2}\sum_{\omega\in\mathrm{supp}(W)}\left|\Pr[X=\omega]-\Pr[Y=\omega]\right|\\
&= \sum_{\omega\in\mathrm{supp}(W)}\frac{\left|\Pr[X=\omega]-\Pr[Y=\omega]\right|}{2\cdot\Pr[W=\omega]}\cdot\Pr[W=\omega]\\
&= \mathbb{E}\left[\left|f(W)\right|\right],
\end{aligned}$$

where $f(\omega) \stackrel{\mathsf{def}}{=} \frac{\Pr[X=\omega]-\Pr[Y=\omega]}{\Pr[X=\omega]+\Pr[Y=\omega]}$ takes values in the interval $[-1,1]$. Hence, we may lower bound the above by

$$\mathbb{E}\left[f(W)^2\right] \geq \frac{\mathbb{E}\left[W \cdot f(W)\right]^2}{\mathbb{E}[W^2]} \qquad \text{(Cauchy–Schwarz)}$$

$$= \frac{1}{\mathrm{Var}(W)} \cdot \left(\sum_{\omega \in \mathrm{supp}(W)} \omega \cdot \Pr[W=\omega] \cdot \frac{\Pr[X=\omega]-\Pr[Y=\omega]}{2 \cdot \Pr[W=\omega]} \right)^2$$

$$(\mathbb{E}[W] = 0 \text{ and definition of } f)$$

$$= \frac{1}{4} \cdot \frac{(\mathbb{E}[X]-\mathbb{E}[Y])^2}{\mathrm{Var}(W)}.$$

□

5 Bonus Material: Proof of Theorem 1.1 in the Case $\epsilon > \epsilon_{\mathsf{c}}$

Now suppose $\epsilon < \epsilon_{\mathsf{c}}$, i.e. $d\theta^2 < 1$. Our goal is to show that $b^*\left(n, \widehat{\mathbb{T}}_d, \epsilon\right) \to 0$ as $n \to \infty$. Towards this, observe that we can apply Bayes' Rule to obtain that

$$\begin{aligned}
&\mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = +1\right), \mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = -1\right)\right) \\
&= \frac{1}{2} \sum_{\tau \in \{\pm 1\}^{L(n)}} \left|\Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = +1\right] - \Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = -1\right]\right| \\
&= \frac{1}{2} \sum_{\tau \in \{\pm 1\}^{L(n)}} \left|\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right] \frac{\Pr[\sigma_{L(n)} = \tau]}{\Pr[\sigma_r = +1]} - \Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = -1\right] \frac{\Pr[\sigma_{L(n)} = \tau]}{\Pr[\sigma_r = -1]}\right| \\
&= \mathbb{E}\left|\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)}\right] - \Pr\left[\sigma_r = -1 \mid \sigma_{L(n)}\right]\right|.
\end{aligned}$$

This suggests that we look at the quantity

$$\begin{aligned}
\mathscr{M}_n(\tau) &\stackrel{\mathsf{def}}{=} \Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right] - \Pr\left[\sigma_r = +1 \mid \sigma_{L(n)}\right] \\
&= \mathbb{E}\left[\sigma_r \mid \sigma_{L(n)} = \tau\right], \qquad \forall \tau \in \{\pm 1\}^{L(n)}.
\end{aligned}$$

We write $\mathscr{M}_n$ for the corresponding random variable, where the randomness comes from drawing $\sigma_{L(n)}$ according to the broadcast process. Due to connections with the *(ferromagnetic) Ising model*, this quantity is sometimes referred to as the *magnetization* of the root vertex. Our goal will be to show that $\mathbb{E}\left|\mathscr{M}_n\right|$ decays as $n \to \infty$. Rather than study $|\mathscr{M}_n|$, we will instead show step-wise decay for the second moment, which is more tractable.

Lemma 5.1. *We have the inequality* $\mathbb{E}\left|\mathscr{M}_n\right| \leq \sqrt{\mathbb{E}\left[\mathscr{M}_n^2\right]}$.

Proof. The claim is equivalent to nonnegativity of the variance of $|\mathscr{M}_n|$. □

Our goal will be to show that

$$\mathbb{E}\left[\mathscr{M}_n^2\right] \leq d\theta^2 \cdot \mathbb{E}\left[\mathscr{M}_{n-1}^2\right], \tag{1}$$

since our assumption that $d\theta^2 < 1$ implies that $\mathbb{E}\left|\mathscr{M}_n\right| \leq \sqrt{\mathbb{E}\left[\mathscr{M}_n^2\right]}$ is decaying to zero exponentially fast as $n \to \infty$.

The benefit of switching to the viewpoint of $\mathscr{M}_n$ is that we can use recursion to understand the probabilities $\Pr[\sigma_r = +1 \mid \sigma_{L(n)} = \tau]$ and $\Pr[\sigma_r = -1 \mid \sigma_{L(n)} = \tau]$. To state this precisely, let $u_1, \ldots, u_d$ be the children of the root r, and for each $i = 1, \ldots, d$, let $L_i(n-1) \subseteq L(n)$ denote the vertices in the subtree rooted at u_i which are at distance $n-1$ from u_i. For $\tau \in \{\pm 1\}^{L(n)}$, we write $\tau_i \in \{\pm 1\}^{L_i(n-1)}$ for the restriction of τ to $L_i(n-1)$. Now let $\mathscr{M}_{n-1}^{(i)}(\tau_i)$ be given by

$$\mathscr{M}_{n-1}^{(i)}(\tau_i) \stackrel{\mathsf{def}}{=} \Pr\left[\sigma_{u_i} = +1 \mid \sigma_{L_i(n-1)} = \tau_i\right] - \Pr\left[\sigma_{u_i} = -1 \mid \sigma_{L_i(n-1)} = \tau_i\right].$$

Note the corresponding random variables $\mathscr{M}_{n-1}^{(1)}, \ldots, \mathscr{M}_{n-1}^{(d)}$ are correlated through the random root assignment σ_r if $(\tau_1, \ldots, \tau_d) \sim \mu_n$. However, they become independent if we fix the value of

σ_r, i.e. if $(\tau_1, \ldots, \tau_d) \sim \mu_n^{\pm 1}$. What will be important for us is that $\mathsf{Law}\left(\mathscr{M}_{n-1}^{(i)}\right) = \mathsf{Law}\left(\mathscr{M}_{n-1}\right)$ individually for each $i = 1, \ldots, d$. This is because marginally, σ_{u_i} is distributed as $\mathsf{Unif}\{\pm 1\}$ and the assignment $\sigma_{L_i(n-1)}$ is independent of σ_r given σ_{u_i}, so $\mathsf{Law}\left(\sigma_{L_i(n-1)}\right) = \mathsf{Law}\left(\sigma_{L(n-1)}\right)$.

Finally, let $f(x) \stackrel{\mathsf{def}}{=} \frac{1-x}{1+x}$, and for $d \in \mathbb{N}$, define

$$g_d\left(x_1, \ldots, x_d\right) \stackrel{\mathsf{def}}{=} f\left(\prod_{i=1}^{d} f(x_i)\right).$$

We have the following theorem.

Theorem 5.2. *For any $\tau \in \{\pm 1\}^{L(n)}$,*

$$\mathscr{M}_n(\tau) = g_d\left(\theta \cdot \mathscr{M}_{n-1}^{(1)}(\tau_1), \ldots, \theta \cdot \mathscr{M}_{n-1}^{(d)}(\tau_d)\right). \tag{2}$$

In particular, we have the distributional recursion

$$\mathscr{M}_n \stackrel{\mathsf{d}}{\equiv} g_d\left(\theta \cdot \mathscr{M}_{n-1}^{(1)}, \ldots, \theta \cdot \mathscr{M}_{n-1}^{(d)}\right). \tag{3}$$

Let us content ourselves for the moment with an informal explanation of how this recursion is derived; we formally prove this in Appendix A. The idea is that because $\sigma_{L_i(n-1)}$ is independent of σ_r given the value of σ_{u_i}, we can express $\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right]$ as a function of the probabilities $\Pr\left[\sigma_{u_i} = \pm 1 \mid \sigma_{L_i(n-1)} = \tau_i\right]$, for any $\tau \in \{\pm 1\}^{L(n)}$. The specific function is known in the literature as *belief propagation*, the *tree recursion*, or the *sum-product algorithm*, and is extremely well-studied. Typically, it is discussed in the context of spin systems like the Ising model; we derive it using Bayes' Rule in Appendix A and postpone a more in-depth treatment of spin systems to a future lecture.

Taking Theorem 5.2 as a given, to establish an inequality like Eq. (1), our goal is now to bound the rather complicated right-hand side of Eq. (3) by something like

$$\mathbb{E}\left[f\left(\prod_{i=1}^{d} f\left(\theta \cdot \mathscr{M}_{n-1}^{(i)}\right)\right)^2\right] \leq \theta^2 \sum_{i=1}^{d} \mathbb{E}\left[\left(\mathscr{M}_{n-1}^{(i)}\right)^2\right] = d\theta^2 \cdot \mathbb{E}\left[\mathscr{M}_{n-1}^2\right]. \tag{4}$$

With this as our aim, let us begin by massaging Eq. (3). We build up g_d inductively.

Lemma 5.3. *For every $d \in \mathbb{N}$, we have that*

$$g_d\left(x_1, \ldots, x_d\right) = g_2\left(g_{d-1}\left(x_1, \ldots, x_{d-1}\right), x_d\right),$$

where g_2 is explicitly given by the expression

$$g_2(x, y) = \frac{1 - \frac{1-x}{1+x} \cdot \frac{1-y}{1+y}}{1 + \frac{1-x}{1+x} \cdot \frac{1-y}{1+y}} = \frac{x+y}{1+xy}.$$

Proof. This is a straightforward calculation, taking advantage of the fact that $f(f(x)) = x$. $\square$

The combinatorial interpretation of the conjunction of Theorem 5.2 and Lemma 5.3 is the following. Imagine we inductively build up our infinite d-ary tree $\widehat{\mathbb{T}}_d$ rooted at r by first building d trees $T_1, \ldots, T_d$, where each tree T_i is an infinite d-ary tree rooted at u_i along with a single edge joining u_i to a special leaf vertex r_i. Independently in each T_i, we can run the same broadcast process initialized from r_i. A direct calculation reveals that $\theta \cdot \mathbb{E}\left|\mathscr{M}_{n-1}^{(i)}\right|$ is precisely the total variation distance/optimal advantage $b^*\left(n, T_i, \epsilon\right)$ with respect to T_i. If we merge the vertices $r_1, \ldots, r_d$ into a single vertex, then we recover $\widehat{\mathbb{T}}_d$. The purpose of the function g_d is to compute the effect this merge operation induces on the reconstruction probability for the root r. Lemma 5.3 captures the step-by-step effect of merging r_1, r_2, then merging r_1, r_2, r_3, etc.

Towards an inequality analogous to Eq. (4), we would like to establish that the factor $\frac{1}{1+xy}$ is at most 1 in absolute value, at least in expectation when we plug in our random variables. This expression looks unwieldy, so let us attempt to "linearize" g_2. Using the identity $\frac{1}{1+r} = 1 - r + \frac{r^2}{1+r}$, we have that

$$\begin{aligned} g_2(x,y) &= (x+y)\left(1 - xy + \frac{x^2y^2}{1+xy}\right) = (x+y) - x^2y - xy^2 + x^2y^2 \cdot g_2(x,y) \\ &\leq (x+y) - x^2y - xy^2 + x^2y^2, && \text{(For } x, y \geq -1\text{)} \end{aligned}$$

where in the final step, we used the fact that $f([-1,+\infty]) = \mathbb{R}_{\geq 0} \cup \{+\infty\}$; in particular, -1 is not in the range of f in our context. Our goal will be to ensure that the terms $-x^2y - xy^2 + x^2y^2$ are negative, at least in expectation, when we plug in our random variables. For this, we need one final rather curious set of identities.

Lemma 5.4. *Recalling the notation $\mu_n^+ = \mathsf{Law}\left(\sigma_{L(n)} \mid \sigma_r = +1\right)$ and $\mu_n = \mathsf{Law}\left(\sigma_{L(n)}\right)$, we have the identities*

$$\begin{aligned} \mathbb{E}_{\sigma_{L(n)} \sim \mu_n^+}\left[\mathscr{M}_n\right] &= \mathbb{E}_{\sigma_{L(n)} \sim \mu_n}\left[\mathscr{M}_n^2\right] \\ \mathbb{E}_{\sigma_{L(n)} \sim \mu_n^+}\left[\mathscr{M}_{n-1}^{(i)}\right] &= \theta \cdot \mathbb{E}_{\sigma_{L(n-1)} \sim \mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right] \\ \mathbb{E}_{\sigma_{L(n)} \sim \mu_n^+}\left[\left(\mathscr{M}_{n-1}^{(i)}\right)^2\right] &= \mathbb{E}_{\sigma_{L(n-1)} \sim \mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right], \end{aligned}$$

for any $i = 1, \ldots, d$.

We prove Lemma 5.4 at the end of this section. With these tools in hand, we prove our final contraction inequality. We do the case $d = 2$ here (with suppressed notation); the general case follows by a straightforward induction, using Lemma 5.3 to add one child subtree at a time. We have

$$\begin{aligned} \mathbb{E}_{\mu_n}\left[\mathscr{M}_n^2\right] &= \mathbb{E}_{\mu_n^+}\left[\mathscr{M}_n\right] && \text{(Lemma 5.4)} \\ &= \mathbb{E}_{\mu_n^+}\left[g_2\left(\theta \cdot \mathscr{M}_{n-1}^{(1)}, \theta \cdot \mathscr{M}_{n-1}^{(2)}\right)\right] && \text{(Using Eq. (2))} \\ &\leq \theta \cdot \mathbb{E}_{\mu_n^+}\left[\mathscr{M}_{n-1}^{(1)}\right] + \theta \cdot \mathbb{E}_{\mu_n^+}\left[\mathscr{M}_{n-1}^{(2)}\right] \\ &\quad - \theta^3 \cdot \mathbb{E}_{\mu_n^+}\left[\left(\mathscr{M}_{n-1}^{(1)}\right)^2\right] \cdot \mathbb{E}_{\mu_n^+}\left[\mathscr{M}_{n-1}^{(2)}\right] - \theta^3 \cdot \mathbb{E}_{\mu_n^+}\left[\mathscr{M}_{n-1}^{(1)}\right] \cdot \mathbb{E}_{\mu_n^+}\left[\left(\mathscr{M}_{n-1}^{(2)}\right)^2\right] \\ &\quad + \theta^4 \cdot \mathbb{E}_{\mu_n^+}\left[\left(\mathscr{M}_{n-1}^{(1)}\right)^2\right] \cdot \mathbb{E}_{\mu_n^+}\left[\left(\mathscr{M}_{n-1}^{(2)}\right)^2\right] && \text{(Independence under } \mu_n^+\text{)} \\ &= 2\theta^2 \cdot \mathbb{E}_{\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right] - \theta^4 \cdot \mathbb{E}_{\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right]^2 && \text{(Lemma 5.4)} \\ &\leq 2\theta^2 \cdot \mathbb{E}_{\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right]. \end{aligned}$$

All that remains to complete the proof of Theorem 1.1 in the $\epsilon > \epsilon_{\mathsf{c}}$ regime is to prove Lemma 5.4 and Theorem 5.2. The former is proved here, while the latter is relegated to Appendix A.

Proof of Lemma 5.4. For this first identity, observe that

$$\begin{aligned} \mathbb{E}_{\sigma_{L(n)} \sim \mu_n^+}\left[\mathscr{M}_n\right] &= \mathbb{E}_{\sigma_{L(n)} \sim \mu_n}\left[\frac{\Pr\left[\sigma_{L(n)} \mid \sigma_r = +1\right]}{\Pr\left[\sigma_{L(n)}\right]}\left(\Pr\left[\sigma_{L(n)} \mid \sigma_r = +1\right] - \Pr\left[\sigma_{L(n)} \mid \sigma_r = -1\right]\right)\right] \\ & && \text{(Change of densities)} \\ &= \mathbb{E}_{\sigma_{L(n)} \sim \mu_n}\left[2\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)}\right]\left(\Pr\left[\sigma_{L(n)} \mid \sigma_r = +1\right] - \Pr\left[\sigma_{L(n)} \mid \sigma_r = -1\right]\right)\right] \\ & && \text{(Bayes' Rule)} \\ &= \mathbb{E}_{\sigma_{L(n)} \sim \mu_n}\left[\mathscr{M}_n^2 + \mathscr{M}_n\right] && (*) \\ &= \mathbb{E}_{\sigma_{L(n)} \sim \mu_n}\left[\mathscr{M}_n^2\right]. && \text{(Using } \mathbb{E}_{\sigma_{L(n)} \sim \mu_n}\left[\mathscr{M}_n\right] = 0 \text{ by symmetry)} \end{aligned}$$

For $(*)$, we used the fact that $\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)}\right] + \Pr\left[\sigma_r = -1 \mid \sigma_{L(n)}\right] = 1$ implies

$$2\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)}\right] = 1 + \left(\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)}\right] - \Pr\left[\sigma_r = -1 \mid \sigma_{L(n)}\right]\right).$$

This proves the first identity. For the second identity, we use the Law of Total Expectation to obtain

$$\begin{aligned}\mathbb{E}_{\sigma_{L(n)}\sim\mu_n^+}\left[\mathscr{M}_{n-1}^{(i)}\right] &= \theta\cdot\mathbb{E}\left[\mathscr{M}_{n-1}^{(i)} \mid u_i \text{ copies } r \text{ directly}\right] + (1-\theta)\cdot\mathbb{E}\left[\mathscr{M}_{n-1}^{(i)} \mid \sigma_{u_i}\sim \mathsf{Unif}\{\pm1\}\right]\\ &= \theta\cdot\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}^+}\left[\mathscr{M}_{n-1}\right] + (1-\theta)\cdot\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}}\left[\mathscr{M}_{n-1}\right]\\ &= \theta\cdot\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right]. \qquad \text{(First identity)}\end{aligned}$$

For the final identity, we again use the Law of Total Expectation, similar to the above, to obtain

$$\mathbb{E}_{\sigma_{L(n)}\sim\mu_n^+}\left[\left(\mathscr{M}_{n-1}^{(i)}\right)^2\right] = \theta\cdot\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}^+}\left[\mathscr{M}_{n-1}^2\right] + (1-\theta)\cdot\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right].$$

But $\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}^+}\left[\mathscr{M}_{n-1}^2\right] = \mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}^-}\left[\mathscr{M}_{n-1}^2\right] = \mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right]$ by symmetry of the assignments ±1, and so the above is equal to $\mathbb{E}_{\sigma_{L(n-1)}\sim\mu_{n-1}}\left[\mathscr{M}_{n-1}^2\right]$. □

A Proof of Theorem 5.2

We express $\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right]$ in terms of the quantities $\Pr\left[\sigma_{u_i} = \pm1 \mid \sigma_{L_i(n-1)} = \tau_i\right]$. To do this, first observe that because the assignments $\sigma_{L_1(n-1)},\ldots,\sigma_{L_d(n-1)}$ are independent conditioned on the value of σ_r, we have

$$\Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = +1\right] = \prod_{i=1}^{d}\Pr\left[\sigma_{L_i(n-1)} = \tau_i \mid \sigma_r = +1\right].$$

Now, using the Law of Total Probability,

$$\begin{aligned}&\Pr\left[\sigma_{L_i(n-1)} = \tau_i \mid \sigma_r = +1\right]\\ &= (1-\epsilon)\cdot\Pr\left[\sigma_{L_i(n-1)} = \tau_i \mid \sigma_{u_i} = +1\right] + \epsilon\cdot\Pr\left[\sigma_{L_i(n-1)} = \tau_i \mid \sigma_{u_i} = -1\right]\\ &= 2\cdot\Pr\left[\sigma_{L_i(n-1)} = \tau_i\right]\cdot\left(1+\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right).\end{aligned}$$

These calculations all hold mutatis mutandis for the case $\sigma_r = -1$. Since

$$\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right] = \frac{\Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = +1\right]}{\Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = +1\right] + \Pr\left[\sigma_{L(n)} = \tau \mid \sigma_r = -1\right]},$$

which follows by Bayes' Rule and $\Pr[\sigma_r = +1] = \Pr[\sigma_r = -1] = 1/2$, plugging in the above calculations and canceling out the factors of $2\cdot\Pr\left[\sigma_{L_i(n-1)} = \tau_i\right]$ yield

$$\begin{aligned}\Pr\left[\sigma_r = +1 \mid \sigma_{L(n)} = \tau\right] &= \frac{\prod_{i=1}^d\left(1+\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right)}{\prod_{i=1}^d\left(1+\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right) + \prod_{i=1}^d\left(1-\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right)}\\ \Pr\left[\sigma_r = -1 \mid \sigma_{L(n)} = \tau\right] &= \frac{\prod_{i=1}^d\left(1-\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right)}{\prod_{i=1}^d\left(1+\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right) + \prod_{i=1}^d\left(1-\theta\cdot\mathscr{M}_{n-1}^{(i)}(\tau_i)\right)}.\end{aligned}$$

This is one way of expressing the *belief propagation* equations, although they are more commonly written in terms of ϵ and the conditional marginal probabilities $\Pr\left[\sigma_{u_i} = \pm1 \mid \sigma_{L_i(n-1)} = \tau_i\right]$ rather than θ and $\mathscr{M}_{n-1}^{(i)}(\tau_i)$. Subtracting these two expressions and dividing both the numerator and denominator by common factors then yield the recursion for $\mathscr{M}_n(\tau)$ stated in Theorem 5.2.