Mixing Rates for the Gibbs Sampler over Restricted Boltzmann Machines

Christopher Tosh

ctosh@cs.ucsd.edu

Abstract: the Gibbs sampler mixes rapidly (sometimes)

Gibbs sampling is a Markov chain method used to sample from a variety of complicated distributions. We show that in the case of Restricted Boltzmann Machines, the mixing rate of the Gibbs sampler can be bounded both above and below.

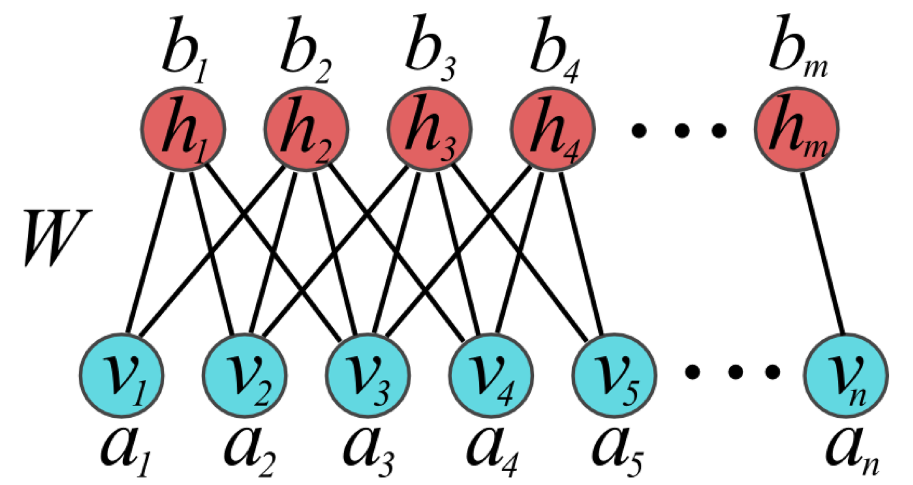
Restricted Boltzmann Machines (RBMs)

Restricted Boltzmann Machines (RBMs):

- Class of undirected bipartite graphical model
- Nodes partitioned into visible layer (n nodes) and hidden layer (m nodes)
- Probability of $(v, h) \in \{0,1\}^{n+m}$:

$$\pi(v,h) = \frac{1}{Z} \exp\left(\sum_{i=1}^{n} a_i v_i + \sum_{j=1}^{m} b_j h_j + \sum_{i,j} v_i W_{ij} h_j \right)$$

where the a_i's and b_j's are biases, the W_{ij}'s are the interaction strengths or weights, and Z is the normalization constant to make π sum to one.



Approximately sampling from π:

- Hard for general weight matrices (Long and Servedio 2010)
- Easy for restricted class of weight matrices (this work)

The Gibbs Sampler

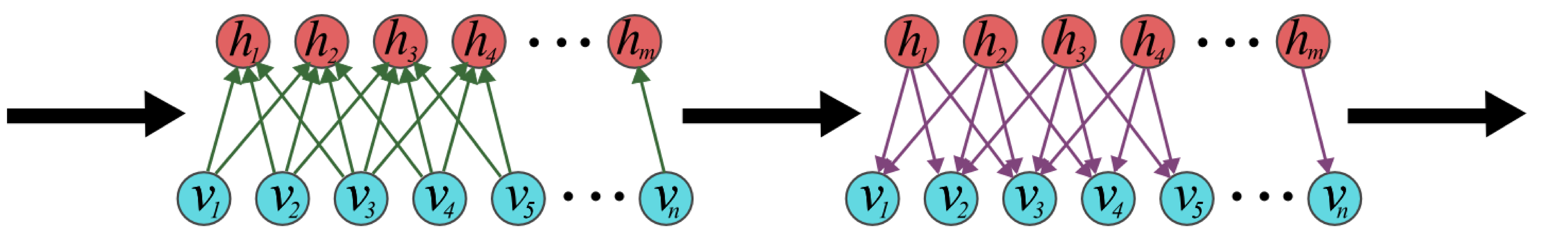
The Gibbs sampler is a Markov chain whose stationary distribution is the distribution $\pi(\cdot)$ from above. It proceeds by starting from an arbitrary configuration $Z_0 \in \{0,1\}^{n+m}$ and then repeating the following.

Sample hidden nodes:
For all hidden h_j, independently draw

$$Z_{t+1}(h_j) \sim \text{Ber}\left(\sigma\left(b_j + \sum_{i=1}^{n} W_{ij}\, Z_t(v_i)\right)\right)$$

Sample visible nodes:
For all visible v_i, independently draw

$$Z_{t+1}(v_i) \sim \text{Ber}\left(\sigma\left(a_i + \sum_{j=1}^{m} W_{ij}\, Z_{t+1}(h_j)\right)\right)$$



Where $\sigma(x) = 1/(1+\exp(-x))$ and $\text{Ber}(p)$ is the Bernoulli distribution with success probability p.
Fact: $Z_t \xrightarrow{d} \pi(v,h)$ as $t \to \infty$.

Mixing Rates

For two measures μ, ν over a discrete state space Ω, the *total variation distance* is half the ℓ_1-distance:

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

For a Markov chain Z_t with stationary distribution π, the *mixing rate* is the minimum number of steps τ_{mix} to lower the total variation distance between the distribution of Z_t and π below 1/4.

Upper Bounds

Theorem 1. *Let a, b, W be an RBM's parameters s.t. $\|W\|_1 \|W^T\|_1 < 4$, then for the Gibbs sampler:*

$$\tau_{mix} \le 1 + \frac{\ln(4n)}{\ln(4) - \ln(\|W\|_1 \|W^T\|_1)}$$

where $\|W\|_1 := \max_j \sum_{i=1}^{n} |W_{ij}|$.

Proof technique: coupling

A *Markovian coupling* of a Markov chain Z_t over Ω with transition matrix P is a Markov chain (X_t, Y_t) over $\Omega \times \Omega$ satisfying

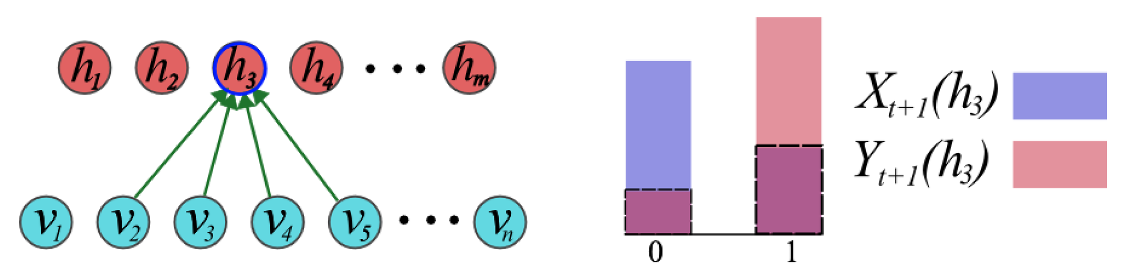
$$Pr(X_{t+1} = x' \mid X_t = x, Y_t = y) = P(x, x'),$$
$$Pr(Y_{t+1} = y' \mid X_t = x, Y_t = y) = P(y, y').$$

Aldous (1983) showed that for any coupling (X_t, Y_t) such that there exists an integer-valued function τ satisfying for all $x, y \in \Omega$ and $\epsilon > 0$,

$$Pr(X_{\tau(\epsilon)} \ne Y_{\tau(\epsilon)} \mid X_0 = x, Y_0 = y) \le \epsilon$$

then Z_t's mixing rate satisfies $\tau_{mix} \le \tau(1/4)$.

Our approach is to couple each node independently, hidden nodes before visible ones. For the example below, the probability that $X_{t+1}(h_3) = Y_{t+1}(h_3)$ is the area of the purple region.



This strategy gives us the following lemma.

Lemma. *There exists a coupling (X_t, Y_t) of the Gibbs sampler such that*

(a) $\mathbb{E}[d_h(X_t', Y_t') \mid X_t, Y_t] \le \frac{1}{2} \|W^T\|_1 d_v(X_t, Y_t)$ *and*
(b) $\mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_t', Y_t'] \le \frac{1}{2} \|W\|_1 d_h(X_t', Y_t')$.

Where (X_t', Y_t') denotes the state immediately after the hidden nodes have been updated; and $d_h(\cdot,\cdot)$ and $d_v(\cdot,\cdot)$ denote Hamming distances over the hidden and visible nodes, respectively.

By the law of total expectation, we have

$$\mathbb{E}\left[d_v(X_{t+1}, Y_{t+1}) \mid X_t, Y_t)\right] \le \frac{\|W\|_1 \|W^T\|_1}{4} d_v(X_t, Y_t).$$

When $\|W\|_1 \|W^T\|_1 < 4$, this distance shrinks in expectation. Markov's inequality finishes the proof.

Lower Bounds

Theorem 2. *Pick any $T > 0$ and $n, m \in \mathbb{N}$ even positive integers. Then there exists $W \in \mathbb{R}^{n \times m}$ s.t.*

$$\|W^T\|_1, \|W\|_1 \le \frac{2\max(n,m)}{\min(n,m)} \ln(4T(n+m))$$

such that the Gibbs sampler over the RBM with no bias and weight matrix W has mixing rate $\tau_{mix} \ge T$.

Proof technique: conductance

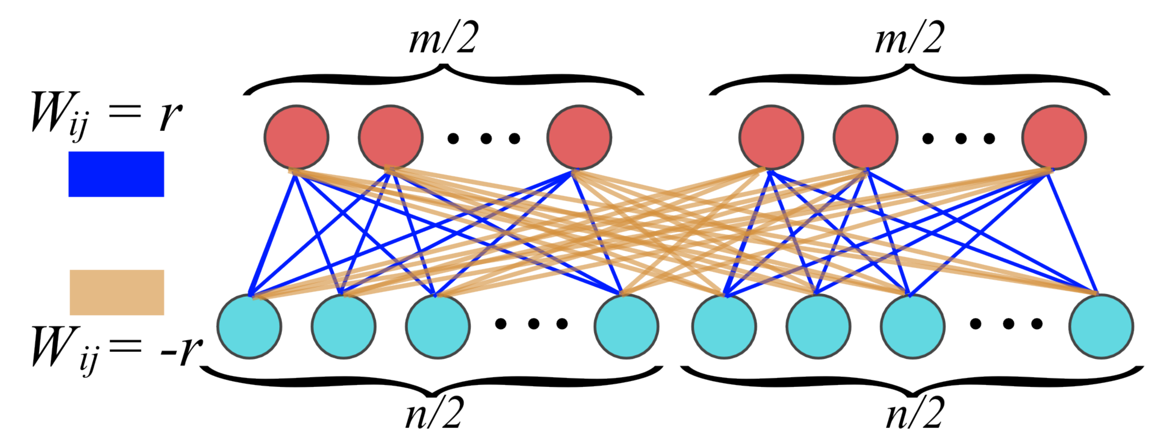
Given a Markov chain P with stationary distribution π and $S \subset \Omega$, the *conductance* of S is

$$\Phi(S) := \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x) P(x,y).$$

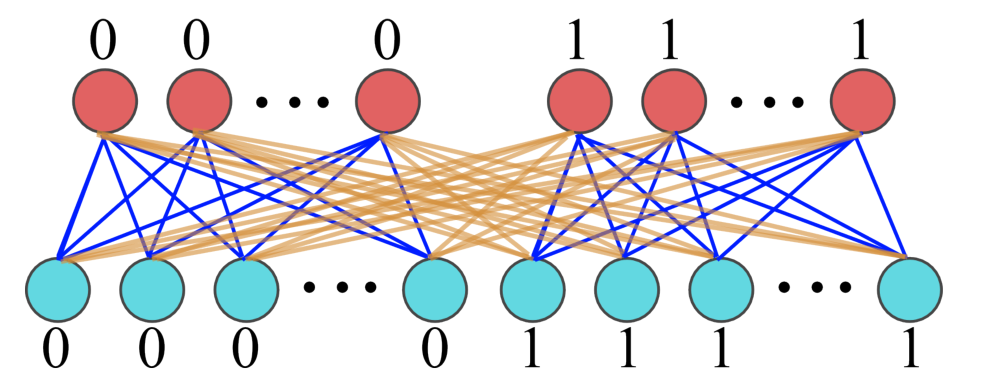
Sinclair (1988) outlines the relationship of conductance and mixing rates as

$$\tau_{mix} \ge \max_{\substack{S \subset \Omega:\\ \pi(S) \le 1/2}} \frac{1}{4\Phi(S)}.$$

Let $r = \frac{2\ln(4T(n+m))}{\min(n,m)}$. We consider an RBM with no bias and weight matrix illustrated below.



When S is the singleton set consisting of the configuration below, we show $\pi(S) \le 1/2$ and $\Phi(S) \le \frac{1}{4T}$.



The conductance theorem gives a lower bound of T on the mixing rate.

Acknowledgements

The author is grateful for generous support from the NSF Graduate Research Fellowship Program under grant DGE-1144086 and for helpful feedback from the anonymous reviewers.