Mixing Rates for the Gibbs Sampler over Restricted Boltzmann Machines

Christopher Tosh
Department of Computer Science and Engineering
University of California, San Diego
ctosh@cs.ucsd.edu

Abstract

The *mixing rate* of a Markov chain $(X_t)_{t=0}^{\infty}$ is the minimum number of steps before the distribution of X_t is close to its stationary distribution with respect to total variation distance. In this work, we give upper and lower bounds for the mixing rate of the Gibbs sampler over Restricted Boltzmann Machines.

1 Introduction

Restricted Boltzmann Machines (RBMs) are an important class of undirected graphical model, the learning of which is integral in many approaches to building deep belief networks [1, 2]. RBMs can be viewed as a fully connected bipartite graph with the two sets of nodes representing a visible layer and a hidden layer, all taking values in $\{0, 1\}$. A configuration $(v, h) \in \{0, 1\}^{n+m}$ has energy

$$E(v, h) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i,j} v_i W_{ij} h_j$$

where the a_i's and b_j's are biases and the W_{ij}'s are the interaction strengths or weights. This energy function induces the Gibbs distribution over configurations: $P(v, h) = \frac{1}{Z} e^{-E(v,h)}$, where Z is the normalizing constant to make the total probability of the distribution one.

For some applications, such as learning the parameters of an RBM from data, we need to compute expectations with respect to the Gibbs distribution, which is a hard task [3]. To overcome this barrier, the approach of *contrastive divergence* [4] is to approximate the expectation by performing Markov chain Monte Carlo (MCMC) integration. The accuracy of this method can be guaranteed if the chosen Markov chain converges rapidly to the Gibbs distribution [5, Chapter 12.6].

1.1 Markov chain theory

A Markov chain is a sequence of random variables $(X_t)_{t=0}^{\infty}$ taking values in some space Ω and satisfying the *Markov property*: $Pr(X_t = x \,|\, X_{t-1}, \ldots, X_0) = Pr(X_t = x \,|\, X_{t-1})$. The transition probabilities can be viewed as a matrix Q indexed by elements of Ω s.t.

$$Q(x, y) = Pr(X_t = y \,|\, X_{t-1} = x).$$

Q is *irreducible* if, for all $x, y \in \Omega$, there exists a $t > 0$ s.t. $Q^t(x, y) > 0$. It is *aperiodic* if

$$\gcd(\{t : Q^t(x, y) > 0\}) = 1 \text{ for all } x, y \in \Omega.$$

A distribution π over Ω is a *stationary distribution* of Q if, when π and P are viewed as matrices indexed by Ω, then $\pi = \pi Q$. A fundamental result of Markov chain theory says that if a Markov chain Q is irreducible and aperiodic, then it has a unique stationary distribution. Furthermore, the distribution of X_t converges to π, regardless of initial distribution [5, Theorem 4.9].

Unfortunately, convergence is only guaranteed in the limit. To get samples in a finite time, we must settle for approximation. Given measures μ, ν over Ω, the *total variation distance* is half the ℓ_1-norm of their difference: $\|\mu - \nu\|_{TV} := \frac{1}{2}\sum_{\omega\in\Omega} |\mu(\omega) - \nu(\omega)|$. The *mixing rate* of a Markov chain Q with unique stationary distribution π is the function $\tau(\epsilon) = \min\{t : \max_{x\in\Omega} \|Q^t(x,\cdot) - \pi\|_{TV} < \epsilon\}$ for $\epsilon \in (0,1)$. Taking ϵ to be any constant less than $1/2$ gives us nontrivial results, but by convention ϵ is often taken to be $1/4$. Thus, where it will cause no confusion, we refer to the mixing time interchangeably with the quantity $\tau_{mix} := \tau(1/4)$.

1.2 Gibbs sampling

In this paper, we will focus on a particular Markov chain whose limiting distribution is the Gibbs distribution above: the Gibbs sampler. The state space $\Omega = \{0,1\}^{n+m}$ is the set of configurations. For convenience, it will be useful to also think of a configuration $X \in \Omega$ as a binary-valued (or binary vector-valued) function, where for example $X(v_i) \in \{0,1\}$ is the setting of the i-th visible node in the RBM and $X(h) \in \{0,1\}^m$ is the setting of all the hidden nodes in the RBM.

Due to the bipartite nature of the RBM, the visible units are independent of each other given the hidden units and vice versa. This makes the implementation of the Gibbs sampler appealingly simple. Denoting $\sigma(x) = 1/(1+\exp(-x))$ as the logistic sigmoid, one step of the Gibbs sampler from state X_t to X_{t+1} can be described as follows.

1. For each hidden node h_j, set $X_{t+1}(h_j)$ to 1 with probability $\sigma(b_j + \sum_{i=1}^n W_{ij} X_t(v_i))$.
2. For each visible node v_i, set $X_{t+1}(v_i)$ to 1 with probability $\sigma(a_i + \sum_{j=1}^m W_{ij} X_{t+1}(h_j))$.

Standard arguments show that the Gibbs sampler is irreducible and aperiodic; moreover, the stationary distribution π is our Gibbs distribution. It will be useful for our purposes to explicitly designate the intermediate state of the Gibbs sampler, after the hidden nodes have been updated but before the visible nodes are updated. Denote this state intermediate state between X_t and X_{t+1} by $X_{t+1/2}$.

Long and Servedio [3] gave strong evidence that even approximately sampling from the Gibbs distribution is hard in general for RBMs. It is therefore unlikely that the Gibbs sampler would mix rapidly in general. However, Long and Servedio's reduction relies on constructing weight matrices whose entries can be quite large. In particular, they allow for the weight matrices W whose max-norm, $\|W\|_{max} := \max_{i,j} |W_{ij}|$, is allowed to grow super-linearly in n and m. This raises the natural question of whether anything can be said for sampling from the Gibbs distribution for RBMs whose weight matrices are smaller in magnitude.

In this paper, we make positive progress in this direction. We are able to show if the weight matrix W and its transpose W^T are sufficiently bounded with respect to its ℓ_1-norm, defined as $\|W\|_1 = \max_j \sum_{i=1}^n |W_{ij}|$, then the Gibbs sampler mixes rapidly. Formally, we demonstrate the following.

Theorem 2.1 *Let $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $W \in \mathbb{R}^{n\times m}$ be the parameters for an RBM s.t. $\|W\|_1\|W^T\|_1 < 4$, then the Gibbs sampler satisfies $\tau_{mix} \le 1 + \frac{\ln(4n)}{\ln(4) - \ln(\|W\|_1\|W^T\|_1)}$.*

The ℓ_1-norm bound on W in Theorem 2.1 can be related to a max-norm bound if we add a sparsity condition. Consider the bipartite graph that W induces and suppose every node has degree bounded by d, then we need the non-zero entries of W to grow like $2/d$ to guarantee rapid mixing. In the extreme case where $d = 1$, we get a max-norm bound of 2. On the other hand, the max-norm bound degenerates to $2/\max(n,m)$ for unrestricted d.

In the other direction, we give lower bounds on the mixing rate for a family of weight matrices.

Theorem 3.1 *Pick any $T > 0$ and $n, m \in \mathbb{N}$ even positive integers. Then there is a weight matrix $W \in \mathbb{R}^{n\times m}$ satisfying $\|W\|_{max} \le \frac{2}{\min(n,m)} \ln(4T(n+m))$ such that the Gibbs sampler over the RBM with zero bias and weight matrix W has mixing rate bounded as $\tau_{mix} \ge T$.*

An immediate consequence of this is that if $n = \Theta(m)$ and $T = 2^{\min(n,m)/2}/4(n+m)$, then there is a weight matrix whose max-norm is 1, but the Gibbs sampler still mixes in time $T = 2^{\Omega(n)}$.

The rest of the paper is organized as follows. In Section 2, we give a proof for Theorem 2.1. In Section 3, we give a proof of Theorem 3.1. We conclude with a discussion in Section 4.

2 Upper bounds for the Gibbs sampler

The goal of this section is to prove the following theorem.

Theorem 2.1. *Let $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $W \in \mathbb{R}^{n \times m}$ be the parameters for an RBM s.t. $\|W\|_1 \|W^T\|_1 < 4$, then the Gibbs sampler satisfies*

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

We will prove Theorem 2.1 via a coupling argument. A *Markovian coupling* of a Markov chain Z_t over Ω with transition matrix Q is a Markov chain (X_t, Y_t) over $\Omega \times \Omega$ whose transitions satisfy

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

The following lemma relates couplings of a Markov chain to the mixing time. It dates back at least to Aldous [6] and can be found in the form we present, for example, in Jerrum [7, Lemma 4.7].

Lemma 2.2. *Let (X_t, Y_t) be a Markovian coupling for Markov chain Z_t such that there exists a function $\tau_{couple} : (0,1) \to \mathbb{N}$ satisfying that for all $x, y \in \Omega$ and $\epsilon > 0$, $Pr(X_{\tau_{couple}(\epsilon)} \neq Y_{\tau_{couple}(\epsilon)} \mid X_0 = x, Y_0 = y) \leq \epsilon$. Then the mixing rate for Z_t satisfies $\tau_{mix} \leq \tau_{couple}(1/4)$.*

Let us now specialize to the Gibbs sampler. Let Ω denote as before the space of configurations. To apply Lemma 2.2 to our situation, it will be useful to define some distances over Ω. For $X, Y \in \Omega$, let $d_v(X, Y) = |\{v_i : X(v_i) \neq Y(v_i)\}|$ denote the visible Hamming distance. Similarly, define $d_h(X, Y) = |\{h_j : X(h_j) \neq Y(h_j)\}|$ as the hidden Hamming distance.

Recall that for the Gibbs sampler, $X_{t+1/2}$ denotes the intermediate state after the hidden nodes have been updated but before the visible nodes have been updated. The following lemma demonstrates that there exists a coupling whose Hamming distance shrinks in expectation after one step.

Lemma 2.3. *Let a, b, W satisfy the conditions of Theorem 2.1. There exists a Markovian coupling (X_t, Y_t) of the Gibbs sampler such that*

1. $\mathbb{E}[d_h(X_{t+1/2}, Y_{t+1/2}) \mid X_t, Y_t] \leq \frac{1}{2}\|W^T\|_1 d_v(X_t, Y_t)$ *and*
2. $\mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_{t+1/2}, Y_{t+1/2}] \leq \frac{1}{2}\|W\|_1 d_h(X_{t+1/2}, Y_{t+1/2})$.

The proof of Lemma 2.2 is left to the appendix. We are now ready to prove Theorem 2.1 .

Proof of Theorem 2.1. Let (X_t, Y_t) be the coupling from Lemma 2.3. Note that this implies that if $d_v(X_s, Y_s) = 0$, then $X_t = Y_t$ for all $t \geq s + 1$. Thus, for any $t \geq 1$,

$$Pr(X_t \neq Y_t \mid X_0, Y_0) \leq Pr(d_v(X_{t-1}, Y_{t-1}) \geq 1 \mid X_0, Y_0).$$

By conditioning on the intermediate state $(X_{t+1/2}, Y_{t+1/2})$ and using the law of total expectation,

$$\mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_t, Y_t] = \mathbb{E}[\mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_{t+1/2}, Y_{t+1/2}] \mid X_t, Y_t] \leq \frac{1}{4}\|W\|_1 \|W^T\|_1 d_v(X_t, Y_t).$$

By applying Markov's inequality and iterating the law of total expectation, we have

$$Pr(d_v(X_{t-1}, Y_{t-1}) \geq 1 \mid X_0, Y_0) \leq \mathbb{E}\left[d_v(X_{t-1}, Y_{t-1}) \mid X_0, Y_0\right] \leq \left(\frac{1}{4}\|W^T\|_1 \|W\|_1\right)^{t-1} d_v(X_0, Y_0).$$

For any $\epsilon > 0$, taking $t \geq 1 + \ln(n/\epsilon)(\ln(4) - \ln(\|W\|_1 \|W^T\|_1))^{-1}$ makes the above less than ϵ. Lemma 2.2 completes the proof. □

3 Lower bounds for the Gibbs sampler

We next turn to giving lower bounds on the mixing rate for the Gibbs sampler. These lower bounds are of a much different flavor than the upper bounds from Section 2. Specifically, while Theorem 2.1 tells us that *any* RBM with a weight matrix satisfying certain conditions has a rapidly mixing Gibbs sampler, the following theorem tells us there *exists* an RBM meeting some condition for which the Gibbs sampler is torpidly mixing.

Theorem 3.1. *Pick any $T > 0$ and $n, m \in \mathbb{N}$ even positive integers. Then there is a weight matrix $W \in \mathbb{R}^{n \times m}$ satisfying $\|W\|_{max} \leq \frac{2}{\min(n,m)} \ln(4T(n+m))$ such that the Gibbs sampler over the RBM with zero bias and weight matrix W has mixing rate bounded as $\tau_{mix} \geq T$.*

To prove Theorem 3.1, we will utilize the method of conductance. Given a Markov chain Q, its stationary distribution π, and a subset $S \subset \Omega$, the *conductance* of S is

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

and the conductance of P, denoted by Φ^*, is the minimum conductance of any set S with $\pi(S) \leq 1/2$. The following theorem, due to Sinclair [9], relates mixing and conductance of a Markov chain.

Theorem 3.2 (Sinclair [9])**.** *For any aperiodic, irreducible, and reversible Markov chain with conductance Φ^* and mixing time τ_{mix}, $\tau_{mix} \geq \frac{1}{4\Phi^*}$.*

We are now ready to prove Theorem 3.1. In the proof, we construct a weight matrix W such that the energy function associated with W has two antipodal absolute minima. The fact that there are two minima means that the singleton set consisting of one of the minima has probability mass less than 1/2 under the Gibbs distribution. Thus, in order to mix rapidly, the Gibbs sampler will need to visit both minima. We will then show that escaping from one of these minima is a very unlikely event.

Proof of Theorem 3.1. Let $r = \frac{2}{\min(n,m)} \ln(4T(n+m))$. Choose a canonical configuration $(v, h) \in \{0,1\}^{n+m}$ such that exactly half of the v_i's are 1 and exactly half of the h_j's are 1. Now let $W \in \mathbb{R}^{n \times m}$ such that $W_{ij} = r$ if $v_i = h_j$ and $-r$ otherwise. Let $\pi(\cdot)$ denote the Gibbs distribution for the RBM with weight matrix W and zero bias and let $S = \{(v, h)\}$ be the singleton set containing only the canonical configuration. Note that if $(\bar{v}, \bar{h}) \in \{0,1\}^{n+m}$ satisfies that $\bar{v}_i = 1$ iff $v_i = 0$ and $\bar{h}_j = 1$ iff $h_j = 0$, then $\pi(v, h) = \pi(\bar{v}, \bar{h})$. Thus, $\pi(S) \leq 1/2$.

It is not hard to see $Pr(\text{change } h_j \mid v) = \sigma\left(-\frac{nr}{2}\right)$ for all $j \in [m]$, where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic sigmoid as before. Similarly, for any $i \in [n]$, $Pr(\text{change } v_i \mid h) = \sigma\left(-\frac{mr}{2}\right)$. Thus,

$$Pr(\text{leave state } (v, h)) \leq \frac{m}{1 + \exp\left(\frac{nr}{2}\right)} + \frac{n}{1 + \exp\left(\frac{mr}{2}\right)} \leq \frac{1}{4T}$$

Thus the conductance of S (and therefore Φ^*) is upper bounded as

$$\Phi(S) = \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x) Pr(\text{we transition from } x \text{ to } y) = Pr(\text{leave state } (v, h)) \leq \frac{1}{4T}$$

Theorem 3.2 completes the proof. □

4 Discussion

We have presented both upper and lower bounds on the mixing rates of the Gibbs sampler for RBMs. In the case of upper bounds, we demonstrated that a simple coupling argument suffices whenever the ℓ_1-norm of the weight matrix is suitably bounded. By way of lower bounds, we gave a particular family of RBMs for which the Gibbs sampler performs poorly even when the weight matrix has a moderately sized max-norm.

One important takeaway of this work is the impact of the ℓ_1-norm of the weight matrix on the mixing rate of the Gibbs sampler. Indeed, when n and m differ by only a constant factor Theorem 3.1 gives a lower bound of $2^{\Omega(\|W\|_1)}/n$. Unfortunately, this lower bound is no longer meaningful when $\|W\|_1 = o(\log(n))$. On the other hand, Theorem 2.1 gives meaningful upper bounds when $\|W\|_1 \|W^T\|_1 < 4$, but does not apply for weight matrices with larger ℓ_1-norms. Closing the gap between what we can prove to mix rapidly and what we can prove to mix torpidly remains an interesting research direction.

Acknowledgments

The author is grateful for support through the NSF Graduate Research Fellowship Program under grant DGE-1144086 and also for the feedback given by the anonymous reviewers.

References

[1] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009.

[2] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[3] Philip M. Long and Rocco A. Servedio. Restricted boltzmann machines are hard to approximately evaluate or simulate. In *ICML*, pages 703–710. Omnipress, 2010.

[4] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[5] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[6] David Aldous. Random walks on finite groups and rapidly mixing markov chains. In *Séminaire de Probabilités XVII 1981/82*, pages 243–297. Springer, 1983.

[7] Mark Jerrum. *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media, 2003.

[8] Geoffrey Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.

[9] Alistair Sinclair. *Randomised Algorithms for Counting and Generating Combinatorial Structures*. PhD thesis, University of Edinburgh, 1988.

[10] Omar Rivasplata. Subgaussian random variables: An expository note. *Internet publication, PDF*, 2012.

A Proofs

Lemma 2.3. *Let a, b, W satisfy the conditions of Theorem 2.1. There exists a Markovian coupling (X_t, Y_t) of the Gibbs sampler such that*

1. $\mathbb{E}[d_h(X_{t+1/2}, Y_{t+1/2}) \mid X_t, Y_t] \leq \frac{1}{2}\|W^T\|_1 d_v(X_t, Y_t)$ *and*
2. $\mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_{t+1/2}, Y_{t+1/2}] \leq \frac{1}{2}\|W\|_1 d_h(X_{t+1/2}, Y_{t+1/2})$.

Proof. Let X, Y be two configurations. Recall the visible and hidden distances are defined as

$$d_v(X, Y) = |\{v_i \,:\, X(v_i) \neq Y(v_i)\}|,$$
$$d_h(X, Y) = |\{h_j \,:\, X(h_j) \neq Y(h_j)\}|.$$

Because we are giving a Markovian coupling, we need to only describe one transition of our coupling. Let us see how we transition from (X_t, Y_t) to (X_{t+1}, Y_{t+1}).

1. For all h_j:
   (a) Say w.l.o.g. $p_X^{(j)} = Pr(h_j = 1 \mid X_t(v)) \geq Pr(h_j = 1 \mid Y_t(v)) = p_Y^{(j)}$.
   (b) Set $Y_{t+1}(h_j)$ to 1 w.p. $p_Y^{(j)}$ and to 0 o/w.
   (c) If $Y_{t+1}(h_j) = 1$, set $X_{t+1}(h_j)$ to 1. Else set $X_{t+1}(h_j)$ to 1 w.p. $p^{(j)} = \frac{p_X^{(j)} - p_Y^{(j)}}{1 - p_Y^{(j)}}$ and 0 o/w.
2. For all v_i:
   (a) Say w.l.o.g. $q_X^{(i)} = Pr(v_i = 1 \mid X_{t+1}(h)) \geq Pr(v_i = 1 \mid Y_{t+1}(h)) = q_Y^{(i)}$.
   (b) Set $Y_{t+1}(v_i)$ to 1 w.p. $q_Y^{(i)}$ and to 0 o/w.
   (c) If $Y_{t+1}(v_i) = 1$, set $X_{t+1}(v_i)$ to 1. Else set $X_{t+1}(v_i)$ to 1 w.p. $q^{(i)} = \frac{q_X^{(i)} - q_Y^{(i)}}{1 - q_Y^{(i)}}$ and 0 o/w.

To see that this is a valid coupling, we need to verify that the marginal distributions are indeed correct. We will only look at the hidden variable updates since the visible variable updates are symmetric. Fix an index h_j. It is clear that for the Y chain (or whichever chain has lower probability of setting $h_j = 1$), the marginal distribution is correct. To see that the X chain follows the correct distribution, note

$$\begin{aligned} Pr(X_{t+1}(h_j) = 1) &= Pr(Y_{t+1}(h_j) = 1) + (1 - Pr(Y_{t+1}(h_j) = 1))p^{(j)} \\ &= p_Y^{(j)} + (1 - p_Y^{(j)})\frac{p_X^{(j)} - p_Y^{(j)}}{1 - p_Y^{(j)}} \\ &= p_X^{(j)}. \end{aligned}$$

Thus we have a valid coupling. We will show that (X_t, Y_t) satisfies inequality (2); inequality (1) follows by symmetrical arguments. By utilizing the independence of visible nodes given hidden nodes, we have

$$\begin{aligned} \mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_{t+1/2}, Y_{t+1/2}] &= \sum_{i=1}^{n} Pr(X_{t+1}(v_i) \neq Y_{t+1}(v_i) \mid X_{t+1/2}, Y_{t+1/2}) \\ &= \sum_{i=1}^{n} \left| q_X^{(i)} - q_Y^{(i)} \right|. \end{aligned}$$

Through simple algebraic manipulations, we have

$$\begin{aligned}
\left|q_X^{(i)} - q_Y^{(i)}\right| &= \left| \frac{1}{1+\exp\left(-a_i - \sum_{j=1}^m W_{ij} X_{t+1/2}(h_j)\right)} - \frac{1}{1+\exp\left(-a_i - \sum_{j=1}^m W_{ij} Y_{t+1/2}(h_j)\right)} \right| \\
&\leq \left| \frac{1-\exp\left(\sum_{j=1}^m W_{ij}\left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right)\right)}{1+\exp\left(\sum_{j=1}^m W_{ij}\left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right)\right)} \right| \\
&= \left| \tanh\left(\frac{\sum_{j=1}^m W_{ij}\left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right)}{2} \right) \right| \\
&\leq \frac{1}{2} \left| \sum_{j=1}^m W_{ij}\left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right) \right|.
\end{aligned}$$

Summing over the visible nodes v_i, we have

$$\begin{aligned}
\mathbb{E}[d_v(X_{t+1}, Y_{t+1}) \mid X_{t+1/2}, Y_{t+1/2}] &\leq \frac{1}{2} \sum_{i=1}^n \left| \sum_{j=1}^m W_{ij}\left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right) \right| \\
&\leq \frac{1}{2} \sum_{j\,:\,Y_{t+1/2}(h_j) \neq X_{t+1/2}(h_j)} \sum_{i=1}^n |W_{ij}| \\
&\leq \frac{1}{2} \|W\|_1 d_h(X_{t+1/2}, Y_{t+1/2}).
\end{aligned}$$

Inequality (1) can be proven symmetrically. □