Conductance and Rapidly Mixing Markov Chains

Jamie King
james.king@uwaterloo.ca

March 26, 2003

Abstract

Conductance is a measure of a Markov chain that quantifies its tendency to circulate around its states. A Markov chain with low conductance will tend to get 'stuck' in a subset of its states whereas one with high conductance will jump around its state space more freely. The mixing time of a Markov chain is the number of steps required for the chain to approach its stationary distribution. There is an inverse correlation between conductance and mixing time. Rapidly mixing Markov chains have very powerful applications, most notably in approximation schemes. It is therefore desirable to prove certain Markov chains to be rapidly mixing, and bounds involving conductance can help us do that. This survey covers many useful bounds involving conductance and gives several specific examples of applications of rapidly mixing Markov chains.

1 Introduction

In this survey we discuss the conductance of Markov chains and its relationship to rapid mixing. The *conductance* of a Markov chain is a scalar measure of the tendency of a Markov chain to move out of a subset of states. The *mixing time* of a Markov chain is the number of steps the Markov chain must make before its probability distribution reaches the stationary distribution (also commonly referred to as the steady-state distribution). Intuitively, high conductance in a Markov chain would imply that the chain has a relatively low mixing time, i.e. it is *rapidly mixing*. There are certain combinatorial tasks that can be computed more easily given a certain Markov chain at its stationary distribution. This makes rapidly mixing Markov chains useful combinatorial tools and gives motivation for proving that certain chains are rapidly mixing. Many of the early results in the field of rapidly mixing Markov chains are due to Aldous in the early to mid 80's [Ald82], [Ald83].

At this point we will introduce the notation used in this survey. We will use V to denote the set of states of a Markov chain. We also have the transition matrix $P = \{p_{ij}\}_{i,j \in V}$ of a Markov chain such that the probability of transition from state i to state j is equal to p_{ij}.

We only consider *ergodic* Markov chains for a reason that is clear given the definition of ergodicity.

Ergodicity Theorem A Markov chain is *ergodic* if and only if there exists a distribution π such that $\pi P = \pi$ and, for any distribution τ,

$$\lim_{n \to \infty} \tau P^n = \pi.$$

In other words, a Markov chain is ergodic if and only if there is a stationary distribution that will always be reached eventually. Note that a Markov chain at its stationary distribution will never achieve any other distribution. Also note that the stationary distribution of any ergodic Markov chain is unique. We can denote the smallest value in π as π_0. In an ergodic Markov chain, $\pi_0 > 0$.

We also only consider Markov chains that are *time-reversible*. A Markov chain is time-reversible if and only if

$$\pi_i p_{ij} = \pi_j p_{ji}, \forall i, j \in V.$$

In other words, in a time-reversible Markov chain at its stationary distribution, the probability of the next transition being from state i to state j is the same as the probability of the next transition being from state j to state i. We will call this probability Q_{ij}. Here we will also define the *ergodic flow* of two subsets of vertices $S_1,\ S_2$:

$$Q(S_1, S_2) = \sum_{i \in S_1, j \in S_2} Q_{ij}.$$

This is the probability that, in a time-reversible Markov chain at its stationary distribution, the next transition is from an element in S_1 to an element in S_2.

We can consider a time-reversible ergodic Markov chain in terms of its underlying graph G, a complete directed weighted graph including a loop from each vertex to itself. Each vertex in G corresponds to a state in the Markov chain and each edge e_{ij} has weight $w_{ij} = Q_{ij}$. This is a very useful representation of a time-reversible ergodic Markov chain, as it allows us to utilise graph theoretic techniques and ideas. For example, $Q(S_1, S_2)$ can intuitively be defined as the weight of the coboundary of S_1 and S_2.

Now we can define the *conductance* of a proper subset of states of a Markov chain. The conductance Φ of a subset S of states in a Markov chain is

$$\Phi(S) = \frac{Q(S, \bar{S})}{\pi_S},$$

where $\bar{S}$ is equal to $V - S$ and $\pi(S)$ is the probability that, at the stationary distribution, the Markov chain will be at some state in S.

The conductance of a Markov chain is defined as the minimum conductance over all subsets S with $\pi_S \leq 1/2$, i.e.

$$\begin{aligned}\Phi &= \min_{S \subset V,\ \pi_S \leq 1/2} \Phi(S) \\ &= \min_{S \subset V,\ \pi_S \leq 1/2} \frac{Q(S, \bar{S})}{\pi_S}.\end{aligned}$$

It is this value Φ that has been used to bound the mixing time of Markov chains. The motivation for giving bounds (especially upper bounds) on the mixing time of a Markov chain is to determine how efficiently a Markov chain can be prepared for use as a tool for solving other problems. If a Markov chain's mixing time is too great, it is likely that a problem can be solved more efficiently without using the chain as a tool.

2 Mixing Times

As stated before, the mixing time of a Markov chain is the number of steps the Markov chain must take before its state distribution reaches its stationary distribution. However, this may not always be such a useful definition. Many Markov chains will asymptotically approach the stationary distribution without actually reaching it in a finite number of steps, let alone a small number of steps. Sometimes it is acceptable for the distribution to be simply *close enough* to the stationary distribution. Therefore it is sometimes useful to define the mixing time of a Markov chain as the number of steps required in order for the Markov chain to come close enough to its stationary distribution.

Jerrum and Sinclair [JS88] define the relative pointwise distance after t steps by

$$\Delta(t) = \max_{i,j \in V} \left\{ \frac{|p_{ij}^{(t)} - \pi_j|}{\pi_j} \right\}.$$

where $p_{ij}^{(t)}$ is the t-step transition probability from i to j, equal to $[P^t]_{ij}$. $\Delta(t)$ therefore gives the largest relative difference between the probability of being at state j after t steps and the probability of being at state j in the stationary distribution π, maximised over initial states i. $\Delta(t)$ is a very

useful measure of the distance from the stationary distribution. In particular, giving an upper bound for $\Delta(t)$ can demonstrate that a Markov chain mixes fast enough for certain purposes.

Sinclair [Sin92] formally defines a related rate of convergence

$$\tau_i(\epsilon) = min\{t : \forall\ t', t' \geq t \rightarrow \Delta_i(t') \leq \epsilon\}$$

where a subscript i denotes that i is the initial state. The more general and useful rate of convergence τ is then defined as

$$\tau = \max_i \tau_i(1/4)$$

where the $1/4$ is somewhat arbitrary and could really be any sufficiently small constant. Essentially, τ is the number of steps required to get 'close enough' to the stationary distribution.

Another way to get around the possibly infinite number of steps required to reach the exact stationary distribution is by using a nondeterministic *stopping rule*. A stopping rule is a rule that observes the walk on a Markov chain and, at each step, decides whether or not to stop the walk based on the walk so far. Stopping rules can make probabilistic decisions and are therefore very powerful tools. Usually the random stopping time of a stopping rule is denoted by Γ and the distribution of the final state is denoted by σ^Γ. Ideally, the expected stopping time $E[\Gamma]$ is small. Here we provide an example of a stopping rule for stopping at the stationary distribution π, assuming π is known: stop at a state i the first time it is visited with probability $\pi_i/(1-\pi_S)$ where S is the set of states (excluding i) that have been visited; do not stop at a state that has been previously visited. With this fairly simple rule, we are guaranteed to stop in the distribution π. Unfortunately, this stopping rule is not very efficient and would therefore never be used in practice. Moreover, we do not always know the value of π before we begin the walk on the Markov chain. Sometimes we don't even know the transition probabilities. The task of finding efficient stopping rules for particular graphs is a large area in Markov chain analysis. Some efficient rules are discussed by Lovász and Winkler in [LW95].

The hitting time from two state distributions σ and τ of a Markov chain is the minimum expected stopping time over all stopping rules that, beginning at σ, stop in the exact distribution of τ. In other words, it is the expected number of steps that the optimal stopping rule takes to move from σ to τ. The hitting time from σ to τ is denoted $\mathcal{H}(\sigma, \tau)$. The mixing time $\mathcal{H}$ of a Markov chain is defined as

$$\mathcal{H} = \max_{\sigma} \ \mathcal{H}(\sigma, \pi).$$

This 'improved' definition of mixing time that utilises stopping rules is much better than the more basic definition that it is the number of steps required to reach the stationary distribution. In a way, it is really the same definition with the addition that we can use stopping rules as 'shortcuts' to reach the stationary distribution.

The term 'rapidly mixing' must now be defined more formally. First of all, to be considered rapidly mixing, a Markov chain must approach its stationary distribution rapidly regardless of the initial state distribution. This is evident in the definitions of both relative pointwise distance ($\Delta(t)$) and mixing time ($\mathcal{H}$) in that they are maximised over all initial state distributions. 'Rapidly' means in a number of steps polynomial in the input size. This does not sound so rapid until you consider that the number of states can be (and typically is) exponential in the input size. Therefore, a rapidly mixing Markov chain typically approaches its stationary distribution after visiting only a small fraction of its states. Consider, for example, a Markov chain for which every entry in the matrix P is the same. This chain is rapidly mixing and will reach its stationary distribution after only one step, at which point it will have visited at most two states.

3 Conductance and Bounds on Mixing Time

The simplest inequality that should be noted is

$$0 < \Phi \leq 1.$$

It is clear that $0 < \Phi$. To see that $\Phi \leq 1$, remember that, for our conductance-minimising set S, $\pi(S) \leq 1/2$. So we have

$$\begin{aligned} Q(S, \bar{S}) &= \sum_{i \in S, j \in \bar{S}} \pi_i p_{ij} \\ &= \sum_{i \in S} \left(\pi_i \sum_{j \in \bar{S}} p_{ij} \right) \\ &\leq \sum_{i \in S} \pi_i \\ &= \pi(S). \end{aligned}$$

Our upper bound follows. These bounds are tight and can only be improved if we know more about the Markov chain.

There is a relationship between the mixing rate of a Markov chain and its eigenvalues, as well as between the conductance and the eigenvalues. The eigenvectors and eigenvalues of a Markov chain are that of its transition probability matrix P. By the definition of the stationary distribution vector π, it is an eigenvector with associated eigenvalue 1. Since P is a doubly stochastic matrix, all eigenvalues are less than or equal to 1. We denote these eigenvalues $\lambda_1 = 1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_N \geq -1$, where N is the number of states in our Markov chain. As a consequence of dealing only with ergodic Markov chains, we also know that $\lambda_N > -1$.

It is the eigenvalue with second greatest magnitude, i.e. $max\{\lambda_2, |\lambda_N|\}$, that is of particular interest. In practice, there are ways to tinker with a Markov chain to ensure that $\lambda_2 > |\lambda_N|$, so in general we can just assume that λ_2 is the eigenvalue of interest. The greater the distance between λ_2 and 1, the faster a Markov chain mixes. This distance is often referred to as the *spectral gap*. The following inequality is due to Sinclair and Jerrum [SJ89]:

$$\Delta(t) \leq \frac{\lambda_2^t}{\pi_0}.$$

The second eigenvalue λ_2 of a Markov chain is also guaranteed to satisfy the following bound involving its conductance:

$$(1 - 2\Phi) \leq \lambda_2 \leq (1 - \frac{\Phi^2}{2}).$$

From these inequalities we can see that a large spectral gap leads to high conductance which implies, as we will show later, faster mixing. This is in accordance with the result we saw before that a second eigenvalue far from 1 leads to a small relative pointwise distance and therefore faster mixing. Alon [Alo86] discusses bounds involving eigenvalues in much greater detail.

Combining and manipulating our bounds, we can obtain the following characterisation of $\Delta(t)$ in terms of Φ:

$$(1 - 2\Phi)^t \leq \Delta(t) \leq \frac{(1 - \Phi^2/2)^t}{\pi_0}.$$

From these results, Lovász and Kannan [LK99] derive the bound on mixing time of

$$\mathcal{H} \leq 32 \log(1/\pi_0) \frac{1}{\Phi^2},$$

where they specifically attribute the factor of $\log(1/\pi_0)$ to the starting configuration and note that, if the initial distribution σ is close enough to π such that $max_i\{\sigma_i/\pi_i\} \leq 2$, then $\mathcal{H}$ is in fact bounded by $O(1/\Phi^2)$.

By our definition, a small conductance implies that there is a bottleneck in the graph, i.e. a subset of states from which it is difficult to escape and circulate around the chain. However, conductance is a measure regarding the tightest bottleneck in the chain, and one bottleneck does not necessarily imply slow mixing or bad circulation. Lovász and Kannan [LK99] show that using the notion of *average conductance* of a Markov chain results in a better bound for mixing time.

First we must define the conductance function

$$\Phi(x) = \min_{S \subset V,\ \pi_S \leq x} \frac{Q(S, \bar{S})}{\pi_S},$$

noting that, for $x \geq 1/2$, $\Phi(x) = \Phi$. The mixing time of a Markov chain is proven to satisfy the inequality

$$\mathcal{H} \leq 32 \int_{\pi_0}^{1/2} \frac{dx}{x\Phi(x)^2}.$$

This is a better bound than the one proven by Jerrum and Sinclair. Instead of using the conductance of the worst subset of states, it uses a weighted average of worst conductances from different sized subsets.

A very strong tool for bounding Markov chain conductance introduced by Jerrum and Sinclair in several papers including [JS88], [JS89], and [SJ89], is the notion of *canonical paths*. A set of canonical paths is essentially a family of simple paths in the underlying graph of a Markov chain that includes a path between each pair of distinct states i and j. What we attempt to do is build this family of paths $\Gamma = \gamma_{ij}$ such that no edge is 'overloaded', i.e. we attempt to minimise the maximum edge loading ρ defined as

$$\rho = \max_e \frac{1}{Q(e)} \sum_{\gamma_{ij} \ni e} \pi_i \pi_j$$

where $Q(e)$ is the weight of edge e, i.e. if u and v are the endpoints of e, then $Q(e) = Q(u, v) = w_{uv}$. Essentially, an overloaded edge is a bottleneck in the Markov chain.

Sinclair [Sin92] proves the following lower bound on Φ:

$$\Phi \geq \frac{1}{2\rho}$$

This bound holds for any choice of canonical paths, which means that finding a decent set of canonical paths can improve our lower bound of Φ, thus improving our upper bound on mixing time. This bound leads directly to the bound

$$\lambda_2 \leq 1 - \frac{1}{8\rho^2}.$$

Diaconis and Stroock [DS91] had previously obtained a better bound for λ_2 by taking into account the lengths of the paths γ_{ij} in their calculation of ρ; Sinclair improved on their bound by changing the way path length was considered:

$$\lambda_2 \leq 1 - \frac{1}{\bar{\rho}},$$

where, if $|\gamma_{ij}|$ is the edge length of path γ_{ij},

$$\bar{\rho} = \max_e \frac{1}{Q(e)} \sum_{\gamma_{ij} \ni e} \pi_i \pi_j |\gamma_{ij}|.$$

Sinclair offers yet more bounds that show an extremely close link between the values Φ, λ_2, and ρ:

$$\rho = O\left(\frac{\log N}{\Phi}\right)$$

$$\lambda_2 \geq 1 - O\left(\frac{\log N}{\rho}\right)$$

Note that the $\log N$ factor is polynomial in the input size. As such, by our definition of rapidly mixing, it cannot 'make or break' a rapidly mixing Markov chain. A stronger result from the same paper is that

$$\rho \leq 16\tau$$

as it implies $\tau = \Omega(\rho)$. Unfortunately, these three results are not particularly positive in that they give a lower bound for the mixing time rather than an upper bound.

4 Applications of Rapidly Mixing Markov Chains

Here we give three examples of applications of rapidly mixing Markov chains: random sampling, approximating the permanent, and volume estimation. Really the random sampling example is more like a stronger tool that is bootstrapped from Markov chains, and is only efficient if the chains are rapidly mixing. Random sampling is the key tool used in the examples of approximating the permanent and volume estimation.

Random Sampling

In some cases it is desirable to take a random sample from a class of combinatorial objects according to some distribution. Motivation for this may be statistical analysis of a class of objects, for example. Often it is possible to construct a Markov chain corresponding to the set of objects, where each state represents an object and the stationary state distribution π is exactly the distribution from which you wish to sample. Once such a Markov chain has been devised, it is simply a matter of running the chain until the state distribution approaches π. This is known as the Markov chain Monte Carlo method.

Propp and Wilson [PW96] provide several techniques for reducing the number of steps τ for which the Markov chain must run, even when τ is initially unknown. Efficient stopping rules such as those discussed in [AD86] and [LW95] can be applied to reduce the expected number of steps.

Approximating the Permanent

The permanent of a matrix is a measure somewhat like the determinant, and only subtly different in definition. Of particular interest is the fact that computing the number of perfect matchings in a bipartite graph is equivalent to computing the permanent of that graph's 0-1 matrix, a type of adjacency matrix particular to bipartite graphs. Unfortunately, unlike the determinant, the permanent of a matrix is very difficult to compute.

Rapidly mixing Markov chains can be used to approximate the permanent. In fact, this technique is the only known method of obtaining a fully polynomial randomised approximation scheme (fpras) for this problem. There are other problems for which the same is true. The approximation of the permanent is due to Jerrum and Sinclair [JS88], [JS89]. The technique first reduces the approximation of the permanent to uniform random sampling from all the perfect matchings of the graph. They then proceed to construct a near-uniform generator for perfect matchings using the Markov chain Monte Carlo method.

Volume Estimation

Computing the volume of a convex n-dimensional body K is an extremely difficult task. For any fixed n it can be done in polynomial time, but the time grows extremely fast with regard to n. It is not even easy to approximate the volume. The only known approximation scheme that is polynomial in both n and $1/\epsilon$, where ϵ is the maximum acceptable error, comes from the Markov chain Monte Carlo technique.

Dyer, Frieze, and Kannan [DFK91] give an approximation algorithm that runs in $O(n^{19})$ time. Kannan, Lovász, and Simonovits [KLS97] give a much more efficient algorithm that runs in $O(n^5)$ time. Both algorithms use the Markov chain Monte Carlo method to sample points likely to be within the

body by randomly walking in the body, preferably near the perimeter. They use an oracle that says whether each point is in the body or not, the output of which helps determine the next move of the random walk.

5 Conclusion

As shown, rapidly mixing Markov chains can be extremely powerful algorithmic tools. The Markov chain Monte Carlo method has been used to develop extremely good approximation algorithms for problems that were previously thought to be impossible to approximate quickly and accurately.

There are many known bounds for the mixing times of Markov chains. Many of the best bounds involve the conductance of the Markov chain, a measure of how well the chain circulates around its states. There are many open problems in the field of rapidly mixing Markov chains, especially regarding the mixing times of specific classes of chains. As more classes of Markov chains are proven to be rapidly mixing, more approximation schemes based on the Markov chain Monte Carlo method will be discovered. Finding tighter bounds for the conductance of certain classes of graphs will continue to be a key area of this ongoing research.

References

[AD86] David Aldous and Persi Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, 93(5):333–348, 1986.

[Ald82] D.J. Aldous. Some inequalities for reversible Markov chains. *J. London Math. Soc. (2)*, 25:564–576, 1982.

[Ald83] D. Aldous. Random walks on finite groups and rapidly mixing markov chains. *Lecture Notes in Mathematics*, 986:243–297, 1983.

[Alo86] Alon. Eigenvalues and expanders. *COMBINAT: Combinatorica*, 6, 1986.

[DFK91] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. *Journal of the ACM*, 38(1):1–17, January 1991.

[DS91] P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of markov chains. *Annals of Applied Probability*, 1(1):36–61, 1991.

[JS88] Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for Markov chains: the approximation of the permanent resolved (preliminary version). In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 235–244, Chicago, Illinois, 2–4 May 1988.

[JS89] M. R. Jerrum and A. Sinclair. Approximating the permanent. *SIAM Journal on Computing*, 18(6):1149–1178, 1989.

[KLS97] Kannan, Lovász, and Simonovits. Random walks and an O*(n^5) volume algorithm for convex bodies. *RSA: Random Structures Algorithms*, 11, 1997.

[LK99] László Lovász and Ravi Kannan. Faster mixing via average conductance. In ACM, editor, *Proceedings of the thirty-first annual ACM Symposium on Theory of Computing: Atlanta, Georgia, May 1–4, 1999*, pages 282–287, New York, NY, USA, 1999. ACM Press.

[LW95] László Lovász and Peter Winkler. Efficient stopping rules for Markov chains (extended abstract). In *Proceedings of the Twenty-Seventh Annual ACM Symposium on the Theory of Computing*, pages 76–82, Las Vegas, Nevada, 29 May–1 June 1995.

[MR97] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1997.

[PW96] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1 & 2):223–252, 1996.

[Sin92] A. Sinclair. Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Probabilty and Computing*, 1:351–370, 1992.

[SJ89] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, July 1989.