Markov Chain Monte Carlo, Mixing, and the Spectral Gap

Luke Vilnis

May 12, 2013

1 Introduction

Modern statistical applications often require sampling from large multivariate probability distributions. This is often made difficult by the presence of the normalizing constant of the distribution, which requires an intractable sum or integral over all possible assignments to the random variable. To work around this, the technique of Markov Chain Monte Carlo allows us to sample from a probability distribution without ever computing the normalizing constant by only considering the ratio of probabilities between neighboring configurations.

MCMC is also useful when the posterior may not be complicated, but the number of variables is too large to be dealt with all at once. Random walks on very large graphs are an example of this. These sorts of graphs arise in internet-scale applications as representations of web pages or social networks. MCMC schemes that sample uniformly over these graphs are of practical interest to obtain unbiased samples. Even though the uniform distribution does not require computing a complex normalizing constant, the number of nodes in the graph can be unknown and intractably huge.

These MCMC schemes require an iterated sampling procedure before they produce samples from the distribution of interest. This is because each successive sample from the chain is correlated with the previous one, but by waiting long enough between samples that correlation becomes negligible - an initial waiting phase is required to render our choice of starting point irrelevant. Furthermore, if we would like to draw multiple independent samples from the distribution, we must give the chain time to "wander" between samples.

A natural question then, is how many steps of the chain are required between samples to ensure independence? Worse yet, what if this number is so large that our MCMC schemes are not even producing a single unbiased sample from the distribution? This property is called the "mixing time" of the chain, and in this paper we will look at the use of the spectrum of the transition matrix for estimating this quantity.

Fair warning though - bounding mixing times is very difficult, and doing so for even seemingly simple schemes requires heavy analysis, with bounds for more complicated but still common uses of MCMC lying seemingly out of reach of current analysis. Furthermore, the spectrum is far from the only way to produce bounds on mixing time - and other (mostly probabilistic) methods, such as concentration inequalities, can sometimes yield tighter bounds or easier proofs.

Some proofs are included in this text and others are omitted for brevity. The reader is encouraged to consult the excellent "Markov Chains and Mixing Times" [2] which is available for free online. It was the primary reference for this survey.

2 Markov Chains and Stationary Distributions

Definition. A *finite Markov Chain* with state space Ω and transition matrix T is a sequence of random variables $\{X_i\}$ on Ω such that

$$P(X_t = x_1 | X_{t-1} = x_2, X_{t-2} = x_3, ...) = P(X_t = x_1 | X_{t-1} = x_2) = T(x_2, x_1)$$

That is, the distribution of an X_i is independent of the history of the chain given the previous value. Furthermore, because of this independence property and the discreteness of the state space, we can completely describe the Markov chain by means of the $|\Omega| \times |\Omega|$ transition matrix T. For this reason we will often identify the Markov chain with its transition matrix and refer to simply the Markov chain T. Note that this definition implies that each row of T is a normalized categorical distribution - T is referred to as a *stochastic matrix.*

Definition. An *irreducible* Markov chain is one where for all $x, y \in \Omega$ there exists a positive integer n such that $T^n(x, y) > 0$ - that is, we can get from any state to any other state with positive probability given the right number of iterations of the chain. Note that the exponent n can depend on the specific choice of x and y.

A Markov chain can be thought of as describing a weighted directed graph, where edges between states are weighted according to the transition probability. In this view, irreducibility is the requirement that this graph be connected - any state must be reachable from any other state in a finite number of transitions.

Definition. An *aperiodic* Markov chain is one for which there exists a positive integer n such that for all $n' \geq n$ we have $T^{n'}(x, x) > 0$ for all $x \in \Omega$ - that is, we can return to a state with positive probability given the enough iterations of the chain. This n need not depend on the specific choice of x since we can take the max over the state space.

Even if every state is reachable from every other state, it may still only be reachable at time steps that are multiples of some number - for example, take the two-state Markov chain that switches back and forth between states deterministically at each time step. An easy way to ensure aperiodicity is to have a positive chance at remaining at each state - that is, the diagonal of the transition matrix must have positive entries. This guarantees that if we ever reach a state, we have a chance of remaining there an arbitrary number of times and "breaking" the periodicity.

Proposition 1. *If a Markov chain T is irreducible and aperiodic, then there exists a positive integer r such that $T^{r'}(x, y) > 0$ for all $x, y \in \Omega$ and all $r' \geq r$ - note that this differs from the definition of irreducibility because r may not depend on the specific choice of x or y.*

Proof. By irreducibility, we know that for all $x, y \in \Omega$ there is some n such that $T^n(x, y) > 0$. By aperiodicity, we know there is a m such that for all $m' > m$, we have $P^{m'}(x, x)$. So we know that $T^{n+m'}(x, y) > 0$ for all $m' > m$, and since we have finite state space we can take the max over all m and n and end up with $r = \max n + \max m$ and $T^r(x, y) > 0$ for all $x, y \in \Omega$. □

Since the definitions of irreducibility and aperiodicity refer only to entries of (powers of) the transition matrix T, we can omit Markov chains from the definition entirely and speak directly about *aperiodic* and *irreducible* matrices.

Note that this last proposition means that an irreducible and aperiodic matrix is a generalization of a positive matrix - it becomes positive and stays positive after enough iterations.

Definition. A probability distribution π is called a *stationary distribution* of the Markov chain with transition matrix T and state space Ω if it satisfies

$$\pi = \pi T$$

or equivalently,

$$\pi(y) = \sum_{x \in \Omega} \pi(x) T(x, y)$$

Definition (Detailed balance)**.** A Markov chain T and distribution π are said to obey *detailed balance* if for all $x, y \in \Omega$, we have:

$$\pi(x) T(x, y) = \pi(y) T(y, x)$$

Proposition 2. If a Markov chain T and distribution π obey detailed balance, then T has stationary distribution π.

Proof. If we sum both sides of the equation over all $y \in \Omega$, we can see that

$$\sum_{y \in \Omega} \pi(y)T(y,x) = \sum_{y \in \Omega} \pi(x)T(x,y) = \pi(x)$$

since rows of T add up to 1. This satisfies the definition of a stationary distribution. □

Detailed balance is a somewhat restrictive assumption. By looking at the spectrum of th transition matrix, we can relax this assumption and provide a more general story about existence of and convergence to stationary distributions.

Definition. Define the *spectral radius* of a linear operator T as

$$R(T) = \sup_{\lambda \in \sigma(T)} |\lambda|$$

where $\sigma(T)$ is the spectrum of T.

Theorem 1 (Perron-Frobenius)**.** Let T be a linear transformation from a finite dimensional vector space $V \to V$. If T is irreducible, aperiodic, and contains all real entries, then $R(T) = \sigma$ is an eigenvalue with a positive eigenvector, and if $\mu \neq \sigma$ is an eigenvalue then $|\mu| < \sigma$.

Proof. A proof of this can be found in Chapter 9 of [4]. □

Proposition 3. If λ is an eigenvalue of a stochastic matrix, $|\lambda| \leq 1$. Further, any stochastic matrix has an eigenvalue $\lambda = 1$.

Proof. Note that the spectrum of the transpose is the same as the spectrum of the matrix. Assume there is some eigenvalue λ greater than 1 and we have $Tx = \lambda x$. We know that each element of Tx is a convex combination of elements of x since each row of T is normalized and positive. However if $\lambda > 1$ then at least one element of λx must have larger modulus than the biggest element of x, which is a contradiction. To see that it always has 1 as an eigenvalue, note that the vector of all 1s is a right eigenvector with eigenvalue 1. □

Proposition 4. An irreducible, aperiodic Markov chain T has a unique stationary distribution π with $\pi(x) > 0$ for all $x \in \Omega$.

Proof. Since the transition matrix is irreducible and aperiodic, we can apply Perron-Frobenius to see that T must have a positive, dominant eigenvalue and associated unique positive left eigenvector, π. Since the matrix has largest eigenvalue 1 and a stationary distribution must be an eigenvector with eigenvalue 1, this shows uniqueness. □

This is a very general result, and is less restrictive than requiring a detailed balance condition - however detailed balance is very easy to check and so when only existence of the stationary distribution must be established, it is still a very useful tool.

3 Convergence

We start by introducing a common linear algebra technique for computing the dominant eigenvector which we will use to show convergence of the Markov chain.

Theorem 2 (Power iteration)**.** Given a matrix A, and a starting vector b_0, if A has a positive eigenvalue with magnitude greater than all its other eigenvalues, and b_0 is not orthogonal to the eigenvector v_1 corresponding to that eigenvalue, then

$$\lim_{n\to\infty} \frac{Ab_n}{\|Ab_n\|} = v_1$$

furthermore, this sequence converges exponentially - that is, for some $\alpha \in (0,1)$ and $C > 0$, we have

$$\|b_n - v_1\| \leq C\alpha^n$$

Proof. A proof of this can be found in any book on numerical linear algebra such as [5]. □

Proposition 5. An irreducible, aperiodic Markov chain T converges to its stationary distribution π - that is, for all $z \in \Omega$,

$$\lim_{n\to\infty} T^n(z,\cdot) = \pi$$

and specifically, this convergence is exponential, that is for some $\alpha \in (0,1), C > 0$:

$$\max_{z\in\Omega} \|T^n(z,\cdot) - \pi\| \leq C\alpha^n$$

where $\|\cdot\|$ is some norm on $\mathbb{R}^{|\Omega|}$.

Proof. By Perron-Frobenius, our irreducible and aperiodic transition matrix meets the conditions required of the matrix for convergence of power iteration, which occurs at an exponential rate.

The max over $z \in \Omega$ represents the starting vector b_0 (a deterministic initial state), which is guaranteed not to be orthogonal to π since it has nonnegative entries, meeting the conditions for power iteration. □

4 Mixing and the spectral gap

Mixing is the property of a Markov chain converging to its stationary distribution. Related properties, like *ergodicity* (roughly, the equivalence between averages over time and averages over the state space in a Markov chain), also fall under the umbrella of mixing but we will not address them.

We showed convergence using the regular l^2 norm on the vector space in the previous section (and indeed by equivalence of norms on finite dimensional vector spaces, that convergence result is unaffected by the choice of norm). But we might want to use a norm more suited to probability distributions. The *total variation distance* (and associated norm) is a common choice:

Definition (Total variation distance)**.** Define the *total variation distance* between two discrete probability distributions as

$$\|\pi - \mu\|_{TV} = \sum_{x\in\Omega} |\pi(x) - \mu(x)|$$

And we can define the maximal distance $d(t)$ of T from π at time t as

$$d(t) = \max_{x\in\Omega} \|T^t(x,\cdot) - \pi\|_{TV}$$

Definition (Mixing time)**.** Define the *mixing time* t_{mix} as

$$t_{mix}(\epsilon) = \min_{d(t)\leq\epsilon} t$$

Remember that we earlier calculated an exponential rate of convergence for mixing. However it was not clear the base of the exponent was. Since our fixed-point iteration is towards the dominant eigenvector, it makes sense that this rate be proportional to the eigenvalue of the second-most dominant eigenvector - a measure of how much the chain “pulls toward” the stationary distribution.

Definition (Spectral gap)**.** Let λ_1 be the leading eigenvalue of the reversible transition matrix T, and $\sigma(T)$ be its spectrum. Define the *spectral gap* as

$$\gamma = \sup_{\lambda \in \sigma(T), \lambda \neq \lambda_1} 1 - |\lambda|$$

To see an example of this in action, consider the Markov chain with uniform transition probability from any state to any other state. Then the transition matrix T will have rank 1, with a single eigenvector of all 1's and the spectral gap will be maximized. This is the fastest possible mixing Markov chain as every step is completely independent from every other step.

We can use the spectral gap γ to bound mixing time. Let $t_{rel} = \frac{1}{\gamma}$

Theorem 3. If T is an irreducible, aperiodic, reversible Markov chain, and $\pi_{min} = \min_{x \in \Omega} \pi(x)$, then

$$t_{mix}(\epsilon) \leq log(\frac{1}{\epsilon \pi_{min}})t_{rel}$$

$$t_{mix}(\epsilon) \geq (t_{rel} - 1)log(\frac{1}{2\epsilon})$$

Proof. These bounds are proved as theorems 12.3 and 12.4 in [2]. □

It bears repeating that the spectral gap is often not the best tool by which to prove mixing results, and probabilistic techinques such as the Markov inequality can give tighter bounds. However the relationship between the spectral gap and mixing is of intrinsic interest.

5 Markov Chain Monte Carlo

As promised in the introduction, we can use the machinery of Markov chains to devise schemes for sampling from complex probability distributions.

Definition (Metropolis chain)**.** Given a distribution of interest π and transition matrix A, define the *Metropolis chain* to be the Markov chain corresponding to the following transition matrix:

$$T(x,y) = \begin{cases} A(x,y)(1 \wedge \frac{\pi(y)A(y,x)}{\pi(x)A(x,y)}) & \text{if } y \neq x \\ 1 - \sum_{z, z \neq x} T(x,z) & \text{if } y = x \end{cases}$$

Proposition 6. The Metropolis chain has the stationary distribution π.

Proof. It is easy to check that detailed balance holds for T and π. □

The transition matrix A defines at each state a *proposal distribution*, and the Metropolis chain can be thought of as accepting or rejecting proposed moves in such a way as to reach the stationary distribution of interest.

Further restrictions are required to show that the Metropolis chain is irreducible and aperiodic. However, our discussion of these properties from earlier gives the guideline that our chain should be able to reach a state from any other state, and should have a positive probability of remaining in the same state, and this will suffice.

Irreducibility is the particularly difficult condition to prove, because it is not clear if making the local moves described in a Metropolis chain can get you from any state to any other state - so we can not in general be sure that we will converge to our stationary distribution. This generally depends on the problem structure and the proposal chain.

6 Random walks on graphs

MCMC can be used to devise a scheme for sampling uniformly from a graph using only local information at each vertex - even when the complete structure of the graph is unknown. Naive random breadth-first search is biased - it visits each node proportional to its degree.

If we desire to sample uniformly from a graph, than π is the uniform distribution. Our proposal distribution A can move uniformly from a vertex to any of its adjacent vertices, so we have $A(i,j) = 1/d_i$ where d_i is the degree of vertex i and we have an edge (i,j). We can write the Metropolis chain (where E is the set of edges) as:

$$T(x,y) = \begin{cases} \frac{1}{d_y} \wedge \frac{1}{d_x} & \text{if } y \neq x, (i,j) \in E \\ 1 - \sum_{z, z \neq x} T(x,z) & \text{if } y = x \\ 0 & \text{if } (i,j) \not\in E \end{cases}$$

This is interesting because using only local degree information we can generate a chain that samples uniformly over the whole graph.

However, the Metropolis chain empirically mixes very slow. What if we wanted to generate a faster mixing random walk on a graph? [1] take the following interesting approach to the problem, even though it is mostly impractical because frequently in applications the graph structure will not be totally known and the size will be extremely large.

Let $\mu(T) = 1 - \lambda^*$ be modulus of the second-largest eigenvalue.

$$\begin{aligned} \text{minimize } & \mu(T) = \|T - (1/n)\mathbf{1}\mathbf{1}^\top\|_2 \\ \text{subject to } & T \geq 0, T\mathbf{1} = \mathbf{1}, T = T^\top \\ & T_ij = 0, (i,j) \not\in \mathcal{E} \end{aligned}$$

This is a convex optimization problem and can be solved with for example, the projected subgradient method. This approach can scale to hundreds of thousands of edges, so one could imagine some applications where this could be used as a fast way to (perhaps) optimize message passing. It is also a nice example of the power of the spectral approach - convexity is a very nice property in an optimization problem as we are guaranteed to get the largest spectral gap possible subject to those constraints.

7 Graph coloring

Graph coloring is a hard combinatorial problem for which we can devise an MCMC scheme. A proper q-coloring is an assignment of a one of q colors to each vertex of the graph in such a way that no two adjacent nodes share the same color. Our posterior distribution of interest is the uniform distribution over all proper q-colorings of the graph G.

To sample from this distribution, we can use the following MCMC chain: pick a vertex v uniformly at random, and a color q uniformly at random. If this color q is valid for v, we change the color of v, otherwise we make no change.

Note that while this MCMC chain has the right stationary distribution, it is far from clear that it is irreducible and convergent to that stationary distribution - in fact, it is easy to devise counterexample graphs and starting states that can not reach a proper coloring by means of these local moves in the general case.

This MCMC chain implicit defines a massive transition matrix of size $q^n \times q^n$ between all possible colorings. Now we can show how to get a bound on its spectral gap, which is pretty exciting. We have to introduce some new tools to tackle this problem.

Definition (Coupling)**.** A *coupling* of two probability distributions μ and ν is a pair of random variables X and Y such that $P(X = x) = \mu(x)$ and $P(Y = y) = \nu(y)$.

Definition (Grand Coupling)**.** A *grand coupling* is a collection of random variables indexed by $x \in \Omega$ and $t \geq 0$, $\{X_t^x\}$. For all $x \in \Omega$, the sequence $\{X_t^x\}_{t=0}^{\infty}$ is a Markov chain starting at x with transition matrix T.

Define a grand coupling for the graph coloring Metropolis chain: at each move, generate a vertex and color pair (v, q) uniformly at random, just as in the single chain case. However, for each element of the grand coupling, we propose the same pair (v, q). Our grand coupling is indexed by every member $x \in \Omega$, that is, every possible coloring (proper or not) of the graph.

Definition (Hamming distance)**.** For two colorings x and y, define the *Hamming distance*

$$\rho(x, y) = \sum_{v \in V} \mathbf{1}_{x(v) \neq y(v)}$$

that is, the number of vertices where the two colorings disagree. ρ is a metric on Ω.

First we need a lemma establishing a form of contraction for our grand coupling.

Lemma 1. *Let Δ be the maximum degree of the graph and $c_{met}(\Delta, q) = 1 - (3\Delta/q)$. If $q > 3\Delta$, and $\rho(x, y) = 1$ then*

$$E(\rho(X_1^x, X_1^y)) \leq (1 - \frac{c_{met}(\Delta, q)}{n})\rho(x, y)$$

Proof. Let v_0 be the vertex at which the two chains disagree, and $\mathcal{N}$ be the set of colors used by the neighbors. After updating the two chains, the Hamming distance will be 0 only if we select v_0, and the proposed color is not used by any of the neighboring vertices. This happens with probability

$$P(\rho(X_1^x, X_1^y) = 0) = (\frac{1}{n})(\frac{q - |\mathcal{N}|}{q}) \geq \frac{q - \Delta}{nq}$$

In order for the Hamming distance after the move to be 2, we examine two cases:

- If some vertex w which is a neighbor of v_0 is picked, note that the set of colorings used by the adjacent vertices excluding v_0 is the same for x and y.
- If $x(v_0)$ and $y(v_0)$ do not belong to this set of colors, and we propose $x(v_0)$ as a new color for w, then y will accept and x will not accept.
- Likewise, if we propose $y(v_0)$, then x will accept and y will not.

In any other case our Hamming distance will be less than 2 after the move. So we have

$$P(\rho(X_1^x, X^y)) \leq (\frac{\Delta}{n})(\frac{2}{q})$$

Combining these two bounds, we have

$$E(\rho(X_1^x, X_1^y) - 1) \leq \frac{2\Delta}{nq} - \frac{q - \Delta}{nq} = \frac{3\Delta - q}{nq}$$
$$E(\rho(X_1^x, X_1^y) - 1) \leq 1 - \frac{q - 3\Delta}{nq}$$

and since $q > 3\Delta$ and $c_met(\Delta, q) = 1 - (3\Delta/q) > 0$ we have

$$E(\rho(X_1^x, X_1^y)) \leq 1 - \frac{c_{met}(\Delta, q)}{n} < 1$$

This handles the case when $\rho(x, y) = 1$.

Now for the general case when $\rho(x, y) = r$. There is a series of r (possibly improper) colorings, $x_0...x_r$ such that $x_0 = x$ and $x_r = y$ and $\rho(x_k, x_{k-1}) = 1$ (changing each disagreeing vertex sequentially). We can apply the bound above to each pair of colorings (x_k, x_{k-1}), and by the triangle inequality for ρ and linearity of expectation we have

$$E(\rho(X_1^x, X_1^y)) \leq \sum_{k=1}^{r} E(\rho(X_1^{x_k}, X_1^{x_{k-1}})) \leq \rho(x, y)(1 - \frac{c_{met}(\Delta, q)}{n}$$

as required.

□

Now that we have established this contraction property of our chain, the following theorem gives us a lower bound on the spectral gap:

Theorem 4 (M. F. Chen (1998))**.** Ω is a finite metric space with metric ρ, T is a Markov chain. If we have a constant $\theta < 1$ and a coupling (X_1, Y_1) of $T(x, \cdot)$ and $T(y, \cdot)$ for each $x, y \in \Omega$ with

$$E(\rho(X_1, Y_1)) \leq \theta\rho(x, y)$$

then $\theta \geq |\lambda|$ for all eigenvalues $\lambda \neq 1$.

Proof. For any function f, we have

$$|Tf(x) - Tf(y)| = |E(f(X_1) - f(Y_1))| \leq E(|f(X_1) - f(Y_1)|)$$

Since Ω is finite, we can define a Lipschitz constant L_f using the metric ρ in the usual way. Then by the hypothesis we have

$$|Tf(x) - Tf(y)| \leq L_f E(\rho(X_1, Y_1)) \leq \theta L_f \rho(x, y)$$

So $L_{Tf} \leq \theta L_f$. Combining these things and taking some eigenvector φ with eigenvalue $\lambda \neq 1$, we have

$$|\lambda| L_\varphi = L_{\lambda\varphi} = L_{T\varphi} \leq \theta L_\varphi$$

so $|\lambda| \leq \theta$ as required. □

Now that we have this bound on the spectral gap, we can go back to our lemma we proved for graph coloring. This is not the tightest bound we can get for this graph coloring scheme - there is a probabilistic argument that can improve on this bound by removing the dependence on the number of proper colorings.

Proposition 7. If $q > 3\Delta$, where Δ is the maximum degree of the graph, then the Metropolis chain defined for random graph coloring has absolute spectral gap

$$\gamma^* \geq \frac{1}{3n\Delta}$$

Furthermore, this gives an upper bound on the mixing time

$$t_{mix}(\epsilon) \leq log(\frac{p}{\epsilon})3n\Delta$$

(where $1/\pi_{min} = p$ is the number of proper colorings of the graph).

Proof. Our contraction on the grand coupling from before gave us:

$$E(\rho(X, Y)) \leq (1 - \frac{1}{3n\Delta})\rho(x, y)$$

This satisfies the contraction assumptions of the previous theorem, so we have

$$\gamma^* \geq \theta = \frac{1}{3n\Delta}$$

The bound on the mixing time comes from the theorem relating spectral gap and mixing time from earlier. □

So we can use coupling and the spectral gap to derive an upper bound on mixing time for graph coloring. Of course our restriction that the number of colors be more than 3 times the maximum node degree is extremely restrictive, as many graphs can be colored with far fewer colors than the maximum (or even minimum) node degree. This should give the reader some appreciation for how difficult it is to derive these sorts of bounds.

8 A few more interesting things

Before concluding this survey, we will introduce a new property of the chain relating to mixing, called the bottleneck ratio, and show how it relates to the two properties we've defined so far, the mixing time and the spectral gap. The interplay between these three quantities and the topology of the graph structure induced by the chain is very appealing.

Definition (Bottleneck ratio)**.** A *bottleneck* in the state space of a Markov chain is a consequence of the fact that in many chains, some states can only be reached from other states by way of some restricted set of states. If there are many such states that require passing through a restricted set of states to reach each other, we call this restricted set of states a bottleneck.

Given a Markov chain T with stationary distribution π, define the *edge measure* Q:

$$Q(x,y) = \pi(x)T(x,y)$$
$$Q(A,B) = \sum_{x\in A, y\in B} Q(x,y)$$

Define the *bottleneck ratio* $\Phi(S)$ of a set $S \in \Omega$ and Φ_* of the whole chain as

$$\Phi(S) = \frac{Q(S,S^C)}{\pi(S)}$$
$$\Phi_* = \min_{S,\pi(S)\leq\frac{1}{2}} \Phi(S)$$

Theorem 5.

$$t_{mix}(\epsilon) \geq \frac{1-2\epsilon}{2\Phi_*}$$

Proof. This is proved as theorem 7.3 in [2]. □

The bottleneck ratio is related to the Cheeger constant of a manifold in differential geometry. The Cheeger constant of a compact Riemannian manifold is the ratio between the volume of the manifold and the area of the smallest hypersurface that divides the manifold in half - the similarity between this and a "bottleneck" is obvious. The Cheeger inequality relates this constant to the second eigenvalue of the Laplacian operator on the manifold, which is similar to the transition matrix of a Markov chain in that it is related to the behavior of a random walk on the manifold (as well as heat diffusion).

The following inequalities are related to the Cheeger inequality:

Theorem 6 (Jerrum and Sinclair (1989))**.**

$$\frac{\Phi_*^2}{2} \leq \gamma \leq 2\Phi_*$$

Proof. This is proved as theorem 13.14 in [2]. □

So the three constants, t_{mix}, γ, and Φ_* are related by a series of inequalities.

The connection between the Cheeger constant and the Markov chain reveals some fascinating interactions between geometry, probability, and analysis. This is also related to techniques that use eigenvalues and eigenvectors of the graph Laplacian to create basis functions that respect global geometry. These are used in computer science for geometrically aware compression and function approximation, such as in [3].

9 Conclusion

In this survey we have reviewed some of the basic ideas involved in analysis of Markov chain mixing and presented connections to applications and other areas of mathematics and computer science. The study of Markov chains is a rich and ongoing research area which has benefitted greatly from cross-pollination from the mathematics, statistics, computer science, and physics communities.

References

[1] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM REVIEW*, 46:667–689, 2003.

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

[3] Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, 8:2007, 2006.

[4] S. Sternberg. *Dynamical Systems.* Dover Books on Mathematics Series. Dover Publications, Incorporated, 2010.

[5] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra.* SIAM: Society for Industrial and Applied Mathematics, June 1997.