Lifting Markov Chains to Speed up Mixing

Fang Chen
Department of Mathematics
Yale University
fchen@math.yale.edu

László Lovász *
Department of Computer Science
Yale University
lovasz@cs.yale.edu

Igor Pak †
Department of Mathematics
Yale University
paki@math.yale.edu

Abstract

There are several examples where the mixing time of a Markov chain can be reduced substantially, often to about its square root, by "lifting", i.e., by splitting each state into several states. In several examples of random walks on groups, the lifted chain not only mixes better, but is easier to analyze.

We characterize the best mixing time achievable through lifting in terms of multicommodity flows. We show that the reduction to square root is best possible. If the lifted chain is time-reversible, then the gain is smaller, at most a factor of $\log(1/\pi_0)$, where π_0 is the smallest stationary probability of any state. We give an example showing that a gain of a factor of $\log(1/\pi_0)/\log\log(1/\pi_0)$ is possible.

1 Introduction

The estimation of the mixing time of finite Markov chains (the time needed for the chain to become approximately stationary) has emerged as a major issue in the design and analysis of various algorithms for sampling, enumeration, optimization, integration etc.

The research presented in this paper was motivated by the work of Diaconis, Holmes and Neal [5], who observed that certain non-reversible chains mix substantially faster than closely related reversible chains. We view their example in a different way: we represent a given chain as the "projection" of another chain, and analyze how this improves the mixing time.

Example 1.1 It is easy to see that for the random walk on an n-path, the mixing time is $\Theta(n^2)$. We can consider this path as a "projection" of the directed $(2n-2)$-cycle: if we generate a random node from the stationary distribution on the cycle (which is uniform), the projection will be a node from the stationary distribution on the path. The mixing time of the random walk on this cycle is $\Theta(n)$. (To be precise, we consider a bidirected cycle with probability 2/3 of going "clockwise" and probability 1/3 of going counterclockwise.)

*Supported by NSF grant No. CCR-9712403.
†Supported by the NSF Postdoctoral Research Fellowship.

STOC '99 Atlanta GA USA

A more interesting example is provided by the random walk on a grid.

Example 1.2 Let G be the $\sqrt{n} \times \sqrt{n}$-grid on the torus. It is easy and well known that the mixing time of the random walk on G is $O(n)$.

Now give the random walk some "inertia". More exactly, we define a walk as follows: if we enter a node, we are most likely to exit over the opposite edge; we turn left or right with probability $1/100\sqrt{n}$; we never turn back. If we start the walk in a given direction, say "North", we are likely to continue going North until we circle the torus about a 100 times. Then we turn; the node where we turn off is essentially uniformly distributed over the cycle we started on. Now we circle the torus about 100 times, going either East or West. When we turn again, our E-W coordinate will also be essentially uniformly distributed, in other words, the second turning point will be essentially uniformly distributed over all nodes of G. And to get to this second turning point, we only needed $O(\sqrt{n})$ steps!

Of course, the walk we defined above is not a random walk on G; rather, it is a random walk on a "lifted" graph $\widehat{G}$, defined as follows. For every node $v \in V(G)$, we introduce four copies, corresponding to the direction we were traveling when we entered the node. Let v be any "old" node, and let v_N, v_E, v_S, v_W be its neighbors to the North, East, South and West. From the new node $(v, South)$ (say), we go to $(v_S, South)$ with probability $1 - 1/\sqrt{n}$, and to $(v_E, East)$ and $(v_W, West)$ with probability $1/(2\sqrt{n})$ (the factor of 100 was only needed in the very informal analysis above).

By symmetry it is clear that the stationary distribution on $\widehat{G}$ is also uniform. Of course, our analysis above showing that the random walk on $\widehat{G}$ mixes fast was not precise: an argument will be given after the exact definition of the mixing time, at the end of section 2.

Further, more involved examples of improving mixing properties of random walks on groups through "lifting" will be given in section 4.

In this paper we study the possibilities and limitations of lifting.

The details of the results depend on the exact notion of mixing time we use. A standard way of defining this would be to consider the number of steps after which the total variation distance of the current distribution from the stationary becomes less than (say)1/4; however, we prefer to work with two quantities whose definition is free of arbitrary parameters, and which allow us to state clean inequalities like proposition 3.3 below.

Still informally, the most basic version is the time $\mathcal{H}$ needed to get the stationary distribution from the worst starting state, using the best stopping rule. Let π_0 be the smallest stationary probability of any state. We show that the mixing time of any lifting is at least $\Omega(\sqrt{\mathcal{H}}/\log\frac{1}{\pi_0})$, and if both the original and the lifted chains are time-reversible, then the mixing time of the lifting is at least $\Omega(\mathcal{H}/\log\frac{1}{\pi_0})$. We also give an example showing that even in this time-reversible case, the mixing time can improve by a factor of $(\log\frac{1}{\pi_0})/\log\log\frac{1}{\pi_0}$.

We also consider the *average set-hitting time*, or *set-time* $\mathcal{A}$, which is the expected number of steps needed to hit the "most remote" set from a random starting node (drawn from the stationary distribution), weighted with the stationary probability of the set. While seemingly it is not a mixing measure, set-time behaves very well and is closely related to $\mathcal{H}$ and other mixing parameters. We show that the set-time of any lifting is at least $\sqrt{\mathcal{A}}$. Results of Aldous and Fill [3] imply that for the time-reversible case, lifting gives no improvement at all in the set-time. This implies easily the results about mixing time.

The numbers $\mathcal{H}$ and $\mathcal{A}$ relate to the standard measures intimately. For example, if we add loops with transition probability 1/2 at each node to get rid of periodicities (lazy walk), then after $O(\mathcal{H})$ steps the total variation distance from the stationary distribution drops to less than 1/4. Furthermore, if we start from a distribution such that the probability of each node i is at most $(1+C)\pi_i$, then after $O(\mathcal{A})$ steps, this bound drops to $(1+C/2)\pi_i$ (see [9] for a survey).

The conductance Φ of the chain, which has been used since the work of Jerrum and Sinclair [6] to estimate its mixing time, plays an important role in these results. We also need a related value, the *flow parameter* $\mathcal{C}$; this satisfies $\mathcal{C} \geq 1/\Phi$ and the celebrated Leighton–Rao theorem implies that $\mathcal{C} = O(\log(1/\pi_0)/\Phi)$ for time-reversible chains. Our main result is that the smallest mixing time achievable through lifting is $\Theta(\mathcal{C})$.

These results will be stated formally in section 3, following preliminaries about mixing times and multicommodity flows.

2 Preliminaries

2.1 Mixing times

We consider a finite irreducible Markov chain M, with set of states V, transition probabilities p_{ij} and stationary distribution π. Often it is convenient to describe the Markov chain in terms of the *ergodic flow* $Q_{ij} = \pi_i p_{ij}$. These values satisfy $\sum_i Q_{ij} = \sum_i Q_{ji}$ and $\sum_{i,j} Q_{ij} = 1$, and every non-negative matrix Q with these properties defines a Markov chain. The stationary distribution can be recovered by $\pi_j = \sum_i Q_{ij}$.

We can also think of a Markov chain as a random walk on a strongly connected digraph $G = (V, A)$, with the p_{ij} considered as weights of the edges.

Recall that the Markov chain is *time-reversible* if $Q_{ij} = Q_{ji}$ for every $i, j \in V$. We can define the *reverse chain* of any Markov chain by $\widehat{Q}_{ij} = Q_{ji}$. A chain is time reversible iff it is equal to its own reverse chain.

The *hitting time* $\mathcal{H}(i,j)$ is the expected number of steps before node j is reached, starting from node i. We define the *access time* $H(\pi, S)$ of a set S of nodes as the expected number of steps before the set S is reached, starting from a random node drawn from the stationary distribution. The *set-time* of the chain is defined by

$$\mathcal{A} = \max_S \pi(S)\mathcal{H}(\pi, S).$$

A *stopping rule* Γ is a map from V^* (the set of finite strings of states) to $[0,1]$, such that for $w \in V^*, w = w^0 w^1 \ldots w^t$ the value of $\Gamma(w)$ is the probability of continuing given that w is the walk so far traversed.

It is often useful to regard a stopping rule Γ as a stopping time, i.e., a random variable whose value is the actual number of steps made before stopping (so we stop at w^Γ). We assume that with probability 1 the rule stops the chain, i.e., the stopping time is finite.

Assume that the starting node w^0 is drawn from some distribution σ. The distribution of w^Γ is denoted by σ^Γ. If $\sigma^\Gamma = \tau$, then we say that Γ is a stopping rule from σ to τ. We say a stopping rule Γ from σ to τ is *optimal* if the mean length of Γ is infimum of all the stopping rules from σ to τ. We define the *access time* $\mathcal{H}(\sigma,\tau)$ from σ to τ to be the mean length of an optimal rule.

The *mixing time from a state* i is $\mathcal{H}(i,\pi)$, the mean length of an optimal rule from the distribution concentrated at i to the stationary distribution. The *mixing time* of the random walk on the graph G is $\mathcal{H} := \max_{i\in V}\mathcal{H}(i,\pi)$.

The set time is related to mixing in an intimate way (cf. [9]):

$$\mathcal{A} \leq \mathcal{H} \leq 128\mathcal{A}\log\frac{1}{\pi_0}. \tag{1}$$

We conclude with showing that the mixing time of the random walk on $\widehat{G}$ in our second introductory example is $\Theta(\sqrt{n})$. Start from any state in $\widehat{G}$. Consider the following stopping rule Γ: with probability 1/2, we stop after the second change of direction, and with probability 1/2, we stop after the third change of direction. A simple calculation shows that when we stop, the probability of being at any given node of $\widehat{G}$ is more than $1/8n$ (the reason for going either 2 or 3 turns is to get the direction of entrance also right: after an even number of turns, we are always going North or South).

Thus we have described a (randomized) stopping rule that, for any starting node, stops in an expected number of $\Theta(\sqrt{n})$ steps, and the probability of stopping at each node is at least 1/2 its stationary probability. By standard results (see e.g. [9]), this implies that the mixing time is $\Theta(\sqrt{n})$.

2.2 Exit frequencies

We'll need the following notion from [9]. Let Γ be a stopping rule from σ to τ. We denote by $x_i(\Gamma)$ the expected number of times we leave node i before stopping. It is easy to see that for every node i, the *conservation equation*

$$\sum_j p_{ji}x_j(\Gamma) - x_i(\Gamma) = \tau_i - \sigma_i$$

holds. It will be convenient to introduce the *scaled exit frequencies* $y_i(\Gamma) = x_i(\Gamma)/\pi_i$.

It was shown in [8] that Γ is an optimal rule from σ to τ if and only if it has a *halting state*, i.e., a node i with $x_i(\Gamma) = 0$. In this case, the conservation equation implies that the exit frequencies depend only on σ and τ, and they will be denoted by $x_i(\sigma,\tau)$ (and $y_i(\sigma,\tau)$). Another consequence of the conservation equation we need is the fact that if Γ is a stopping rule from σ to σ, then $x_i(\Gamma) = \pi_i \mathsf{E}(\Gamma)$.

In the case of time-reversible chains, and stopping rules achieving the stationary distribution π, more can be said

about halting states. Let us say that z is a halting state for node a if z is a halting state of an optimal stopping rule from a to π (this does not depend on which optimal rule we consider). Then the halting states for a are characterized as the nodes maximizing the hitting time to a. Furthermore, if a is a node with $\mathcal{H}(a,\pi)$ maximum, then so is z, and a is the halting state for z.

2.3 Conductance and flows

An $s-t$ flow ($s,t \in V$) in a digraph $G=(V,E)$ is a non-negative valued function f defined on the edges such that

$$\sum_j f(ji) = \sum_j f(ij)$$

for every node $i \neq s,t$. The *value* of the flow is defined by

$$\mathrm{val}(f) = \sum_j f(sj) - \sum_j f(js) = \sum_j f(jt) - \sum_j f(tj),$$

and *cost* is defined by

$$\mathrm{cost}(f) = \sum_{i,j} f(ij)$$

A flow with value 0 is called a *circulation*. The ergodic flow Q_{ij} is a circulation of cost 1.

A *multicommodity flow* is a collection $f=(f^{st})$ of flows, one $s-t$ flow f^{st} for each pair of nodes s and t. In this paper, we need to consider only multicommodity flows with $\mathrm{val}(f^{st}) = \pi_s\pi_t$ for all s and t, so this will be automatically assumed. We also assume that the flows are minimal, i.e., no directed cycle has positive flow value on each edge.

We define the *congestion* of a multicommodity flow as the least K such that

$$\sum_{s,t} f^{st}(ij) \le KQ_{ij}$$

for every edge ij. We define the *local cost* of a multicommodity flow as the least K such that

$$\sum_t \mathrm{cost}(f^{st}) \le K\pi_s, \qquad \sum_s \mathrm{cost}(f^{st}) \le K\pi_t$$

for every s and t. Finally, let $\mathcal{C}$ denote the the smallest K such that there exists a multicommodity flow with congestion and local cost at most K. Since $\mathcal{C}$ can be written as the optimum of a linear program, it is polynomial time computable.

Sometimes it is more convenient to think of a multicommodity flow as a weighted collection of directed paths $\{(P_r,w_r) : 1 \le r \le N\}$, where the total weight of paths from node i to node j is $\pi_i\pi_j$.

The *conductance* of a Markov chain is defined by

$$\Phi = \min_{\substack{S\subseteq V,\\ 0<\pi(S)<1}} \frac{Q(S,V\setminus S)}{\pi(S)\pi(V\setminus S)}, \tag{2}$$

where $Q(A,B)$ is shorthand for $\sum_{i\in A, j\in B} Q_{ij}$.

It is easy to see that $\mathcal{C} \ge 1/\Phi$. For time-reversible chains, it is easy to derive from a well-known theorem of Leighton and Rao [7] the following reverse bound:

$$\mathcal{C} = O\left(\log\frac{1}{\pi_0}\frac{1}{\Phi}\right).$$

2.4 Conductance and mixing

Standard results, first obtained by Jerrum and Sinclair [6], use conductance to bound mixing parameters. For reversible chains, the following bounds are well known. Let $1-\lambda_2$ be the eigenvalue gap, and define the *relaxation time* to be $\mathcal{L} = \frac{1}{1-\lambda_2}$. Then

$$\frac{1}{\Phi} \le \mathcal{L} \le \frac{8}{\Phi^2}, \tag{3}$$

and

$$\frac{1}{\Phi} \le \mathcal{H} \le \log\frac{1}{\pi_0}\frac{10}{\Phi^2}. \tag{4}$$

For general chains, there does not seem to be a standard way to define the eigenvalue gap, but similar bounds can be proved for the set-time $\mathcal{A}$.

Lemma 2.1 *For every Markov chain,*

$$\frac{1}{4\Phi} \le \mathcal{A} \le \frac{20}{\Phi^2} \tag{5}$$

and

$$\frac{1}{\Phi} \le \mathcal{H} \le \frac{3000}{\Phi^2}\log\frac{1}{\pi_0} \tag{6}$$

The proofs of these lemmas are omitted. The connection between multicommodity flows and mixing, in one direction, is established by the following lemma:

Lemma 2.2 *For every Markov chain,*

$$\mathcal{H} \ge \frac{1}{2}\mathcal{C}.$$

Proof. Consider the multicommodity flow $f=(f^{st})$, where

$$f^{st}(ij) = \pi_t\pi_s Q_{ij}\left(y_i(s,\pi) + \overleftarrow{y}_j(t,\pi)\right),$$

where $\overleftarrow{y}_j$ denote the scaled exit frequencies in the reverse chain. It is easy to check, using the conservation equation, that f^{st} is indeed a flow of value $\pi_s\pi_t$ from s to t. The cost of this flow is

$$\sum_{ij} f^{st}(ij) = \pi_t\pi_s\left(\sum_i \pi_i y_i(s,\pi) + \sum_j \pi_j \overleftarrow{y}_j(t,\pi)\right)$$

$$= \pi_s\pi_t(\mathcal{H}(s,\pi) + \overleftarrow{\mathcal{H}}(t,\pi)) \le 2\pi_s\pi_t\mathcal{H}.$$

(Here $\overleftarrow{\mathcal{H}}(t,\pi)$ denotes mixing times for the reverse chain; we use that $\overleftarrow{\mathcal{H}} = \mathcal{H}$.) Moreover,

$$\sum_{s,t} f^{st}(ij) = Q_{ij}\left(\sum_s \pi_s y_i(s,\pi) + \sum_t \pi_t \overleftarrow{y}_j(t,\pi)\right).$$

Now here $\sum_s \pi_s y_i(s,\pi)$ can be considered as the scaled exit frequency of the following (non-optimal) stopping rule from π to π: "choose a random starting point s from π and then follow an optimal stopping rule to π". It follows that these exit frequencies are the same for each i, and hence

$$\sum_{s,t} f^{st}(ij) = Q_{ij}K$$

for some constant K. To obtain K, we can sum over all edges ij, which gives that $K = \sum_{s,t}\mathrm{cost}(f^{st}) \le 2\mathcal{H}$. □

3 Lifting and Collapsing

Let M and $\widehat{M}$ be two finite Markov chains with underlying sets V and $\widehat{V}$, respectively. We denote by $\widehat{\pi}$, $\widehat{p}$ etc. the stationary distributions, transition probabilities etc. in $\widehat{M}$.

We say that M is a *collapsing* of $\widehat{M}$, if there is a mapping $\widehat{V} \to V$ such that

$$\pi_v = \widehat{\pi}(f^{-1}(v)) = \sum_{i \in f^{-1}(v)} \widehat{\pi}_i$$

for every $v \in V(G)$, and

$$p_{vu} = \sum_{i \in f^{-1}(v), j \in f^{-1}(u)} \frac{\widehat{\pi}_i}{\widehat{\pi}(f^{-1}(v))} \widehat{p}_{ij}$$

for every pair $v, u \in V(G)$. We also say that $\widehat{M}$ is a *lifting* of M. For the random walk on an undirected graph, collapsing simply means identifying vertices with the same image.

Note that for the ergodic flows we have

$$Q(A, B) = \widehat{Q}(f^{-1}(A), f^{-1}(B)).$$

It is easy to check that every chain obtained by collapsing a reversible chain remains reversible. (More generally, the reverse of the collapsed chain can be obtained by collapsing the reverse chain by the same mapping.) On the other hand, a lifting of a reversible chain may be nonreversible, and we will see that in order to gain significantly in the mixing time, we will have to look for nonreversible liftings.

In their book [3], Aldous and Fill [3] call the collapsed chain the *induced* chain, and prove monotonicity properties of several parameters under collapsing.

It is a trivial observation that the conductance Φ cannot be increased by lifting: if there is a "bad" partition in M, then it lifts to a "bad" partition in $\widehat{M}$. Similarly, the flow parameter $\mathcal{C}$ cannot be decreased by lifting: given a multicommodity flow in the lifted chain, it can be projected to the original chain to get a multicommodity flow with equal or less congestion and local cost.

Lemma 2.1 and the monotonicity of Φ under lifting implies the following.

Theorem 3.1 *Let M be a finite irreducible Markov chain and let $\widehat{M}$ be a lifting of it. Then*

$$\widehat{\mathcal{A}} \geq \frac{\sqrt{5}}{40} \sqrt{\mathcal{A}}, \tag{7}$$

and

$$\widehat{\mathcal{H}} \geq \frac{1}{10\sqrt{30}} \sqrt{\frac{\mathcal{H}}{\log(1/\pi_0)}}. \tag{8}$$

Recall that in the case of the random walk on an n-path, $\frac{1}{\Phi} = \Theta(n)$, $\mathcal{H} = \Theta(n^2)$, $\mathcal{A} = \Theta(n^2)$, and we lifted it to a dicycle whose mixing time is $\Theta(n)$. This shows that the bound in (7) is best possible.

We also know that a factor of $\log\log(1/\pi_0)$ is needed in (8). Consider a random walk on a S_n generated by all $\frac{n}{2}$ cycles. The mixing time is $\Theta(\log n)$. Indeed, consider a set B of permutations with no fixed point. It is known that $|B| \sim \frac{n!}{e}$ (see e.g. [4]). After each step the number of fixed points decreases by a factor of about 2. Thus the hitting time to B is $\Theta(\log n)$, which gives a lower bound on the mixing time (see [11]). The upper bound is also $O(\log n)$ (see [13]).

On the other hand, diameter of the Cayley graph is $O(1)$ (see [15]). This is easy to see because in one step one can cyclically permute any $\frac{n}{2}$ elements. Now by a later result, Theorem 4.3, there exists a lifting with $\widehat{\mathcal{H}} = O(\Delta) = O(1)$. The stationary distribution is uniform here, $\pi_0 = 1/n!$. So in (8), a factor of $\log\log(1/\pi_0) = \log\log n! = \log n$ is needed. We don't know whether there is an example where the factor $\log(1/\pi_0)$ is needed.

3.1 Optimal lifting

There are chains that cannot be lifted to get $\widehat{\mathcal{H}} \approx \frac{1}{\Phi}$ (an example is a "path with a drift", where we step with probability 2/3 to the right and 1/3 to the left). We do not know whether there is always a lifting that makes $\widehat{\mathcal{A}} \approx \frac{1}{\Phi_G}$. But we have the following general theorem about the best mixing time.

Theorem 3.2 *For every chain M,*

$$\frac{1}{2}\mathcal{C} \leq \inf \widehat{\mathcal{H}} \leq 144\mathcal{C},$$

where the infimum is taken over all liftings of M.

Proof. The first inequality is an immediate consequence of lemma 2.2 and the monotonicity of $\mathcal{C}$ under lifting. To prove the second, let $f = \{(P_r, w_r)\}$ be a multicommodity flow from π to π with congestion and local cost at most $\mathcal{C}$, given in the form of a weighted collection of directed paths. It is not hard to see that, using this set of paths, we can construct another set of paths $\{\widehat{P}_r\}$ with weights $\{\widehat{w}_r\}$ such that they define a multicommodity flow from π to π with congestion and path length at most $\widehat{\mathcal{C}}$, where $\mathcal{C} \leq \widehat{\mathcal{C}} \leq 12\mathcal{C}$. To simplify the notations, we will still use $\{P_r\}$ and $\{w_r\}$ to denote $\{\widehat{P}_r\}$ and $\{\widehat{w}_r\}$. Let ℓ_r be the length of P_r. The following equations are obvious but will be important:

$$\sum_r w_r = 1, \qquad \ell_r \leq \widehat{\mathcal{C}}, \tag{9}$$

and

$$\sum_{r:\ P_r \text{ starts at } i} w_r = \pi_i, \qquad \sum_{r:\ P_r \text{ ends at } i} w_r = \pi_i \tag{10}$$

Furthermore,

$$\sum_{r:\ ij \in E(P_r)} w_r \leq \widehat{\mathcal{C}} Q_{ij} \tag{11}$$

Informally, we lift the chain as follows: if we start at a node i, we select one of the paths P_r out of that node at random, with probability proportional to the weight w_r.

Formally, we construct the lifted graph $\widehat{G}$ from G by adding a directed path P'_r of length ℓ_r connecting i to j if P_r goes from i to j (these added paths have no internal node in common with G or with each other).

The ergodic flow on an edge ij of the lifted chain is defined by

$$\widehat{Q}_{ij} = \begin{cases} w_r/2\widehat{\mathcal{C}}, & \text{if } ij \in E(P'_r), \\ Q_{ij} - \displaystyle\sum_{r:\ ij \in E(P_r)} w_r/2\widehat{\mathcal{C}}, & \text{if } ij \in E(G). \end{cases}$$

It is clear that this is a circulation and it is easy to check that $\sum_{ij} \widehat{Q}(ij) = 1$, thus it defines a Markov chain on $\widehat{G}$. The stationary distribution of the lifted chain is given by

$$\widehat{\pi}_i = \begin{cases} w_r/2\widehat{\mathcal{C}}, & \text{if } i \in V(P'_r) \setminus V(G), \\ \pi_i - \sum_{r:\ P_r \text{ thru } i} w_r/2\widehat{\mathcal{C}}, & \text{if } i \in V(G). \end{cases}$$

Now there is a natural way of mapping the paths P'_r onto the paths P_r, which defines a homomorphism of the graph $\widehat{G}$ onto G. It is easy to check that this collapses the random walk on $\widehat{G}$ onto the random walk on G.

Notice for an old node i,

$$\frac{1}{2}\pi_i \le \widehat{\pi}_i \le \pi_i\ , \tag{12}$$

and the probability of getting on a directed path P'_r starting at i is

$$\widehat{P}_{ij} = \frac{\widehat{Q}_{ij}}{\widehat{\pi}_i} = \frac{w_r}{2\widehat{\mathcal{C}}\widehat{\pi}_i}\ .$$

Hence the probability of getting on any directed path starting at i is

$$\sum_{r:\ P'_r \text{ starts at } i} \frac{w_r}{2\widehat{\mathcal{C}}\widehat{\pi}_i} = \frac{1}{2\widehat{\mathcal{C}}\widehat{\pi}_i} \sum_{r:\ P_r \text{ starts at } i} w_r = \frac{\pi_i}{2\widehat{\mathcal{C}}\widehat{\pi}_i}\ ,$$

and by (12), it is bounded between $\frac{1}{2\mathcal{C}}$ and $\frac{1}{\mathcal{C}}$.

We claim that the mixing time of the lifted chain is at most $144\mathcal{C}$.

Consider the following stopping rule. First walk until you see an old node, then keep walking until you reach an old node by going through a directed path. Let the node be X, then the distribution of X is π. With probability $1/2$, stop; otherwise, continue walking until get onto any directed path P'_r. Once on P'_r, choose an interior node Y of P'_r randomly and uniformly, and stop at Y.

We claim that we stop at each node of $\widehat{G}$ with at least $(1/2)$ of its stationary probability. If v is an old node, then $X = v$ with probability π_v, and so we stop there with probability at least $\pi_v/2 \ge \widehat{\pi}_v/2$. If we continue walking after X, let w^k be the kth point in the walk starting from X. Because at any old node i, the probability of getting on any directed path is between $\frac{1}{2\mathcal{C}}$ and $\frac{1}{\mathcal{C}}$, a coupling argument shows that for any old node i,

$$\begin{aligned} \text{Prob}(w^k &= i | w^0, \cdots, w^k \text{ are old points}) \\ &\ge \left(1 - \frac{1}{\widehat{\mathcal{C}}}\right)^k \pi_i \end{aligned} \tag{13}$$

If v is a new point on the directed path P'_r which connects the old nodes i to j. Then

$$\begin{aligned} &\text{Prob(stop at } v) \\ &\ge \frac{1}{2}\sum_{k=0}^{\infty} \text{Prob}(w^k = i | w^0, \cdots, w^k \text{ are old points}) \\ &\times \text{Prob(at } i\text{, get on the path } P'_r) \times \frac{1}{\ell_r} \\ &\ge \frac{1}{2}\sum_{k=0}^{\infty} \left(1 - \frac{1}{\widehat{\mathcal{C}}}\right)^k \pi_i \frac{w_b}{2\widehat{\mathcal{C}}\widehat{\pi}_i} \frac{1}{\widehat{\mathcal{C}}} \\ &\ge \frac{w_b}{4\widehat{\mathcal{C}}^2} \sum_{k=0}^{\infty} \left(1 - \frac{1}{\widehat{\mathcal{C}}}\right)^k \\ &= \frac{w_b}{4\widehat{\mathcal{C}}} = \frac{1}{2}\widehat{\pi}_v \end{aligned}$$

Obviously, this stopping rule takes at most $6\widehat{\mathcal{C}}$ steps. Now a folklore "fill-up" argument (cf. [1]) implies that $\mathcal{H} \le 12\widehat{\mathcal{C}} \le 144\mathcal{C}$. □

3.2 Reversible lifting

Aldous and Fill show that if the lifted chain is time-reversible, then the average set-hitting time $\mathcal{A}$ cannot be decreased by lifting:

Proposition 3.3 *Let M and $\widehat{M}$ be two finite reversible chains, and assume that $\widehat{M}$ is a lifting of M. Then*

$$\mathcal{A} \le \widehat{\mathcal{A}}$$

This implies, by (1):

Corollary 3.4 $\mathcal{H} \le 128 \log(1/\pi_0)\widehat{\mathcal{H}}$.

Thus if the lifting $\widehat{G}$ is reversible, then the gain in mixing time, if any, is marginal. In the case of graphs without multiple edges, $\log \frac{1}{\pi_0} = O(\log n)$, so the gain is only $O(\log n)$.

Aldous and Fill raise question whether collapsing always decreases the mixing time (at least up to a constant factor). Example 3.5 below shows that this is not the case: it can happen that $\widehat{\mathcal{H}}$ is smaller than $\mathcal{H}$ by a factor of $\log(1/\pi_0)/\log\log(1/\pi_0)$, which is almost best possible by corollary 3.4.

Example 3.5 In the following example, we need the following well-known facts: (a) if the underlying graph of a Markov chain is a tree, then it is time-reversible. (b) Furthermore, if ij is an edge, and V_1 is the set of nodes separated from j by ij, than $\mathcal{H}(ij) = \pi(V_1)/Q_{ij}$.

The graph $\widehat{G}$ has $2k+2$ nodes, labeled $v_0, \ldots, v_{k+1}$ and $u_1, \ldots, u_k$. From node v_0, we step to v_1 with probability $1/2$, and stay at v_0 otherwise. From node u_i, we step to v_i with probability $1/(2i)$, and stay at u_i otherwise. From node v_i $(1 \le i \le k-1)$, we step to v_{i-1} with probability $1/(2i+6)$, to u_i, with probability $1/(2i+6)$, and to v_{i+1}, with probability $(i+2)/(i+3)$. From v_k, with step to each of v_{k-1}, v_{k+1} and u_k with probability $1/3$. From v_{k+1}, we step to v_k with probability $1/2k$, otherwise stay at v_{k+1}

The graph G is obtained by identifying u_i with v_{i-1}, for $i = 1, ..., k$.

By the introductory remark, both of these chains are time-reversible. To explain the structure of $\widehat{G}$, we note that the transition probabilities were chosen recursively so that $\widehat{\mathcal{H}}(v_{i-1}, v_i) = 2$ and $\widehat{\mathcal{H}}(u_i, v_i) = 2i$ for $i = 1, \ldots, k$. The node v_{k+1} is just a twin of u_k. It is not difficult to compute that $\widehat{\mathcal{H}}(v_k, v_{k+1}) = 4k+8$. One can check then that $\widehat{\mathcal{H}}(u_i, v_{k+1}) = 6k+8$ for every leaf u_i, and also $\widehat{\mathcal{H}}(v_0, v_{k+1}) = 6k+8$. Most of the stationary probability is concentrated on v_{k+1} and u_k, and hence if we start at v_0, it takes asymptotically $6k$ steps to mix. This turns out to be also true for every other leaf, while we need less from internal nodes. So $\widehat{\mathcal{H}} \sim 6k$.

On the other hand, on the graph G we have $\mathcal{H}(v_{i-1}, v_i) = i+1$ for $i = 1, \ldots, k$, and $\mathcal{H}(v_k, v_{k+1}) = 4k+8$, hence

$\mathcal{H}(v_0, v_k) = k(k+3)/2$, and $\mathcal{H}(v_0, v_{k+1}) = (k^2+11k+16)/2$. As before, this implies that the mixing time from v_0 (which is trivially a pessimal node) is asymptotically $k^2/2$.

To conclude, note that the sum of degrees in G (and $\widehat{G}$) is $2^{k-1}(k+1)!(3k+4)$, and the minimum degree is 2, whence $\log(1/\pi_0) \sim k \log k$, and so the factor $k/12$ we gain by lifting is indeed $\Theta(\log(1/\pi_0)/\log\log(1/\pi_0))$.

4 More examples: random walks on groups

Let G be a finite group, S be its set of generators, Q be a probability distribution on S. Define an irreducible Markov chain $M = M(G, S, \mathrm{Q})$, with set of states G, transition probabilities $p_{g,h} = \mathrm{Q}(g^{-1}h)$ given $g^{-1}h \in S$, and $p_{g,h} = 0$ otherwise. Observe that M has a uniform stationary distribution $\pi = U$. One can think of M as a random walk on a weighted Cayley graph $\Gamma = \Gamma(G, S)$. Note also that M is reversible if $S = S^{-1}$, and $\mathrm{Q}(s) = \mathrm{Q}(s^{-1})$ for all $s \in S$.

The problem of estimating the mixing time $\mathcal{H}$ for this special class of Markov chains is classical (see [3, 4, 11]) and is often very delicate. The problem is that M is often rapidly mixing, say $\mathcal{H} = O(\log^c |G|)$ for some constant $c > 0$. Thus we cannot even ignore a factor of $\log(1/\pi_0) = \log|G|$. We will show that sometimes nonreversible liftings of reversible random walks mix better and are easier to analyze.

Example 4.1 Let $G = S_n$ be a symmetric group. Consider a generating set of *Coxeter* (adjacent) transpositions $R = \{(1,2), (2,3), \ldots, (n-1,n), (n,1)\}$, and let Q be uniform on S. It is easy to see that the corresponding random walk $M(S_n, R, \mathrm{Q})$ mixes very slowly. Indeed, we move element 1 with probability $1/n$ in each direction on a circle, and leave it with probability $1-2/n$. Thus we need about $O(n^3)$ steps just to have element 1 "mixed". A coupling argument in [2] shows that $\mathcal{H} = O(n^3 \log n)$.

Now notice that the diameter $\Delta = \Theta(n^2)$, and a constant portion of the group lies at distance $> \Delta/2$ from identity. Thus $\mathcal{C} = \Omega(n^2)$ and a lifted Markov chain must have $\widehat{\mathcal{H}} = \Omega(n^2)$. Let us construct a lifting with $\widehat{\mathcal{H}} = \Theta(n^2)$.

Let $\widehat{G}$ be the set of pairs (σ, i), $\sigma \in S_n$, $1 \le i \le n$. Using analogy with example 1.2, let the walk move from (σ, i) to $(\sigma \cdot (i, i+1), i+1)$ with probability $1-1/100n$ and to $(\sigma \cdot (j, j+1), j+1)$, $j = \sigma^{-1}(\sigma(i)+1)$ with probability $1/100n$. The idea is to give the random walk some "inertia": if we just moved element k clockwise, we are likely to keep moving it in the this direction. We do it for about $100n$ steps, so then the element k gets into nearly uniform position. Then we start moving element $k+1$, etc. After about $O(n^2)$ steps, when we have finished moving all the elements, we get a nearly uniform permutation, which gives us $\widehat{\mathcal{H}} = O(n^2)$.

Example 4.2 Let $G = U(n+1, \mathbf{F_p})$ be the group of upper triangular matrices over the finite field with ones on the diagonal. Let the generating set S be the set of all matrices with one nonzero entry right above diagonal. Thus applying a generator is the same as adding to the (uniformly chosen) i-th row the $(i+1)$-th row multiplied by a uniform element $a \in \mathbf{F_p}$. As before, let Q be uniform and assume p is large enough (say, $p > 2n^3$). This random walk has been extensively studied before as an example of a random walk on a nilpotent group (see [12, 14]). The best known upper bound in [12] gives $\mathcal{H} = O(n^{2.5})$ while only $\mathcal{H} = \Omega(n^2)$ is known for the lower bound.

We will present a construction of a lifting with a mixing time $\widehat{\mathcal{H}} = \Theta(n^2)$. Simply let the row i be chosen not uniformly, but with high probability *right above the previous row*. Formally, let $\widehat{G}$ be the set of pairs (g, i), $1 \le i \le n$, $g \in G$. At (g, i), with probability $1 - 1/100n^2$ we move to $(g', i-1 \bmod n)$, where g' is obtained by adding row i times uniform $a \in \mathbf{F_p}$ to row $i-1 \bmod n$, and with probability $1/100n^3$ we move to (g, j), $1 \le j \le n$. Note that we will switch from a cycle after about $O(n^2)$ steps. After the first $O(n)$ steps we get a nearly uniform first row, after the next $O(n)$ steps we get a nearly uniform second row, etc. When we switch we get to a uniform j and uniform $g \in G$. Stop then. This defines a stopping time and an easy computation gives us $\widehat{\mathcal{H}} = O(n^2)$ (cf. [12]).

Let us now state a general result for this case. Let $d(g, h)$ be the distance between $g, h \in G$ in the Cayley graph $\Gamma(G, S)$. Let

$$\overline{D} = \frac{1}{|G|} \sum_{g \in G} d(e, g) < \Delta$$

be the average distance from the identity element. Denote $q_0 = \min_{s \in S} \mathrm{Q}(s)$.

Theorem 4.3 *For every finite group G, a set of generators S, and a probability distribution Q on S,*

$$C_1 \, \overline{D} \le \inf \widehat{\mathcal{H}} \le \frac{1}{q_0} \, C_2 \, \overline{D},$$

where the infimum is taken over all possible liftings, and C_1,C_2 are universal constants.

The proof follows easily from Theorem 3.2. If we allow flexibility in the choice of Q one can sharpen the theorem by deleting a factor of $1/q_0$ on the right hand side. Surprisingly, together with (8) this gives a solution of the Conjecture 7.3 in [3].

Theorem 4.4 *For every finite group G, a set of generators S, there exists a probability distribution Q on S, such that for a corresponding random walk we have:*

$$\mathcal{H} < C \, (\overline{D})^2 \log |G|,$$

where where C is a universal constant independent of G.

A different proof of this result was independently discovered by Jim Fill (personal communication).

5 Concluding remarks

The key open question is: how to use the lifting of Markov chains in sampling algorithms? Many of the current applications (for example, to volume computation) use the "conductance-squared" bound (Lemma 2.1 or its variations) to estimate the mixing time (and often this is best possible). Theorem 3.2 shows that if we use an appropriate lifting, then we can reduce the mixing time to about the square root in such cases!

The catch is to construct the appropriate lifting. Our results in section 3 show that this is closely related to explicitly constructing multicommodity flows with minimum congestion. Section 4 shows that this is possible in some cases, namely for groups; hopefully, it is also possible in other cases like random walks in convex bodies.

References

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

[2] D. J. Aldous, Random walks on finite groups and rapidly mixing Markov chains, *Lecture Notes in Mathematics* **986** (1983), 243–297.

[3] D. J. Aldous and J. Fill, *Time-reversible Markov chains and random walks on graphs* (book in preparation)

[4] P. Diaconis, *Group Representations in Probability and Statistics*, IMS, Hayward, California (1988)

[5] P. Diaconis, S. Holmes and R. Neal, Analysis of a non-reversible Markov chain sampler, TR BU-1385-M, Biometric Unit, Cornell University (1997)

[6] M. Jerrum and A. Sinclair, Conductance and the rapid mixing property for Markov chains: the approximation of the permanent resolved, *Proc. 20nd Annual ACM Symposium on Theory of Computing.* (1988), 235–243.

[7] F.T. Leighton and S. Rao, An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms, *Proc. 29th Annual Symposium on Foundations of Computer Science.* (1988), 422–431.

[8] L. Lovász and P. Winkler, Efficient stopping Rules for Markov Chains, *Proceedings of the 1995 ACM Symposium on the Theory of Computing,* 76–82.

[9] L. Lovász and P. Winkler, Mixing Times, in *Microsurveys in Discrete Probability,* (ed. D. Aldous and J. Propp), DIMACS Series, AMS, 1998

[10] L. Lovász and P. Winkler, Reversal of Markov chains and the forget time, *Combinatorics, Probability and Computing,* to appear.

[11] I. Pak, *Random walks on groups: Strong uniform time approach*, Ph.D. thesis, Harvard University (1997)

[12] I. Pak, Two random walks on upper triangular matrices, preprint (1998)

[13] Y. Roichman, Upper bound on the characters of the symmetric groups, *Invent. Math.*, **125** (1996), no. 3, 451-485

[14] R. Stong, Random walk on the upper triangular matrices, *Ann. Prob.*, **23** (1995), 1939–1949

[15] U. Vishne, Mixing and covering in the symmetric groups, *J. Algebra*, **205** (1998), no. 1, 119-140