Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications 136 (2021) 145–191
stochastic processes and their applications
www.elsevier.com/locate/spa

Characterizing limits and opportunities in speeding up Markov chain mixing

Simon Apers[a,b,*], Alain Sarlette[c,d], Francesco Ticozzi[e,f]

[a] *CWI, The Netherlands*
[b] *QuIC, Université Libre de Bruxelles, Belgium*
[c] *Inria Paris, France*
[d] *Department of Electronics and Information Systems, Ghent University, Belgium*
[e] *Dipartimento di Ingegneria dell'Informazione, Università di Padova, Italy*
[f] *Department of Physics and Astronomy, Dartmouth College, NH, USA*

Received 29 September 2019; received in revised form 8 March 2021; accepted 14 March 2021
Available online 22 March 2021

Abstract

A variety of paradigms have been proposed to speed up Markov chain mixing, ranging from non-backtracking random walks to simulated annealing and lifted Metropolis–Hastings. We provide a general characterization of the limits and opportunities of different approaches for designing fast mixing dynamics on graphs using the framework of "lifted Markov chains". This common framework allows to prove lower and upper bounds on the mixing behavior of these approaches, depending on a limited set of assumptions on the dynamics. We find that some approaches can speed up the mixing time to diameter time, or a time inversely proportional to the graph conductance, while others allow for no speedup at all.

Keywords: Markov chains; Mixing time; Algorithm design and analysis; Network theory (graphs)

1. Introduction

The importance of algorithms based on Markov chains is widely appreciated. In computer science, random walks and Markov chain Monte Carlo form the backbone of many randomized algorithms to solve tasks such as approximating the volume of convex bodies [21] or the

* Corresponding author. Most of the work was done while the author was at Ghent University.
E-mail addresses: smgapers@gmail.com (S. Apers), alain.sarlette@inria.fr (A. Sarlette), ticozzi@dei.unipd.it (F. Ticozzi).

https://doi.org/10.1016/j.spa.2021.03.006
0304-4149/

permanent of a non-negative matrix [32], or to solve combinatorial optimization problems using simulated annealing methods [37]. In physics, Markov chain Monte Carlo is an indispensable tool for sampling and simulation of many-body systems. Some examples are the use of Glauber dynamics to simulate the Ising model [44] or the Metropolis–Hastings algorithm [27,45] to sample from the Gibbs distribution.

In general, a Markov chain can be used to sample from a probability distribution π that is not directly available nor explicitly known. Instead, known properties of the target distribution are translated into a stochastic evolution that is engineered to converge, or *mix*, to an equilibrium distribution which coincides with the target one. In various contexts, such Markov chain is easier to obtain or to implement rather than direct sampling from π. Under rather mild conditions, running this Markov chain from any starting position and stopping after a sufficient number of steps T, the resulting state will be approximately distributed according to the target equilibrium distribution π. Critical to these applications is how fast the stochastic process converges, and estimating this convergence speed or mixing time is often a difficult task [3,40]. Approaches include estimating the spectral radius of the transition map [20,39], or using advanced coupling and stopping time arguments [43].

1.1. Speeding up Markov chains

In order to decrease the number of required steps T, thus accelerating the convergence towards π, a wide range of approaches has been proposed. The following are some examples, which will be treated in more detail in Section 4. All approaches describe local dynamics over the node set of a graph, in which the dynamics can only move from a node to any of its neighboring nodes (i.e., nodes that have an edge to the present node).

- *Stopping rules:* The simple Markov chain scheme has a deterministic stopping time T, i.e. the transitions specified by the Markov chain are run for a fixed number of steps T, upon which its state is returned. As an extension, one can choose the stopping time randomly, according to some predefined distribution or dependent on the nodes that have been visited [34,40,43]. Such choice is formally described by a stopping rule. For instance, if the stopping time is uniformly distributed over some fixed time interval $[0, T]$, then the output is called the Cesaro average. A more advanced stopping rule could say, e.g. go on until you have seen each node at least once (this relates to the Markov chain *cover time* [40]). By returning a sample obtained through a stopping rule, it is possible to converge faster to the target distribution π. More precisely, one specifies a stopping rule such that the distribution over nodes, conditioned on having stopped, is ϵ-close to π; and the mixing performance is measured by the expectation of the stopping time.
- *Non-backtracking random walks:* Consider that we want a sample from π, which is the stationary distribution of a random walk on an undirected graph, moving from a given node to any neighbor with equal probability. When applying the random walk from a given starting node, there is a probability that the walker moves from node a to node b and then directly back to a; such a move is generally detrimental for spreading on the graph. A non-backtracking random walk therefore assigns a decreased probability $\alpha \ll 1$ of traversing the same edge twice in a row. That is, the probability of choosing at time $t+1$ the same node where one was at time $t-1$ is decreased to $\alpha \ll 1$ with respect to a uniform choice among available neighbors, and the probability of choosing any of the other available neighbors uniformly is accordingly increased. This process has the same stationary distribution π over the nodes, and [4,19,22,36] have shown that this approach generally speeds up mixing as compared to a simple random walk.

- *Simulated annealing or slowly-varying Markov chains:* Markov chains are sometimes used to find the minimum of a function g over the graph nodes. The target stationary distribution should thus be much larger at a minimum than at other places, and this can be achieved by choosing the probability of jumping from i to j larger than the probability of jumping from j to i if $g(j) < g(i)$. If a jump towards higher values of g is assigned a tiny probability, then this Markov chain has a high probability to get stuck for a long time in local minima that are not the global one. In contrast if the influence of g on jumping probabilities is made too weak, the stationary distribution is essentially random over all nodes. As a remedy, a time-dependent sequence of Markov chains can be proposed, whose transition probabilities and stationary distributions converge gradually to the "irregular" goal distribution concentrated on local minima, such that during the early steps of the sequence one can efficiently jump out of local minima. See for instance [37].
- *Gather-and-distribute strategies:* This method was originally proposed in a consensus setting [24], where a given load must be distributed as fast as possible over nodes of a network in a distributed way. The gather-and-distribute strategy is a time-varying procedure, using a sequence of two time intervals. In terms of Markov chain mixing, during the first intervals, transition probabilities are chosen so as to move all the probability mass to a single predefined node. Thus after this first time interval, whatever the initial distribution, the Markov chain ends up in a single predefined situation. Knowing that the second time interval starts from this particular probability distribution, its transition probabilities are then designed to redistribute the probability according to the goal distribution. As an example, on a complete binary tree of depth D, one could choose the transition probabilities during the first D time steps so that all the probability mass moves onto the root node. From this known situation, common to all initial distributions, it is not hard to design transition probabilities over the next D time steps (e.g. using a "stochastic bridge", see Section 5.2) in order to redistribute the probability mass towards the target stationary distribution. The mixing time of this approach is thus $2D$ on the binary tree example, exponentially improving over the simple random walk mixing time $\Omega(2^D)$.
- *Data-augmented and lifted Markov chains:* Consider again the problem of obtaining a sample from some (indirectly specified) goal distribution π over the nodes. Sometimes this distribution is easier to obtain as the marginal of a distribution on a larger sample space. In this case, one can obtain a sample by using a Markov chain that evolves on an augmented state space, consisting of the original variable and some latent variables, and just discarding the value of the latent variable in the obtained sample, see e.g. [60,62]. A related strategy is to use latent variables and augmented graphs not only for an easier specification of the target, but also to possibly accelerate the convergence thanks to memory or momentum effects, see e.g. [12,15,18,19] and the next sections for details on Markov chains on lifted graphs.

This list of seemingly distinct ideas provides a range of speedups over the use of a simple random walk. From the literature cited above, non-backtracking random walks are shown to provide (at least) a constant factor speedup, lifted Markov chains can provide up to a quadratic speedup, and gather-and-distribute strategies can provide exponential speedup or even more, on certain graphs.

Example 1.1 (*Mixing on a Cycle*)**.** For illustration we consider the toy example of sampling from the uniform distribution on a cycle graph. This example lies at the basis of several speedup

ideas like [15,18,35]. The possible walker positions are the integers $0, 1, 2, \ldots, n-1$, and at each discrete time step, the walker can decide to stay put, to add 1 or to subtract 1 to its position (modulo n).

A standard *random walk* would add $+1$ or -1 to the current position, each with probability $1/2$. After $t \gg 1$ steps the standard deviation from the original position of the walker is of order $\sqrt{(t)}$. As a consequence, it will take approximately n^2 steps for the random walk to converge to the uniform distribution.

A *stopping rule* could say: always add $+1$ to your position, but at time $t \geq 0$ stop the process there with probability $\min(\frac{\delta}{(1-\delta)^t}, 1)$ for $\delta = 1/n$. Effectively this means that there is a probability δ to stop the process at each time $t \in \{0, 1, \ldots, n-1\}$ and hence after $n-1$ time steps we have a uniform distribution. (Note that the time-dependence of the stopping rule is needed: a constant probability to decide and stop once having performed any t time steps would not achieve the same distribution, not even approximately.)

A *non-backtracking walk*, and also the *lifted Markov chains* introduced in [15,18], would say: start with a given value v *and* a given sign $\circ \in \{+, -\}$. At each time-step, add $\circ 1$ to the current position, and change $\circ$ with a probability $\delta \ll 1$. In this way, the walker will preferably keep moving in the same direction, while occasionally turning around and then keep moving in the other direction. This can be loosely viewed as a random walk with effective moves of order $1/\delta$ over time steps $1/\delta$, and therefore its long-term diffusive behavior rather involves a Gaussian of standard deviation $\frac{1}{\delta}\sqrt{\delta t} = \sqrt{t/\delta} \gg \sqrt{t}$. It is shown in [18] that for $\delta \sim 1/n$ this walk has mixing time $O(n)$.

A *slowly varying Markov chain* could start with adding $+1$ at each position v. This introduces a deterministic drift on the cycle, but it will not lead to a uniform distribution. To resolve this, the chain is slowly varied towards a standard random walk. Similarly to the lifted Markov chain, the state will have a tendency to explore more of the circle thanks to the initial deterministic walk dynamics, but ultimately it will converge to the uniform distribution thanks to the final random walk dynamics.

A *gather-and-distribute* strategy could implement the following moves. During the first n steps, gather all probability mass on position 0. After this the state of the system is perfectly known, and we can efficiently disperse it uniformly,[1] independent of the initial distribution. These dynamics exactly map any initial distribution to the uniform distribution after $O(n)$ steps. □

1.2. Aim and contributions of the present paper

One must note that not all the above speed-up approaches build on the same prior knowledge of the graph and target distribution, e.g. gather-and-distribute appears to require more prior insight to design an efficient algorithm — at least in its most basic implementation. In the present paper, we set this point aside and instead ask the question: *what speedup can we ultimately expect from each one of these approaches?*

To our knowledge, indeed, no classification or clear comparison of the achievable speedups for this variety of approaches is known, so that it is not exactly clear which ones are more

[1] Here is a detailed example of how to do this. First, over the time steps $t = 1, 2, \ldots, n-1$ make move -1 from every position except for position 0, and at position 0 just stand still. Then at times $t = n, n+1, \ldots, 2n-2$, at position 0 make move $+1$ with probability $(2n-t)/n^2$ and otherwise stand still, and at all other positions make move $+1$.

promising to pursue towards more advanced e.g. adaptive design versions, and which ones will quickly hit a hard limit.

In this paper, we show how the attainable speedups can be categorized on the basis of fundamental properties of the stochastic processes resulting from such algorithms, like invariance of the target distribution or the initialization of auxiliary variables. This can allow to quickly assess the potential of a technique before digging into its deeper details and possible variations. Such results add to the recently revived interest in irreversible and beyond-Metropolis–Hastings techniques, some interesting results of which are presented in [11,12,54,55,61]. Our analysis builds on the fact that a wide set of speedup approaches, including all the ones listed above, can be cast into the overarching framework of lifted Markov chains (LMCs). This allows us to use mixing time bounds in the LMC framework, in order to derive bounds on the speedup achievable within each individual approach. The translation into an LMC is not unique, and therefore it is important to consider the best achievable performance under more abstract properties of LMC classes. The results are also of independent interest for the LMC literature itself, as they clarify how the mixing time bound depends precisely on the assumptions of the setting, and how some traditional assumptions relate to existing algorithmic approaches. Finally, one of the motivations for the present work is to pave the way towards a quantitative and fair comparison between lifted chains and quantum walks: although very different from an internal dynamics perspective, both these models rely on non-Markovian effects and share many similarities that can be made precise within the proposed framework [8].

The remainder of the paper can be summarized as follows. We first introduce the LMC model in Section 2, along with several particular LMC constructions which will be used in our discussion. The associated properties or constraints that we consider on an LMC are presented in Section 3. Section 4 is dedicated to translate into LMCs and to discuss the above defined properties for some key classes of algorithms. This will allow us to properly collocate our results with respect to existing ones and to highlight among others the importance of two aspects that have a key role: the ability of locally *initializing* the latent variables of the lifted chain, and whether or not we impose *invariance* of the target distribution during the whole evolution of the LMC. These two properties allow us to immediately identify some "extreme" scenarios, described in Section 5, where either lifting does not yield any advantage over a standard Markov chain, or it potentially allows for reaching the target in the trivial minimum time, corresponding to the diameter of the graph. For the latter, we explicitly provide an LMC that mixes in diameter time, building on *stochastic bridges* [23,52] which can be efficiently set up provided one has a full knowledge of the graph. This is still different from a practical construction based on local knowledge only, yet we recall that both in [15] and in our paper, the purpose is to research the ultimate potential of a method, not to design new practical algorithms. A summarized version of the results of Section 5 can be found in the conference proceedings [7].

In Section 6, we establish lower bounds on the mixing time in intermediate scenarios, i.e. when only one of the above constraints is requested, together with constraints on the reducibility of the LMC, the mixing of the LMC towards its own lifted stationary distribution, and its ergodic flows. These bounds depend on the *conductance* of the graph, which provides a richer description of the graph topology than the diameter. In particular, we show that a conductance bound for the mixing time of lifted Markov chains holds under either of two seemingly unconnected constraints: (i) if we impose that the lifted chain mixes from any initial state in the entire lifted state space, i.e. without allowing to choose the initial values of the latent variables (this is essentially extending the scope of the result in [15]); or (ii) if we impose that

the lifted dynamics keeps the target distribution invariant, i.e. when the system starts well-mixed it must stay so for all times. Conductance bounds are typically stricter than the diameter-time bound, yet examples show how they still allow for the lifted chains to significantly outperform the best possible standard Markov chain. Furthermore, we show that the other constraints – i.e. obeying imposed ergodic flows, irreducibility of the LMC, and considering the mixing properties of the lifted distribution vs. the marginal on the original nodes – do not significantly modify the achievable mixing time.

In Section 7 we provide some further observations: we show that most of the bounds that we prove are tight up to log-factors, we further illustrate how particular properties can be deduced indirectly from our scenarios and bounds, and we discuss possible extensions of our results to other settings. To conclude, a summary of the results and a brief outlook on future developments are provided in Section 8.

2. Setting: mixing dynamics on graphs and their lifts

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ a set of N nodes, which we label as $\mathcal{V} = \{1, \dots, N\}$ and $\mathcal{E}$ the set of directed edges, i.e. ordered pairs of nodes. Throughout the paper, graphs are assumed to be connected and a real function on the node space $\mathcal{V}$ will be represented as a vector in $\mathbb{R}^N$. In particular, we denote by e_i the canonical basis vector, with all elements zero except its i'th element equal to one. The notation e_i will also be used more generally for canonical basis vectors whose dimension is clear from the context. We thus write

$$f = \sum_i f(i)\, e_i \in \mathbb{R}^N$$

and the value $f(i)$ of the function f on node i is obtained as $\langle e_i, f\rangle = e_i^\dagger f$, where $\dagger$ indicates the transpose (adjoint) of a vector or matrix. We will use the shorthand notation $f_i = f(i)$ for components of vectors and matrices, except thus for the particular case e_i which denotes the canonical basis vector. For any two vectors v and w, we use the convention that $v\, w^\dagger$ is the outer product, yielding a rank-one matrix, whereas $v^\dagger\, w$ would be the inner product, yielding a scalar. (Usually this is represented by v being a column vector and $v^\dagger$ a row vector.)

Let $\mathbb{P}_N$ be the set of probability vectors in $\mathbb{R}^N$, i.e. each $p \in \mathbb{P}_N$ satisfies $p_i \geq 0$ $\forall i = 1, 2, \dots, N$ and $\sum_{i\in\mathcal{V}} p_i = 1$. Each component p_i represents the probability of node $i \in \mathcal{V}$. In particular, $p = e_i$ denotes a probability distribution with all weight on node $i \in \mathcal{V}$.

Throughout the paper, we analyze dynamics designed for the following task:

Problem 1 (*Design of Mixing Dynamics*)**.** Design a discrete-time stochastic dynamical system that converges towards a target probability distribution π on $\mathcal{V}$, as fast as possible from any initial condition on $\mathcal{V}$, while respecting the locality associated to the graph $\mathcal{G}$.

By respecting the locality of the graph, we mean that the evolution over one time step can only involve transitions between nodes connected by an edge. One may envision to relax this assumption in an algorithmic framework, in association with further constraints on how to navigate between the nodes and at which algorithmic cost. For instance, allowing moves between just a few non-connected nodes can sometimes dramatically improve the mixing time [26]. In the absence of a canonical way to relax the graph locality criterion, we do not explore this here and thus make the central assumption of a fixed graph $\mathcal{G}$ encoding all the allowed transitions. A number of further constraints on the stochastic process, which precisely specify the design problem at hand, will be discussed in Section 3. The main message of the

paper is that, as we shall see, the performance of the best possible solution will critically depend on these assumptions.

A common approach to address the mixing problem is to make π the attractive steady-state distribution of a Markov chain on the graph. More explicitly, consider a Markov discrete-time stochastic process $\{v(t)\}_{t\geq 0}$ on the node space, entirely specified by conditional probabilities $\text{Proba}\big(v(t+1)=i|v(t)=j\big)=P_{i,j}$ for i and $j\in\mathcal{V}$. The locality constraints induced by $\mathcal{G}$ impose $P_{i,j}=0$ if $\mathcal{E}$ contains no edge from j to i. Notice that in probability theory, what we call P is often called $P^{\dagger}$. If $p(t)$ is the probability vector associated to the distribution of state of the Markov chain at time t, then its evolution is generated by its one-step transition matrix $P=(P_{i,j})$, via:

$$p(t+1)=P\,p(t)\,.$$

In order for $p(t+1)$ to be a probability vector we need P to be a column-stochastic matrix, i.e. $\sum_j P_{i,j}=1$ for all i. Furthermore, to solve Problem 1, it should have π as its unique, globally attractive steady-state distribution. For instance, if $\pi_i>0$ for all i, then P should be irreducible in order to allow $p(t)$ to converge to π from any initial $p(0)\in\mathbb{P}_N$. Solving Problem 1 can then be viewed as accelerating the convergence towards the stationary distribution, compared to just iterating P.

The Markov chain convergence can be accelerated by adding memory to the process, beyond just the current position in $\mathcal{V}$, see e.g. [15,18]. Formally, this leads to Markov chains on lifted graphs or, for short, *lifted Markov chains* (LMCs).

Definition 1. A graph $\hat{\mathcal{G}}=(\hat{\mathcal{V}},\hat{\mathcal{E}})$ on $\hat{N}$ nodes is called a *lift of* $\mathcal{G}$ if there exists a surjective map $\mathfrak{c}:\hat{\mathcal{V}}\mapsto\mathcal{V}$, such that:

$$(i,j)\in\hat{\mathcal{E}}\ \text{ implies }\ (\mathfrak{c}(i),\mathfrak{c}(j))\in\mathcal{E}.$$

We denote by $\mathfrak{c}^{-1}$ the map that takes as input a single node $k\in\mathcal{V}$ and outputs the set of nodes $j\in\hat{\mathcal{V}}$ for which $\mathfrak{c}(j)=k$.

We will denote by $x\in\mathbb{P}_{\hat{N}}$ a distribution over the lifted graph nodes $\hat{\mathcal{V}}$. The associated marginal distribution over $\mathcal{V}$ is given by $p_k=\sum_{j\in\mathfrak{c}^{-1}(k)}x_j$. In vector representation, this induces the linear map

$$p=Cx\,, \tag{1}$$

with C a matrix of zeros and ones. In a lifted Markov chain for $\mathcal{G}$, the distribution $p(t)$ on $\mathcal{V}$ at time t is deduced as the marginal of $x(t)$, while the evolution of $x(t)$ is generated by a linear, stochastic, discrete-time map on the lifted graph:

$$x(t+1)=A\,x(t). \tag{2}$$

Here A satisfies the locality constraints on $\hat{\mathcal{G}}$ induced by the underlying $\mathcal{G}$, i.e. $A_{j,\ell}\neq 0$ only if $(\mathfrak{c}(j),\mathfrak{c}(\ell))$ is an edge of $\mathcal{G}$. The locality of the lifted chain equivalently means that for each x there exists a stochastic matrix $P^{(x)}$ satisfying the locality constraints of $\mathcal{G}$ and such that (2) corresponds to $p(t+1)=P^{(x)}\,p(t)$, with $p(t)=Cx(t)$ as in (1). Explicitly, $P^{(x)}=C\,A\,B^{(x)}$ where $B^{(x)}$ is a linear stochastic map from $\mathbb{R}^N$ to $\mathbb{R}^{\hat{N}}$, with

$$B^{(x)}_{i,j}=\frac{x_i(t)}{\sum_{k\in\mathfrak{c}^{-1}(j)}x_k(t)}=\frac{x_i(t)}{p_j(t)}\quad\text{ if }\quad\mathfrak{c}(i)=j,$$

and $B^{(x)}_{i,j} = 0$ otherwise. From the point of view of $\mathcal{V}$, the applied transition matrix changes over time, in a supposedly smart way governed by $x(t)$. The notation (A, C) is taken over from systems theory and hints at $p = Cx$ being the "output" of interest, whose evolution can be induced by the evolution of some larger "state" x. In Markov modeling, A would be called a hidden Markov chain [53].

In algorithmic applications, which motivate our setting, the pair $(\hat{\mathcal{G}}, A)$ is to be designed in order to accelerate the convergence towards π with respect to the (best) Markov chain P on the original graph $\mathcal{G}$. Of course, all algorithms that are meant to achieve a mixing speedup need not be designed as an LMC. However, in most cases, they can be translated (non-uniquely) into an LMC — see e.g. the examples in the introduction and their discussion in Section 4. Bounds on the best mixing time achievable with the set of LMCs associated to a given algorithmic technique, then translate into essential mixing time bounds for the underlying algorithm.

Before pursuing the mixing time analysis, we present some particular lift constructions which will be useful in the proofs and examples.

2.1. LMCs with product graphs

From $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ two graphs with node sets of cardinality N_1 and N_2 respectively, we can construct a graph on the cartesian product $\hat{\mathcal{V}} = \mathcal{V}_1 \times \mathcal{V}_2$, whose nodes are pairs (i, j), with $i \in \mathcal{V}_1$ and $j \in \mathcal{V}_2$. The edges $\hat{\mathcal{E}}$ will be all the quadruples $((i, j), (k, \ell))$ such that $(i, k) \in \mathcal{E}_1$ and $(j, \ell) \in \mathcal{E}_2$. Recall that, in agreement with the mixing task, we assume that the graph edges always contain all self-loops. If $e_{1,i}, e_{2,j}$ are the canonical basis vectors associated to $i \in \mathcal{V}_1$ and $j \in \mathcal{V}_2$, respectively, we must associate to the corresponding product node (i, j) the Kronecker product vector $e_{1,i} \otimes e_{2,j}$ (see e.g. [31] for more details). Those Kronecker products form a basis for the real space associated to $\hat{\mathcal{V}}$, that is $\mathbb{R}^{N_1 N_2} = \mathbb{R}^{N_1} \otimes \mathbb{R}^{N_2}$, and we denote by $x \in \mathbb{P}_{N_1 N_2}$ a probability distribution over $\hat{\mathcal{V}}$. The product graph $\hat{G}$ of $\mathcal{G}_1$ with any other graph $\mathcal{G}_2$ is always a valid lifted graph of $\mathcal{G}_1$, with the associated surjective map $\mathfrak{c} : \hat{\mathcal{V}} \to \mathcal{V}_1$ where $\mathfrak{c}(v_1, v_2) = v_1$. Indeed by construction, an edge $((i, j), (k, l))$ can be present in $\hat{\mathcal{G}}$ only if (i, k) is an edge of $\mathcal{G}_1$, so graph locality is respected. In particular, the product graph of $\mathcal{G}_1$ with a complete graph is a valid lift of $\mathcal{G}_1$.

The product graph $\hat{\mathcal{G}}$ of $\mathcal{G}$ with an auxiliary graph $\mathcal{G}_2$ is a valuable tool towards designing lifted Markov chains. Indeed, a time-homogeneous Markov chain on $\hat{\mathcal{G}}$ can be defined by specifying any $N_1 N_2 \times N_1 N_2$ stochastic matrix A, with nonzero elements only in positions corresponding to pairs of nodes $\in \hat{\mathcal{V}}$ connected by an edge $\in \hat{\mathcal{E}}$. Replacing the dynamics of P on p by the dynamics of any such A on x, yields a valid LMC on $\mathcal{G}$. Note that in general, x and A take a more general form than a Kronecker product over $\mathcal{G}$ and $\mathcal{G}_2$. Indeed, we a priori allow the step taken on the $\mathcal{V}$ coordinate to depend on the value of the $\mathcal{V}_2$ coordinate (sometimes called the "coin coordinate"). This conditioning should add memory to the process, possibly accelerating its mixing behavior.

The following constructions of LMCs with product graphs are used as building blocks in our examples and in the proofs of the main results.

2.2. Some particular LMC tools

Clock lift: The following construction, which we call a *clock-lift*, allows us to construct a *time-homogeneous* lifted chain, whose marginal follows the evolution of some specified *time-inhomogeneous* Markov chain represented by a finite sequence of T stochastic matrices

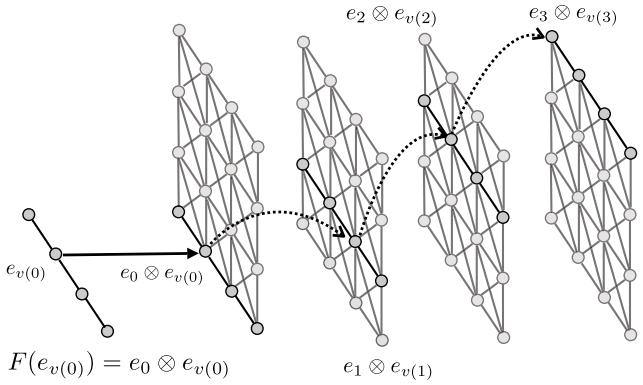

Fig. 1. Illustrating the clock-lift: The initial graph $\mathcal{G}$, here a path on 4 nodes depicted on the left, is lifted to a 4×4 grid $\hat{\mathcal{G}}$ via a product with the path on $\{0, 1, 2, 3\}$. The latter is meant to allow applying a sequence of transition matrices $P(1), P(2), P(3)$ consecutively to the nodes of $\mathcal{G}$. We have depicted 4 copies of $\hat{\mathcal{G}}$ along the times axis (i.e. towards the right) to illustrate the evolution associated to an initial distribution $e_{v(0)}$ on $\mathcal{V}$. This distribution is first mapped to $\hat{\mathcal{V}}$ in a natural way by F, with the first coordinate associated to the sequence $\{0, 1, 2, 3\}$ (vertical axis) and the second coordinate associated to the original $\mathcal{V}$ (depth axis). At the kth time-step, the transition matrix A implies that the first coordinate on $\hat{\mathcal{V}}$ is deterministically incremented by +1 (vertical motion on the figure), while an update $v(k)$ for the second coordinate is selected probabilistically by applying $P(k)$ to $v(k-1)$.

$P(1), P(2), \ldots, P(T)$. This is attained by including the time variable in the node space, in a way that is reminiscent of the inclusion of time as a state variable in dynamical systems theory, in order to replace time-dependence by dependence on this additional coordinate.

Explicitly, consider the product graph $\hat{\mathcal{G}}$ of the original graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the path graph associated to the time interval $[0, T]$. The latter thus has node set $\{0, 1, \ldots, T\}$ and edges $(s, s+1)$ for $s = 0, 1, \ldots, T-1$. The product graph produces a lift that effectively introduces T additional copies of each original node, indexed by time. The basic idea is depicted in Fig. 1 for $\mathcal{G}$ a path graph of 4 nodes. As a lifted node space we thus consider $\hat{\mathcal{V}} = \{(s, v) : s \in \{0, 1, \ldots, T\} \text{ and } v \in \mathcal{V}\}$, and as edges the ones of the product graph. This yields a valid lift of $\mathcal{G}$ with the surjective map $\mathfrak{c} : \hat{\mathcal{V}} \to \mathcal{V}$ defined as $\mathfrak{c}((s, v)) = v$. We then construct a lifted Markov chain on $\hat{\mathcal{G}}$ by choosing a matrix on $\mathbb{R}^{T+1} \otimes \mathbb{R}^N$ of the form:

$$A = \sum_{s=1}^{T} e_s e_{s-1}^\dagger \otimes P(s) + e_T e_T^\dagger \otimes I_{\mathcal{V}} \,,$$

where $I_{\mathcal{V}}$ is the identity on $\mathbb{R}^{|\mathcal{V}|}$. When this LMC is initialized with

$$x(0) = e_0 \otimes p(0) \,,$$

the lifted distribution indeed follows

$$x(t) = A^t x(0) = e_t \otimes P(t)P(t-1)\ldots P(1)p(0) \,,$$

so that $Cx(t) = P(t)P(t-1)\ldots P(1)p(0) = p(t)$ for all $0 \le t \le T$ and $Cx(t) = p(T)$ for all $t \ge T$. We note that the clock-lift is reducible: all initial distributions are eventually mapped

to the set $\{(T, v) : v \in \mathcal{V}\}$, and remain there. The stationary distribution is hence supported only on that set.

Periodic clock lift: A *periodic clock-lift* is a variant of the clock-lift where the product graph is constructed by replacing the path in time by a cycle, obtained by connecting the time-index e_{T-1} back to e_0. The corresponding transition matrix is now:

$$A = \sum_{s=1}^{T-1} e_s e_{s-1}^\dagger \otimes P(t) \; + \; e_0 e_{T-1}^\dagger \otimes P(T)\,.$$

For an initial state $x(0) = e_0 \otimes p(0)$, the output $p(t)$ is then given by periodically applying the time-varying Markov chain transitions $P(1), P(2), \ldots, P(T), P(1), P(2), \ldots$. Sometimes, towards further modifications, it is handy to introduce a "special" step after the T first transition matrices and before re-applying the sequence again. This formally corresponds to applying the above periodic clock-lift to the sequence $P(1), P(2), \ldots, P(T), P(T+1)$, with $P(T+1)$ encoding the "special" step. In contrast to the regular clock-lift, the periodic clock-lift can be irreducible, and have a stationary distribution that is nonzero on the full (lifted) node set.

Node-clock lift: Assume that for each initial node $p(0) = e_i$ with $i \in \mathcal{V}$, we have built target stochastic evolutions $p^{(i)}(t) = P^{(i)}(t)P^{(i)}(t-1)...P^{(i)}(1)\,e_i$, with all the sequences $\{P^{(i)}(k)\}$ satisfying the locality constraints of $\mathcal{G}$. We would then like to merge these N independent evolutions such that, starting from an initial distribution $p(0)$, the system follows the whole trajectory of the particular chain $p^{(i)}(t)$ with a probability $p_i(0)$. In other words, we know what to do when we start at any individual node – it is not always the same – and we want to formally build a *single LMC* which indeed ensures that the appropriate sequences are followed.

We can indeed combine these target sequences for different i independently, into an LMC which we call a *node-clock lift*. As depicted on Fig. 2, the lifted graph is now the product between a path encoding the time index, like in the clock-lift, a complete graph on $\mathcal{V}$ encoding the index i from the above set of sequences, and the original graph. The lifted node space becomes $\hat{\mathcal{V}} = \{(s, v_0, v) : s \in \{0, 1, \ldots, T\}$ and both $v_0, v \in \mathcal{V}\}$. This is a valid lift of $\mathcal{G}$ with the surjective map $\mathfrak{c} : \hat{\mathcal{V}} \to \mathcal{V}$ where $\mathfrak{c}((s, v_0, v)) = v$. To ensure the target evolution, we propose the transition matrix on the lifted graph:

$$A = \sum_{s=1}^{T-1} \sum_{i \in \mathcal{V}} e_t e_{t-1}^\dagger \otimes e_i e_i^\dagger \otimes P^{(i)}(s) + \sum_{i \in \mathcal{V}} e_T e_{T-1}^\dagger \otimes \bar{p}\, e_i^\dagger \otimes P^{(i)}(T) + e_T e_T^\dagger \otimes I_{\mathcal{V}} \otimes I_{\mathcal{V}}\,, \quad (3)$$

with the same notation as for the clock-lift. Here $\bar{p}$ denotes any fixed distribution over $\mathcal{V}$; the motivation for the corresponding term will be explained shortly. When the lift is initialized with

$$x(0) = e_0 \otimes \sum_{v \in \mathcal{V}} p_v(0)(e_v \otimes e_v)\,,$$

the system indeed follows the target stochastic evolution $p^{(i)}(t)$ with a probability $p_i(0)$. In particular, for a distribution $p(0) = e_i$ concentrated on a node i of the original graph, the initial state for the lift is $x(0) = e_0 \otimes e_i \otimes e_i$. By applying A, the state then follows $x(t) = e_t \otimes e_i \otimes p^{(i)}(t)$ over the first $T-1$ time steps. The operation with $\bar{p}$ serves to induce convergence on the v_0 coordinate of (s, v_0, v): when sitting at a lifted node of type $(T-1, v_0, v)$, the corresponding term in A ensures that besides applying the correct transition $P^{(v_0)}(T)$ to v, we also "erase" the no longer necessary information of v_0, towards $\bar{p}$ for any v_0. Note that on the v_0 coordinate of $\hat{\mathcal{G}}$ we have a complete graph so this operation can be implemented at will. Thanks to this

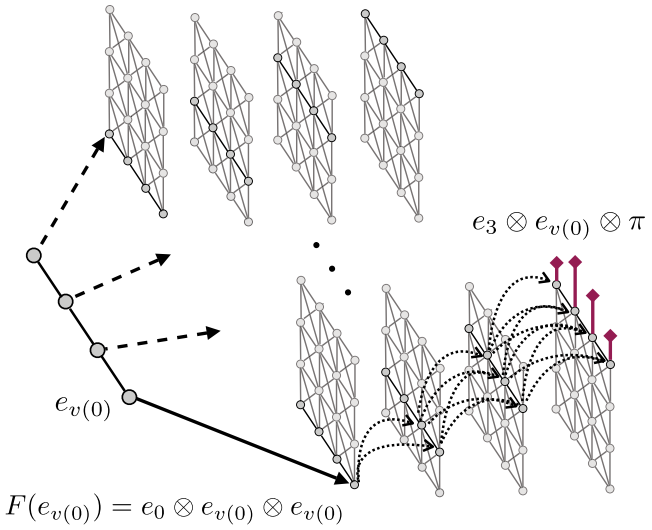

Fig. 2. Illustrating the node-clock-lift, in which the assigned map sequences $P^{(i)}(1), \dots, P^{(i)}(T)$ that were given for each extreme initial distribution $p(0) = e_i$, are combined into a single time-homogeneous lifted chain by essentially combining N clock lifts. The same initial graph $\mathcal{G}$ as on Fig. 1 (a path on 4 nodes) is depicted on the left. The lifted graph now corresponds to the product of the 4×4 grid, depicted on Fig. 1, with a complete graph on 4 nodes; to avoid clutter, we have just represented this in the depth direction as 4 versions of the grid, and subsumed all the possible edges for jumping between those versions. We have depicted 4 copies of this $\hat{\mathcal{G}}$ along the times axis (i.e. towards the right) to illustrate the evolution associated to an initial distribution on $\mathcal{V}$. This distribution is first mapped to $\hat{\mathcal{V}}$ in a natural way, with weight $p(0)_i$ associated to each node $(0, i, i) \in \hat{\mathcal{V}}$, where the first coordinate is associated to the time sequence $\{0, 1, 2, 3\}$ (vertical axis), the second coordinate is associated to the map sequence index (here grid-version on the depth axis) and the third coordinate is associated to the original $\mathcal{V}$ (here position along depth axis on the selected grid). At the kth time-step, the transition matrix A implies that the first coordinate of $(s, v_0, v) \in \hat{\mathcal{V}}$ is deterministically incremented by +1 (vertical motion on the figure), the second coordinate i.e. selection of grid version does not change (except possibly via $\bar{p}$ on the very last step, not represented here), while an update $v(k)$ for the third coordinate is selected probabilistically by applying $P^{(v_0)}(k)$ to $v(k-1)$.

operation, if $p^{(i)}(T) = \pi$ for all i and provided the lift is accordingly initialized, the node-clock-lift converges to a unique distribution not only over $\mathcal{V}$ (this would be π) but also over $\hat{\mathcal{V}}$ (this would be $\hat{\pi} = e_T \otimes \bar{p} \otimes \pi$). Note that some of the lifted nodes will never be populated (e.g. $(0, v_0, v) \in \hat{\mathcal{V}}$ with $v_0 \neq v$), so in fact $\hat{\mathcal{G}}$ can be slightly reduced at the cost of losing the compact description as a product graph.

Similarly to the periodic clock-lift, we can construct a *periodic node-clock-lift*, where we identify $e_T \otimes e_v \otimes e_v$ with $e_0 \otimes e_v \otimes e_v$. I.e., the nodes of type (T, v_0, v) are dropped and each node of type $(T-1, v_0, j)$ now jumps to $(0, v(j), v(j))$, where $v(j)$ is selected among the neighbors of j in $\mathcal{V}$ according to the transition matrix $P^{(j)}(T)$. This ensures that after such step the sequences $P^{(i)}(t)$ are applied repeatedly. The "erasure" operation is dropped in this construction, but the new transition matrix applies another convergence on the v_0 coordinate of a lifted node (s, v_0, v), namely making it equal to the v coordinate whenever jumping from $s = T-1$ to $s = 0$. Thus if the v coordinate converges, the v_0 coordinate ought to as well.

Example 1.1 mentions mixing speedups which can be obtained for instance with time-dependent Markov chains; those can thus be reformulated as an LMC using a clock-lift. This

hints at how the LMC tools introduced here will allow to study mixing time bounds for general algorithms, reformulated as LMCs.

3. Mixing time and design scenarios

The overarching message of this paper is that the achievable mixing performance is critically dependent on some constraints and insensitive to some others, besides the locality associated to the graph $\mathcal{G}$. We now specify those constraints and emerging design scenarios considered here, in the LMC framework. It will appear that even the definition of mixing time depends on the imposed constraints.

3.1. Initialization of the lift

When a stochastic dynamics is seen as an algorithm, one must specify how to initialize it. The unmovable part of the initialization is a distribution $p(0) = p_0$ over the nodes of $\mathcal{V}$. We follow the standard mixing literature by assuming that the designer has no insightful control on this distribution, so the algorithm has to treat all distributions $p(0) \in \mathbb{P}_N$. For further inputs to the algorithm, we will consider two possible scenarios and the associated sets of initial distributions $\mathcal{S}$:

(S) In a first scenario, the algorithm designer is allowed to tune the initial values of the latent variables in the LMC. More precisely, in addition to $\hat{\mathcal{G}}$ and A, the algorithm designer can choose how to lift the weight $p_k(0)$ attributed to each node $k \in \mathcal{V}$ of the original graph $\mathcal{G}$, onto a distribution over its associated lifted nodes $x_{\mathfrak{c}^{-1}(k)}(0)$ in agreement with the locality constraints. We further require that the designed initialization is a *linear* map $F : p(0) \mapsto x(0)$. These constraints imply that

$$CFp(0) = p(0) \quad \text{for all} \quad p(0)\,.$$

This last condition avoids exchanges among nodes through F before starting the actual algorithm, namely imposing $F_{k,j} = 0$ whenever $\mathfrak{c}(k) \neq j$. Such initialization map is compatible with the clock-lift and node-clock-lift constructions proposed in Section 2.2. In this scenario, the set $\mathcal{S}$ of relevant initial conditions for the LMC does not comprise all possible distributions $x(0)$ on $\hat{\mathcal{V}}$, but only those of the form $x(0) = Fp(0)$, for all initial distributions $p(0)$ on $\mathcal{V}$ and a fixed designed map F.

(s) In other cases, there might be no insightful control over the initialization of the lifted dynamics. The set of relevant initial conditions $\mathcal{S}$ for the LMC is then the whole $\mathbb{P}_{\hat{N}}$.

3.2. Invariance of the target marginal

For a Markov chain on $\mathcal{G}$, mixing is necessarily towards its unique *invariant* distribution, i.e. $P\pi = \pi$. For a lifted Markov chain, however, even if the marginal converges to π, having $Cx(t) = \pi$ at some time does not necessarily imply $Cx(t+1) = \pi$. While imposing such property at all intermediate times turns out to restrict the potential role of any additional memory slots too much, one may reasonably request that at least the system does not leave the target π when it starts there at $t = 0$. It arguably imposes to "avoid unnecessary work"; it can also play an essential role towards interfacing the Markov chain with other algorithmic elements, in particular implementing the key task of *amplification*, i.e., boosting the success probability of a randomized algorithm (in our case the closeness to the stationary distribution)

by rerunning the algorithm on its own output, see e.g. [48]; and in a similar spirit it allows to ensure that the LMC *stabilizes* the system towards π, even in presence of perturbations (see [8]). We thus identify two possible scenarios:

(i) We impose $Cx(t) = \pi$ for all $t > 0$ whenever $Cx(0) = \pi$, for all $x(0) \in \mathcal{S}$.
(I) We allow $Cx(t) \neq \pi$ even when $Cx(0) = \pi$.

3.3. Marginal vs lift mixing time

In the examples from the introduction, we have encountered two different ways to define the mixing time of a Markov chain or LMC. The "Monte-Carlo" interpretation considers the smallest possible T that guarantees we are close to the target distribution over the nodes. On the other hand, the "Las-Vegas" viewpoint considers the *expectation* of some probabilistic stopping time t such that, conditioned on having stopped, we are close to the target distribution over the nodes. We refer the interested reader to the book [40]. In this paper, we analyze the mixing time in the Monte-Carlo sense, which we define next. An equivalence exists between the two settings, as we discuss at the end of Example 4.1.

A Markov chain on $\mathcal{G}$ associated to a transition matrix P is said to mix to π if $P\pi = \pi$ and if for all $\epsilon > 0$ there exists $\tau(\epsilon) > 0$ such that, for all $p \in \mathbb{P}_N$, we have:

$$\|P^t p - \pi\|_{TV} \leq \epsilon \quad \text{for all} \quad t \geq \tau(\epsilon).$$

We call $\tau(\epsilon)$ its ϵ-mixing time.[2] It is typical to consider $\tau(1/4)$ as a reference mixing time. Since LMCs are themselves Markov chains, this can be directly translated to the mixing time of x on the lifted space towards its stationary value $\bar{x}$ — with the obvious slight modification that in scenarios with initialization (S) the convergence property must hold only for all $x \in \mathcal{S}$, instead of all $x \in \mathbb{P}_{\hat{N}}$. Thus, the LMC represented by A mixes to $\bar{x}$ with mixing time $\tau(\epsilon)$ if and only if, for all $x \in \mathcal{S}$ and all $\epsilon > 0$, we have:

$$\|A^t x - \bar{x}\|_{TV} \leq \epsilon \quad \text{for all} \quad t \geq \tau(\epsilon).$$

Most papers which propose bounds on LMC mixing time (see Section 3.7) do indeed consider this convergence criterion. Our original algorithmic task however is to accelerate convergence of the *marginal* $p(t) = Cx(t)$, compared to the performance of the original chain P. To this aim, we define the marginal mixing time.

Definition 2 (*Marginal Mixing Time*)**.** A lifted chain on $\hat{\mathcal{G}}$ associated to a transition matrix A is said to mix to the marginal π on $\mathcal{G}$ from initial conditions $\mathcal{S}$, if for all $\epsilon > 0$ there exists $\tau_M(\epsilon) > 0$ such that for all $x \in \mathcal{S}$ we have:

$$\|CA^t x - \pi\|_{TV} \leq \epsilon \quad \text{for all} \quad t \geq \tau_M(\epsilon).$$

We call $\tau_M(\epsilon)$ its ϵ-*marginal mixing time*.

Of course, $\tau_M(\epsilon) \leq \tau(\epsilon)$ for all ϵ. While the convergence of x is a sufficient proxy for the convergence of $p = Cx$, it is not truly necessary. The distinction could be especially relevant because LMCs specifically designed to speed up convergence may be all but generic. For instance, some typical designs that involve constructions where the lifted Markov chain x

[2] We recall that the total variation distance between two distributions p and p' is $1/2$ times the 1-norm of their difference $\|p - p'\|_1 = \sum_{i=1}^{N} |p_i - p'_i|$.

in fact does not even converge to a stationary value, but the projected state p does — see the clock lifts in 2.2. Furthermore, it is easy to find lifted walks where p converges much faster than x, see e.g. Example 4.1. One could wonder whether conversely, a Markov chain with quickly converging p can always be adapted to also have quickly converging x, or whether sometimes there is a strict advantage to be gained when we are only interested in p. We therefore distinguish which type of convergence we are requesting:

(M) The aim is to optimize convergence of the marginal $p(t)$ towards π, as measured by $\tau_M(\epsilon)$.
(m) The aim is to optimize convergence of $x(t)$ towards $\bar{x}$, as measured by $\tau(\epsilon)$.

We will focus mainly on the mixing time for $\epsilon = 1/4$. This usually represents a good quantification of the general ϵ-mixing time, and is therefore commonly referred to as the main mixing time parameter in the literature [40]. For completeness we do mention that this restriction is not perfect. Indeed, the motivation is that typically the bound $\tau(\epsilon) \in O(\tau(1/4)\log(1/\epsilon))$ holds for any $\epsilon > 0$. On the one hand, the actual ϵ-mixing time might be better (by a log-factor). For instance on the cycle graph, we can have that actually $\tau(\epsilon) \in O(\tau(1/4))$ (see Section 6). On the other hand, as we will see in Section 6.4, the bound does not always hold for LMCs. Initial dynamics may bring the LMC quickly to $\epsilon = 1/4$ distance, but after this the convergence speed may slow down. Nonetheless, in general the $\tau(1/4)$ mixing time remains a good indicator, and we will use it as a reference throughout the paper.

3.4. Reducibility of the lift

Irreducibility means that there cannot exist a partition of $\mathcal{V}$ into subsets $\mathcal{X}$ and $\mathcal{V} \setminus \mathcal{X}$ such that $P_{i,j} = 0$ for all (i, j) with $i \in \mathcal{X}$ and $j \in \mathcal{V} \setminus \mathcal{X}$. Several related robustness properties can motivate the use of irreducible Markov chains instead of reducible ones. For instance, irreducible processes keep mixing all nodes for all times, such that the effect of an occasional erroneous transition at some time step could always be corrected in the future.

In order to converge, starting from any initial state, to a unique stationary distribution π with $\pi_i > 0\ \forall i$, the original Markov chain P must be irreducible. However, the same need not necessarily apply to a lifted Markov chain A, depending on the rest of the setting. Hence two scenarios emerge, when P is irreducible:

(R) The lifted Markov chain A is allowed to be reducible.
(r) The lifted Markov chain A must be irreducible (if P is).

From a mixing viewpoint, one could wonder whether reducible lifts could lead to singularly fast behavior.

When the graph associated to P is itself reducible, the lifted graph will always be reducible too.

3.5. Matching ergodic flows

When $\mathcal{G}$ and π are given, one can think about the optimization problem of computing the compatible Markov chain P with fastest convergence towards π, see for instance [14] for symmetric P. For the lifted Markov chain, one could perform a similar optimization on A — with the added difficulties that (i) the discrete structure $\hat{\mathcal{G}}$ is to be designed too and (ii) accelerating convergence has been shown to require irreversibility (thus not symmetric

A) [15]. Yet in some cases, one may be given not only $\mathcal{G}$ and π but also a reference P, whose associated typical or "ergodic flows" between distinct nodes, for some reason, should not be exceeded by the transitions induced by the lifted chain [15]. This could be relevant when for instance certain transitions should not be overloaded. Note that the flow from a node to itself is usually not constrained; therefore, when considering ergodic flows, we discard what happens for $i = j$.

More precisely, for a given Markov chain P, the associated ergodic flows are defined by $Q^{(P)}_{i,j} = P_{i,j}\pi_j$, i.e. the weight that flows from j to i when the system is on the steady state π. For a lifted chain A and (one of) its steady state(s) $\hat{\pi}$, one can similarly define ergodic flows $\hat{Q}^{(A;\hat{\pi})}_{i,j} = A_{i,j}\hat{\pi}_j$. To compare ergodic flows in A and P, one then compares $Q^{(P)}_{i,j}$ and

$$\hat{Q}^{(A;\hat{\pi})}_{\mathfrak{c}^{-1}(i),\mathfrak{c}^{-1}(j)} = \sum_{\ell,k:\mathfrak{c}(\ell)=i,\mathfrak{c}(k)=j} \hat{Q}^{(A;\hat{\pi})}_{\ell,k}.$$

The Markov chain $\tilde{P}^{(A;\hat{\pi})}$ on $\mathcal{G}$ defined by $\tilde{P}^{(A;\hat{\pi})}_{i,j}\pi_j = \hat{Q}^{(A;\hat{\pi})}_{\mathfrak{c}^{-1}(i),\mathfrak{c}^{-1}(j)}$ is called the *induced chain on $\mathcal{G}$ by the lift A and distribution $\hat{\pi}$* [3]. The ergodic flows of A with respect to the stationary distribution $\hat{\pi}$ are equal to the ergodic flows of P if the induced chain $\tilde{P}^{(A;\hat{\pi})} = P$; they do not exceed the ergodic flows of P if $\tilde{P}^{(A;\hat{\pi})}_{i,j} \leq P_{i,j}$ for all $j \neq i$. This leads to the following scenarios.

- (e) A reference Markov chain P is given and the ergodic flows of A cannot exceed the ergodic flows of P. When A has several steady states $\hat{\pi}$ and hence several ergodic flows $\hat{Q}^{(A;\hat{\pi})}$ corresponding to the same π and accessible from $\mathcal{S}$, this must hold for each of them.
- (E) No constraint is imposed on the ergodic flows of the lifted Markov chain.

Note that the definition of ergodic flows for the case of non-unique $\hat{\pi}$ is an extension of the traditional definition, see for instance [3].

3.6. On combinations of constraints and relations between lift design problems

In the following, in order to compactly refer to a set of requirements in the statements and comparisons, we shall specify an alternative (upper- or lower-case letter) for each of the properties described in the previous subsections, and denote the set of dynamics allowed by a design scenario by the corresponding string of 5 letters, e.g. (sImrE). A shorter string may be used to indicate properties that hold for all the compatible alternatives: e.g. (sImE) includes the scenarios (sImrE) and (sImRE).

The lower- and upper-case letters in the previous subsections have been chosen so that the upper-case are *less constraining* than the corresponding lower-case letter. This implies that a scenario associated to some capital letters always includes all lifts available in a scenario where some of those letters are substituted by their lowercase versions, i.e.:

$$\begin{aligned} &(\mathtt{S})|_x \supseteq (\mathtt{s})|_x \ ; \ (\mathtt{I})|_x \supseteq (\mathtt{i})|_x \\ &(\mathtt{M})|_x \supseteq (\mathtt{m})|_x \ ; \ (\mathtt{R})|_x \supseteq (\mathtt{r})|_x \ \ ; \ (\mathtt{E})|_x \supseteq (\mathtt{e})|_x. \end{aligned} \tag{4}$$

Here $|_x$ denotes arbitrary choices, equal on the left and right hand sides of the inclusion symbol, for the other four letters. The inclusion symbol expresses that all algorithms allowed by the right-hand side, are also allowed by the left-hand side. From this it follows that we obtain lower or at most equal optimal mixing times when relaxing the constraints from lowercase to uppercase. The comparison of E, e scenarios with non-lifted chains needs particular care.

Indeed, requiring no particular ergodic flows allows to select better lifted walks, but it simultaneously gives more freedom to optimize P for the non-lifted process.

In regard of these scenarios, our investigation thus aims at answering questions like: how much can be gained by relaxing a constraint, e.g. from (r) to (R) or from (e) to (E)? Conversely, how constraining are such requirements, usually made as basic assumptions of the setting in the literature like [15], on ultimately achievable mixing speed? How much are differences in mixing speedup, found in the literature, tied to the constraints/properties imposed sometimes implicitly on the setting?

In this paper, we will a priori not pose any definite judgment on the relevance of one or the other combination of constraints, and thus just report the bounds on the mixing speed for all scenarios.

Remark. In addition to general properties on the process as we list above, an important aspect of engineered mixing dynamics is the amount of resources they require for implementation. Following other abstract work on mixing bounds, we will not consider this aspect in detail, as it further depends on the specific application and the information available at the algorithm design stage. All the constructions we employ in the proofs, while having mostly existential rather than practical value, have a number of lifted nodes polynomial in N. Their main goal however is not so much constructive, but rather to assess the ultimate potential of various algorithms for accelerated mixing on the basis of their fundamental properties.

3.7. Directly related previous work

Previous work by [18,19] and [15] have considered the (sImr) or (sImre) context. In particular, [15] has established the conductance bound that we derive in Theorem 3, for the particular scenario (sImre). More recently, this result was extended to the continuous-time case [54], which we will not treat here. In the following sections, however, we will prove that such bounds can be derived from just the (s) and (se) constraints, and that (s) is also in some sense necessary: if we relax the setting to any scenario of (SI), then the diameter becomes the trivial yet tight lower bound.

Our upcoming Theorem 2 on diameter-time mixing is reminiscent of finite-time consensus results such as those in [29]. Here we have a more restricted setting, including positivity constraints and time-invariant dynamics. The lift takes care of the time-dependence of the algorithm, as we shall see, while positivity is intrinsically built in our framework, apparently without affecting the fast convergence. Our results show that diameter-time convergence is obtained at the cost of loosening some of the above properties with respect to [15,18,19].

4. LMCs for existing algorithms and their properties

Before getting to the results, we illustrate the different LMC constraints by showing how the speedup approaches mentioned in the introduction can be cast into the LMC framework. The resulting LMCs will be used as a basis to apply the bounds that we derive in the upcoming sections, in order to deduce mixing time properties for the algorithms themselves. Note that the translation of a given algorithm into an LMC is not unique. A trivial example is that one can always construct a new LMC by adding, to an existing LMC, dummy nodes which the lifted walker never reaches but which could change some lift properties. Less trivial variations can arise though. In fact, this ambiguity goes back to Markov chain modeling itself. We try

below to give, for each example, a lift with reasonably small number of lifted nodes and which follows in a reasonably direct way from the algorithm description.

When investigating the constraints satisfied by an algorithm, the translation into the LMC framework merits two remarks.

- Some LMC properties are a direct and unambiguous consequence of the algorithm itself: e.g. whether the algorithm keeps the target invariant when it starts there, thus property (i).
- The satisfaction of other constraints might depend on the aforementioned non-unique translation into an LMC. In this case, the different LMC translations will give different lower bounds on the algorithm's mixing time; the algorithm itself would of course satisfy the most stringent of these bounds.

The selection of examples provided here is nowhere close to exhaustive, and serves mainly to illustrate the variety of general ideas.

Example 4.1 (*Blind Stopping Rules and Cesaro Mixing*)**.** In this example, we explicitly construct a lift for algorithms with blind stopping rules, and further discuss the difference between (m) and (M) in a concrete case.

Consider a stopping rule associated to a Markov chain P that chooses a stopping time according to some distribution ν over a finite interval $[0, T-1]$. We call this a *blind stopping rule*. Cesaro averaging corresponds to a special blind stopping rule where ν is the uniform distribution. We implement such rule as an LMC by adding to the original node set a binary register $\{\texttt{run}, \texttt{stop}\}$ and a register keeping the timestep $\{0, \dots, T\}$ (see Section 2.2 for details on the standard *clock-lift* construction), resulting in the lifted node set

$$\hat{\mathcal{V}} = \{\texttt{run}, \texttt{stop}\} \times \{0, \dots, T\} \times \mathcal{V}.$$

By linearity, the associated LMC is entirely defined by its action on $x = e_i$ for all $i \in \hat{\mathcal{V}}$. It can then be described as follows:

> "When $x = e_{\texttt{run},t,i}$ for some $t < T$ and $i \in \mathcal{V}$, go to $e_{\texttt{stop},t+1,i}$ with probability $\tilde{\nu}(t)$ and to $e_{\texttt{run},t+1,j}$ with probability $(1-\tilde{\nu}(t))P_{j,i}$. Otherwise, stand still".

Here $\tilde{\nu}$ is a rescaled probability distribution over $[0, T]$ defined such that $(1-\tilde{\nu}(t+1))(1-\sum_{t'=0}^{t}\nu(t')) = 1-\sum_{t'=0}^{t+1}\nu(t')$, i.e. it expresses the conditional probability of stopping at t, knowing that we have not stopped before. We can describe this LMC by the transition matrix below, where we "$\otimes$" denotes the Kronecker product (see Section 2.1):

$$\begin{aligned} A = \sum_{t=0}^{T-1} \Big[& (1-\tilde{\nu}(t)) e_{\texttt{run},t+1} e_{\texttt{run},t}^{\dagger} \otimes P \\ & + \tilde{\nu}(t) e_{\texttt{stop},\texttt{t}+1} e_{\texttt{run},\texttt{t}}^{\dagger} \otimes I_{\mathcal{V}} + e_{\texttt{stop},\texttt{t}+1} e_{\texttt{stop},\texttt{t}}^{\dagger} \otimes I_{\mathcal{V}} \Big] \\ & + (e_{\texttt{stop},\texttt{T}} e_{\texttt{stop},\texttt{T}}^{\dagger} + e_{\texttt{stop},\texttt{T}} e_{\texttt{run},\texttt{T}}^{\dagger}) \otimes I_{\mathcal{V}} \ , \end{aligned}$$

initialized with $F_{(\texttt{run},0,i),i} = 1$, and zero elsewhere. Note that some choices are arbitrary in this construction, e.g. how the t part of the lifted nodes evolves when we are on stop and what happens in the case run, T, which has no incoming edges and thus will never be populated with initialization F.

This procedure has a finite running time T and consequently, it will only exceptionally reach a unique final distribution for all initial conditions. A standard procedure to induce convergence,

as proposed in [43], is to apply *amplification*, that is, after having run the algorithm once, it is run again on its own output. This amplification is easy to incorporate into the above LMC with the following modification:

$$\begin{aligned} A = \sum_{t=0}^{T-1} \Big[& (1-\tilde{\nu}(t)) e_{\mathtt{run},t+1} e^\dagger_{\mathtt{run},t} \otimes P \\ & + \tilde{\nu}(t) e_{\mathtt{stop},\mathtt{t}+1} e^\dagger_{\mathtt{run},\mathtt{t}} \otimes I_{\mathcal{V}} + e_{\mathtt{stop},\mathtt{t}+1} e^\dagger_{\mathtt{stop},\mathtt{t}} \otimes I_{\mathcal{V}} \Big] \\ & + ((1-\gamma) e_{\mathtt{stop},\mathtt{T}} e^\dagger_{\mathtt{stop},\mathtt{T}} + e_{\mathtt{stop},\mathtt{T}} e^\dagger_{\mathtt{run},\mathtt{T}}) \otimes I_{\mathcal{V}} \\ & + \gamma e_{\mathtt{run},0} e^\dagger_{\mathtt{stop},\mathtt{T}} \otimes I_{\mathcal{V}} \ , \end{aligned} \tag{5}$$

for some $\gamma \in [0, 1)$.

The algorithm clearly falls under the (S) scenarios, i.e. it corresponds to an imposed initialization of the LMC. The LMC is still reducible (R) because of the nodes of the form $(\mathtt{run}, T, v)$. By dropping those unnecessary nodes, it becomes irreducible (r) provided P was irreducible, for all $\gamma \in (0, 1]$.

For $\gamma = 1$ the chain is periodic and x does not converge to a unique steady state (consider the t part of the lifted nodes), but the marginal $p = Cx$ in general does converge to the unique stationary distribution π of P, possibly faster than with the Markov chain P alone, so this is (M). For $\gamma < 1$, the whole LMC is irreducible and aperiodic, and thus it converges to a unique $\hat{\pi}$ (m), but this is not really the initial focus of the algorithm. Since only P or $I_{\mathcal{V}}$ are applied altogether to the $\mathcal{V}$ component of the lifted nodes, it is easy to check that with the initialization F the situation $x(0) = CF\pi$ maintains $Cx(t) = \pi$ for all $t > 0$, i.e. invariance holds (i); this also confirms that $C\hat{\pi} = \pi$, i.e. the algorithm reaches the correct limit distribution. Finally, the ergodic flows of A do not exceed the ergodic flows of P since, by construction, $\tilde{P}^{(A;\hat{\pi})}$ will in fact correspond to $(1-\alpha)P + \alpha I$ for some $\alpha \in [0, 1]$; in other words, the lift only contains additional weight on the self-loops, caused by the probability of stopping for some steps before running P again.

Interestingly, the convergence rates for p and x with this algorithm can be very different indeed. For instance for $(1-\gamma) \ll 1$, the marginal p converges quite fast to π, according to the stopping time algorithm with amplification [40,43]; but the t component of the lifted nodes in x evolves essentially periodically, with only order $(1-\gamma)$ mixing at each step, and thus it converges slowly. The (m) scenario as considered in [15] and related papers, would thus suggest to avoid $\gamma \simeq 1$ with this LMC. But this, and more importantly the lower bounds of [15] and related papers on the mixing time for scenarios with (m), might just be related to the particular LMC translation. In particular, *spectral approaches* to estimate the LMC convergence rate would automatically examine (m). We thus have a first concrete question: may the focus on (M) instead of (m) allow to break some known bounds on mixing time?

A more concrete example with Cesaro mixing on the cycle graph is discussed in Example 6.2.

To finalize this example we note that more general (non-blind) stopping rules also exist to reach the Markov chain stationary distribution. These may depend on the specific nodes that have been visited, rather than only on the number of steps. However, in [43, Theorem 4.26] and related material it is shown that an equivalence can be established, *at least as long as the stopping condition does not modify the limit distribution.* Namely, if such a general stopping rule has an expected stopping time T, then the amplified Cesaro average of the Markov chain over $O(T \log 1/\epsilon)$ steps will also be ϵ-close to the stationary distribution. As

a consequence, we can achieve the same mixing performance using the blind Cesaro stopping rule. The lower bounds on the Cesaro mixing time will therefore carry over to lower-bound the expected stopping time for general stopping rules. When the stopping rule is instrumental in modifying the limit distribution, our bounds would not carry over though. □

Example 4.2 (*Non-Backtracking Random Walks*)**.** Non-backtracking random walks on undirected graphs, as in [4,19,22,36], can be translated rather directly to the following LMC. The lifted nodes $\hat{\mathcal{V}} = \mathcal{E}$ correspond to the directed edges of the original graph, such that $(i, j) \in \hat{\mathcal{V}}$ means that the walker is on node $i \in \mathcal{V}$, and coming from node $j \in \mathcal{V}$. Then the non-backtracking transition matrix is defined by $A_{(i,k),(k,l)} = (1-\alpha)/(d_k - 1)$ for $i \neq l$, and $A_{(i,k),(k,i)} = \alpha$, where d_k is the degree i.e. the number of edges incident on node k, and α is supposed to be small.

In the literature, this algorithm has been considered with two types of initialization.

- In [4,22,36], the lift is initialized with "all nodes treated equally" in the graph. This corresponds to the initialization map $F_{(i,j),i} = 1/d_i$ for $(i, j) \in E$ and zero otherwise, i.e. a uniform superposition over all possible "previous node" assumptions. It is straightforward to check that this lift is invariant, thus resulting in the setting (`Si`).
- In [19], they do not specify an initialization, i.e. we are in a situation (`s`). However, in that case the LMC does not satisfy invariance, so we are in (`sI`). Indeed, consider for instance π being the uniform distribution over the cycle on N nodes, and an initialization where $Cx(0) = \pi$ and $x_{(1,2)}(0) = x_{(3,2)}(0) = 1/N$, i.e. by some chance we assume initially that both nodes 1 and 3 were reached coming from node 2. In the next step, the algorithm will avoid going back to node 2, from any of its neighbors, so we will have $p_2(1) = 2\alpha/N \ll 1/N = \pi_2$, implying $Cx(1) \neq \pi$.

One could thus wonder which setting gives the most promising convergence speed, more generally: can we go faster with an algorithm that satisfies (`Si`), or one that satisfies (`sI`)? Apart from this distinction, the LMC associated to the non-backtracking random walk is irreducible and by keeping the symmetry on graphs it also matches the ergodic flows of the simple random walk. The object of interest is clearly the marginal over $\mathcal{V}$. □

Example 4.3 (*Time-Dependent Markov Chains, Simulated Annealing and Gather-and-Distribute Strategies*)**.** Applying a different transition matrix $P^{(t)}$ at each time step can speed up convergence. To cast this into the LMC setting, we use a clock-lift (see Section 2.2): the lifted nodes are taken of the form (i, s) with $i \in \mathcal{V}$ and $s \in \{0, 1, \ldots, T\}$. The transition matrix is defined by $A_{(i,s+1),(j,s)} = P^{(s)}_{i,j}$ for $0 \leq s \leq T-1$, the action of A on $e_{i,T}$ to be defined, and zero elsewhere. The LMC is initialized with $F_{(i,0),i} = 1$, and zero elsewhere. The LMC leaves some freedom about the action of A on $e_{i,T}$, i.e. the course of action after the end of the provided sequence $P^{(s)}$. The properties of the related LMC can depend on this choice and on the sequence $P^{(t)}$.

As an example, choose the action of A on $e_{i,T}$ to be $A_{(j,0),(i,T)} = P^{(T)}_{j,i}$. This choice would imply that effectively the LMC periodically applies the sequence $P^{(1)}, \ldots, P^{(T)}$. I.e., it corresponds to a time-dependent Markov chain $P^{(t)}$ with $P^{(t)} = P^{(t \bmod (T+1))}$. This yields a scenario (`Mr`). If furthermore each $P^{(t)}$ has the same stationary distribution π, then we can be

in a scenario (Si). An example of such algorithm on the cycle graph is

$$\begin{aligned} P^{(0)} &= P^{+} = \sum_{k=1}^{N} e_{(k+1)\mathrm{mod}(N)} e_k^{\dagger}, \\ P^{(1)} &= (1-\alpha) \sum_{k=1}^{N/2} e_{(2k+1)\mathrm{mod}(N)} e_{2k}^{\dagger} + (1-\alpha) \sum_{k=1}^{N/2} e_{2k} e_{(2k+1)\mathrm{mod}(N)}^{\dagger} + \alpha \sum_{k=1}^{N} e_k e_k^{\dagger}, \end{aligned} \tag{6}$$

for some parameter $\alpha \in [0, 1]$. This procedure is inspired by the LMC of [15,18], further discussed in Example 6.1. It improves the mixing time to $O(N)$, compared to the random walk $O(N^2)$.

With the choice of $A_{(j,0),(i,T)} = P^{(T)}_{j,i}$, i.e. periodically repeating the sequence of transition matrices, one can relax the initialization $F_{(i,0),i} = 1$ and instead allow to start with any distribution over the lifted nodes (i, s). In this case, even if all the $P^{(k)}$ have the same stationary distribution π, invariance is in general lost. Indeed, consider e.g. the periodically repeating LMC associated to (6) and initialized with $x(2k, 0) = x(2k + 1, 1) = \frac{1}{N}$, thus satisfying $p(j) = \pi(j) = 1/N$ for all nodes j; then after the first time step, we have $p(2k) = x(2k, 1) = \frac{1-\alpha}{N}$ while $p(2k + 1) = x(2k + 1, 0) = \frac{1+\alpha}{N}$, thus $p \neq \pi$. Thus there appears to be a choice between (Si) and (sI), to which we will come back. The following are general and concrete algorithms that fall under the framework of time-varying Markov chains.

Both simulated annealing and gather-and-distribute schemes over any finite time T can also be represented by a time-dependent sequence of Markov chains $P^{(t)}$, but now typically with each a different stationary distribution $\pi(t)$. For simulated annealing, $P^{(t)}$ and $\pi(t)$ converge progressively towards the desired P and π. For gather-and-distribute, the sequence results from an "intuitively simple" way of moving probability mass around, mostly in a deterministic way, during two consecutive time intervals: (i) during the "gather" time interval, we choose $P^{(1)}, \ldots, P^{(D_{\mathcal{G}})}$ so that all the probability mass moves onto a single designated root node, $p(D_{\mathcal{G}}) = e_r$ (e.g., using a shortest-path tree of the graph, with all transition probabilities aimed towards this designated root node r); (ii) then, during the "distribute" time interval, we have the advantage to start from a known distribution $p(D_{\mathcal{G}}) = e_r$ and we can therefore easily design a sequence $P^{(D_{\mathcal{G}}+1)}, \ldots, P^{(2D_{\mathcal{G}})}$ from e_r to π to redistribute this probability mass according to the target stationary distribution π. Note that unlike in the previous paragraph or in Example 4.1, simulated annealing and gather-and-distribute do not come with the idea of re-applying the same algorithm on its own output. The algorithmic idea is thus rather to take $A_{(j,T),(j,T)} = 1$ for all j. Then the LMC is reducible, because lifted nodes of the form (i, s_1) cannot reach any lifted nodes of the form (j, s_2) with $s_2 < s_1$. Due to the time-varying stationary distribution $\pi(t)$ of $P^{(t)}$ in the original algorithm, it is clear that $CAF\pi = P^{(1)}\pi \neq \pi$ in most cases, so we are in a scenario with (SIM). Since the steady state in fact involves no probability flows, we trivially do not exceed the ergodic flows of P. □

Example 4.4 (*Lifted and Data-Augmented Markov Chains*)**.** Lifted Markov chains, in the restricted sense of [15,18], enlarge the state space to introduce memory or momentum effects into the dynamics of some Markov chain P, without assuming further control elements. The mixing time of this enlarged chain is therefore determined without initialization, (s), so with respect to arbitrary initial states on the lifted state space, and mixing is required on the entire lifted state space (m). In addition the lifted Markov chain is required to be irreducible (r), to respect the ergodic flows (e) and can violate invariance of π (I). Other designs of lifted

Markov chains in this explicit sense have been proposed in [17,28,41] on the basis of spatial directionality on the graph, and in [10,41] inspired by network communication ideas.

Data-augmentation, as in [60], is implemented in situations where a Markov chain P over a node set $\mathcal{V}$ is more naturally or easily formulated in a factorized way: $P_{i,j} = \sum_z P''_{i,(j,z)} P'_{(j,z),j}$ with $z \in \mathcal{V}'$ some latent variable. Thus first P' maps j to some (j,z), with z a random auxiliary state value, then P'' maps back (j,z) to a random node $i \in \mathcal{V}$, the original node set.[3] For examples in the settings of genetics or statistical inference we refer the interested reader to [60]. Although this formulation will of course converge on $\mathcal{V}$ with exactly as many iterations of the product $P''P'$ as the underlying random walk with iterations of P itself, we show for illustration how it can be reformulated as an LMC. In some contexts, this analysis might also suggest how the original data-augmentation algorithm could be modified towards speeding up convergence. For the LMC, we introduce the lifted nodes $\hat{\mathcal{V}} = \mathcal{V} \otimes \mathcal{V}'$. The lifted transition matrix A satisfies $A_{(i,z),(j,z')} = P''_{i,(j,z)} P'_{(j,z),j}$ (so the arrival lifted nodes distribution does not depend on z'). The output map C corresponds to summing over z, and the induced Markov chain will be equal to P. Data-augmented Markov chains fall in the class (`siMRe`). Indeed, the algorithm comes with no clear way to initialize the z latent variables of the lift. The lift keeps an initial state $Cx(0) = \pi$ invariant, since the component i of CAx is given by

$$\begin{aligned}\sum_z \sum_{j,z'} A_{(i,z),(j,z')} x_{j,z'} &= \sum_{j,z'} \sum_z P''_{i,(j,z)} P'_{(j,z),j} x_{j,z'} \\ &= \sum_j P_{i,j} \sum_{z'} x_{j,z'} = \sum_j P_{i,j} \pi_j = \pi,\end{aligned}$$

whenever $\sum_{z'} x_{j,z'} = \pi_j$ i.e. whenever $Cx = \pi$. One is interested only in the marginal output after discarding the $\mathcal{V}'$ register. Irreducibility might depend on the factorization $P''P'$ and may not be known a priori, so let us say that it is not imposed. By an argument similar to the one for invariance, the lift does respect the ergodic flows of P. □

The following table summarizes the constraints satisfied by the examples. We recall that these properties are those of the LMC that we have associated above to each algorithm; while some properties are unambiguous, others can thus depend on the considered LMC as we have explained. The table illustrates that the variety of proposed algorithms also satisfy different constraints around the basic Problem 1. The ergodic flows appear to be matched (`e`) in all cases, although for some approaches, like gather-and-distribute, this may look somewhat artificial. We will rigorously establish how relaxing (`e`) to (`E`) may or may not lead to more efficient algorithms.

random walk	`s`	`i`	`m`	`r`	`e`
stopping rules	`S`	`i`	`M`	`r`	`e`
non-backtracking RW [4]	`S`	`i`	`M`	`r`	`e`
non-backtracking RW [19]	`s`	`I`	`M`	`r`	`e`
simulated annealing	`S`	`I`	`M`	`R`	`e`
gather-and-distribute	`S`	`I`	`M`	`R`	`e`
lifted MC à la [15,18]	`s`	`I`	`m`	`r`	`e`
data-augmented MC [60,62]	`s`	`i`	`M`	`R`	`e`

[3] To gain some intuition, imagine that $\mathcal{V}$ is a set of train stations and $\mathcal{V}'$ enumerates the train platforms. A random walk between train stations, starting from train station j, can then be factorized as follows. At each train station j, first pick a random departing train $z \in \mathcal{V}'$ with probability $P'_{(j,z),j}$; then pick a random stop i along the train line (j,z) with probability $P''_{i,(j,z)}$.

In the following sections, we investigate the mixing properties of LMCs as a function of the associated constraints scenario. This allows us to conclude with implications for the mixing times of the algorithms themselves.

5. Minimal and maximal acceleration of mixing: invariance and initialization, both or none

We start by identifying the scenarios for which the lift cannot provide any advantage in mixing time with respect to a simple Markov chain, and those that allow for the fastest (diameter time) mixing. Remarkably, the only constraints that are relevant to determine these "extreme" behaviors concern the capability of initializing the lift, and the invariance of π.

5.1. Scenarios where lifting does not speed up mixing

We first show that, with (si), the lifted Markov chain cannot go faster than the best non-lifted chain P compatible with the graph, even if we relax the other constraints, e.g. only looking at the marginal mixing time.

Theorem 1. *In all scenarios featuring* (si)*, for any lifted Markov chain* $(\hat{\mathcal{G}}, A)$ *whose marginal* $p(t) = Cx(t)$ *mixes to* π*, there exists a (time-invariant) stochastic matrix* $P^{\mathfrak{q}}$ *such that* $p(t+1) = P^{\mathfrak{q}} p(t)$ *for all* t*.*

Proof. Since we have (s), the lift can start from any distribution x over $\hat{\mathcal{V}}$. Invariance (i) then requires that for any x for which $Cx = \pi$, we have $CAx = \pi$. The main idea of the proof is that, with these constraints, it is necessary that any two $x^{(1)}$, $x^{(2)}$ for which $Cx^{(1)} = Cx^{(2)}$, induce the same flow on $\mathcal{G}$.

Given a lifted Markov chain satisfying (si), consider a map $\mathfrak{q} : \mathcal{V} \mapsto \hat{\mathcal{V}}$ that maps every $j \in \mathcal{V}$ to a single node $k_j \in \hat{\mathcal{V}}$ for which $\mathfrak{c}(k_j) = j$. Let $x = q(p)$ denote the distribution with $x_{\mathfrak{q}(j)} = p_j$ for all $j \in \mathcal{V}$, and $x_i = 0$ for all remaining $i \in \hat{\mathcal{V}}$. Defining

$$P^{\mathfrak{q}}_{i,j} = \sum_{\ell \in \mathfrak{c}^{-1}(i)} A_{\ell, \mathfrak{q}(j)} \,,$$

we will show that for any $x(t)$ with $p(t) = Cx(t)$, the lifted Markov chain satisfies

$$p(t+1) = P^{\mathfrak{q}} p(t) \,, \tag{7}$$

i.e. the LMC behaves like the non-lifted Markov chain $P^{\mathfrak{q}}$. Proving (7) amounts to proving that

$$CAx = P^{\mathfrak{q}} Cx \tag{8}$$

for all $x \in \mathbb{P}_{\hat{N}}$. For any x of the form $x = q(p)$, with $p \in \mathbb{P}_N$, we indeed have (8) by construction. For any other x, defining $x^{(q)} = q(Cx)$, there remains to show that $CAx = CAx^{(q)}$. To do so, select some $a > 0$ such that $a\pi_j > p_j$ for all $j \in \mathcal{V}$ and define $\pi' = \eta\,(a\pi - p) \in \mathbb{P}_N$, with $1/\eta = \sum_{j \in \mathcal{V}} (a\pi_j - p_j) = a - 1$ i.e. $a = 1 + 1/\eta$. Now select any distribution x' over $\hat{\mathcal{V}}$ such that $Cx' = \pi'$ and let $x^{(1)} = (x + x'/\eta)/a$, $x^{(2)} = (x^{(q)} + x'/\eta)/a$, which are properly normalized distributions. We then have by construction $a(x^{(1)} - x^{(2)}) = x - x^{(q)}$, and with $Cx^{(1)} = Cx^{(2)} = \pi$. Invariance (i) requires that $CAx^{(1)} = CAx^{(2)} = \pi$, which readily implies $CA(x - x^{(q)}) = 0$. □

The data-augmented Markov chains in the sense of [60,62] fall into this category. We had already anticipated that these chains do not converge faster than the simple Markov chain would. We here can see that this is implied by the properties (si) and that any attempt at proposing accelerated versions of this algorithm must consider breaking one of these two constraints at least.

5.2. Scenarios where lifting allows for diameter-time mixing

A basic bound on mixing time is that, under locality constraints, the equilibrium distribution cannot be reached in a time that is shorter than the graph diameter $D_{\mathcal{G}}$. This directly implies a similar bound for the formal definition of ϵ mixing time, e.g. $\tau(1/4) \geq D_{\mathcal{G}}/2$ [40]. The diameter bound holds for time-inhomogeneous Markov chains too, and it is easy to see that lifted dynamics must satisfy it as well, even in absence of any constraints besides graph locality. We next characterize a class of scenarios that do allow for mixing in diameter time. Remarkably, this is possible for any graph, as soon as we are allowed to smartly initialize the additional degrees of freedom introduced by the LMC and we do not impose invariance of $Cx(0) = \pi$.

Theorem 2. *For any given $\epsilon > 0$, all scenarios in* (SI) *admit a lifted Markov chain for which $\tau_M(\epsilon) \leq \tau(\epsilon) \leq D_{\mathcal{G}} + 1$, with $D_{\mathcal{G}}$ the graph diameter; the associated lifted graph has of order $D_{\mathcal{G}} N^2$ nodes.*

Proof. The proof uses a node-clock-lift construction, see Section 2.2, where each sequence $\{P^{(i)}(t)\}$ is designed to induce fast convergence from $p = e_i$ towards $p = \pi$ for an initial node i.

More precisely, consider any two distributions p and p' over the nodes $\mathcal{V}$ of a graph $\mathcal{G}$ with diameter $D_{\mathcal{G}}$. There always exists a time-varying Markov chain $\{P(t)\}_{t=1}^{D_{\mathcal{G}}}$ such that $p' = P(D_{\mathcal{G}})P(D_{\mathcal{G}} - 1)\dots P(1)\,p$, where all the $P(t)$ satisfy the locality constraints imposed by $\mathcal{G}$, see e.g. [8,16,23,52]. We call this sequence $\{P(t)\}_{t=1}^{D_{\mathcal{G}}}$ a *stochastic bridge* from p to p'. The existence of such a bridge, as well as its construction, can be derived using a max-flow min-cut argument [8]. A concrete example of such construction can be found in Example 1.1 in relation with the gather-and-distribute technique: during the time steps $T/2$ to T, we use a bridge that maps e_0 onto the target uniform distribution.

Now given any $\mathcal{G}$ and π, for each node $i \in \mathcal{V}$ we start by building such a stochastic bridge from $p(0) = e_i$ towards the target $p(D_{\mathcal{G}}) = \pi$. We then combine these bridges via a node-clock-lift into a single LMC. More precisely, the construction of Section 2.2 is applied as follows. Denoting $P^{(s,i)}$ the matrix $P(s)$ of the stochastic bridge associated to e_i, we define the lifted nodes $\hat{\mathcal{V}} = \{(s, i, v) : s = 0, 1, 2, \dots, D_{\mathcal{G}}; i \in \mathcal{V}; v \in \mathcal{V}\}$ where the original graph corresponds to marginalizing over the indices (s, i). Then we let the (s, i) act as conditional variables to apply the transition matrix $P^{(s+1,i)}$ to the original nodes, for all $s < D_{\mathcal{G}}$; the i variable does not change during this step, while the s variable is incremented by $+1$ like in the construction of Example 4.3 for instance. For the moment we leave open what happens to nodes of the form $(D_{\mathcal{G}}, i, v)$. This LMC, in the (S) scenarios, is associated to the initialization map $F_{(s=0,v_0=i,v=i),i} = 1\ \forall i \in \mathcal{V}$ and all other $F_{i,j} = 0$. By construction, the weight $p_k(0)$ associated to $k \in \mathcal{V}$ then just follows the associated stochastic bridge and gets distributed into a fraction $p_k(0) \cdot \pi$ of the distribution $Cx(D_{\mathcal{G}})$. Thus, the marginal converges exactly ($\epsilon = 0$) to π, within $D_{\mathcal{G}}$ time steps. There remains, at least, to specify the action of the LMC on nodes of the form $(D_{\mathcal{G}}, i, v)$ and to analyze the related constraints.

If we just impose $A\, e_{(D_{\mathcal{G}},i,v)} = e_{(D_{\mathcal{G}},i,v)}$ for all i, v, then all the probability weight stops moving and we have indeed $p(t) = \pi$ for all $t \geq D_{\mathcal{G}}$ in the scenario (SIMRE). This however is the weakest in terms of constraints, and we must show that we can do better. The issue for (m) with the constructed lift is that $x(t) = e_{D_{\mathcal{G}}} \otimes p(0) \otimes \pi$ for all $t \geq D_{\mathcal{G}}$, i.e. it depends on p_0. This can be solved at least formally in a single additional step by applying an *erasure operator** locally at each node $v \in \mathcal{V}$: slightly enlarge $\hat{\mathcal{V}}$ by adding nodes of type (s, i, v) with $s = D_{\mathcal{G}} + 1$, and let $A\, e_{(D_{\mathcal{G}},i,v)} = e_{D_{\mathcal{G}}+1} \otimes \pi \otimes e_v$ with $A\, e_{(D_{\mathcal{G}}+1,i,v)} = e_{(D_{\mathcal{G}}+1,i,v)}$ for all $i, v \in \mathcal{V}$. With this LMC we have $x(t) = e_{D_{\mathcal{G}}+1} \otimes \pi \otimes \pi$ for all $t \geq D_{\mathcal{G}} + 1$ in the scenario (SImRE). Next, it is easy to treat the satisfaction of ergodic flows, thanks to their singular definition in (R) scenarios. Indeed, ergodic flows only depend on the action of A on $\hat{\pi}$ and $\hat{\pi}$ has support on nodes of type (s, i, v) with $s = D_{\mathcal{G}} + 1$ only. With this, it is sufficient for (e) to let A act like $I_{\mathcal{V}} \otimes P$ instead of $I_{\mathcal{V}} \otimes I_{\mathcal{V}}$ on the subspace spanned by $\{e_{(D_{\mathcal{G}}+1,i,v)} : i, v \in \mathcal{V}\}$, proving the theorem for (SIR).

When irreducibility is required, a first point is to drop the nodes which are obviously never populated in the full node-clock-lift, like e.g. the nodes $e_{(0,i,v)}$ with $i \neq v$. In addition, to ensure that the LMC is irreducible (provided P was), the main idea is to link each node of type $e_{(D_{\mathcal{G}}+1,i,v)}$ back to an associated node of type $e_{(0,v,v)}$. However, because invariance of π is not ensured, we must do this with a lifted edge of *small* probability weight $\gamma \ll 1$. Furthermore, again due to (I), this modification possibly leads to a modified steady state with $C\hat{\pi} \neq \pi$ as soon as $\gamma \neq 0$. The technical point then is to slightly adapt some transition probabilities in the LMC such that (i) the new steady-state does satisfy $C\hat{\pi} = \pi$ exactly; (ii) for small γ the corrections are so small that $\hat{\pi}$ remains $\epsilon/2$-close to the former stationary distribution $e_{D_{\mathcal{G}}+1} \otimes \pi \otimes \pi$; (iii) the modifications are made such that ergodic flows of the new construction do not exceed those of P; and (iv) the state of the LMC, after having followed a path essentially equal to the reducible one for the $D_{\mathcal{G}} + 1$ first steps, remains $\epsilon/2$ close to $e_{D_{\mathcal{G}}+1} \otimes \pi \otimes \pi$ (and thus ϵ-close to $\hat{\pi}$) for all $t > D_{\mathcal{G}} + 1$. The technical *irreducible approximation Lemma*, detailed in Appendix, proves that such construction is possible. □

Examples of existing strategies in the (SI) setting are simulated annealing and gather-and-distribute approaches. Indeed, the latter was designed so as to allow for diameter-time mixing. A priori, simulated annealing schemes could allow for the same speedup. However, simulated annealing type algorithms of course correspond to a subclass of these (SI) scenarios, designed with a particular structure, mostly a sequence of *reversible* Markov chains $P^{(t)}$ derived from limited knowledge/analysis, e.g. not knowing the target π. Our analysis does certainly not answer the question whether simulated annealing, with its design constraints, can be efficiently tuned to converge in the order of the diameter; but at least a priori such appreciable speedup is not excluded.

6. Results on conductance bounds

In this section we discuss conductance bounds for LMC scenarios. Section 6.1 discusses the known conductance bounds which, as we show in Section 6.2, lies intermediate to the diameter and random walk bounds derived in the previous section. In Section 6.3 we prove that the bound holds for a number of scenarios beyond what was known, and in Section 6.4 we relativize the relevance of ergodic flows in the light of these results.

6.1. Existing conductance bounds on the mixing time

A key quantity, widely used in obtaining bounds on the mixing time [2,39,46], is the conductance of a stochastic transition matrix P on $\mathcal{G}$. For a subset $\mathcal{X} \subseteq \mathcal{V}$ let $\pi(\mathcal{X}) = \sum_{i\in\mathcal{X}} \pi_i$, where we recall that π is the stationary distribution under P. The *conductance* $\Phi(P)$ *of* P is defined as [40]:

$$\Phi(P) = \min_{\mathcal{X}\subset\mathcal{V}:0<\pi(\mathcal{X})\leq\frac{1}{2}} \frac{\sum_{i\in\mathcal{X}, j\notin\mathcal{X}} P_{j,i}\,\pi_i}{\pi(\mathcal{X})}.$$

This characterizes the minimal steady-state probability flow that is cut when separating the nodes into two disjoint sets. Given only a graph $\mathcal{G}$ and a target stationary distribution π over $\mathcal{V}$, we define *the conductance* Φ *of* $\mathcal{G}$ *towards* π as

$$\Phi = \max_{P:P\sim\mathcal{G}, P\pi=\pi} \Phi(P),$$

where the maximization runs over all stochastic P satisfying the locality constraints of $\mathcal{G}$ (denoted by $P \sim \mathcal{G}$) and whose stationary distribution is π. If π is the uniform distribution, then Φ is upper bounded by the edge expansion of $\mathcal{G}$. If a graph family has bounded degree, then $\Phi \in \Theta(\Phi(P))$ for P the simple random walk, i.e., the simple random walk gives approximately the best conductance. The Barbell graph (see Example 6.3) proves that this does not hold if the degree is unbounded.

The conductance can be used to bound how the minimum time for the convergence of a mixing process is constrained. Loosely speaking, it is known that $\tau(1/4)$ is of the order of $1/\Phi$ or larger, for any Markov chain P (*Conductance bound*, see e.g. [40]); and [15] among others have proved that the same bound holds for lifted Markov chains. However, these bounds are proven only in the scenario (slmre), which is a quite restrictive setting among those we consider; for instance, one might wonder why a lifted walk designed to help converge on $\mathcal{V}$ would really care about (m). Clarifying whether such a bound, or a variation thereof, would hold for other scenarios was one of the main motivations for this paper.

Before going on, let us briefly comment on the conductance Φ when no P is imposed. This expresses how the mixing time is constrained by the *graph topology* and the stationary distribution π alone and it is natural to anticipate that it will play a role in lift scenarios with (E). In this context, the same caveat as after (4) is in order: when relaxing the scenario from (e) to (E), better lifts are admitted but also possibly a more favorable conductance, since dropping (e) implies dropping any reference to a particular P. In particular, a fair treatment in scenario (E) shall compare for each graph $\mathcal{G}$, the fastest possible lifted Markov chain A (in terms of $\tau(1/4)$) with the best possible conductance $\Phi(P)$ over all admissible P, where the optimal P *may differ* from the $\tilde{P}^{(A;\hat{\pi})}$ obtained as an induced Markov chain of the fastest lift.

6.2. Conductance: examples

The following examples illustrate that the conductance bound sits typically in between the two cases considered in Section 5. On the one hand, the very reason why LMCs were introduced in the line of work related to [15], is that the conductance bound is not strict for random walks: on some graphs and π, the mixing time of any compatible P is in fact $\geq 1/\Phi^2$, while an LMC in the sense of [15] can reach $\tau(1/4) = \Theta(1/\Phi)$; see [15] for details, and Example 6.1. On the other hand, the diameter $D_{\mathcal{G}}$ can sometimes be much smaller than $1/\Phi$, see e.g. Example 6.3.

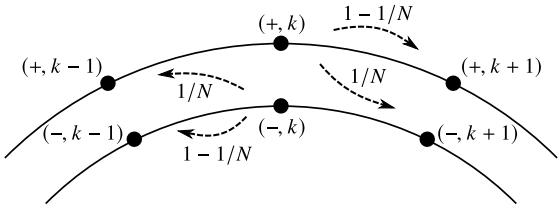

Fig. 3. The transitions for the non-backtracking RW [18] on the cycle with N nodes, see Example 6.1.

Example 6.1 (*Non-Backtracking and LMC on the Cycle*)**.** Consider a stochastic process on the finite cycle graph, i.e. the graph with nodes $\mathcal{V} = \{1, \dots, N\}$, and where node k is connected by an edge to nodes $k \pm_N 1$, where $\pm_N$ denotes addition or subtraction modulo N. As introduced in Section 4, a non-backtracking random walk on the cycle can be described as an LMC with lifted nodes $\hat{\mathcal{V}} = \{(s, i) : s \in \{+, -\} \text{ and } i \in \mathcal{V}\}$, with transition matrix $A_{(i,\pm),(i\pm_N 1,\pm)} = 1 - \alpha$ and $A_{(i,\pm),(i\pm_N 1,\mp)} = \alpha$. The allowed transitions and the relative probabilities are depicted on Fig. 3 for $\alpha = 1/N$. We initialize the walk with the map $F_{(i,+),i} = F_{(i,-),i} = 1/2$ for all $i \in \mathcal{V}$, and zero otherwise.

This construction is introduced in [18], without the initialization map. They show that: if α is of order $1/N$, so keeping rotating around the cycle in the same direction is preferred, but the probability to reverse direction at least once per round around the cycle is significant; then the mixing time $\tau(1/4)$ of x on $\hat{\mathcal{V}}$ is of order N. In contrast, the mixing time $\tau(1/4)$ towards the uniform π of any non-lifted walk on the cycle would be of order N^2 (see e.g. [25], which proves that the optimum is achieved with the simple walk, i.e. probability $1/2$ to take each edge).

For large N, this speedup becomes significant. The LMC of [18] thus achieves accelerated mixing in the (`sImrE`) scenario, or (`sImre`) under the reasonable constraint of imposing ergodic flows with circular symmetry. Yet without initialization map the lift does not satisfy invariance, i.e. starting at x with $Cx = \pi$ would not necessarily imply $CAx = \pi$. Indeed, consider $x_i(0) = 1/N$ for $i = (3, +)$ and $i = (1, -) \in \hat{\mathcal{V}}$; and for each $k \in \mathcal{V} \setminus \{1, 3\}$, any distribution satisfying $x_{(k,+)}(0) + x_{(k,-)}(0) = 1/N$. When applying one step of the LMC to this $x(0)$, no weight can flow to the nodes $\{(2, -), (2, +)\}$, while the latter lose their own weight to neighbors $(1, \pm)$ and $(3, \pm)$. I.e., we have $Cx(0) = \pi$ but $(CAx(0))_2 = 0$. This loss of invariance is consistent with the fact that else, i.e. with (`si`), the setting would have to satisfy Theorem 1 and hence could not feature any speedup over the best random walk.

Alternatively, the non-backtracking construction *with a circular-symmetric initialization map* [4] achieves the same speedup in the (`SiMrE`) or (`SiMre`) scenario. We thus find an interplay between the (`sI`) and the (`Si`) scenarios. □

Example 6.2 (*Blind Stopping Rule on the Cycle*)**.** For mixing on the cycle, a similar speedup as in Example 6.1 can be attained by using a blind stopping rule or Cesaro mixing. This corresponds to a concrete illustration of the construction described in Example 4.1.

The idea is to let the walk turn deterministically around the cycle, say clockwise at each step; and stop the process at a random time t sampled uniformly over $1, 2, \dots, T$. In the language of Section 4, the original Markov chain follows $P = P^+$, where P^+ as in (6) is the clockwise permutation matrix over N nodes. If we stop this Markov chain at a uniformly chosen timestep in $[0, N - 1]$, it is clear that the final state is distributed uniformly over the cycle. In a less

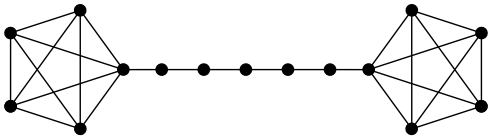

Fig. 4. The Dumbbell graph on $3n = 15$ nodes, see Example 6.3. For this graph the diameter is $\Theta(n)$, both the random walk conductance $\Phi(P)$ and the graph conductance $\Phi_{\mathcal{G}}$ are $\Theta(n^{-2})$ and the random walk mixing time is $\Theta(n^3)$. This graph clearly separates the lower bounds determined by the diameter, the conductance and the random walk mixing time.

stripped-down setting where the exact value of N is unknown at the design stage, one could sample the stopping time uniformly over $[0, T-1]$ for some guess T, and apply amplification as explained in Example 4.1. Interestingly, this maintains accelerated convergence. For instance, after the first T steps when starting say from node 1 with $N < T < 2N$, the distribution is $p_k(T) = 2/T$ for all $k \leq T-N$ and $p_k(T) = 1/T$ for all $k > N$, thus having already distributed a weight $1/T$ uniformly on each node. Thanks to invariance (i) of this algorithmic procedure (see Section 4), this success is booked for all future times and by applying amplification we get a $\tau(\epsilon)$ of order $N \log(1/\epsilon)$. In general, with such stopping rule, p converges asymptotically to the uniform distribution over the cycle at a rate of order T, in a scenario from (SiMrE), while x converges more slowly, with characteristic time T^2, as a scenario from (SimrE). One might thus wonder whether this is a fundamental difference between scenarios, or only due to the particular algorithm. Example 6.1 suggests the latter, since on the cycle the algorithm of [18] converges with a fast rate even as a scenario from (SimrE). In the following we provide the answer to this question for general graphs. □

For the cycle graph, because $D_{\mathcal{G}} = 1/\Phi = N/2$, the diameter and conductance bounds on mixing time are of the same order. The next example shows that for some graphs, $1/\Phi$ can significantly differ from $D_{\mathcal{G}}$.

Example 6.3 (*Dumbbell and Barbell Graph*)**.** The $2n$-node Barbell graph $K_n - K_n$ consists of two completely connected subgraphs on n nodes, connected to each other by a single "central" edge $(n, n+1) \in \mathcal{E}$. For visualization, it is a particular case of Fig. 4 where the central path would reduce to a single edge. This graph is a notable example in mixing time studies because of the clear bottleneck behavior of this central edge [3,13]. As a consequence, we can show that the inverse conductance $1/\Phi$ associated to the uniform distribution π on the Barbell graph will be significantly larger than the diameter $D_{\mathcal{G}} = 3$. To see this, first consider the random walk P on this graph (i.e. when sitting at a node, we have equal probabilities to take any of its edges). For technical reasons, we add self-loops to all nodes except for the central nodes — this ensures the graph is regular. Then the stationary distribution is uniform, and the central edge $(n, n+1)$ has a transition probability $1/n$. Now consider the cut $\mathcal{X} = \mathcal{X}_n$ where $\mathcal{X}_n$ contains all the nodes on one side of the central edge $(n, n+1)$; the latter is thus the only one to be cut. Then this results in a random walk conductance $\Phi(P) \leq 1/(n^2 - n + 1)$, so that $1/\Phi(P)$ is significantly larger than $D_{\mathcal{G}}$. Towards estimating the optimal conductance Φ over all Markov chains P, we note that we can actually improve the conductance by choosing smarter weights, as mentioned earlier. To this end, add self-loops of strength $1/2$ to all nodes except for the central 2 nodes, and increase the transition probability of the central edge to $1/2$. Then the stationary distribution remains uniform (the transition matrix is symmetric), yet

the conductance is improved to $1/(2n)$. We can easily show that this is nearly optimal: for the aforementioned central cut $\mathcal{X} = \mathcal{X}_n$,

$$\Phi = \max_P \Phi(P) = \max_P \min_{\mathcal{X}\subset\mathcal{V};0<\pi(\mathcal{X})\leq\frac{1}{2}} \frac{\sum_{i\in\mathcal{X}, j\notin\mathcal{X}} P_{j,i}\,\pi_i}{\pi(\mathcal{X})}$$
$$\leq \max_P \frac{\sum_{i\in\mathcal{X}_n, j\notin\mathcal{X}_n} P_{j,i}\,\pi_i}{\pi(\mathcal{X}_n)} = \max_P \frac{P_{n,n+1}\,(1/2n)}{1/2} \leq \frac{1}{n}\,.$$

From first to second line, we have replaced the minimum by the particular cut $\mathcal{X} = \mathcal{X}_n$, and to conclude we have used that the central edge can at most have $P_{n,n+1} = 1$. This bound clearly shows that also $1/\Phi$ can be unboundedly larger than $D_{\mathcal{G}}$.

The Dumbbell graph, see Fig. 4, is a variation of the above, where two completely connected subgraphs K_n on n nodes are connected by a path of length n. Its diameter is $n + 2$. The random walk P on the Dumbbell graph has a steady state π with weight of order $1/n$ on each node of the completely connected subgraphs K_n and a weight of order $1/n^2$ on each node of the central path. This random walk is known to converge towards π in order n^3 [40]. The associated conductance $\Phi(P)$, obtained by cutting in the middle of the central path, is of order $1/n^2$. Thus in this example, $1/\Phi$ associated to the stationary distribution of the random walk, is both strictly larger than the diameter and strictly smaller than the mixing time of the random walk. □

The previous examples put mixing times of order $1/\Phi$ strictly between the convergence speed of non-lifted walks, as in Theorem 1, and the diameter bound of Theorem 2. The aim of this section is precisely to identify scenarios where, while a lifted Markov chain could outperform the non-lifted chains, it can never significantly beat the conductance bound. The present section focuses on establishing lower bounds on the mixing time. The next section will comment on the speedup, from Theorem 1 to conductance bound, indeed being attainable for those scenarios.

6.3. Identifying scenarios which provide an advantage within the conductance bound

We start with some preliminary results. We note that the gist of the following lemma is well-known, and appears in for instance [40,59].

Lemma 1. *Consider a stochastic matrix P, not necessarily irreducible, on a node set $\mathcal{V}$ and one of its stationary distributions π. Take any $\mathcal{X} \subseteq \mathcal{V}$ such that $\pi(\mathcal{X}) \neq 0$ and define the distribution $\tilde{\pi}^{(\mathcal{X})}$ by*

$$\tilde{\pi}_i^{(\mathcal{X})} = \begin{cases} \eta\,\pi_i \text{ for } i \in \mathcal{X} \\ 0 \text{ for } i \notin \mathcal{X} \end{cases} \qquad \text{with } \tfrac{1}{\eta} = \sum_{i\in\mathcal{X}} \pi_i = \pi(\mathcal{X})\,.$$

Then for all $t \geq 1$ we have

$$\sum_{j\notin\mathcal{X}} (P^t \tilde{\pi}^{(\mathcal{X})})_j \;\leq\; t\;\Phi_{\mathcal{X},\pi}(P)\,,$$

where $\Phi_{\mathcal{X},\pi}(P) = (\sum_{i\in\mathcal{X}, j\notin\mathcal{X}} P_{j,i}\,\pi_i)/\pi(\mathcal{X})$, can be viewed as a conductance associated to π and the particular subset $\mathcal{X}$.

Proof. We will first prove and later use the following facts:

$$\sum_{j\notin\mathcal{X}}(P\tilde{\pi}^{(\mathcal{X})})_j = \left\|P\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\right\|_{TV} \tag{9}$$

$$\sum_{j\notin\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_j \leq \left\|P^t\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\right\|_{TV}\ ,\ \forall t\geq 0.$$

To obtain the equality in (9), we rewrite the total variation distance:

$$\left\|P\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\right\|_{TV} = \sum_{u\in\mathcal{V}:(P\tilde{\pi}^{(\mathcal{X})})_u\geq\tilde{\pi}_u^{(\mathcal{X})}} (P\tilde{\pi}^{(\mathcal{X})})_u - \tilde{\pi}_u^{(\mathcal{X})}\ .$$

We then observe that $(P\tilde{\pi}^{(\mathcal{X})})_u \geq \tilde{\pi}_u^{(\mathcal{X})} = 0$ trivially for all $u\notin\mathcal{X}$, while the following computations yield the opposite conclusion for all $u\in\mathcal{X}$. Indeed, by the definition of $\tilde{\pi}^{(\mathcal{X})}$, we have

$$\begin{aligned}(P\tilde{\pi}^{(\mathcal{X})})_u &= \sum_{j\in\mathcal{V}} P_{u,j}\tilde{\pi}_j^{(\mathcal{X})} = \sum_{j\in\mathcal{X}} P_{u,j}\frac{\pi_j}{\pi(\mathcal{X})}\\ &\leq \frac{\sum_{j\in\mathcal{V}} P_{u,j}\pi_j}{\pi(\mathcal{X})} = \frac{\pi_u}{\pi(\mathcal{X})} = \tilde{\pi}_u^{(\mathcal{X})}.\end{aligned}$$

Thus the rewritten total variation distance reduces to $\sum_{j\notin\mathcal{X}}\big((P\tilde{\pi}^{(\mathcal{X})})_j - 0\big)$.

To obtain the inequality in (9), we expand the total variation distance:

$$\begin{aligned}\|P^t\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\|_{TV} &= \frac{1}{2}\sum_{u\in\mathcal{V}}|(P^t\tilde{\pi}^{(\mathcal{X})})_u - \tilde{\pi}_u^{(\mathcal{X})}|\\ &= \frac{1}{2}\sum_{u\notin\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_u + \frac{1}{2}\sum_{u\in\mathcal{X}}|(P^t\tilde{\pi}^{(\mathcal{X})})_u - \tilde{\pi}_u^{(\mathcal{X})}|\\ &\geq \frac{1}{2}\sum_{u\notin\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_u + \frac{1}{2}\left|\sum_{u\in\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_u - \tilde{\pi}_u^{(\mathcal{X})}\right|\\ &= \frac{1}{2}\sum_{u\notin\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_u + \frac{1}{2}\left|\left(1 - \sum_{u\notin\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_u\right) - 1\right|\\ &= \sum_{u\notin\mathcal{X}}(P^t\tilde{\pi}^{(\mathcal{X})})_u\ ,\end{aligned}$$

thus proving (9).

Next we obtain (see justifications below):

$$\begin{aligned}&\|P^t\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\|_{TV}\\ &\leq \|P^t\tilde{\pi}^{(\mathcal{X})} - P^{t-1}\tilde{\pi}^{(\mathcal{X})}\|_{TV} + \|P^{t-1}\tilde{\pi}^{(\mathcal{X})} - P^{t-2}\tilde{\pi}^{(\mathcal{X})}\|_{TV} + \cdots + \|P\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\|_{TV}\\ &\leq t\ \|P\tilde{\pi}^{(\mathcal{X})} - \tilde{\pi}^{(\mathcal{X})}\|_{TV} = t\ \Phi_{\mathcal{X},\pi}(P).\end{aligned}$$

From the first to second line we have used the triangle inequality on the ℓ_1 norm; from second to third line, we have used repeatedly that any stochastic matrix P contracts the ℓ_1 norm [40], i.e. for arbitrary distributions $p^{(1)}$ and $p^{(2)}$ we have $\|Pp^{(1)} - Pp^{(2)}\|_{TV} \leq \|p^{(1)} - p^{(2)}\|_{TV}$; for the last equality, we have used the equality in (9) and the explicit computation $\sum_{j\notin\mathcal{X}}(P\tilde{\pi}^{(\mathcal{X})})_j = \sum_{u\in X, j\notin X} P_{j,u}\tilde{\pi}_u^{(\mathcal{X})} = \sum_{u\in X, j\notin X}\frac{P_{j,u}\pi_u}{\pi(\mathcal{X})}$ where the last expression is the definition of $\Phi_{\mathcal{X},\pi}(P)$. The stated result then follows by using inequality (9) in front of the first line. □

Lemma 2. *Consider a lifted Markov chain A on $\hat{\mathcal{V}}$ and one of its steady states $\hat{\pi}$. Then the mixing time $\tau_M(1/4)$ for scenarios with* (sM) *satisfies $\tau_M(1/4) \geq 1/(4\,\Phi(\tilde{P}^{(A;\hat{\pi})}))$, where $\tilde{P}^{(A;\hat{\pi})}$ is the induced chain on $\mathcal{V}$ associated to A and $\hat{\pi}$.*

Proof. Take a subset $\mathcal{X} \subseteq \mathcal{V}$ such that $\pi(\mathcal{X}) = \hat{\pi}(\mathfrak{c}^{-1}(\mathcal{X})) \leq 1/2$ and denote $\hat{\mathcal{X}} = \mathfrak{c}^{-1}(\mathcal{X})$. Define $\hat{\pi}^{(\hat{\mathcal{X}})}$ similarly to $\tilde{\pi}^{(\hat{\mathcal{X}})}$ in Lemma 1, except now we are constructing it on the lifted space $x, \hat{\pi}, \hat{\mathcal{X}} = \mathfrak{c}^{-1}(X)$ instead of on $y, \pi, \mathcal{X}$. We then get:

$$\begin{aligned}
\|CA^t\hat{\pi}^{(\hat{\mathcal{X}})} - C\hat{\pi}\|_{TV} &= \frac{1}{2}\sum_{i\in\mathcal{V}}\left|\sum_{j\in\mathfrak{c}^{-1}(i)}(A^t\hat{\pi}^{(\hat{\mathcal{X}})})_j - \hat{\pi}_j\right| \\
&\geq \frac{1}{2}\left|\sum_{i\in\mathcal{X}}\sum_{j\in\mathfrak{c}^{-1}(i)}(A^t\hat{\pi}^{(\hat{\mathcal{X}})})_j - \hat{\pi}_j\right| + \frac{1}{2}\left|\sum_{i\notin\mathcal{X}}\sum_{j\in\mathfrak{c}^{-1}(i)}(A^t\hat{\pi}^{(\hat{\mathcal{X}})})_j - \hat{\pi}_j\right| \\
&=: \|C'A^t\hat{\pi}^{(\hat{\mathcal{X}})} - C'\hat{\pi}\|_{TV}
\end{aligned}$$

where we define C' by $C'_{1,v} = 1$ if $\mathfrak{c}(v) \in \mathcal{X}$, $C'_{0,v} = 1$ if $\mathfrak{c}(v) \notin \mathcal{X}$, all other $C'_{i,j} = 0$. That is, C' projects the distribution x on $\hat{\mathcal{V}}$ onto a distribution among the two options, $\hat{\mathcal{X}}$ and its complementary. With the reverse triangle inequality and using the same notation $\Phi_{\mathcal{X}}$ as in Lemma 1, we further develop:

$$\begin{aligned}
\|C'A^t\hat{\pi}^{(\hat{\mathcal{X}})} - C'\hat{\pi}\|_{TV} &\geq \|C'\hat{\pi}^{(\hat{\mathcal{X}})} - C'\hat{\pi}\|_{TV} - \|C'A^t\hat{\pi}^{(\hat{\mathcal{X}})} - C'\hat{\pi}^{(\hat{\mathcal{X}})}\|_{TV} \qquad (10)\\
&= (1 - \pi(\mathcal{X})) - \sum_{j\notin\hat{\mathcal{X}}}(A^t\hat{\pi}^{(\hat{\mathcal{X}})})_j \\
&\geq \frac{1}{2} - t\,\Phi_{\hat{\mathcal{X}},\hat{\pi}}(A) = \frac{1}{2} - t\,\Phi_{\mathcal{X}}(\tilde{P}^{(A;\hat{\pi})}).
\end{aligned}$$

The first equality follows by noting that $C'\hat{\pi}^{(\hat{\mathcal{X}})}$ and $C'\hat{\pi}$ have weight 1 resp. $\pi(\mathcal{X})$ on $\mathcal{X}$, whereas $C'A^t\hat{\pi}^{(\hat{\mathcal{X}})}$ has weight $1 - \sum_{j\notin\hat{\mathcal{X}}}(A^t\hat{\pi}^{(\hat{\mathcal{X}})})_j$ on $\mathcal{X}$. To get the last inequality, for the first term we have used our assumption that $\pi(\mathcal{X}) \leq 1/2$. For the second term we have used Lemma 1, for which the only condition was that $\hat{\pi}(\hat{\mathcal{X}}) > 0$; the latter holds since $\hat{\pi}(\hat{\mathcal{X}}) = \pi(\mathcal{X})$ and the paper throughout assumes $\pi(i) > 0$ for all i. The final equality is obtained by recalling that the induced chain is defined precisely such that $\tilde{P}^{(A;\hat{\pi})}_{i,j}\pi_j = \sum_{u\in\mathfrak{c}^{-1}(i),v\in\mathfrak{c}^{-1}j} A_{u,v}\hat{\pi}_v$; this readily implies that for the particular preimage subsets $\hat{\mathcal{X}} = \mathfrak{c}^{-1}(\mathcal{X}) \subset \hat{\mathcal{V}}$ we do have $\Phi_{\hat{\mathcal{X}},\hat{\pi}}(A) = \Phi_{\mathcal{X}}(\tilde{P}^{(A;\hat{\pi})})$.

If the mixing time $\tau(1/4)$ is equal to T, then in particular the initial condition $\hat{\pi}^{(\hat{\mathcal{X}})}$ must converge close enough to π within T steps, i.e. we need

$$\|CA^T\hat{\pi}^{(\hat{\mathcal{X}})} - C\hat{\pi}\|_{TV} \leq 1/4\ .$$

By (10) this requires $1/2 - T\,\Phi_{\mathcal{X}}(\tilde{P}^{(A;\hat{\pi})}) \leq 1/4$ i.e. $T \geq 1/(4\Phi_{\mathcal{X}}(\tilde{P}^{(A;\hat{\pi})}))$. Since this is true for all $\mathcal{X} \subset \mathcal{V}$ with $\pi(\mathcal{X}) \leq 1/2$, it is in particular true for the $\mathcal{X}$ for which the minimum value of $\Phi_{\mathcal{X}}(\tilde{P}^{(A;\hat{\pi})})$, i.e. the conductance $\Phi(\tilde{P}^{(A;\hat{\pi})})$, is attained. □

Note that the above results hold irrespective of having invariance property (i) or (I). The following result extends the bound from [15] in the (sImre) setting to all the (s) and (se) settings, i.e., allowing for reducible lifts and for situations where we only care about the mixing time τ_M of the *marginal* distribution.

Theorem 3. *The scenarios including* (s) *satisfy a conductance bound of the form* $\tau(1/4) \geq \tau_M(1/4) \geq 1/(4\Phi)$ *, or* $\tau(1/4) \geq \tau_M(1/4) \geq 1/(4\Phi(P))$ *in scenarios with* (e) *and ergodic flows specified by* P.

Proof. Lemma 2 applies directly to (sMe), where (e) imposes that $\tilde{P}^{(A;\hat{\pi})} = P$ for all $(A;\hat{\pi})$ and thus $\Phi(\tilde{P}^{(A;\hat{\pi})}) = \Phi(P)$. Regarding (sME), assume that the claim would not hold i.e. a particular $A, \hat{\pi}$ would allow $\tau_M(1/4) < 1/(4\Phi)$ with Φ the largest $\Phi(P)$ among all admissible P. This particular lift satisfies, as any other, that $\tau_M(1/4) \geq 1/(4\Phi(\tilde{P}^{(A;\hat{\pi})}))$, where the induced chain $\tilde{P}^{(A;\hat{\pi})}$ is by construction an admissible P, yielding an admissible conductance $\Phi(P)$ on $\mathcal{G}$. This directly gives a contradiction. The corresponding scenarios with (m) instead of (M) follow by (4). □

There remains to treat the scenarios with (Si). We next show that, when looking at the convergence of the marginals in absence of requirements about irreducibility and ergodic flows, as in (MRE), one can trade the constraint (i) for the constraint (s) by paying a small price in the mixing time.

Lemma 3. *Let* A_1 *be a lift in* (SiMRE) *with mixing time* $\tau_M(1/4) = \tau$. *Then there exists a lift* A_2 *in* (sIMRE) *that has* $\tau_M(1/4) = 2\tau$.

Proof. Consider the given lift A_1 in (SiMRE): we can construct the periodic node-clock lift (see Section 2.2) which, modulo proper initialization, first follows A_1's evolution from $t = 0$ to $t = T = \tau_M(1/4)$, and then periodically repeats this evolution. The proper initialization F associates the weight p_i to the lifted node $(t, v_0, v) = (0, i, i)$. To see how this allows to trade (i) for (s), we examine the evolution of an initial state (t_i, v_0, v) with this periodic node-clock lift A_2 (equivalent to considering a Kronecker delta distribution as initial distribution $x(0)$). After $T - t_i$ steps with A_2, each such state is mapped to the set of nodes $\mathcal{F}_0 := \{(0, v, v)\}$. From there, it follows the periodic evolution generated by A_1. This implies that after $T = \tau_M(1/4)$ more steps with A_2, i.e. at $t = t_* = 2T - t_i$, for any v_0, v defining the initial marginal $x(0)$, the distribution $x(t_*)$ satisfies $\|Cx(t_*) - \pi\|_{TV} \leq 1/4$ and has support on the image of the initialization map F. By construction we can thus write $x(t_*) = Fp(t_*)$ for some distribution $p(t_*)$ over $\mathcal{V}$. For the next T steps, we thus have $x(t_* + t) = (A_1)^t Fp(t_*)$. Invariance (I) of π under A_1 means $C(A_1)^t F\pi = CF\pi = \pi$. Therefore we can write:

$$\begin{aligned}\|Cx(t_* + t) - \pi\|_{TV} &= \|C(A_1)^t Fp(t_*) - C(A_1)^t F\pi\|_{TV} \\ &\leq \|(A_1)^t Fp(t_*) - (A_1)^t F\pi\|_{TV} \\ &\leq \|Fp(t_*) - F\pi\|_{TV} = \|p(t_*) - \pi\|_{TV}.\end{aligned}$$

The second inequality holds because the stochastic matrix A_1 contracts the ℓ_1 norm. The last equality holds since the F associated to our bridge is just a relabeling of nodes, from $\mathcal{V}$ to a subset of $\hat{\mathcal{V}}$. Since we had $\|Cx(t_*) - \pi\|_{TV} \leq 1/4$ already, we have just shown that $\|Cx(t) - \pi\|_{TV} \leq 1/4$ for all $t \geq t_*$, when we start A_2 with all the weight of $x(0)$ concentrated on a single node $\in \hat{\mathcal{V}}$.

For an arbitrary $x(0)$, the state after $2T$ steps is a convex combination of cases with $x_i(0) = 1$, so the property $\|Cx(t) - \pi\|_{TV} \leq 1/4$ for all $t \geq 2T \geq t_*$ is maintained. The periodic node-clock-lift A_2 is thus a chain in (sIMRE), whose $\tau_M(1/4)$ mixing time is at most twice the one of the chain A_1 in (SiMRE). □

We note that we already encountered this interplay between (sI) and (Si) scenarios in the two formulations of the LMC on the cycle in Example 6.1.

Remark. The lift constructed in the proof (potentially) loses invariance. This is the case because A_2 is built to follow A_1, and thus leave π invariant, only if initialized in the set $\mathcal{F}_0 := \{(0, v, v)\}$; as soon as in trading (S) towards (s) we relax this initialization and allow starting at nodes (t_i, v_0, v) with $t_i \neq 0$, we have no guarantee anymore about what happens when $Cx(0) = \pi$. We also note that through the periodic node-clock-lift construction, we could lose any matching ergodic flows that were present in A_1, so scenarios with (Sie) are not covered by this result.

Theorem 4. *Settings with* (SiE) *satisfy* $\tau(1/4) \geq \tau_M(1/4) \geq 1/(8\Phi)$.

Proof. Since we have shown in Theorem 3 that $\tau_M(1/4) \geq 1/(4\Phi)$ for (sIMRE), by Lemma 3 we have $\tau_M(1/4) \geq 1/(8\Phi)$ for (SiMRE). The same bound then holds for all scenarios with (SiE), since they are all more constrained than (SiMRE) and with the same conductance. □

We cannot strengthen (E) to (e) directly in the above proof, since this would affect both sides of the inequality: the allowed lifts, and the allowed P for computing the conductance.

6.4. On an equivocal role of ergodic flows

We conclude our characterization of lower bounds on mixing time by treating the scenarios of type (Sie) on their own. We argue that this setting appears like one of the most natural scenarios: a tailored initialization of auxiliary states is allowed, and invariance of the target under such initialization is required. Caring about irreducibility or about the full distribution including auxiliary states will play no key role. There thus remains to investigate how ergodic flows fare under this setting.

In accordance with Theorem 3, and keeping in mind the (i) $\leftrightarrow$ (s) duality, one could expect a $\Omega(1/\Phi(P))$ lower bound similar to the (se) setting. In the following example however we show that this does not hold.

Example 6.4 (*Dumbbell Graph*)**.** Starting with a random walk on the dumbbell graph $K_n - K_n$ on $2n$ nodes, we define P by adding a self-loop of weight $1/n$ to all nodes except the two middle nodes — this ensures that the stationary distribution π of P is the uniform one. The conductance $\Phi(P) = 1/n^2$ is obtained by considering a cut over the middle edge. We will show that a lifted walk satisfying (Sie), with ergodic flows as imposed by the random walk, can significantly beat the associated conductance bound $1/\Phi(P)$, and in fact reach $1/\Phi$ where Φ corresponds to another Markov chain that converges to the uniform distribution π on the Dumbbell graph; in particular, this Markov chain would have a much higher weight on the middle edge. This shows that a statement similar to Theorem 3 in the (se) setting does not hold in this (ie) setting.

To construct the LMC, we start from a sequence of transition matrices $\{P_t\}_{t=1}^T$ that induces convergence in finite time and leaves the target distribution invariant, i.e. such that:

$$P_T P_{T-1} \dots P_2 P_1 e_v = \pi \quad \forall v \in \mathcal{V},$$
$$P_t P_{t-1} \dots P_2 P_1 \pi = \pi \quad \forall 1 \leq t \leq T.$$

An easy solution goes as follows:

1. Mix the probability mass on the left cluster and right cluster individually, corresponding to the transition matrix
$$P_1 = \frac{1}{n}\begin{bmatrix} \vec{1}\vec{1}^T & \\ & \vec{1}\vec{1}^T \end{bmatrix},$$
with $\vec{1}$ the all-ones vector;
2. Equilibrate the two middle nodes, corresponding to the transition matrix P_2 defined by
$$P_2 = \begin{bmatrix} I_{n-1} & & & \\ & 0 & 1/2 & \\ & 1/2 & 0 & \\ & & & I_{n-1} \end{bmatrix};$$
3. Perform a clockwise permutation on the left and on the right cluster individually, i.e.,
$$P_3 = \begin{bmatrix} P^+ & \\ & P^+ \end{bmatrix},$$
with P^+ the clockwise permutation matrix over n nodes;

4.-T. Repeat steps 2. and 3. another $n-1$ times; thus $T = 2n+1$.

It is straightforward to check that this sequence of transition matrices satisfies invariance and that indeed $(P_3P_2)^n P_1 e_v = \pi$ for all $v \in \mathcal{V}$.

Building a clock-lift on basis of this sequence $\{P_t\}_{t=1}^T$, the ergodic flows reduce to self-loops and thus (e) is trivially satisfied, such that we obtain an LMC in the (SiRe) setting that exactly converges towards π in $2n+1$ steps. We here see how the setting is "cheated": while possibly large flows occur during the transient phase, in stationary regime they vanish such that any imposed ergodic flows are not exceeded. The same idea can be pursued for the other settings.

Indeed, consider the LMC that modifies the previous construction by replacing $e_T e_T^\dagger \otimes I$ in the transition matrix A by
$$(1-\gamma) e_T e_T^\dagger \otimes I + \gamma e_0 e_T^\dagger \otimes I \ ,$$
for some small probability $\gamma > 0$ to restart the sequence. The resulting LMC is irreducible, and it still satisfies invariance of π when initialized with S as previously. In addition, we can prove that for sufficiently small γ it will not violate the ergodic flows, and the $(1/4)$-mixing time will remain unchanged. For any choice of γ, the stationary distribution over lifted nodes takes the form $\hat{\pi} = \tilde{c} \otimes \pi$, where $\tilde{c}$ is the distribution over $0, 1, \ldots, T$ with weight $\frac{1}{1+T\gamma}$ on T and $\frac{\gamma}{1+T\gamma}$ on $0, 1, \ldots, T-1$. When $\gamma \ll 1/T$, the stationary distribution is essentially localized on the Tth level, with a total probability of order $\gamma T \ll 1$ on the other levels. Since the ergodic flows on the Tth level are zero (we apply the identity on the node space), we get that the total amount of ergodic flow through any edge is at most γT, which will obey the ergodicity condition provided that we choose γ sufficiently small. To analyze the mixing time, notice that we initialize the chain on the 0-th level. By construction of our sequence $\{P_t\}_{t=1}^T$, the chain will be exactly distributed as $e_T \otimes \pi$ after T steps, after which the $\mathcal{V}$ degree of freedom will remain in the distribution π (thanks to the invariance property of the sequence $\{P_t\}_{t=1}^T$).

What remains to be proven is that the clock degrees of freedom will also mix in time $O(T)$, for any $\gamma \ll T$. The clock degree of freedom evolves on a cycle with $T+1$ nodes, with transitions $s \mapsto s+1$ with probability 1 for all $s = 0, 1, \ldots, T$ and a probability $(1-\gamma)$ to

stay at node T and probability γ to move from node T to node 0. The corresponding mixing time easily follows from a coupling argument [40].[4] To summarize, this construction yields an LMC with mixing time $T = 2n+1$ towards the uniform distribution, in the setting (Simre) and thus in all the settings (Sie), while satisfying the ergodic flows of a random walk for which $1/\Phi(P) = n^2$. □

The above example shows that when ergodic flows are imposed, *the lift in a scenario* (Sie) *can nevertheless significantly beat the conductance bound associated to these ergodic flows.* The actual bound becomes the conductance bound associated to Φ, the highest conductance over all possible P compatible with $\mathcal{G}, \pi$. In this sense, in (Sie) scenarios the lifts can serve as a way to circumvent the limitations imposed by (overly-constraining) ergodic flows.

Theorem 5. *Any lifted Markov chain satisfying* (Sie)*, has its mixing time bounded by* $\tau(1/4) \geq \tau_M(1/4) \geq 1/(8\Phi)$*, the conductance bound associated* to the graph $\mathcal{G}$ *and* π.

Proof. This follows from Theorem 4. Indeed, assume the opposite i.e. that we can make $\tau_M(1/4)$ arbitrarily smaller than $1/(8\Phi)$ by choosing an appropriate lift. This particular lift A of course remains a valid option when relaxing the ergodic flow constraint, i.e. for the corresponding setting in (SiE). Therefore we would also have $\tau(1/4) \leq 1/(8\Phi)$ for (SiE) which is a contradiction with Theorem 4. □

The most important message for (Sie) is not the bound of Theorem 5, but rather the fact that *it cannot be tightened by replacing* Φ *with* $\Phi(P)$. The latter is indeed the point of the academic Example 6.4, and is further elaborated in the next section. By comparing this result with Theorem 3, it becomes apparent that trading (s) for (i) does have an effect when (e) is also part of the scenario.

Imposing ergodic flows (e) is rather standard in the literature on lifted Markov chains, like in the (sImre) scenario of [15]. In this scenario, like for the original random walk P, ergodic flow constraints may directly impose slow mixing. However, when initialization is allowed for the lifted Markov chain, which after all appears natural in algorithmic settings, it appears that matching ergodic flows do not impose a hard limit on the mixing time. From the construction of Example 6.4, this happens by inducing fast mixing during a transient phase, before reaching the subspace where ergodic flows do constrain the convergence. Such transient may arguably not be the behavior one wishes to encourage in practice. Note that when excluding the overly bad (si) scenarios, it is not exactly clear how one could specify that "transients are not what we want to look at" in a stabilizing lifted Markov chain (i.e. when relaxing (S) to maintain the constraint (i)). Also, the academic example above might not be the only way to break the conductance bound associated to a particular ergodic flow. From this insight, we would not claim to have found a particularly relevant speed-up strategy. However, this observation does point to the fact that issues with ergodic flows, and more generally transient behaviors, should be treated with caution in the (Si) scenarios.

From the examples of Section 4, the blind stopping rule and non-backtracking algorithms fall under these (Si) scenarios and thus have a mixing time bounded by $1/\Phi$. By keeping the symmetry of the graph, the non-backtracking algorithms match the ergodic flows of the simple

[4] Consider a pair of independent runs of this Markov chain, starting from nodes $x \neq y$. The probability that each chain is at node T after T steps is at least $(1-\gamma)^T$, and hence the probability that they collide after T steps is at least $(1-\gamma)^{2T}$. This will be at least 3/4 if for instance $\gamma \leq 1/(10T)$, and by a standard argument [40, Theorem 5.2] this implies that $\tau(1/4) \leq T$.

random walk $\Phi(P)$. Note that for bounded degree graphs and taking for P the random walk, $\Phi(P)$ is of the same order as Φ so the difference just mentioned has no effect; while for graphs with large degrees, the effect of simple non-backtracking as described in Example 4.2 almost vanishes.

7. Tightness and complementary observations

In this section we present some further results and observations, related to the lower bounds on mixing time. The mixing time bounds for the scenarios of Section 5 are obviously tight, i.e. they can indeed be achieved by an appropriate LMC on any graph. In Section 7.1 we establish tightness as well for most of the scenarios involving a conductance bound; this builds rather directly on the result of [15]. In Section 7.2, we observe how some graph properties can also be directly derived from our bounds on the mixing performance. In Section 7.3 we mention some extensions of our results.

7.1. On the tightness of the conductance bounds

Theorems 3–5 provide lower bounds for the scenarios of type (sI) and (Si). In order to complete the picture and claim a truly relevant classification, we have to establish how tight these bounds can be. This question can be settled to a great extent rather quickly. For random walks, and thus the (si) scenarios, we have already mentioned that the conductance bound is not tight: the mixing time can be quadratically larger. This gap was the main motivation for introducing LMCs in [15], which allowed them to prove the following:

> *Given an irreducible P (and thus $\pi_i > 0$ for all i), one can construct an LMC A_1 in* (sImre) *with a stopping rule whose expected running time reaches the conductance bound up to a factor* $\log \frac{1}{\min_i \pi_i}$, *i.e.,* $\tau(1/4) \in O\Big(\frac{1}{\Phi(P)} \log \frac{1}{\min_i \pi_i}\Big)$.

By our discussion on general stopping rules at the end of Example 4.1, we can replace the stopping rule by a Cesaro average. Combined with the LMC construction in the same example for simulating Cesaro averaging, this allows us to prove tightness of certain conductance bounds. We are in fact convinced that all the conductance bounds are tight (up to a log-factor), yet for brevity and clarity we limit the proofs to the most direct cases. In particular, the results are restricted to irreducible P and the bounds get worse as some π_i get close to zero. We do not believe that this feature is essential, but developing tighter bounds as some π_i go to zero is beyond the scope of this work.

We start with the (sI) case.

Lemma 4. *Given a graph $\mathcal{G}$, an irreducible Markov chain P and a stationary distribution π,*

- *for scenarios in* (sIe) *there exists an LMC with a mixing time*

$$\tau(1/4) \in O\Big(\frac{1}{\Phi(P)} \log \frac{1}{\min_i \pi_i}\Big);$$

- *for scenarios in* (sIE) *there exists an LMC with a mixing time*

$$\tau(1/4) \in O\Big(\frac{1}{\Phi} \log \frac{1}{\min_i \pi_i}\Big).$$

Proof. We prove the first bullet by constructing an LMC in (sImre), which is the most constrained setting. As mentioned before the Lemma, we start with transition matrix A_1 over node set $\hat{\mathcal{V}}$ corresponding to the LMC constructed in [15]. This LMC is in (sImre) and, with a particular stopping rule, has a mixing time equal to the promised one. We replace this particular stopping rule by the amplified Cesaro average stopping rule, for $T \in O\left(\frac{1}{\Phi(P)} \log \frac{1}{\min_i \pi_i}\right)$. By our remark in Example 4.1, for $\gamma = 1$, the mixing time associated to A_1 with Cesaro average stopping rule equals the mixing time associated to A_1 with the original stopping rule. Then we can cast this Cesaro average as a new LMC, as we do in Section 4. 4.1 but now using A_1 instead of P in the construction. This yields a larger LMC A' over $\hat{V}' = \{\texttt{run}, \texttt{stop}\} \times \{0, \dots, T\} \times \hat{\mathcal{V}}$. If now the LMC associated to A' starts from any node in $\{\texttt{run}\} \times \{0\} \times \hat{\mathcal{V}}$, then the marginal evolution correctly follows the evolution of the amplified Cesaro average of the LMC A_1, hence inheriting the promised mixing time. To see that the same holds when starting from an arbitrary node of the LMC associated to A', note that after at most $T \in O\left(\frac{1}{\Phi(P)} \log \frac{1}{\min_i \pi_i}\right)$ steps the LMC will necessarily visit a node in $\{\texttt{run}\} \times \{0\} \times \hat{\mathcal{V}}$, after which it will mix as promised, thus adding at most a factor 2 to the mixing time. This proves the claim for the scenario (sIMe).

To prove the claim for (sIme), we modify the LMC similarly to Example 6.4 by adding a holding probability $(1 - \gamma)$ at the Tth clock level. As argued in that example, if we set $\gamma = 1/(10T)$, then the mixing time over the node and clock degrees-of-freedom remains $O(T)$. We must discuss two adaptations compared to this example. A first difference is that here we do not necessarily start from the 0-th clock level. However, after at most T steps we pass through the Tth level, from which we jump to the 0-th level with probability $1/(10T)$. Hence after $O(T)$ steps we will pass through the 0-th clock level with high probability, so that we are in the setting of Example 6.4. A second difference is that we must also mix on the $\{\texttt{run}, \texttt{stop}\}$ degree-of-freedom. This is just a binary variable, but coupled to the other degrees of freedom. However, both the stopping rule transitions and the clock transitions are independent of the $\hat{V}$ degree of freedom, and furthermore the clock transitions are independent of the stopping rule. This allows to efficiently check that A' indeed does mix towards a stationary distribution including the $\{\texttt{run}, \texttt{stop}\}$ degree-of-freedom.

To prove the second bullet, we instead start from the LMC constructed according to [15] for the Markov chain P' maximizing the conductance $\Phi(P') = \Phi$. On this LMC we can then apply the same reasoning as above. □

Next we study tightness of the conductance bounds for the scenarios (Si).

Lemma 5. *Given a graph $\mathcal{G}$, an irreducible Markov chain P and a stationary distribution π, for every scenario in* (Si) *there exists an LMC with a mixing time* $\tau(1/4) \in O\left(\frac{1}{\Phi} \log \frac{1}{\min_i \pi_i}\right)$.

Proof. It suffices to prove the statement for an LMC in (Simr). Similarly to Lemma 4, we start from the LMC A_1 constructed in [15]. In the proof of Lemma 4 we show how to incorporate this LMC into a larger LMC in (sIMe), described by a transfer matrix A' having the appropriate mixing time. To obtain the same mixing time with A' in a scenario (SimrE), we make two adaptations.

First, we choose a suitable initialization that would satisfy invariance. Thereto let $\hat{\pi}$ be the stationary distribution of the LMC A' from the proof of Lemma 4; by irreducibility of A', we know that such a unique distribution exists. Now we can simply choose the unique local and linear initialization map F which maps π to the distribution $\{\texttt{run}\} \times \{0\} \times \hat{\pi}$ over the lifted node space, and this will necessarily result in an invariant LMC.

Next, we change the amplification parameter $\gamma = 1$ to some $\gamma \ll 1/T$ with T as in the proof of Lemma 4. This adaptation is equal to the adaptation in Example 6.4. As is argued in that example, if we pick γ sufficiently small then (i) the LMC will also mix over the clock degree-of-freedoms, (ii) the mixing time is unchanged (even with the additional `run/stop` degrees-of-freedom, see proof of Lemma 4), and (iii) the ergodic flows will not be violated. □

7.2. Deriving other properties from mixing scenarios bounds

We conclude this section with a few examples of reasonings based on our classification and mixing time bounds. We do not claim that this is the easiest way to obtain the following properties, but at least once our results are known they follow as rather direct implications.

Example 7.1 (*Symmetry Implications*)**.** If a fast design that ensures diameter-time convergence in an (`SI`) scenario also gives (`i`) for free, e.g. thanks to some symmetry property, then due to (`Si`) satisfying the conductance bound, we deduce in an indirect way that the diameter $D_{\mathcal{G}} \in \Theta(1/\Phi)$. This happens for instance on the cycle graph, as follows. Having developed a sequence of transition matrices $\{P_t\}_{t=1}^{T}$ to reach diameter-time convergence when starting from node 1, it is clear that the corresponding sequence for node k can be obtained just by translation, i.e. whenever the sequence for starting on node 1 has a transition probability $p_{(i+1,j+1)} = \alpha$ between two nodes $i+1, j+1$, the sequence for starting on node k has transition probability $p_{(i+k,j+k)} = \alpha$ (all additions of node labels are modulo N of course). With this we build an LMC as for the standard (`SI`) scenario, but it is obvious that when starting from $p(0) = \pi$, by symmetry, the LMC will keep $p(t) = \pi$ at all times, i.e. we satisfy in fact (`Si`). In this case the conductance bound should hold, but the LMC by construction converges in diameter time. This is consistent with $D_{\mathcal{G}} = N/2$ and $\Phi = 2/N$ for the uniform distribution over the cycle.

Conversely, on graphs where $D_{\mathcal{G}} \in \Theta(1/\Phi)$, one does not lose in mixing speed potential by imposing a constraint like (`s`) or (`i`) a priori; this enables e.g. to keep design possibilities in check. As a concrete example, on the cycle, the mixing algorithms of Examples 6.1 and 6.2 already achieve diameter-time mixing. □

Finally, we can formulate a trade-off where allowing an error on the steady state enables accelerated mixing. This is reminiscent of a result by Batu et al. [9] about possible improvement of the mixing time when considering "ϵ-close Markov chains" with respect to some metric. Such ϵ-close convergence is also playing a role more recently in examining the speedup enabled by quantum algorithms as compared to stochastic non-quantum algorithms for sampling from distributions with specified properties [1].

Example 7.2 (*Changing π to $\pi + \epsilon$*)**.** For any graph $\mathcal{G}$ and target distribution π, consider the LMC construction as we use in the proof of Theorem 2 for the scenario (`SImRE`): a node-clock-lift with erasure operator (see Section 2.2). Starting from $\mathcal{S}$, the probability distribution thus follows some unknown trajectory until reaching $x(t) = e_{D_{\mathcal{G}}+1} \otimes \bar{p} \otimes \pi$ after $D_{\mathcal{G}} + 1$ steps, and then staying there (we recall that $\bar{p}$ is an arbitrary distribution over $\mathcal{V}$ imposed via the erasure operator). If we start on a node outside $\mathcal{S}$, we are not guaranteed much.

The goal is to construct a related LMC where we do know, or bound, what happens when starting at any node. For this we first drop from the above LMC all nodes that would never be reached when starting from $\mathcal{S}$. Next, *without any other modification*, we assign that each lifted node $e_{D_{\mathcal{G}}+1} \otimes e_{v0} \otimes e_v$ jumps to $e_0 \otimes e_v \otimes e_v$ with probability $\gamma \ll 1$, and stays put with probability $1 - \gamma$. This is similar to our construction for (`SIr`) in the proof of Theorem 2,

except that now we will not perform any other tuning. When starting from an arbitrary node $e_d \otimes e_{v0} \otimes e_v$, over the first $D_{\mathcal{G}} + 1 - d$ time steps, the behavior is not really controlled, except: the clock-index in the lifted node gets incremented by 1; and the "starting-node" index remains put during the $D_{\mathcal{G}} - d$ first steps, and in the last step it gets projected to $\bar{p}$ provided $d \neq (D_{\mathcal{G}} + 1)$. After this time, we thus reach a distribution $e_{D_{\mathcal{G}}+1} \otimes \tilde{\pi} \otimes \tilde{\tilde{\pi}}$ with $\tilde{\tilde{\pi}}$ arbitrary and $\tilde{\pi} \in \{e_{v0},\ \bar{p}\}$. From here, at each further time step, a fraction γ will jump back to $\mathcal{S}$ and thus correctly converge towards the original π in $D_{\mathcal{G}} + 1$ additional steps. In the end, the LMC will thus converge towards a distribution

$$x = \epsilon \tilde{x} + (1 - \epsilon) e_{D_{\mathcal{G}}+1} \otimes \pi \otimes \pi \ , \tag{11}$$

where $\tilde{x}$ is a distribution over $\hat{\mathcal{V}}$ orthogonal to $e_{D_{\mathcal{G}}+1}$, and ϵ is a small scalar which we will compute below. The corresponding steady state on $\mathcal{V}$ is thus at a total variation distance ϵ from π. The just constructed Markov chain is ergodic, so this steady state is unique and indeed attracts any initial conditions.

Similarly to Example 6.4, the 'clock' coordinate of the lifted nodes follows a Markov chain that does not depend on the other coordinates, with $\mathcal{V}' = \{0, 1, \ldots, D_{\mathcal{G}} + 1\}$ and transition matrix

$$P' = \sum_{s=0}^{D_{\mathcal{G}}} e_{s+1} e_s^\dagger + \gamma e_0 e_{D_{\mathcal{G}}+1}^\dagger + (1 - \gamma) e_{D_{\mathcal{G}}+1} e_{D_{\mathcal{G}}+1}^\dagger \ .$$

The steady state of this Markov chain is $p'(s) = \frac{\gamma}{1+(D_{\mathcal{G}}+1)\gamma}$ for $s = 0, 1, \ldots, D_{\mathcal{G}}$ and $p'(D_{\mathcal{G}} + 1) = \frac{\gamma}{1+(D_{\mathcal{G}}+1)\gamma}$. This gives the value $\epsilon = \frac{(D_{\mathcal{G}}+1)\gamma}{1+(D_{\mathcal{G}}+1)\gamma}$. The bottleneck in this Markov chain is the jump from $s = D_{\mathcal{G}} + 1$ to $s = 0$ with probability γ. In the LMC, we are ensured to have converged to a steady state like (11), if we condition on having gone at least once through this bottleneck and then again up the chain on $\mathcal{V}'$. A significant part of the trajectories will have done this for $t > 1/\gamma$, and thus taking $t = m/\gamma + 2(D_{\mathcal{G}} + 1)$ for some small integer m, we can guarantee that the LMC will have converged close to its steady state (11), *from any starting node*, and thus $\tau(1/4) < m/\gamma + 2(D_{\mathcal{G}} + 1)$ in the sense (sIm). Taking all things together, we conclude:

> *For any $\mathcal{G}$ and π, there exists an LMC in the sense* (sIm) *converging to a steady state at a total-variation distance ϵ from π with a mixing time $\tau(1/4) < O(\frac{1}{\epsilon} D_{\mathcal{G}})$. By Theorem 3, this implies that for any $\mathcal{G}$ and π, there exists a steady state at a total-variation distance ϵ from π and whose associated conductance Φ satisfies $1/\Phi < O(\frac{1}{\epsilon} D_{\mathcal{G}})$.* □

Before concluding, we mention that the constructive nature of our results may allow them to identify graphs with good mixing properties. For instance, as could already be said with the result of [15] (in adapted form with our mixing time definition), the LMC constructions which achieve tightness to the conductance bound in a (m) setting, actually provide graphs on node set $\hat{\mathcal{V}}$ for which the conductance bound can be reached without lifting. As another example, the LMC constructed in Theorem 2 provides a graph and a subset of nodes such that, when starting from that subset, the stationary distribution is reached in diameter-time. While such observations are elementary consequences of our investigation, they might turn out to be useful in other contexts.

7.3. Extensions of our results' scope

The generality of the LMC framework seems eligible to cover more general dynamics and algorithms. In [6] we had noted that the conductance bound holds for any stochastic process that satisfies graph locality and invariance. In this section we discuss the possible extension of our results to consensus algorithms and pseudo-lifting.

1. Consensus Algorithms: Markov chains can be viewed as the dual of *consensus algorithms* for e.g. load balancing or rumor spreading. Reaching consensus among nodes of a graph means converging towards the same value on each node. Linear algorithms which propose speedups for reaching consensus can be found for instance in [30,33,42,49,51,58]. The presence of local memory bears resemblance to LMCs, and it would be interesting to carry over our results to bound the achievable speedup for consensus. This involves two points.

- **Value knowledge:** In a stochastic process, the probability x_k or p_k of being at node k describe the algorithm's state, but it is a priori not accessible to the walker. In a consensus algorithm instead, p_k represents a workload, sensor measurement, or a similar concrete value known to node k. Decision processes for consensus can thus depend on the value of p_k, i.e. as if in a Markov Chain the decision of the walker sitting at k would depend on its probability to have ended up there. Remarkably, the LMC framework allows to encode such knowledge in the lifted nodes, at least in a (S) setting with large memory. Regarding this point, our LMC results would thus carry over, e.g. for the (SiM) setting in a way similar to [6].
- **Positivity:** Linear consensus algorithms a priori do not have to satisfy positivity: e.g. a memoryless linear consensus update $p_k(t+1) = \sum a_{k,j}\, p_j(t)$ must satisfy $\sum_j a_{k,j} = 1$ for each k, but they do not need to be positive. It has been shown that higher-order consensus algorithms accelerate convergence precisely when the weights attributed are not all positive. Hence one might wonder how the bounds derived in our paper fare when relaxing positivity. A concrete example of a linear algorithm for accelerated consensus that does not obey positivity is proposed in [50], achieving convergence in $O(N)$ steps on any graph. Apart from this positivity condition however, it can be seen to fit the scenario (Si). And, since for any graph the conductance associated to the uniform distribution is at least of order $1/N$, it does satisfy the $O(1/\Phi)$ convergence time bound for LMCs. The following result shows that this is not pure coincidence, as our LMC results do apply to consensus algorithms, at least for the local convergence close to the consensus situation.

Proposition 1. *Consider any bounded consensus algorithm, satisfying invariance (i.e. $p_1 = p_2 = \cdots = p_N = \bar{p}$ at $t = 0$ implies the same property for all $t > 0$) and linearity (if two initial conditions $p(0)$ and $p'(0)$ evolve as $p(t)$ and $p'(t)$ respectively, then for any $a, a' \in \mathbb{R}$ the initial condition $a\, p(0) + a'\, p'(0)$ evolves as $a\, p(t) + a'\, p'(t)$ for all $t \geq 0$). Consider the set of initial conditions $\mathcal{R}_a = \{q \in \mathbb{R}^N : q_k = (1-a) + a\, p_k, \quad p \in \mathbb{P}_N\}$. There exists a value of $a > 0$ such that for all $\alpha \in (0, a]$, the consensus algorithm needs a time lower bounded by $O(1/\Phi)$ to converge from the worst state in $\mathcal{R}_\alpha$ towards a total-variation distance $\alpha/4$ from consensus.*

Proof. By boundedness and linearity of the algorithm, there exists a finite a such that for all initial states in $\mathcal{R}_\alpha$, the evolution under the consensus algorithm satisfies the positive locality constraint of an LMC; this is very clear e.g. from the abstract formulation of locality in [6].

Hence we can write a stochastic bridge (see the proof of Theorem 2) starting at each of the extrema of $\mathcal{R}_\alpha$ (corresponding to $p = e_k$ in its definition), and combine them into an LMC of type (Si). For initial conditions in $\mathcal{R}_\alpha$, the evolution and convergence of the consensus algorithm thus equals the evolution and convergence of the LMC. There remains to bound the convergence of an LMC for restricted initial conditions.

Consider any LMC in the scenario (SiM) and converging towards the uniform distribution. By linearity and invariance, if the mixing time $\tau(1/4) = T$, then the time requested for all states in $\mathcal{R}_\alpha$ to reach a total-variation distance $\alpha/4$ from the uniform equals T. Together with the previous paragraph and the lower bound on the mixing time of an LMC in the scenario (SiM), this implies the announced result. □

This result provides a bound for a range of consensus algorithms, like [50] or so-called *polynomial filters* [5,38,47,56,57], where linearity and invariance are standard. The restriction of initial states to some set is always necessary for consensus algorithms, admittedly we cannot claim to have the least constraining one regarding global convergence speed.

2. Pseudo-Lifting: In [34] the concept of "pseudo-lifting" a Markov chain is introduced. This comes down to creating a lifted Markov chain whose probability distribution does not on itself converge to the appropriate limit distribution, and neither does its marginal. They however use a generalized stopping rule, conditioning on being on a certain subset of the lifted nodes, to ensure a correct limit distribution conditioned on stopping. Although motivated by several other reasons, the name "pseudo-lift" can ultimately be linked to the fact that, unlike in the LMC analysis of [15] and in Example 4.1, the stopping rule does change the limit distribution. Moreover, contrary to the latter's typical conductance bound, the pseudo-lifting algorithm allows to achieve diameter-time convergence on general graphs. This seems to hint at the fact that reformulating so general stopping rules in the pure LMC framework would lead to (SI) scenarios.

8. Summary and perspective

We show how a wide range of approaches to accelerate the mixing time of random walks can be cast as lifted Markov chains (LMCs) in different scenarios — see examples throughout the text and further ones in the appendix. We provide an extensive classification of these scenarios, and show that the limits and opportunities of acceleration approaches feature a subtle yet clear dependency on the scenario in which they can be cast. This allows us to put the specific results on LMCs, such as the conductance bounds of [15], into perspective, and extend them directly to different algorithms. In order to identify the relevant scenarios, we introduce and investigate 5 possible constraints on a LMC, motivated by existing examples, and we show that fundamental bounds on its mixing time can be attributed to merely two of these: (s) — impossibility of locally initializing the lift, or (i) — demanding invariance of the target distribution. Importantly, these are properties which can be directly checked on an algorithm itself — without needing to actually translate it as an LMC.

More precisely, we show that *as soon as either* (s) *or* (i) *are imposed*, the mixing time is subject to a conductance bound; the case with (s) appears closer to existing literature like [15], while the case with (i) we would argue is closer to actual algorithmic formulations. We further invigorate this claim by proving that requiring *neither* (s) *nor* (i) allows to mix in diameter time, whereas *requiring both* (s) *and* (i) is too restrictive: a lift in this scenario cannot accelerate the mixing time as compared to a non-lifted Markov chain. These results are summarized in Fig. 5.

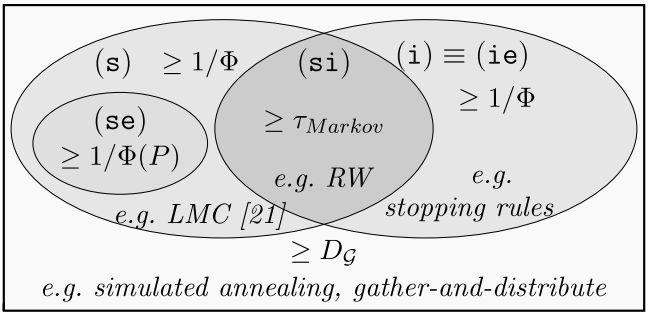

Fig. 5. Summary of the main bounds on mixing time $\tau(1/4)$, mentioning only the most relevant constraints and up to small constant factors. By τ_{Markov} we mean the mixing time of the non-lifted Markov chain $p(t+1) = P\,p(t)$, where in case (E) the transition matrix P can be optimized and in case (e) it is given. **All bounds are — essentially tight**, in the sense that for all scenarios and all graphs there exists an LMC attaining the lower bound up to a log factor; — **essentially significant**, in the sense that there exist graphs where the various bounds differ by polynomial factors.

A main message of this analysis is that one should be careful with specifying scenarios for accelerated mixing. The relevance of imposing ergodic flows for instance, becomes rather questionable when one allows for algorithm initialization. The invariance requirement (i) may appear like a natural "stabilizing" requirement in an algorithmic setting, but it differs from the typical LMC setting in [15], and if added to the (s) assumption of [15] it would allow no speedup at all. Properly identifying such properties gains particular importance when lifted Markov chains are to be compared to other acceleration strategies, like the discrete-time quantum walks [35] which we are addressing elsewhere [8]. On the positive side, our results clarify, with explicit lift constructions, that some properties come essentially for free. For instance, converging on the full state space (m) rather than only on the output variable is not much more demanding. We also clarify an explicit tradeoff where allowing for an error on the steady state, enables to attain a $\tau(1/4)$ closer to the graph diameter despite being in the conditions of the conductance bound.

A question that is not touched by our classification is the complexity of algorithm design. Our proofs are constructive and thus also indicate the potential in the use of lifted Markov chains to speed up mixing. However, as in [15], from an algorithmic viewpoint the value of the processes we construct is more existential than practical. Our proof of Theorem 2 for instance heavily builds on a non-distributed, extensive optimization of the edge weights and, as such, it is certainly not a viable option for e.g. Markov chains used for Monte-Carlo sampling in large systems. In the light of this, a most important open problem regards the development of heuristics or suboptimal versions of our algorithms that would restrict the information used not only for online implementation but also for algorithm design. To the best of our knowledge, except for particular graphs exhibiting strong symmetries, a general way to build such a process and obtain significant mixing speedup remains an open issue for lifted Markov chains, consensus algorithms, and quantum walks. In this sense, our classification provided in Fig. 5 could at least clarify the potential of different scenarios, or suggest a way to conveniently modify existing approaches.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Some partial results of this work regarding the role of initialization and invariance that could be of interest to the system and control community had been presented in the conference paper [7].

Appendix. Irreducible approximation lemma

We here provide a technically rigorous proof that a reducible LMC can be adapted into an irreducible one with the same mixing scenario properties. We specifically address the lift construction of Section 2.2, as used at the end of the proof of Theorem 2, and prove that it can be made to satisfy the four properties stated there. This is essentially a somewhat technical yet uncomplicated ϵ, δ type argument.

Lemma 6. *Consider a node-clock-lift LMC with canonical initialization which converges in finite time T to a unique limit distribution $e_T \otimes \pi \otimes \pi$, on which it acts like $e_T e_T^\dagger \otimes I_{\mathcal{V}} \otimes P$, where P is irreducible and $P\pi = \pi$. Then for any $\epsilon > 0$, there exists a non-empty set of small enough $\gamma \in (0, \bar{\gamma})$ and associated transition matrices $\tilde{P}$ over $\mathcal{V}$ such that: replacing $e_T e_T^\dagger \otimes I_{\mathcal{V}} \otimes P$ by*

$$(1-\gamma) e_T e_T^\dagger \otimes I_{\mathcal{V}} \otimes \tilde{P} + \gamma\, e_0 e_T^\dagger \otimes \sum_{v, v_0 \in \mathcal{V}} e_v e_{v_0}^\dagger \otimes e_v e_v^\dagger \,,$$

yields an irreducible LMC, with unique limit distribution satisfying $C\hat{\pi} = \pi$, mixing time $\tau(\epsilon) \leq T$, and not exceeding the ergodic flows of P.

Proof. In essence the point is to construct an LMC which:

(i) implements a stochastic bridge towards some π' that is ϵ-close to the target π; thus after T steps we have a state $e_T \otimes p(0) \otimes \pi'$ and we have mixed appropriately;
(ii) at the same time, due to $\gamma \neq 0$, the steady state is not $e_T \otimes p(0) \otimes \pi'$ but another close distribution $\hat{\pi}$ which does satisfy $C\hat{\pi} = \pi$ *exactly*.

The proof is organized as follows. *a)* We begin by formally computing the adjusted $\tilde{P}$ to ensure $C\hat{\pi} = \pi$ for any fixed value of $\gamma \in (0, \gamma_0)$, for some $\gamma_0 > 0$; *b)* we prove that there exists $\gamma_1 \in (0, \gamma_0)$ to ensure that $\tilde{P}$ is a proper transition matrix for all $\gamma \leq \gamma_1$, in particular it is positive; *c)* a simple continuity argument ensures that there exists $\gamma_2 \in (0, \gamma_1)$ such that moreover the implied ergodic flows will not exceed those of P, for all $\gamma \leq \gamma_2$; *d)* then we analyze the mixing time and prove that for any $\epsilon > 0$, there exists $\bar{\gamma} \in (0, \gamma_2)$ such that the mixing time result holds for all $\gamma \leq \bar{\gamma}$.

a) To compute $\tilde{P}$, we first note that, with $\tilde{P}$, the steady state $\hat{\pi}$ will be of the form

$$\hat{\pi} = \frac{1}{1+T\gamma} e_T \otimes \pi \otimes \tilde{\pi} + \sum_{t=0}^{T-1} \frac{\gamma}{1+T\gamma} A^t F \tilde{\pi} \,, \tag{A.1}$$

where $\tilde{\pi} = \gamma\pi + (1-\gamma)\tilde{P}\tilde{\pi}$. We will *a1)* first, compute a $\tilde{\pi}$ close to π and such that $C\hat{\pi} = \pi$ in the expression (A.1), without caring about its relation to $\tilde{P}$; *a2)* next, show how to construct a corresponding $\tilde{P}$ (without caring for positivity yet.)

a1) Let $\sum_{t=0}^{T-1} CA^t F =: T\,B$. We thus want that $\left(\frac{1}{1+T\gamma} + \frac{T\gamma}{1+T\gamma} B\right) \tilde{\pi} = \pi$. There obviously exists a $\gamma_0 > 0$ such that this equation is invertible for all $\gamma \in (0, \gamma_0)$, and gives a

solution $\tilde{\pi}$ that equals π up to terms of order γ. In particular, there exist $\alpha, \gamma_0 > 0$ such that $\tilde{\pi}_k \geq \alpha > 0$ for all k, and for all $\gamma > 0$ smaller than γ_0.

a2) There remains to find $\tilde{P}$ such that $\tilde{P}\tilde{\pi} = \frac{1}{1-\gamma}(\tilde{\pi} - \gamma\pi)$. Since P is irreducible, there must exist some $\beta > 0$ such that the edges (i, j) of $\mathcal{G}$ for which $P_{i,j} \geq \beta/(1-\beta)$, contain a rooted spanning tree $\mathcal{G}_\beta$. Then the matrix $P^{(\beta)} = (1-\beta)P + \beta I$ will have self-loops larger than β, i.e. $P^{(\beta)}_{k,k} \geq \beta$ for all k, and also weights $P^{(\beta)}_{i,j} \geq \beta$ for all edges of the spanning tree $\mathcal{G}_\beta$. (The value of β will only be used in part *b)* of the proof.) We write $\tilde{P} = P^{(\beta)} + P'$ with P' to be identified but only covering the spanning tree $\mathcal{G}_\beta$; then defining $y = (\frac{1}{1-\gamma} - P^{(\beta)})(\tilde{\pi} - \pi)$, the equation to solve becomes:

$$P'\tilde{\pi} = y \ , \quad \sum_{k=1}^{N} P'_{k,\ell} = 0 \ \text{ for } \ell = 1, 2, \ldots, N$$

$$P'_{k,\ell} = 0 \text{ for all } k \neq \ell : (k, \ell) \notin \mathcal{G}_\beta \ ,$$

knowing that $\sum_{k=1}^{N} y_k = 0$. We can construct a solution P' with the following algorithm:

1: $\mathcal{G}_{rem} := \mathcal{G}_\beta$; $P' :=$ zeros$(N \times N)$;
2: **while** $|\mathcal{G}_{rem}| > 1$ **do**
3: select $j \in$ leaves of $\mathcal{G}_{rem}$;
4: $k :=$parent(j);
5: $P'_{j,k} := (y_j - P'_{j,j}\tilde{\pi}_j)/\tilde{\pi}_k$;
6: $P'_{k,k} := P'_{k,k} - P'_{j,k}$;
7: $\mathcal{G}_{rem} := \mathcal{G}_{rem} \setminus \{j\}$;
8: **end while**

Line 5 ensures to satisfy $P'\tilde{\pi} = y$ for row j. Indeed, since any j' for which we have added a nondiagonal element on row j' of P' at a previous iteration has been thrown out of $\mathcal{G}_{rem}$, the only nonzero element on row j so far can be $P'_{j,j}$; in the future we will have thrown out j of $\mathcal{G}_{rem}$, so $P'_{j,j}$ will not further change; this makes the solution of the linear equation trivial. Line 6 ensures to maintain $\sum_{\ell=1}^{N} P'_{\ell,k} = 0$ despite the change on $P'_{j,k}$ on line 5. One node is thrown out of $\mathcal{G}_{rem}$ at each iteration, and the algorithm stops when $\mathcal{G}_{rem}$ comes down to a single isolated node r (root of the spanning tree). At this point all rows of $P'\tilde{\pi} = y$ have been treated, except row r. Writing $\sum_{k\neq r}[P'\tilde{\pi}]_k - y_k = 0$ and from our problem setting $\sum_{k=1}^{N} P'_{k,\ell}\tilde{\pi}_\ell = \sum_{k=1}^{N} y_k = 0$, we see that in fact $P'\tilde{\pi} = y$ is satisfied for row r as well, and we have constructed a solution P', i.e. some $\tilde{P}$ such that $C\hat{\pi} = \pi$.

b) To ensure that $\tilde{P}$ is positive, a sufficient condition is to show that the elements of P', in absolute value, are smaller than β. Indeed, $\tilde{P} = P^{(\beta)} + P'$ and $P^{(\beta)}_{j,k} \geq \beta$ by construction for all j, k for which we allow $P'_{j,k} \neq 0$ (i.e., the off-diagonal elements corresponding to edges of $\mathcal{G}_\beta$ and the diagonal elements corresponding to self-loops of $P^{(\beta)}$). For any fixed β, we observe that the components of P' can in fact be made arbitrarily small by taking γ small enough. This follows because

- $P' = 0$ for $\gamma = 0$, as we then have $\tilde{\pi} = \pi$ so $y = 0$ and the above algorithm keeps setting values to 0; and
- P' is continuous in γ around 0: indeed, everything in the construction is linear except the division by $\tilde{\pi}_k$, but we have shown that $\tilde{\pi}_k \geq \alpha$ for all $\gamma \leq \gamma_0$.

Therefore there exists a $\gamma_1 > 0$ such that, for any $\gamma \leq \gamma_1$, there is a (positive) stochastic matrix $\tilde{P}$ yielding an irreducible LMC whose steady state $\hat{\pi}$ satisfies $C\hat{\pi} = \pi$ exactly.

c) For any fixed $\beta > 0$ in *a), b)*, the ergodic flows of the LMC are continuous in γ, and for $\gamma = 0$ they are exactly equal to those of $P^{(\beta)}$ (by the very same argument as in *b)*). Therefore, for any fixed $\beta > 0$ and $\delta > 0$, there exists a $\gamma_\beta \in (0, \gamma_1)$ such that the proposed LMC construction adds no more than δ to the ergodic flows of $P^{(\beta)}$. Note that for two nodes j, k with $P_{j,k} = P^{(\beta)}_{j,k} = 0$ (no edge in P), we will also keep zero ergodic flows in any LMC. For $P_{j,k} \neq 0$, the ergodic flows of $P^{(\beta)}_{j,k}$ are strictly smaller than those of P, and deviating from $P^{(\beta)}$ by no more than a sufficiently small δ will ensure that also the final LMC has ergodic flows at most equal to those of P, for all $\gamma \in (0, \gamma_2)$ with $\gamma_2 < \gamma_1$ guaranteeing a sufficiently small δ for all edges. We have thus constructed an LMC satisfying all the properties of the statement, except possibly the mixing time.

d) To analyze the mixing time, we will further restrict γ, now making it dependent on the value ϵ to which we want to approximate the target $\hat{\pi}$ (recall that we must prove the statement with constraint (m), not just the convergence of the marginal). This might suggest a not very practical construction, but it suffices for the purpose of our existence proof. Since we did not modify the steps that the walk will take during the T first time steps (i.e. before it reaches a node of type (T, i, v)), we still have $x_T = e_T \otimes \pi \otimes \pi$ for any $\gamma > 0$ and all initial states $Fp(0)$. By a similar continuity argument as in *b)*, for any given $\epsilon > 0$, there exists a $\gamma_3 \in (0, \gamma_2)$ such that $\|\pi - \tilde{\pi}\|_{TV} \leq \|\hat{\pi} - e_T \otimes \pi \otimes \pi\|_{TV} < \epsilon/2$. Indeed for $\gamma = 0$ we have $\hat{\pi} = x_T$; and for $\gamma \neq 0$ the isolated eigenvector solution of $Ax = x$ changes continuously, with γ influencing A continuously. So we have proved the mixing time condition for $t = T$, there remains to prove that $x(t)$ will not move away from $\hat{\pi}$ too much for any $t > T$.

d1) This can be seen by first observing that the "clock" degrees of freedom undergo an independent Markov chain on $\mathcal{V}^{(clock)} = \{s = 0, 1, \ldots, T\}$ with transition matrix

$$P^{(clock)} = \sum_{i=0}^{T-1} e_{i+1} e_i^\dagger + (1-\gamma) e_T e_T^\dagger + \gamma e_0 e_T^\dagger. \tag{A.2}$$

This is a particular Markov chain on a path or cycle graph, which is independent of the original problem and $\mathcal{G}$, except for its length T, and with stationary distribution $\pi^{(clock)}$ of the form (this is just the marginal of (A.1)):

$$\pi_0^{(clock)} = \pi_1^{(clock)} = \cdots = \pi_{T-1}^{(clock)} = \gamma/(1+\gamma T)\,, \qquad \pi_T^{(clock)} = 1/(1+\gamma T)\,.$$

Writing the distribution over clock states at time t as $w(t) = \pi^{(clock)} + q(t)$, with q the deviation from the stationary distribution $\pi^{(clock)}$ of $P^{(clock)}$, a few computations give:

$$\|q(t+T)\|_{TV} \leq 2T\gamma \|q(t)\|_{TV}\,.$$

Selecting $\gamma < (\eta\epsilon)/(2T) =: \gamma_\eta$ ensures that the deviation from the stationary distribution $\|q(t+T)\|_{TV} \leq \eta\epsilon$ for all $t > 0$. Moreover, the stationary distribution itself will then satisfy $\pi_T^{(clock)} > 1/(1+\eta\epsilon/2)$.

d2) The other degrees of freedom (v_0, v) of the LMC undergo a motion conditioned on the "clock" evolution. Consider the actions applied on (v_0, v) for a given evolution of the clock value for $t > T$ time steps. The walker has first moved up to $x_T = e_T \otimes \pi \otimes \pi$, then stayed on e_T for some r_0 steps, thus applying $I \otimes \tilde{P}^{r_0}$, then applied J to jump back to a state of type $F\tilde{P}^{r_0}\pi$, then possibly moved up again towards $x_t = e_T \otimes \pi \otimes \pi$, again stayed there for some r_1 steps, and so on. The resulting distribution over (v_0, v) associated to the clock value $s = T$

is thus a convex combination of distributions $\pi \otimes \tilde{P}^r \pi$, for $r = 0, 1, 2, \ldots$. Therefore, for any $t > T$, the state takes the form

$$x(t) = \sum_{k=0}^{T} w_k \, e_k \otimes \bar{y}_k \otimes \bar{y}'_k + \; w_T \, e_T \otimes \pi \otimes \Big(\sum_{r=0}^{+\infty} u_r \tilde{P}^r \pi \Big) \,,$$

for some distributions $\bar{y}_k$ and $\bar{y}'_k$ over $\mathcal{V}$, a distribution w over $\{0, 1, \ldots, T\}$, and a distribution u over $\mathbb{N}$. In part *d1)* we have shown that $\|w(t) - \pi^{(clock)}\|_{TV} \leq \eta\epsilon$ and that $\pi_T^{(clock)} > \frac{1}{1+\eta\epsilon/2}$, which together implies $w(t)_T \geq 1 - \frac{5\eta\epsilon/2+(\eta\epsilon)^2}{1+\eta\epsilon/2} =: 1 - \eta_1$, for all $\gamma \leq \gamma_\eta$ and all $t > T$. Denoting $x_s^{(\mathcal{V}^2)} = \sum_{v_0, v \in \mathcal{V}} x_{(s,v_0,v)}$, we then have:

$$\begin{aligned}
\|x(t) - \hat{\pi}\|_{TV} &= \sum_{s=0}^{T-1} \|x(t)_s^{(\mathcal{V}^2)} - \hat{\pi}_s^{(\mathcal{V}^2)}\|_{TV} + \|x(t)_T^{(\mathcal{V}^2)} - \hat{\pi}_T^{(\mathcal{V}^2)}\|_{TV} \\
&\leq \frac{1}{2} |x(t)_s^{(\mathcal{V}^2)}| + \|x(t)_T^{(\mathcal{V}^2)} - \tfrac{1}{1+\gamma(T)} \pi \otimes \tilde{\pi}\|_{TV} \\
&\leq \frac{\eta_1}{2} + \max_{u \in \mathbb{P}_\mathbb{N}} \Big\| w(t)_T \pi \otimes \Big(\sum_{r=0}^{+\infty} u_r \tilde{P}^r \pi \Big) - \tfrac{1}{1+\gamma T} \pi \otimes \tilde{\pi} \Big\|_{TV} .
\end{aligned}$$

We have here replaced $\hat{\pi}_T^{(\mathcal{V}^2)}$ by its expression $\frac{1}{1+\gamma T} \pi \otimes \tilde{\pi}$ from (A.1), and in the last line $x(t)_T^{(\mathcal{V}^2)}$ by its worst-case expression from our analysis just above. We next decompose as standard, replacing the convex combination of $\tilde{P}^r$ by the symbol W:

$$\begin{aligned}
\|w(t)_T \pi \otimes W\pi - \tfrac{1}{1+\gamma T} \pi \otimes \tilde{\pi}\|_{TV} &\leq |w(t)_T - \tfrac{1}{1+\gamma T}| \, \|\pi \otimes \tilde{\pi}\|_{TV} \\
&\quad + |w(t)_T| \, \|\pi \otimes (W\pi - \tilde{\pi})\|_{TV} \\
&\leq |w(t)_T - \tfrac{1}{1+\gamma T}|/2 + \|W\pi - \tilde{\pi}\|_{TV} \\
&\leq \eta\epsilon + \|W\pi - \tilde{\pi}\|_{TV} \,.
\end{aligned}$$

The last inequality follows from $\|w(t) - \pi^{(clock)}\|_{TV} \leq \eta\epsilon$ proved in *d1)*. We can treat its last terms as

$$\max_{u \in \mathbb{P}_\mathbb{N}} \Big\| \sum_{r=0}^{+\infty} u_r \tilde{P}^r \pi - \tilde{\pi} \Big\|_{TV} = \max_{u \in \mathbb{P}_\mathbb{N}} \Big\| \sum_{r=0}^{+\infty} u_r \tilde{P}^r (\pi - \tilde{\pi}) \Big\|_{TV} \leq \|\pi - \tilde{\pi}\|_{TV} \leq \epsilon/2$$

for all $\gamma < \gamma_3$, as settled at the beginning of point *d)*. Taking all things together, we have shown that for any fixed ϵ and η, there exists $\gamma_3 > 0$ and $\gamma_\eta > 0$ such that:

$$\|x(t) - \hat{\pi}\|_{TV} \leq \frac{5\eta\epsilon/2 + (\eta\epsilon)^2}{2 + \eta\epsilon} + \eta\epsilon + \frac{\epsilon}{2}$$

for all $\gamma \leq \min(\gamma_\eta, \gamma_3)$ and all $t > T$. Selecting η such that $\frac{5\eta\epsilon/2+(\eta\epsilon)^2}{2+\eta\epsilon} + \eta\epsilon \leq 1/2$ ensures that $\|x(t) - \hat{\pi}\|_{TV} \leq \epsilon$ for all $t > T$, thus ensuring a mixing time $\tau(\epsilon) = T$.

This concludes the proof, as it suffices to take $\bar{\gamma}$ smaller than $\min(\gamma_\eta, \gamma_3)$ with the corresponding value of η, to ensure that all statements of the lemma are satisfied. □

References

[1] S. Aaronson, A. Arkhipov, The computational complexity of linear optics, in: Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing, ACM, 2011, pp. 333–342.

[2] D. Aldous, On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing, Probab. Engrg. Inform. Sci. 1 (01) (1987) 33–46.

[3] D. Aldous, J. Fill, Reversible Markov chains and random walks on graphs, 2002, Unfinished monograph. lin k.

[4] N. Alon, I. Benjamini, E. Lubetzky, S. Sodin, Non-backtracking random walks mix faster, Commun. Contemp. Math. 9 (04) (2007) 585–603.

[5] S. Apers, A. Sarlette, Accelerating consensus by spectral clustering and polynomial filters, IEEE Trans. Control Netw. Syst. 4 (3) (2017) 544–554.

[6] S. Apers, A. Sarlette, F. Ticozzi, Bounding the convergence time of local probabilistic evolution, in: International Conference on Geometric Science of Information, Springer, 2017, pp. 754–762.

[7] S. Apers, A. Sarlette, F. Ticozzi, When does memory speed-up mixing?, in: 2017 IEEE 56th Annual Conference on Decision and Control (CDC), IEEE, 2017, pp. 4940–4945.

[8] S. Apers, A. Sarlette, F. Ticozzi, Simulation of quantum walks and fast mixing with classical processes, Phys. Rev. A 98 (3) (2018) 032115.

[9] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, P. White, Testing that distributions are close, in: Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, IEEE, 2000, pp. 259–269.

[10] F. Bénézit, A. Dimakis, P. Thiran, M. Vetterli, Gossip along the way: Order-optimal consensus through randomized path averaging, in: Allerton, LCAV-CONF-2009-004, 2007.

[11] J. Bierkens, Non-reversible metropolis-hastings, Stat. Comput. 26 (6) (2016) 1213–1228.

[12] J. Bierkens, G. Roberts, et al., A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model, Ann. Appl. Probab. 27 (2) (2017) 846–882.

[13] S. Boyd, P. Diaconis, P. Parrilo, L. Xiao, Fastest mixing Markov chain on graphs with symmetries, SIAM J. Optim. 20 (2) (2009) 792–819.

[14] S. Boyd, P. Diaconis, L. Xiao, Fastest mixing Markov chain on a graph, SIAM Rev. 46 (4) (2004) 667–689.

[15] F. Chen, L. Lovász, I. Pak, Lifting Markov chains to speed up mixing, in: Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, ACM, 1999, pp. 275–281.

[16] D. Dervovic, For every quantum walk there is a (classical) lifted Markov chain with faster mixing time, 2017, arXiv preprint arXiv:1712.02318.

[17] P. Diaconis, Some things we've learned (about Markov chain Monte Carlo), Bernoulli 19 (4) (2013) 1294–1305.

[18] P. Diaconis, S. Holmes, R.M. Neal, Analysis of a nonreversible Markov chain sampler, Ann. Appl. Probab. (2000) 726–752.

[19] P. Diaconis, L. Miclo, On the spectral analysis of second-order Markov chains, Ann. Fac. Sci. Toulouse Math. 22 (3) (2013) 573–621.

[20] P. Diaconis, D. Stroock, Geometric bounds for eigenvalues of Markov chains, Ann. Appl. Probab. (1991) 36–61.

[21] M. Dyer, A. Frieze, R. Kannan, A random polynomial-time algorithm for approximating the volume of convex bodies, J. ACM 38 (1) (1991) 1–17.

[22] R. Fitzner, R. van der Hofstad, Non-backtracking random walk, J. Stat. Phys. 150 (2) (2013) 264–284.

[23] T.T. Georgiou, M. Pavon, Positive contraction mappings for classical and quantum Schrödinger systems, J. Math. Phys. 56 (3) (2015) 033301.

[24] L. Georgopoulos, Definitive Consensus for Distributed Data Inference (Ph.D. dissertation), EPFL, 2011.

[25] B. Gerencsér, Markov chain mixing time on cycles, Stochastic Process. Appl. 121 (11) (2011) 2553–2570.

[26] B. Gerencsér, J.M. Hendrickx, Improved mixing rates of directed cycles by added connection, J. Theoret. Probab. 32 (2) (2019) 684–701.

[27] W.K. Hastings, Monte Carlo Sampling methods using Markov chains and their applications, Biometrika 57 (1) (1970) 97–109.

[28] T.P. Hayes, A. Sinclair, Liftings of tree-structured Markov chains, in: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer, 2010, pp. 602–616.

[29] J.M. Hendrickx, R.M. Jungers, A. Olshevsky, G. Vankeerberghen, Graph diameter, eigenvalues, and minimum-time consensus, Automatica 50 (2) (2014) 635–640.

[30] J.M. Hendrickx, G. Shi, K.H. Johansson, Finite-time consensus using stochastic matrices with positive diagonals, IEEE Trans. Automat. Control 60 (4) (2015) 1070–1073.

[31] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 2012.
[32] M. Jerrum, A. Sinclair, E. Vigoda, A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries, J. ACM 51 (4) (2004) 671–697.
[33] B. Johansson, M. Johansson, Faster linear iterations for distributed averaging, IFAC Proc. Vol. 41 (2) (2008) 2861–2866.
[34] K. Jung, D. Shah, J. Shin, Distributed averaging via lifted Markov chains, IEEE Trans. Inform. Theory 56 (1) (2010) 634–647.
[35] J. Kempe, Quantum random walks: an introductory overview, Contemp. Phys. 44 (4) (2003) 307–327.
[36] M. Kempton, Nonbacktracking random walks and a weighted Ihara's theorem, Open J. Discrete Math. 6 (4) (2016) 207–226.
[37] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, et al., Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[38] E. Kokiopoulou, P. Frossard, Polynomial filtering for fast convergence in distributed consensus, IEEE Trans. Signal Process. 57 (1) (2009) 342–354.
[39] G.F. Lawler, A.D. Sokal, Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality, Trans. Am. Math. Soc. 309 (2) (1988) 557–580.
[40] D.A. Levin, Y. Peres, E.L. Wilmer, Markov Chains and Mixing Times, American Mathematical Soc., 2009.
[41] W. Li, H. Dai, Y. Zhang, Location-aided fast distributed consensus in wireless networks, IEEE Trans. Inform. Theory 56 (12) (2010) 6208–6227.
[42] J. Liu, B.D. Anderson, M. Cao, A.S. Morse, Analysis of accelerated gossip algorithms, Automatica 49 (4) (2013) 873–883.
[43] L. Lovász, P. Winkler, Mixing times, Microsurv. Discrete Probab. 41 (1998) 85–134.
[44] F. Martinelli, Lectures on glauber dynamics for discrete spin models, in: Lectures on Probability Theory and Statistics, Springer, 1999, pp. 93–191.
[45] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087–1092.
[46] M. Mihail, Conductance and convergence of Markov chains-A combinatorial treatment of expanders, in: Foundations of Computer Science, 1989., 30th Annual Symposium on, IEEE, 1989, pp. 526–531.
[47] E. Montijano, J.I. Montijano, C. Sagues, Chebyshev Polynomials in distributed consensus applications, IEEE Trans. Signal Process. 61 (3) (2013) 693–706.
[48] R. Motwani, P. Raghavan, Randomized Algorithms, Chapman & Hall/CRC, 2010.
[49] S. Muthukrishnan, B. Ghosh, M.H. Schultz, First-and second-order diffusive methods for rapid, coarse, distributed load balancing, Theory Comput. Syst. 31 (4) (1998) 331–354.
[50] A. Olshevsky, Linear time average consensus and distributed optimization on fixed graphs, SIAM J. Control Optim. 55 (6) (2017) 3990–4014.
[51] B.N. Oreshkin, M.J. Coates, M.G. Rabbat, Optimization and analysis of distributed averaging with short node memory, IEEE Trans. Signal Process. 58 (5) (2010) 2850–2865.
[52] M. Pavon, F. Ticozzi, Discrete-time classical and quantum Markovian evolutions: Maximum entropy problems on path space, J. Math. Phys. 51 (4) (2010) 042104.
[53] L. Rabiner, B. Juang, An introduction to hidden Markov models, IEEE ASSP Mag. 3 (1) (1986) 4–16.
[54] K. Ramanan, A. Smith, Bounds on lifting continuous-state Markov chains to speed up mixing, J. Theoret. Probab. 31 (3) (2018) 1647–1678.
[55] L. Rey-Bellet, K. Spiliopoulos, Improving the convergence of reversible samplers, J. Stat. Phys. 164 (3) (2016) 472–494.
[56] S. Safavi, U.A. Khan, Revisiting finite-time distributed algorithms via successive nulling of eigenvalues, IEEE Signal Process. Lett. 22 (1) (2015) 54–57.
[57] A. Sandryhaila, S. Kar, J.M. Moura, Finite-time distributed consensus through graph filters, in: ICASSP, 2014, pp. 1080–1084.
[58] A. Sarlette, Adding a single state memory optimally accelerates symmetric linear maps, IEEE Trans. Automat. Control 61 (11) (2016) 3533–3538.
[59] D.A. Spielman, S.-H. Teng, A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning, SIAM J. Comput. 42 (1) (2013) 1–26.
[60] M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation, J. Amer. Statist. Assoc. 82 (398) (1987) 528–540.
[61] K.S. Turitsyn, M. Chertkov, M. Vucelja, Irreversible Monte Carlo algorithms for efficient sampling, Physica D 240 (4) (2011) 410–414.
[62] D.A. Van Dyk, X.-L. Meng, The art of data augmentation, J. Comput. Graph. Statist. 10 (1) (2001) 1–50.