Quantum mixing of Markov chains for special distributions

2015 New J. Phys. 17 073004

(http://iopscience.iop.org/1367-2630/17/7/073004)

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
5 March 2015

REVISED
3 May 2015

ACCEPTED FOR PUBLICATION
2 June 2015

PUBLISHED
3 July 2015

PAPER

Quantum mixing of Markov chains for special distributions

V Dunjko[1,2,3] and H J Briegel[1]

[1] Institute for Theoretical Physics, University of Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria
[2] Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia
[3] Author to whom any correspondence should be addressed.

E-mail: vedran.dunjko@uibk.ac.at and hans.briegel@uibk.ac.at

Keywords: quantum walks, quantum mixing, Markov chain Monte Carlo

Abstract

The preparation of the stationary distribution of irreducible, time-reversible Markov chains (MCs) is a fundamental building block in many heuristic approaches to algorithmically hard problems. It has been conjectured that quantum analogs of classical mixing processes may offer a generic quadratic speed-up in realizing such stationary distributions. Such a speed-up would also imply a speed-up of a broad family of heuristic algorithms. However, a true quadratic speed up has thus far only been demonstrated for special classes of MCs. These results often presuppose a regular structure of the underlying graph of the MC, and also a regularity in the transition probabilities. In this work, we demonstrate a true quadratic speed-up for a class of MCs where the restriction is only on the form of the stationary distribution, rather than directly on the MC structure itself. In particular, we show efficient mixing can be achieved when it is known beforehand that the distribution is monotonically decreasing relative to a known order on the state space. Following this, we show that our approach extends to a wider class of distributions, where only a fraction of the shape of the distribution is known to be monotonic. Our approach is built on the Szegedy-type quantization of transition operators.

1. Introduction

Quantum walks have, amongst other reasons, been long investigated for their capacity to speed up mixing processes—that is, speeding up the task of preparing stationary distributions of a given Markov chain (MC). Efficient mixing is a much coveted property in the context Markov Chain Monte Carlo (MCMC) approaches to many algorithmic methods for hard combinatorial problems and problems arising in statistical physics [1]. In the case of time-reversible MCs it is well-known that the bound on mixing times is tight relative to the *spectral gap* δ of the MC—both the lower and the upper bounds on the (approximate) mixing times are proportional to $1/\delta$, whereas other quantities (e.g. the allowed error of the approximation) appear only as logarithmic factors. Improvements in attaining target distributions then, in the classical case, always stem from additional constructions: e.g. by utilizing sequences of slowly evolving MCs in simulated annealing, or by using alternative MCs which mix faster toward the same distribution (e.g. graph lifting [2]). Annealing approaches, which utilize sequences of MCs, instead of just the final sequence, have also been explored in a quantum setting, where speed-ups relative to classical simulated annealing have been reported [3, 4]. However, the approaches based on annealing are not proven to help generically, and their utility is established, essentially, on a case-by-case basis.

Nonetheless, even without resorting to additional structures it is possible that here quantum mechanics may help generically. By employing quantum analogs of transition operators of MCs, speed-ups of mixing times already been proven [5, 6] in the cases where the underlying transition graph corresponds to periodic lattices and the torus. In these works, the quantum operator employed was $U_t = \exp(-\mathrm{i}D_P t)$, where D_P is the discriminant operator of the (time-reversible) transition operator P, which is equal to P itself in the cases when the stationary distribution of P is uniform. The exact definition of D_P will be given later. These results contribute towards a

© 2015 IOP Publishing Ltd and Deutsche Physikalische Gesellschaft

working conjecture that quantum transition operators may offer a generic quadratic speed-up in mixing times [5].

Alternative approaches to quantum mixing, based on the Szegedy quantum transition operator, have been already proposed by Richter[4], based on observations by Childs [5, 7]. In particular, it has been observed that so-called *hitting* algorithms, which attempt to find a particular element in the state space, by starting from the stationary distribution of a MC, and using the transition operator, may be run in reverse to realize a *mixing* algorithm. However, to our knowledge, these approaches were not pursued further due to their inefficiency. In such an approach, the mixing time has a prohibitive dependence on the probabilities occurring in the stationary distribution, which lead to an additional lower bound dependence of $\Omega(\sqrt{N})$ on the state space size N. Nonetheless, Szegedy walk operators have been successfully employed in other contexts, mostly relying on decreasing so-called hitting times of random walks [8–11]. For a recent review on quantum walks see e.g. [12].

In this work we re-evaluate the approach based on the Szegedy quantum transition operator (as outlined by Richter in [5]), and show how the lower bound state-space-size dependence of $\Omega(\sqrt{N})$ can be *exponentially* improved to $O(\log^{3/2}(N))$ in the case when additional knowledge is available on the shape of the stationary distribution. The outline of the paper is as follows: in section 2 we give the preliminaries and set up the notation. Following this, in section 3 we explain the main result for a special case of monotonically decaying distributions. In section 4 we explain when and how the result can be extended to a much wider class of distributions, elaborate why similar techniques cannot be useful in classical mixing problems, and also prove the optimality of our approach. We finish off in section 5 with a brief discussion.

2. Preliminaries

We begin by a brief recap of the basic concepts and results in discrete-time, time-homogeneous MC theory. A discrete-time, time-homogeneous MC is characterized by a transition matrix (operator) P which acts on a state space S with N states. In this work we will represent the transition operators P matrices as left-stochastic matrices, that is, a matrices with non-negative, real entries with columns summing to one. With this convention, the transition matrices act on (column) probability vectors from the left, and the entry P_{ji} denotes the probability of the transition from the state i to the state j. We note that in MC literature, the transition operators are also often represented as right-stochastic matrices, acting on row-vectors from the right. We have opted for the left-stochastic convention as in quantum mechanics, by convention, operators act on state vectors from the left.

The transition matrix P, together with an initial distribution, fully specifies a MC and we will often refer to P as the transition matrix and the MC, interchangeably. If P is irreducible (that is, P is an adjacency matrix of a strongly connected graph) and aperiodic (the greatest common divisor of the periods of all states is 1), then there exists a unique stationary distribution π, such that $P\pi = \pi$. We will represent distributions as a non-negative column vector $\pi = (\pi_i)_{i=1}^N, \pi_i \in \mathbb{R}_0^+$, such that $\sum_i \pi_i = 1$. Irreducible and aperiodic MCs mix: a sequential application of P onto any initial state in the limit yields the stationary distribution. More precisely, it holds that $\lim_{t\to\infty} P^t\sigma = \pi$, for all initial distributions σ. This property is sometimes referred to as *the fundamental theorem of MCs*.

In this work we will focus on time-reversible, irreducible and aperiodic MCs. A MC P with a stationary distribution π is said to be time-reversible if it satisfies detailed balance:

$$\pi_i P_{ji} = \pi_j P_{ij}, \ \forall \ i, j. \tag{1}$$

More generally, for an irreducible, aperiodic MC P, over the state space of size N with a stationary distribution π, we define the time-reversed MC P^* with $P^\star = M(\pi) P^T M(\pi)^{-1}$, where $M(\pi)$ is the diagonal matrix[5] $M(\pi) = \mathrm{diag}(\pi_1, \ldots, \pi_N)$. Then, P is time-reversible if $P = P^\star$. The discriminant matrix D_P is defined as $D_P = M^{1/2}(\pi) P^T M(\pi)^{-1/2}$, and it can be shown that it is always symmetric for time-reversible MCs. Since a time-reversible transition matrix P is similar to a symmetric matrix, its eigenvalues are real, and also by the Perron–Frobenius theorem they are less or equal to 1 in absolute value (value 1 being reserved for the stationary distribution which is also the +1 eigenvector).

If λ_2 denotes the second largest eigenvalue (in absolute value) then $\delta = 1 - |\lambda_2|$ is called the *spectral gap* of the MC P.

[4] The approach to quantum mixing presented here was developed before the authors were aware of the observation by Richter, and independently from [5]. However, in later stages of literature review it became apparent the basic idea behind this approach was already described in Richter's paper, effectively as a side-note in the preliminaries section.

[5] The inverse of D always exists, as stationary distributions of irreducible aperiodic MCs have non-zero support over the entire state space.

Next, the mixing time $\tau(\epsilon)$, within error ϵ, for P is defined as:

$$\tau(\epsilon) = \min\left\{ t \middle| D\left(P^t\sigma, \pi\right) \leqslant \epsilon, \forall\ \sigma \right\}, \tag{2}$$

where $D(\pi, \sigma)$ denotes the total variation distance on distributions π, σ: $D(\pi, \sigma) = 1/2 \sum_j |\pi_j - \sigma_j|$, or, more generally, the trace distance of the density matrices $\hat{\pi}, \hat{\sigma}$: $D(\hat{\pi}, \hat{\sigma}) = 1/2\ \ Tr\,[|\hat{\pi} - \hat{\sigma}|]$.

The mixing time (for the MC P, with a stationary distribution π) has a tight bound, in the time-reversible case, proven by Aldous in [13], but which we present in a more detailed form derived from [14]:

$$|\lambda_2|\ \frac{1}{\delta} \log \frac{1}{2\epsilon} \leqslant \tau(\epsilon) \leqslant \frac{1}{\delta}\left(\log \frac{1}{\pi_{\min}} + \log \frac{1}{\epsilon} \right) \tag{3}$$

for $\pi_{\min} = \min_i \pi_i$.

For more details on MCs, we refer the reader to [15].

Next, we present the basic elements of Szegedy-style approaches to quantum walks. Part of the presentation follows the approach given in [9].

Szegedy-style quantum walks can be viewed as walks over a bipartite graph, realized by duplicating the graph of the original MC, specified by the transition matrix P. The basic building block is a diffusion operator U_P which acts on two quantum registers of N states and satisfies

$$U_P\ |i\rangle_{\mathrm{I}}|0\rangle_{\mathrm{II}} = |i\rangle_{\mathrm{I}} \sum_j \sqrt{P_{ji}}\ \ |j\rangle_{\mathrm{II}}. \tag{4}$$

The criterion above does not uniquely specify a diffusion operator but any operator satisfying the equation above will serve our purpose.

It is easy to see that U_P establishes a walk step from the first copy of the original graph to the second. The operator U_P, and its adjoint are then used to construct the following operator:

$$\mathrm{ref}(A) = U_P\left(\mathbb{1}_{\mathrm{I}} \otimes Z_{\mathrm{II}}\right) U_P^\dagger, \tag{5}$$

where $Z = 2\ |0\rangle\langle 0| - \mathbb{1}$ reflects about the state $|0\rangle$. The operator ref(A) is a reflection about the subspace $A = \mathrm{span}\left(\{U_P\ |i\rangle|0\rangle\}_i\right)$, and is independent of the choice of diffusion operator satisfying equation (4). A second reflection is established by defining a second diffusion operator, realizing a walk step from the second copy of the graph back to the first: $V_P = SWAP_{\mathrm{I,II}}\, U_P\, SWAP_{\mathrm{I,II}}$. From here, we proceed analogously as in the case for the set A, to generate the operator ref(B), reflecting over $B = \mathrm{span}\left(\{V_P\ |0\rangle|j\rangle\}_j\right)$. The Szegedy walk operator is then defined as $W(P) = \mathrm{ref}(B)\mathrm{ref}(A)$. In [8, 9] it was shown that the operator $W(P)$ and P are closely related, in particular in the case P is time-reversible, which we clarify next.

Given a distribution π, we call the state[6] $|\pi\rangle = \sum_{i=0}^{N-1} \sqrt{\pi_{i+1}}\ \ |i\rangle$ the *coherent encoding of the distribution* π. For us it is convenient to define a one-step diffused version of the encoding above, specific to a particular Markov chain: $|\pi'\rangle = U_P\ |\pi\rangle_{\mathrm{I}} \otimes |0\rangle_{\mathrm{II}}$, where U_P is the Szegedy diffusion operator. It is easy to see that $|\pi\rangle$ and $|\pi'\rangle$ are trivially related via the diffusion map (more precisely, the isometry $|\pi\rangle \rightarrow U_P\ |\pi\rangle \otimes |0\rangle$) and moreover that the computational measurement of the first register of $|\pi'\rangle$ also recovers the distribution π. In slight abuse of terminology, we shall refer to both encodings as *the coherent encoding* of the distribution π, and denote them both $|\pi\rangle$, where the particular encoding will be clear from context. Next, we clarify the relationship between the classical transition operator P and the Szegedy walk operator $W(P)$. Let π be the stationary distribution of P so $P\pi = \pi$. Then the coherent encoding of the stationary distribution π of P, given with $|\pi\rangle = U_P \sum_i \sqrt{\pi_i}\ \ |i\rangle|0\rangle$, is also a + 1 eigenstate of $W(P)$, so $W(P)|\pi\rangle = |\pi\rangle$. Moreover, on the subspace $A + B$, so-called *busy subspace*, it is the unique +1 eigenstate. On the orthogonal complement of the busy subspace, $W(P)$ acts as the identity. Moreover, the spectrum of P and $W(P)$ is intimately related, and in particular the spectral gap δ is quadratically smaller than the phase gap

$$\Delta = \min\left\{ 2\ |\theta| :\ \mathrm{e}^{\mathrm{i}\theta} \in \sigma\left(W(P)\right), \theta \neq 0 \right\}, \tag{6}$$

where θ denote the arguments of the eigenvalues, i.e. eigenphases, of $W(P)$. In other words, we have that $1/\Delta \in O\left(1/\sqrt{\delta}\right)$. This relationship is at the very basis of all speedups stemming from employing the Szegedy quantum walk operator. We refer the reader to [8, 9] for further details on Szegedy-style quantum walks.

A useful central tool in the theory quantum walks employing the Szegedy walk operator is the so-called approximate reflection operator ARO $(P) \approx 2\ |\pi\rangle\langle\pi| - \mathbb{1}$, which approximately reflects over the state $|\pi\rangle$. The basic idea for the construction is as follows: By applying Kitaev's phase detection algorithm on $W(P)$ (with precision $O(\log(\Delta))$), applying a phase flip to all states with phase different from zero, and by undoing the phase

[6] Since the labels of the states also denote the rows and columns of the transition matrix, it is customary to start the enumeration from 1. In the quantum case the first state is usually denoted with a zero: $|0\rangle$. For this reason, we shift the indices of π_i by 1 to maintain consistency.

detection algorithm, we obtain an arbitrary good approximation of the reflection operator $R(P) = 2\,|\pi\rangle\langle\pi| - \mathbb{1}$, for any state within $A + B$. The errors of the approximation can again be efficiently suppressed by iteration (by the same arguments as for the $|\pi\rangle$ measurement) [9], so the cost of the ARO is in $\tilde{O}(1/\Delta) = \tilde{O}(1/\sqrt{\delta})$.

Thus, the second gadget in our toolbox is the operator $\mathrm{ARO}(P)$, which approximates a perfect reflection $R(P)$ on $A + B$, while incurring a cost of $\tilde{O}(1/\sqrt{\delta})$ calls to the walk operator $W(P)$.

The $\mathrm{ARO}(P)$, along with the capacity to flip the phase of a chosen subset of the computational basis elements, suffices for the implementation of an amplitude amplification [16] algorithm which allows us to find the chosen elements. To illustrate this, assume we are given the state $|\pi\rangle$, the (ideal) reflection with a transition matrix $R(P)$, and assume we are interested in finding some set of elements $M \subseteq \{1, \ldots, N\}$. The subset M is typically specified by an oracular access to a phase flip operator defined with $Z_M = \mathbb{1} - 2\sum_{i\in S}|i\rangle\langle i|$. The element searching then reduces to iterated applications of $Z_M R(P)$ (which can be understood as a generalized Grover iteration, more precisely amplitude amplification) onto the initial state $|\pi\rangle$. Let $\tilde{\pi}$ denote the conditional probability distribution obtained by post-selecting on elements being in M from π, so

$$\tilde{\pi} = \begin{cases} \dfrac{\pi_i}{\epsilon}, & \text{if } i \in M, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

with $\epsilon = \sum_{j\in M} \pi_j$. Let $|\tilde{\pi}\rangle = U_P \sum_i \sqrt{\tilde{\pi}_i}\;|i\rangle|0\rangle$ denote the coherent encoding of $\tilde{\pi}$. Note that the measurement of the first register of $|\tilde{\pi}\rangle$ outputs an element in M with probability 1. Thus successfully preparing this state implies that we have found a desired element from M.

As it was shown in [9], applications of Z_M, and $R(P)$ maintain the register state in the two-dimensional subspace $\mathrm{span}(\{|\pi\rangle, |\tilde{\pi}\rangle\})$, and moreover, using $\tilde{O}(1/\sqrt{\epsilon})$ applications of the two reflections will suffice to produce a state $|\psi\rangle \in \mathrm{span}\{|\pi\rangle, |\tilde{\pi}\rangle\}$, such that $|\langle\psi|\,\tilde{\pi}\rangle|^2$ is a large constant (say above 1/4). Measuring the first register of such a state will result in an element in M with a constant probability, which means that by iterating this process k times ensures an element in M is found with an exponentially increasing probability in k. However, since the state $|\psi\rangle$ is also in $\mathrm{span}\{|\pi\rangle, |\tilde{\pi}\rangle\}$, it is easy to see that the measured state, conditional on being in M, will also be distributed according to $\tilde{\pi}$. This observation was used in [17], and also in [10] to produce an element sampled from the truncated stationary distribution $\tilde{\pi}$, in time $\tilde{O}(1/\sqrt{\epsilon}) \times \tilde{O}(1/\sqrt{\delta})$ where the δ term stems from the cost of generating the $\mathrm{ARO}(P)$, and $\tilde{O}(1/\sqrt{\delta})$ corresponds to the number of iterations which have to be applied. This is a quadratic improvement relative to using classical mixing, and position checking processes which would result in the same distribution.

However, the same process can be used *in reverse* to generate the state $|\pi\rangle$ starting from some fixed basis state $|i'\rangle = U_P\,|i\rangle|0\rangle$ with cost $\tilde{O}(1/\sqrt{\delta}) \times \tilde{O}(1/\sqrt{\pi_i})$. The resulting state of the reverse process is constantly close to the state $|\pi\rangle$, which is our target state. This basic idea was already observed in [5] by Richter, however, at first glance it seems prohibitive as the resulting mixing time is proportional to $\tilde{O}(1/\sqrt{\pi_i})$. If no assumptions are made on the stationary distribution, this dependency is lower bounded by $\tilde{O}(1/\sqrt{N})$, as the smallest probability in a distribution is upper bounded by $1/N$.

For this work, we point out that the preparation process, starting from an initial basis state is trivially generalized: if $|\psi\rangle$ is any initial state, and we have the capacity to reflect over it, we can reach a state close to the target state $|\pi\rangle$, with an overall cost $\tilde{O}(1/\sqrt{\delta}) \times O(1/\sqrt{F(|\psi\rangle, |\pi\rangle)})$, where $F(|\psi\rangle, |\pi\rangle) = |\langle\psi|\,\pi\rangle|^2$ is the standard fidelity. This follows when realizing that the search/unsearch algorithms are in fact amplitude amplification [16] algorithms.

Finally we point out that having a constant-distance approximation of the stationary distribution is effectively all we need. Given an approximation constantly far from $|\pi\rangle$, an arbitrarily good approximation of distance ϵ can be achieved in time $O(1/\sqrt{\delta} \times \log(1/\epsilon))$, by again running phase estimation (iteratively) of $W(P)$ on the approximate state, and this time measuring the phase-containing register. This $|\pi\rangle$-projective measurement is described in more detail in [18], and it follows from theorem 6 in [9] .

We have previously used this idea in [18] to achieve more efficient mixings in the context of slowly evolving sequences of MCs, where the initial states were either basis states, or a state encoding the uniform distribution.

In this work, we will take this idea significantly further, by intrinsic properties of monotonically decaying distributions. Let Ω be a finite state space, and let $\leqslant_\Omega$ be a total order on Ω. Then a distribution d is monotonically decaying, relative to the order $\leqslant_\Omega$, if for its probability mass function f_d we have: $x, y \in \Omega, (x \leqslant_\Omega y) \Rightarrow (f_d(x) \geqslant f_d(y))$. In this work we will be representing the state space elements with integers, and the order will be the standard order so a distribution π is monotonically decaying if $i \leqslant j \Rightarrow \pi_i \geqslant \pi_j$, monotonically increasing if $i \leqslant j \Rightarrow \pi_i \leqslant \pi_j$. Finally, a distribution π is strongly unimodal if there exists a $k \in \Omega$ such that for $i \leqslant j \leqslant k$ we have $\pi_i \leqslant \pi_j$ and for $k \leqslant i \leqslant j$ we have $\pi_i \geqslant \pi_j$.

3. Main result

The results of the previous section already establish that if we have an (perhaps oracular) access to good approximations of the targeted stationary distribution, arbitrarily good mixing by unsearching is efficient. Here, we will show that for the case of monotonically decaying distributions we can construct good initial states efficiently, in time $O(\log(N))$, independently from the shape of the distribution. The approach is outlined as follows. We will define a particular family of N distributions (*ladder distributions*), which are the extreme points of the convex space of decaying distributions. Then, we show that a particular fixed subset containing $\log(N)$ ladder distributions, for any decaying distribution π, necessarily contains an element which is at most logarithmically far from π. This forms our theorem given below. Following this, we show all distributions in this log-sized subset are efficiently constructed, and present the overall mixing algorithm.

The key result is the following theorem:

Theorem 1. *Let $N = 2^n$, $n \in \mathbb{N}$ be an integer, and let $D_N^{\gtrless} \subseteq \mathbb{R}^N$ be the convex space of monotonically decaying distributions over $\{1, \ldots, N\}$. Then there exists a set of distributions $S \subseteq D_N^{\gtrless}$, $|S| = n$, such that for every $\pi \in D_N^{\gtrless}$, there exists $\nu \in S$ satisfying*

$$D(\pi, \nu) \leqslant 1 - \frac{1}{2(n+1)}. \tag{8}$$

Proof. We begin by constructing an N-sized set $S' = \{\sigma^i\}_{i=1}^N$ of 'ladder' distributions which are extreme points of the convex space $D_N^{\gtrless}$. They are defined as:

$$(\sigma^i)_j = \begin{cases} 1/i & \text{if } j \leqslant i, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Any distribution $\pi \in D_N^{\gtrless}$ can be represented as a convex combination of distributions in S' as follows:

$$\pi = \pi_1 \sigma^1 + \sum_{i=2}^{N} \pi_i \left(i\sigma^i - (i-1)\sigma^{i-1} \right), \tag{10}$$

as the parentheses above contain the ith Kronecker-delta distribution. The expression above can be, by reshuffling, restated as

$$\pi = \left(\sum_{i=1}^{N-1} i(\pi_i - \pi_{i+1})\sigma^i \right) + N\pi_N \sigma^N = \sum_{i=1}^{N} q_i \sigma^i, \tag{11}$$

where, for $1 \leqslant i < N$ $q_i = i(\pi_i - \pi_{i+1})$ and $q_N = N\pi_N$. Since the distribution π is monotonically decaying, we have that $q_i \geqslant 0$, and it is also easy to see that $\sum_i q_i = 1$. In other words, $q = (q_i)_i$ is a distribution as well. Using the representation above, we can express the distance between π and $\sigma^k \in S'$ as follows:

$$D\left(\pi, \sigma^k\right) = D\left(\sum_{i=1}^{N} q_i \sigma^i, \sigma^k \right) \leqslant \sum_{i=1}^{N} q_i D\left(\sigma^i, \sigma^k\right), \tag{12}$$

where the last inequality follows from the strong convexity of the trace distance. The distance between individual distributions in S' is easy to express explicitly:

$$D\left(\sigma^i, \sigma^k\right) = 1 - \frac{\min\{i, k\}}{\max\{i, k\}}. \tag{13}$$

If we define the matrix $V = (V_{ij})_{ij}$ where $V_{ij} := \dfrac{\min\{i, k\}}{\max\{i, k\}}$, we can express the bound on the trace distance between π and σ^k as:

$$D\left(\pi, \sigma^k\right) \leqslant 1 - v_k^T q, \tag{14}$$

that is 1 minus the standard inner product between the kth column of V, denoted v_k, and the probability vector q which uniquely specifies π. Next, we focus on the log-sized subset $S \subseteq S'$ given with $S = \{\sigma^i | i = 2^k, k = 0, \ldots, n-1\}$, and establish a lower bound on $\min_q \max_k (v_{2^k})^T q$ by coarse-graining. This will then yield an upper bound on the distance between an arbitrary decaying distribution π and the set S.

From the definition of the vectors v_2k, $k < n$ it is easy to see that the following holds:

$$(v_{2^k})_j \geqslant 1/2 \text{ for all } 2^k \leqslant j \leqslant 2^{k+1}. \tag{15}$$

We can see this more generally as, per definition, for v_l, $l \leqslant N/2 = 2^{n-1}$ and $l \leqslant j \leqslant 2l$ we have that $(v_l)_j = \dfrac{l}{j} \geqslant \dfrac{l}{2\,l} = 1/2$. Next, we introduce the coarse-graining operator $\mathcal{L}$, mapping real vectors over 2^n elements to real vectors over $n + 1$ elements:

$$\mathcal{L}(q) = \left(\tilde{q}_j\right)_{j=1}^{n+1} \text{ with} \tag{16}$$

$$\tilde{q}_1 = q_1, \text{ and for } j > 1,\ \tilde{q}_j = \sum_{l=2^{j-2}+1}^{2^{j-1}} q_l. \tag{17}$$

It is clear that $\mathcal{L}$ also maps distributions to coarse-grained distributions. By equation (15) we have that

$$v_{2^k} \geqslant w_{2^k}, \text{ for } k < n, \text{ with} \tag{18}$$

$$(w_{2^k})_j = \begin{cases} 1/2 & \text{if } 2^k \leqslant j \leqslant 2^{k+1}, \\ 0, & \text{otherwise,} \end{cases} \tag{19}$$

where the inequality is taken element-wise. The ancillary vectors w_{2^k} just capture the positions where the vector v_{2^k} has entries larger or equal to 1/2, setting those to 1/2 and the rest to zero. But then it follows, for $0 \leqslant k \leqslant n - 1$, that

$$v_{2^k}^T q \geqslant w_{2^k}^T q, \tag{20}$$

where the inequality is taken element-wise. Next, with Δ_k we denote the Kronecker-delta distribution with unit support only over the kth element, and the inequalities are element-wise, on the coarse-grained $n + 1$ element space. It holds that

$$\left(w_{2^0}^T\right) q \geqslant \frac{1}{2} \Delta_k \cdot \mathcal{L}(q), \text{ for } \ 1 \leqslant k \leqslant 2, \text{ and} \tag{21}$$

$$\left(w_{2^k}^T\right) q \geqslant \frac{1}{2} \Delta_{k\ +\ 2} \cdot \mathcal{L}(q), \text{ for } \ 0 < k < n. \tag{22}$$

To see the first claim, note that $(w_{2^0}^T)q \ = \ 1/2(q_1 + q_2)$, whereas $\Delta_1 \cdot \mathcal{L}(q) = q_1$ and $\Delta_2 \cdot \mathcal{L}(q) = q_2$. For the second inequality, note that the left-hand side of the inequality sums up all elements from q which lie between positions 2^k and 2^{k+1} (border points included), and multiplies it with 1/2. The right-hand side of the inequality picks out the $(k + 2)^{nd}$ element of the coarse-grained distribution $\mathcal{L}(q)$, which, by definition, sums the entries of q between the same boundaries, but not including the lower boundary. Then we have that

$$\max_{k \in \{1, \ldots, n+1\}} v_{2^{k-1}}^T q \geqslant \max_{k \in \{1, \ldots, n+1\}} w_{2^{k-1}}^T q \geqslant \frac{1}{2} \max_{k \in \{1, \ldots, n+1\}} \Delta_k \cdot \mathcal{L}(q). \tag{23}$$

Then, the target min-max expression is also lower bounded by

$$\min_q \max_{k \in \{1, \ldots, n+1\}} \frac{1}{2} \Delta_k \cdot \mathcal{L}(q), \tag{24}$$

which is easy to evaluate: $\mathcal{L}(q)$ is an arbitrary distribution over $n + 1$ elements, and we are free to optimize the overlap of this distribution with all Kronecker-delta distributions on the same space. The minimum is attained when all the overlaps are the same, so when $\mathcal{L}(q)$ is uniform over the space of $n + 1$ elements, and we have $\min_q \max_{k \in \{1, \ldots, n+1\}} 1/2 \Delta_k^T \cdot \mathcal{L}(q) = \frac{1}{2(n+1)}$. This also lower bounds our target expression $\min_q \max_k (v_{2^k})^T q$, and proves our claim. □

In the proof above we have explicitly constructed the $\log(N)$ distributions from the set S. They are the $n = \log_2(N)$ distributions $\nu^k := \sigma^{2^k}$, for $0 \leqslant k \leqslant n - 1$ which have uniform support from the first element up to element at the (2^k)th position. For them to be useful for the quantum mixing algorithm the coherent encodings of these distributions have to have an efficient construction, which is the case. Start by initializing the n-qubit register (sufficient for encoding distributions over the $N = 2^n$ state-space) in the 'all-zero' state. Then, the kth distribution is achieved by applying the Hadamard gate to the last k qubits. This realizes the state $|\nu^k\rangle = |0\rangle^{\otimes(n-k)} | + \rangle^{\otimes(k)}$, which encodes the desired distributions, and the reverse of this process, along with the reflection over the 'all zero' state realizes the reflection over $|\nu^k\rangle$ efficiently as well.

A few remarks are in order. First, although we have phrased the result for the case when the state space is a power of 2, this is without loss of generality—any decaying distribution over N elements is trivially a decaying distribution over the larger set of $\left\lceil \log_2(N) \right\rceil$ elements, where we assign zero probability to the tail of the distribution. The trace distance result remains the same, once the ceiling function is applied to the log term, hence it yields the same scaling.

Next, note that theorem 1, along with the given simple method for preparing the log-sized set of initial states, already yields an efficient algorithm for the preparation of decaying stationary distributions. To see this, assume first we know which distribution ν^k out of S minimizes the trace distance, bounded by $1 - \frac{1}{2(\log(N)+1)}$. If $|\pi\rangle$ is the coherent encoding of the target stationary distribution, by the known inequalities between the trace distance and the fidelity[7] we have that:

$$\left|\langle \pi |\, \nu^k \rangle\right|^2 \geqslant \frac{1}{(2(\log(N) + 1))^2}. \tag{25}$$

But then, by the results of section 2, we can attain the stationary distribution in time $\tilde{O}(1/\sqrt{\delta}) \times O(2(\log(N) + 1)) = \tilde{O}(1/\sqrt{\delta}) \times O(\log(N))$, where the soft-O ($\tilde{O}$) part suppresses the logarithmically contributing factor stemming from the acceptable error term. However, since we do not know which distribution minimizes the distance, the trivial solution is to sequentially run the algorithm for each one. This yields an overall $\log(N)$ factor, yielding $\tilde{O}(1/\sqrt{\delta}) \times O(\log^2(N))$ as the total complexity. We can do slightly better by encasing the entire procedure of 'searching for the correct initial distribution' in a Grover-like search algorithm, more precisely, an amplitude amplification prodecure.

To see this is possible, note that whether or not a particular distribution was the correct initial choice can be heralded—the $|\pi\rangle$-projective measurement, for instance, reveals whether we succeeded or did not. Moreover, the ARO(P) operator itself will help realize the Grover oracle, which flips the phase of all states which are not the target distribution. The overall procedure can then be given as follows:

First, we initialize the system in the state $\sum_{j=0}^{n-1} |j\rangle_{\mathrm{I}} |\psi_j\rangle_{\mathrm{II}}$, where $|\psi_j\rangle$ is the coherent encoding of the jth distribution from the set S. This has a complexity of $O(\log^2(N))$ in the state-space size, but is independent from δ. Then, in quantum parallel, we run the quantum mixing algorithm on register II (with complexity $\tilde{O}(1/\sqrt{\delta}) \times O(\log(N))$), followed by one application of the ARO(P), followed by an un-mixing (the running of the mixing algorithm in reverse). This will, approximately (and up to a global phase of -1), introduce a relative phase of -1 at those $|j\rangle$ terms, where the searching procedure yielded a success. This constructs the phase-flip operator.

The remainder is the operator which flips over the state $\sum_{j=0}^{n-1} |j\rangle_{\mathrm{I}} |\psi_j\rangle_{\mathrm{II}}$, which has a cost of $O(\log^2(N))$. Since at least one distribution, by the correctness of our mixing algorithm, yields the target distribution, this extra layer of amplitude amplification needs to be run in a randomized fashion, (since only the lower bound is known) [16], on the order of $\sqrt{\log(N)}$ times, until the correct initial distribution is found. The overall complexity is then given by $O\left(1/\sqrt{\delta} \times \log^{3/2}(N)\right) + O\left(\log^{5/2}(N)\right)$. The error factor (multiplying both additive terms) which we have for simplicity omitted, and which guarantees that the distance from the target distribution is within ϵ in the trace distance, is given with $O(\log(1/\epsilon) + \log\ \log(N))$. The additional $\log\ \log(N)$ term stems from the fact that the ARO(P) operator is applied $O(\log(N))$ many times which accumulates errors. However, since the effective total error is given by the union bound, it will suffice to rescale the target precision to $\epsilon := \epsilon/\log(N)$, which yields the log log term [9]. In practice, $1/\sqrt{\delta}$ tends to dominate $\log(N)$, thus we have the complexity $O\left(1/\sqrt{\delta} \times \log^{3/2}(N)\right)$, omitting the logarithmically contributing error terms. One of the features of our approach is that the actual output of the protocol is a particular coherent encoding of the target probability distribution. The classical probability distribution can then be recovered by a measurement of the output state. Having such a quantum output is desirable if our protocol is to be embedded in a larger algorithm where the preparation is just an initial step. Examples where this is assumed include hitting algorithms [9, 10], and algorithms which aim at sampling from a (renormalized) part of the distribution [10, 17]. We point out that this property is not a necessary feature of all quantum algorithms for mixing—there are promising approaches which utilize decoherence to speed up mixing [5], which may preclude a coherent output. The property that the output is a coherent encoding of the target distribution is also maintained in extensions of our protocol, which we describe in the next section.

4. Lower bounds and extensions

The approach we have described in the previous section trivially extends to monotonically increasing distributions as well—since the trace distance is invariant under the matching permutations of the elements of the two distributions, the same proof holds, where we use 'ladder distributions' which are reversed in the order of the probabilities. However, the approach can be further extended to strongly unimodal distributions, and

[7] Note that the total variation distance on the distributions, corresponds to the trace distance of the incoherent encodings of probability distributions, whereas we are interested in the fidelity of the coherent encodings. However, the Uhlmann fidelities of coherent and incoherent encodings are equal, so the standard bounds do apply.

beyond, if additional knowledge about the target stationary distribution is assumed. At the end of this section, we will explain how such extensions can be obtained. Before this, we will address two natural theoretical questions which arise from our approach.

First, in the previous section we have only provided an upper bound of the distance of the set S and an unknown monotonically decaying distribution. *A priori*, it is not inconceivable that, for the restricted case of monotonically decaying distributions, it may be possible that there exists a significantly better choice, with a better bound—perhaps achieving a constant distance, instead of a $\log(N)$ dependence. Here we will show that a significant improvement of our result is not possible.

Second, in our setting we have assumed a specific type of prior knowledge of the target stationary distribution. It is a fair question whether such knowledge, along with the capacity to prepare particular initial distributions, may already offer a significant speed up in the case of classical mixing. If this were the case, our result should not be considered as a true speed up of classical mixing. However, we show that the type of assumption we impose for the quantum algorithm does not help in the classical approach.

4.1. Lower bounds

The cornerstone of our result relies on the fact that there exists a $\log(N)$-sized set of distributions in $D_N^{\gtrless}$, which has an overlap (fidelity) of no less than $O(\log^{-2}(N))$ from any distribution π in $D_N^{\gtrless}$. It is a fair question whether the $\log(N)$ dependence can be dropped altogether, and be replaced by a constant, in the complexity of the mixing algorithm.

A necessary precondition for this, in the case of our approach, is the following claim:

Claim 1 There exists a constant $0 \leqslant \eta < 1$, and a family of (arbitrary) distributions $\{\mu^{(N)}\}_N$, one for each state space size N, such that for every $N \in \mathbb{N}$, and for every $\pi \in D_N^{\gtrless}$ we have that $D(\mu^{(N)}, \pi) \leqslant \eta$.

If claim 1 were to be true, and if the coherent encodings of distributions $\mu^{(N)}$ were efficiently constructable (say in time $O(\text{polylog}(N))$), then this would constitute a significant improvement over our result. To get a bit of intuition, consider a generalization of claim 1, where π^k are arbitrary distributions. In this case the claim clearly does not hold. Consider any family $\{\mu^{(N)}\}_N$. Then for $N \in \mathbb{N}$, let $\mu_{\min} = \min_j(\mu^{(N)})_j$ be the smallest probability occurring in $\nu^{(N)}$, and let $j_{\min} = \text{argmin}_j(\mu^{(N)})_j$ be the position of the smallest probability. Then we can choose π to be the Kronecker delta distribution at position $j_{\min}$ which yields the distance $1 - \mu_{\min} \geqslant 1 - 1/N$, which converges to 1 with the state space size N. In the case when π is in $D_N^{\gtrless}$, and the proof is a bit more involved, and we provide it next.

Note that claim 1 implies that each member of the family $\left\{\mu^{(N)}\right\}_N$ is, specially, within $\eta < 1$ distance from all the ladder distributions σ^k, for all N. Thus, to negate claim 1, it will suffice to show that this is not possible. In the following, we will, for convenience, use not the trace distance, but rather overlaps (square-roots of fidelities) between the coherent encodings of distributions. Then, we will show that optimal states $|\mu^{(N)}\rangle$ (those which now *maximize* the overlaps with all the corresponding ladder distribution states), have an overlap with the ladder distribution states which decays with N. By the standard inequalities connecting fidelities and trace distances, the negation of claim 1 follows.

Consider an N-dimensional setting, and let $\left\{|\sigma^k\rangle\right\}_{k=1}^N$ be the set of coherent encodings of the ladder distributions, so $|\sigma^k\rangle = \frac{1}{\sqrt{k}}\sum_{i=1}^k |i\rangle$. Let $|\mu\rangle$ be any pure quantum state which maximizes all the overlaps, thus attains the minimal overlap of $o_{\min} = \max_{|\mu\rangle} \min_k |\langle\sigma^k|\,\mu\rangle|$.

We first show that such a state, necessarily, attains the same overlap with all the ladder distribution states. To see this, first note that the vectors $\left\{|\sigma^k\rangle\right\}$ form a (non-orthogonal) basis of $\mathbb{C}^N$, hence they are linearly independent. Then, to each vector $|\sigma^k\rangle$, we can associate a normalized *reciprocal vector* $|\sigma_r^k\rangle$ [19], which is orthogonal to all other ladder distribution vectors, so $\langle\sigma_r^k|\,\sigma^{k'}\rangle = 0$, whenever $k \neq k'$, and $\langle\sigma_r^k|\,\sigma^k\rangle = \gamma_k \neq 0$.

Next, let $o_k = |\langle\sigma^k|\,\mu\rangle|$ denote the overlap between the state $|\mu\rangle$ and the kth ladder distribution state. Suppose that not all overlaps are equal. Then there exist a largest overlap (which need not be unique), and k be the index corresponding to one such overlap.

We can express $|\mu\rangle$ as

$$|\mu\rangle = \alpha\left|\sigma_r^k\right\rangle + \beta\left|\sigma_r^{k\perp}\right\rangle, \tag{26}$$

where $|\sigma_r^{k\perp}\rangle$ is a vector in the subspace orthogonal to $|\sigma_r^k\rangle$. Since $|\sigma_r^k\rangle$ is orthogonal to all other ladder states, we have that

$$o_{k'} = \left|\left\langle \mu \middle| \sigma^{k'} \right\rangle\right| = |\beta| \left|\left\langle \sigma_r^{k\perp} \sigma^{k'} \right\rangle\right|, \text{ for } k' \neq k. \tag{27}$$

Now, consider a parametrized modification of the state $|\mu\rangle$, denoted $|\mu(x)\rangle$, $x \in [1/2,\ 1]$ which, as x increases, increases the absolute value of β, at the expense of the absolute value of the α term:

$$|\mu(x)\rangle = \mathcal{N}\left((1 - x)\alpha \left|\sigma_r^k\right\rangle + x\beta \left|\sigma_r^{k\perp}\right\rangle\right), \tag{28}$$

where $\mathcal{N}$ is the normalization factor $\mathcal{N} = (|\alpha|^2(1 - x)^2 + |\beta|^2 x^2)^{-1/2}$. It is clear from equation (27) that increasing x continuously increases the overlaps of $|\mu(x)\rangle$ with all the ladder states, except perhaps with the kth ladder state. Regarding the overlap with the kth state, denoted $o_k(x)$, one of three things may happen, all of which lead to a contraction with the optimality of the initial state $|\mu\rangle$: (a) $o_k(x)$ may increase, in which case $|\mu\rangle$ was not optimal, as we increase all the overlaps; (b) $o_k(x)$ may remain unchanged, in which case, again, the minimal overlap is increased; (c) the overlap $o_k(x)$ may decrease. However, in the case c, since o_k is strictly larger than all other overlaps, and since any change in o_k is continuous in x, by decreasing this overlap by less than the difference between o_k and the second largest overlap (which is not equal to o_k) we still increase the minimal overlap. Thus $|\mu\rangle$ was not optimal, which proves that the optimal choice must have all overlaps equal.

Suppose now that $|\mu\rangle = \sum_k \mu_k\ |k\rangle$ is an optimal state. Then, all μ_k coefficients must be real and non-negative. This holds as the overlap with any of the ladder states can only increase by placing $|\mu_k|$ instead of μ_k, since all the coefficients in the ladder states are non-negative reals.

Finally, we explicitly compute the optimal overlaps. Since all the ladder states have real non-negative coefficients, and since the optimal $|\mu\rangle$ has all real non-negative coefficients, all the inner products between the ladder states and the state μ are real, non-negative, and coincide with the overlaps o_k. Then, we can express the vector of overlaps $\mathbf{o} = [o_1,\ \ldots,\ o_N]^T$ with the following matrix expression:

$$\mathbf{o} = M\left[\mu_1, \ldots, \mu_N\right]^T, \tag{29}$$

where the kth row of the $N \times N$ matrix M collects the coefficients of the kth ladder distribution vector. The kth row of M is then given with $1/\sqrt{k}\ [\underbrace{1, \ldots, 1}_{k},\ 0\ldots,\ 0]$. This is a lower-triangular matrix whose inverse M^{-1} is easy to give explicitly. The kth diagonal element of M^{-1} is given with $\sqrt{k}$, the lower sub-diagonal is given with $(M^{-1})_{k,k-1} = -\sqrt{k - 1}$ (for $k > 1$), and the other entries are zero. Next, note that since all the overlaps are equal, the vector of overlaps is proportional to the 'all ones' vector, thus $\mathbf{o} = \omega\,[1,\ \ldots,\ 1]^T$, for $\omega \in [0,\ 1]$. Thus we have

$$\omega M^{-1}[1,\ \ldots,\ 1]^T = [\mu_1,\ \ldots,\ \mu_N]^T. \tag{30}$$

The right-hand side of the expression above is a normalized vector in the Euclidean norm (since it is a quantum state), so ω, which equals the optimal overlap, is equal to the inverse of the Euclidean norm of the vector $M^{-1}[1,\ \ldots,\ 1]^T$. By computing this norm, we obtain that

$$1/(\omega^2) = \sum_{k\ =\ 1}^{N}\left(\sqrt{k} - \sqrt{k - 1}\right)^2. \tag{31}$$

To find an upper bound on ω, we need to lower bound the sum on the right-hand side above. It is easy to see that $\left(\sqrt{k} - \sqrt{k - 1}\right)\left(\sqrt{k} + \sqrt{k - 1}\right) = 1$, so $\left(\sqrt{k} - \sqrt{k - 1}\right)^2 = 1/\left(\sqrt{k} + \sqrt{k - 1}\right)^2$. Then we have

$$1/(\omega^2) = \sum_{k=1}^{N}\frac{1}{2\,k + 2\sqrt{k(k - 1)} + 1} \geqslant \sum_{k=1}^{N}\frac{1}{2\,k + 2\sqrt{k^2} + 1} = \sum_{k=1}^{N}\frac{1}{4k + 1} \geqslant 1/4\sum_{k=1}^{N}\frac{1}{k} = 1/4H_N, \tag{32}$$

where H_N denotes the Nth harmonic number. Thus we have:

$$\omega^2 \leqslant 4\frac{1}{H_N} \leqslant 4\log^{-1}(N), \tag{33}$$

where the last inequality holds as $H_N \geqslant \log(N)$. By the standard inequalities between fidelities and trace distances, this is in contradiction with claim 1 as the optimal trace distance (and fidelity) then converges to zero in the state space size N with a logarithmic rate. This also shows that our protocol is near-optimal.

4.2. Classical mixing for decaying distributions

The results of section 2 show that, in the case of quantum mixing through un-searching, the overall mixing time strongly depends on the initial state—the mixing time is proportional to the inverse of the square-root of the fidelity between the initial state and the targeted stationary distribution state. A similar statement holds when we consider the trace distances between the initial distribution (encoded by the quantum state) and the stationary distribution. This is clear as the trace distance (of classical distributions), and fidelity (of their coherent encodings) are tightly connected. Moreover, this dependence of the mixing time on the distance between the initial and target state is robust—regardless of what particular initial state we pick, the mixing time just depends on the distance.

In the classical case, intuitively it is also clear that starting from a distribution close to the target distribution, must speed up mixing. As an extreme example, if we wish to achieve mixing within ϵ, and we are given an initial state which is already within ϵ from the target, the mixing time (in the sense of the number of required applications of the walk operator) is zero. Is the improvement as robust in the classical case, as it is in the quantum scenario? Here we show that in the classical case it is not, and being close to the target distribution helps just moderately. To show this we first clarify a fact about classical mixing times, the definition of which we repeat for the benefit of the reader. The mixing time $\tau(\epsilon)$, within error ϵ, for MC P, with a stationary distribution π is defined as:

$$\tau(\epsilon) = \min \left\{ t \middle| D(P^t\sigma, \pi) \leqslant \epsilon, \forall\ \sigma \right\}, \tag{34}$$

where $D(\pi, \sigma)$ denotes the total variation distance on distributions π, σ: $D(\pi, \sigma) = 1/2 \sum_j |\pi_j - \sigma_j|$. The mixing time definition requires the state $P^t\sigma$ to be ϵ close to π for all initial states σ, that is, it looks for the worst case initial σ. By convexity, and the triangle inequality, the worst case initial state σ will be a Kronecker-delta distribution with total mass at some state space element.

We can introduce an analogous mixing time quantity, *relative mixing*, which extends the standard mixing time in that the initial state is guaranteed to be within η from the target state:

$$\tau_\eta(\epsilon) = \min \left\{ t \middle| D(P^t\sigma, \pi) \leqslant \epsilon, \forall\ \sigma\ s.\ t.\ D(\sigma, \pi) \leqslant \eta \right\}. \tag{35}$$

Now, suppose we are given a MC P, and we wish to evaluate a bound on $\tau_\eta(\epsilon)$ for this MC. In order to capture robust properties, the definition above asks for the worst case as well (as the distance requirement must hold for *all* $\sigma\ s.t.\ D(\sigma, \pi) \leqslant \eta$), so to bound the relative mixing times, we can construct the following distribution ρ:

$$\rho = (1 - \eta)\pi + \eta\sigma_{\text{worse}}, \tag{36}$$

where we choose σ_{worse} to be the worst-case initial state for the MC P if we wish to mix it within ϵ/η. Then we have:

$$\frac{1}{2}\ \|(1 - \eta)\pi + \eta\sigma_{\text{worse}} - \pi\| = \eta\frac{1}{2}\ \|\sigma_{\text{worse}} - \pi\| \leqslant \eta, \tag{37}$$

so ρ is within η distance from π, as required. Now, we are looking for an integer $t \geqslant 0$, such that:

$$\frac{1}{2}\ \|P^t\rho - \pi\| \leqslant \epsilon. \tag{38}$$

Then we have:

$$\frac{1}{2}\ \|P^t\rho - \pi\| = \frac{1}{2}\ \|(1 - \eta)P^t\pi + \eta P^t\sigma_{\text{worse}} - \pi\| = \frac{1}{2}\ \|(1 - \eta)\pi + \eta P^t\sigma_{\text{worse}} - \pi\| \tag{39}$$

$$= \eta D(P^t\sigma_{\text{worse}}, \pi), \tag{40}$$

hence, we require a t such that

$$D(P^t\sigma_{\text{worse}}, \pi) \leqslant \frac{\epsilon}{\eta}. \tag{41}$$

However, since σ_{worse} was chosen to be the worst case state for mixing within $\frac{\epsilon}{\eta}$, we have that

$$\tau_\eta(\epsilon) \geqslant \tau(\epsilon/\eta). \tag{42}$$

Thus the lower bound of the relative mixing time is just the standard mixing time, where ϵ is replaced with ϵ/η. Thus, we get the following lower bounds for relative mixing:

$$|\lambda_2|\ \frac{1}{\delta}\ \log\frac{1}{2(\epsilon/\eta)} \leqslant \tau(\epsilon/\eta) \leqslant \tau_\eta(\epsilon). \tag{43}$$

It is now clear that relative mixing has the same dependence on $1/\delta$, hence the improvement is only marginal. To make a fair comparison to the quantum mixing case we have shown, we can set $\eta = 1/2$ (for our algorithm, the trace distance is always larger than this), and see that the lower bound for classical mixing is lower bounded by $O(1/\delta\ \log(1/(4\epsilon)))$, which is essentially the same scaling as for standard mixing time.

For completeness, we point out that prior works have asked a related question, observing that the mixing time is an essentially robust quantity, independent from the setting (that is, in what context) the mixing is applied. We refer the reader to [20] for a collection of such results.

4.3. Extensions

While the main theorem we have used in our approach assumes monotonic (decaying or increasing) distributions, this can be easily extended further. For instance, assume that we know that π, the target distribution over $N = 2^n$ elements only decays to some element k, its behavior is unknown from that point on, and the total support up to element k is $p = \sum_{i=1}^{k} \pi_i$. Consider for the moment the truncated distribution $\tilde{\pi}$, obtained by setting all probabilities after k to zero and re-normalizing (by multiplying with $1/p$.) By theorem 1, we know that there exist an efficiently constructable log-sized set S of 'ladder' distributions over N, such that for at least one of them, σ, it holds that $D(\sigma, \tilde{\pi}) \leqslant 1 - (n+1)^{-1}/2$. But then we have, by the triangle inequality, homogeneity of the trace norm, and the fact that the maximal distance is unity, that:

$$D(\sigma, \pi) \leqslant pD(\sigma, \tilde{\pi}) + (1-p) = 1 - \frac{p}{2(n+1)}. \tag{44}$$

This implies that the total complexity of the mixing algorithm we have described, applied to this setting will be multiplicatively increased by p^{-1}. Note that the same reasoning will hold in the mirrored case, where we know that π is increasing from some element k, with corresponding support of p.

This simple observation already allows us to efficiently prepare target distributions whose probability mass functions are convex (decaying to some element, and increasing from that element). To see this note that either the mass of the distribution prior its minimum, or after, must be above or equal to 1/2. Thus, we can simply run the algorithm assuming both options which then yields just a constant multiplicative overhead of 4 (two runs $\times(1/2)^{-1}$).

Another extension of this observation is the case where relative to a known order, the distribution is, for a known contiguous subset (say, from element indexed with k to element indexed with $k+l$, including all in between) of the state space elements (with total weight p), decaying or increasing.

In this case as well, the mixing time only suffers a $1/p$ pre-factor. In particular, this implies that distributions which are strictly unimodal (meaning increasing to some element, and decreasing from that element) can also be efficiently prepared, provided the mode k is known. To see why this holds, note that in the strictly unimodal case, the total mass of the probability distribution either up to the kth element, or after, must be above 1/2. Then, the ladder distributions need to be constructed only up to the kth element, which again can be done with polylog(N) overhead. Unfortunately, for this case, the knowledge of the position of the mode k is neccessary—the Kronecker-delta distribution is also unimodal. For our approachm the capacity to efficiently mix to (a distribution arbitrarily close to) an arbitrary Kronecker-delta distribution would immediately imply efficient mixing for all distributions. This is beyond what we can claim.

5. Discussion

In this work, we have addressed the problem of attaining stationary distributions of MCs using a quantum approach. We have built on observations, originally made by Richter and Childs, that quantum *hitting* algorithms run *in reverse* can serve as mixing algorithms. These observations initially received little attention due to their apparent inefficiency—an *a priori* square-root scaling with the system size N. We have shown, in contrast, that in the cases when it is known beforehand that the target distribution is decaying relative to a known order on the state space, the dependency on the system size is only $\log^{3/2}(N)$. We have also shown the essential optimality of this bound for our approach. In particular, an explicit dependence on the system size is unavoidable and logarithmic. Following this, we have shown how our approach easily extends to a much wider class of distributions, including concave distributions, but also strictly unimodal distributions, where the position of the mode is known. Unfortunately, such assumptions are often not satisfied in many physics-inspired applications which require mixing of MCs. For instance, in statistical physics, the mode is often the quantity explicitly sought, when the distribution is known to be unimodal. In other uses, e.g. the computation of a permanent of the matrix, the underlying state space is not simply characterized at all, and knowing the order would already imply the solution to the problem.

Nonetheless, other applications involving MC mixing, such as artificial intelligence [17] and applications relying on bayesian inference (which often rely on MCMC) may have more instances where our approach may yield a genuine quantum speed-up. Moreover, the quantum algorithm we have provided realizes a coherent encoding of the stationary distribution, which can be used as a fully quantum subroutine, for instance in the preparation of initial states in e.g. hitting algorithms [9, 10].

Another possibility includes settings where the shape (and mode) of the target distribution is known (say Gaussian) and we are interested in learning higher moments of the distribution by mixing and sampling. For instance, all correlations of Gaussian states in quantum optics are captured by the second moments, whereas the mode and mean coincide, and reveal nothing about correlations. We leave the applications of our results for future work. From a more theoretical point of view, the results of this work highlight another difference between classical and quantum mixing, in particular, the approaches which rely on reversing hitting algorithms. In the classical mixing case, the choice of the initial state does not substantially contribute to the overall mixing efficiency. In contrast, in the quantum case, improvements in the choice of the initial state can, as we have shown, radically alter the overall performance.

While the conjecture that quantum approaches to mixing can yield a generic quadratic speed-up in all cases remains open, our approach extends the class of MCs for which such a speed up is possible. Notably, unlike in other studied cases where speed up has been shown, our assumptions lay only on the structure of the stationary distribution and the state space of the MC, rather directly on the structure (underlying digraph) of the MC itself.

Acknowledgments

The authors acknowledge the support by the Austrian Science Fund (FWF) through the SFB FoQuS F 4012, and the Templeton World Charity Foundation grant TWCF0078/AB46. VD thanks Markus Heyl and Nicolai Friis for many insightful discussions. The authors also thank an anonymous referee for pointing out a mistake in an earlier version of the proof of optimality in section 4.1 which was circumvented by a new proof.

References

[1] Newman M E J and Barkema G T 1999 *Monte Carlo Methods in Statistical Physics* (Oxford, UK: Oxford University Press)
[2] Chen F, Lovász L and Pak I 1999 Lifting Markov Chains to speed up mixing *Proc. 31st Annual ACM Symp. on Theory of Computing* pp 275–81
[3] Somma R D, Boixo S, Barnum H and Knill E 2008 Quantum simulations of classical annealing processes *Phys. Rev. Lett.* **101** 130504
[4] Wocjan P and Abeyesinghe A 2008 Speedup via quantum sampling *Phys. Rev.* A **78** 042336
[5] Richter P C 2007 Quantum speedup of classical mixing processes *Phys. Rev.* A **76** 042306
[6] Richter P C 2007 Almost uniform sampling via quantum walks *New J. Phys.* **9** 73
[7] Childs A 2004 Quantum information processing in continuous time *PhD Thesis* Massachusetts Institute of Technology
[8] Szegedy M 2004 Quantum speed-up of markov chain based algorithms *Proc. 45th Annual IEEE Symp. on Foundations of Computer Science* pp 32-41
[9] Magniez F, Nayak A, Roland J and Santha M 2011 Search via quantum walk *SIAM J. Comput.* **40** 142–64
[10] Krovi H, Magniez F, Ozols M and Roland J 2015 Quantum walks can find a marked element on any graph *Algorithmica* 1–57
[11] Magniez F, Santha M and Szegedy M 2005 Quantum algorithms for the triangle problem *Proc. 16th Annual ACM-SIAM Symp. on Discrete Algorithms SODA '05* (Philadelphia: SIAM) pp 1109–17
[12] Reitzner D, Nagaj D and Bužek V 2011 Quantum Walks *Acta Phys. Slovaca* **61** 603–725
[13] Aldous D 1982 Some inequalities for reversible Markov chains *J. London Math. Soc.* **25** 564–7
[14] Levin D A, Peres Y and Wilmer E L 2006 *Markov Chains and Mixing Times* (Providence, RI: American Mathematical Society)
[15] Norris J R 1998 *Markov Chains* (Cambridge: Cambridge University Press)
[16] Brassard G, Høyer P, Mosca M and Tapp A 2002 Quantum amplitude amplification and estimation *Quantum Computation and Quantum Information* ed S J Lomonaco Jr, vol 305 (AMS Contemporary Mathematics) pp 53–73
[17] Paparo G D, Dunjko V, Makmal A, Matrin-Delgado M A and Briegel H J 2014 Quantum speedup for active learning agents *Phys. Rev.* X **4** 031002
[18] Dunjko V and Briegel H J 2015 Quantum mixing for slowly evolving sequences of Markov chains arXiv: 1503.01334
[19] Chefles A and Barnett S 1998 A Optimum unambiguous discrimination between linearly independent symmetric states *Phys. Lett.* A **250** 223–9
[20] Aldous D, László L and Winkler P 1995 Mixing times for uniformly ergodic Markov chains *Stoch. Process. Appl.* **2** 165–85