Quantum Markov Chains and Quantized Metropolis-Hastings

Daniel Mandragona
Department of Mathematics
Texas A&M University
College Station, Texas
dandragona@tamu.edu

Abstract

Markov chain based and Monte Carlo methods are widely used in all scientific disciplines. The framework of modeling a problem as a probabilistic distribution is a robust solution in many scenarios due to its ability to circumvent potentially non-computable complexities. As such, much effort has been placed on developing novel probabilistic algorithms and improving existing ones. The heart of these algorithms involve computing an estimate distribution that is to resemble the unknown one. A popular method for this is the Metropolis Hasting's Algorithm which creates a Markov chain utilizing Monte Carlo sampling to rapidly approach a good estimate of the unknown distribution. In this paper we will be discussing a framework for the quantization of random walks outlined by [1] and then also a quantum adapted Metropolis-Hastings algorithm as constructed by [2].

I. Introduction

Monte Carlo methods are computational algorithms that rely upon random sampling in an effort to achieve information about a particular system. They are particularly helpful when the system under question is computationally intractable such as in many-body problems, and as such this class of algorithms can be seen across all science disciplines. One such important question to answer about a system is to view the system as a probability space. For example if we have a box filled with labeled particles each interacting with each other we may ask the question *what is the probability that I find particle, p_i, in region R?* As the number of particles increase this question quickly becomes computationally expensive. The Monte Carlo answer to this would be to repetitively sample the position of the i-th particle and estimate the ground-truth probability.

The Metropolis-Hastings algorithm builds a Markov chain from a sequence of random samples coming from the underlying distribution. This sequence is a random walk and under certain conditions on the Markov process will converge to a stationary distribution equivalent to that of the underlying distribution. This is wonderful when direct sampling of the underlying distribution is a difficult task. An example of such a problem is the protein folding question, i.e., given a protein sequence what 3D structure will it fold to? There are many possible 3D structures that the sequence can fold to, hence the search space is computationally intractable. Randomly sampling from this space of structures is computationally difficult as well. If instead we start with a 3D structure and randomly perturb atomic positions or bond angles, i.e., making small random moves then we can compute such changes. This would in turn give us a random walk through the search space. Metropolis-Hastings tells us that we may use this walk to obtain a good estimate of the true distribution.

In this paper we will first discuss the framework and results of Szegedy on adapting random walks to the quantum setting, then we will move on towards the classical Metropolis-Hastings algorithm, and then finally we will adapt this algorithm for the quantum setting following the work of (*Lemieux, et al., [2]*).

II. Markov Theory

Put simply, a **discrete time Markov process (chain)** is a sequence of random variables, $X_1, X_2, \ldots$, such that given a "state" on the first t-many steps, $(x_1, x_2, \ldots, x_t)$ the probability that we move to state: x_{t+1} depends only on our current state, x_t. Mathematically this is the following property:

$$\begin{aligned} P(X_{t+1} = x_{t+1} | X_1 = x_1, \ldots, X_t = x_t) \\ = P(X_{t+1} = x_{t+1} | X_t = x_t) \end{aligned} \tag{1}$$

This property is known as the **memoryless** property of Markov chains. Now looking at random walks, if we have that a random walk's transition probabilities only depend on which vertex one is at, then the random walk is memoryless and hence a Markov chain.

For the remainder of this section we mainly summarize some basis Markov chain theory as outlined by [3]. Suppose now that the state space, Ω, for the Markov chain is finite. Then given any state $x \in \Omega$ there are only a finite number of possible states that we can go to starting at x. Then let T be the matrix whose x-th row and y-th column is the probability $T_{x,y} = P(y|x)$. This is known as a transition matrix. Note that T is square. An important concept that we will be using heavily is that of the stationary distribution. A **stationary distribution**, π, in this context is a vector whose entries are non-negative and sum to 1 that is also unchanged when applied to T, that is:

$$\pi^T = \pi^T T$$

We may think of π_i as the probability of state $i \in \Omega$. Given an initial distribution vector, ν, (for example the probability to get to integer z starting at 0 in a simple random walk) we can say:

$$\nu_i = P(X_0 = i)$$

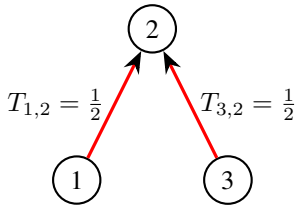
So then ν is a vector that describes a probability distribution and that:

$$(\nu^T T)_x = \sum_{y \in \Omega} \nu_y T_{y,x}$$

This can be interpreted as the total probability reaching the state x after a transition. Then we evolve the state by repetitively applying T and observe that after n-many steps that:

$$\begin{aligned} P(X_n = j) &= \sum_{i \in \Omega} P(X_0 = i)(T^n)_{i,j} \\ &= \sum_{i \in \Omega} \nu_i (T^n)_{i,j} \\ &= (\nu^T T^n)_j \end{aligned}$$

We say that a Markov chain is **irreducible** if there exists for any states $i, j \in \Omega$ a path: $(i, k_1, k_2, \ldots, k_s, j)$ which has a positive probability of occurring. This is not always guaranteed, for example:



States 1 and 3 do not have paths to each other with positive probability.

We now define another important property of Markov chains. We say that the **period** of a state, i, is the GCD of the following set: $C_i := \{n \in \mathbf{N} : (T^n)_{i,i} > 0\}$. Conceptually the set (for which we will take the GCD over) contains all steps, n, such that the probability to start at i and come back to i is positive for a path of n-many steps. The period of i is then the GCD of all such times.

Example 1. *Consider the simple random walk on the integers. If you start at $z \in \mathbb{Z}$, then you can only get back to z with paths of even length. It then follows that the period of all integers under a simple random walk is just 2 since $2 \in C_i$ and 2 must divide any other cycle length.*

A few observations can be made, first being that an irreducible Markov chain features all states having the same period. This follows from the fact that if there is a path from state i to j with positive probability and one from j to i with positive probability then they must have the same period.

Proof. Let m be the path length of the positive probability path from i to j and n for the path from j to i. Then let d_i be the period of i (similarly for d_j). We have that $m + n$ is the path length of a path from i to i. Hence d_i divides $m + n$. Let $s \in \{n \in \mathbf{N} : (T^n)_{j,j}\}$. Then $n + m + s$ is again a path from i to i. Hence we have that:

$$d_i | (m + n + s) \implies d_i | s$$

Since $\frac{m+n+s}{d_i} = \frac{m+n}{d_i} + \frac{s}{d_i}$ are all integers. Thus, by definition of GCD we have $d_j \geq d_i$. Reversing the roles gives us that $d_i \geq d_j$. Hence they must be equal. □

So now we have that an irreducible Markov chain can be described by a single period. We say that a Markov chain is aperiodic if all of its periods are 1. An irreducible, aperiodic Markov chain is called **Ergodic**. Ergodic Markov chains are special because they have well-defined long term behavior. The importance of aperiodicity in this conclusion is that when a random walk is aperiodic then there must be some amount of steps, N, such that for any $t \geq N$ we have that there is a cycle for any state in Ω of length exactly t. This follows from the fact that given any two coprime integers, then there is a largest number that can't be formed by their positive span. This is proved in Appendix [A]. We now arrive at an important result:

Theorem 1. *An Ergodic Markov chain on a finite state space has a stationary distribution, π^T, and given any initial distribution, ν, we have that:*

$$\lim_{n \to \infty} \nu^T T^n = \pi^T \tag{2}$$

Proof. We first observe that for all $n \in \mathbf{N}$ that $\nu^T T^n$ is a probability distribution. This is clear though since:

$$\begin{aligned} \sum_{j \in \Omega} (\nu^T T^N)_j &= \sum_{j \in \Omega} \sum_{i \in \Omega} \nu(i)(T^N)_{i,j} \\ &= \sum_{i \in \Omega} \nu(i) \sum_{j \in \Omega} (T^N)_{i,j} \\ &= \sum_{i \in \Omega} \nu(i) \cdot 1 \\ &= 1 \end{aligned}$$

Thus, we have that the set of probability distributions on Ω, denoted by $\mathcal{P}$, is closed under the action of T. It is also easy to see that the set of probability distributions on Ω is the intersection of the closed upper right quadrant of R^Ω (probability distributions must be nonnegative) with the hyperplane $x_1 + \cdots + x_{|\Omega|} = 1$ (the entries must sum to 1). Since Hyperplanes are closed and bounded we have that $\mathcal{P}$ is also closed and bounded and hence compact by Heine-Borel. Therefore the sequence of probability distributions given to us by (2) must have a convergent subsequence. Using *Cesaro Averages* we can arrive at the fact that tail-behavior of the sequence does not depend on ν, and so π^T always exists. □

III. Quantum Random Walk Framework

We now introduce the Szegedy operator which quantizes a classical random walk as outlined in [1]. We then show that the phase gap for this operator is quadratically larger than its classical counterpart. This shows that we will experience a quadratic speed-up in regards to the mixing (meaning convergence to stationary distribution) of our Markov process from a well-known result.

Let T be the transition matrix for a classical random walk. The quantization process by Szegedy involves constructing a

Bipartite random walk, in our case the two bipartite states are both Ω i.e., $X = \Omega$ and $Y = \Omega$, then let $P_{x,y} = T_{x,y}$ and let $Q_{y,x} = T_{y,x}$. This gives us the basis for its Hilbert space:

$$\{|x\rangle |y\rangle : x \in X, y \in Y\}$$

Define the following operators (which may be thought of as column vectors):

$$\phi_x = \sum_{y \in Y} \sqrt{P_{x,y}} |x\rangle |y\rangle \quad \psi_y = \sum_{x \in X} \sqrt{Q_{y,x}} |x\rangle |y\rangle \tag{3}$$

Now let A be all such columns (ϕ_x) and B be the columns (ψ_y). The Szegedy's walk operator, W, is defined as:

$$\begin{aligned} W &= \text{ref}_B \text{ref}_A \\ \text{ref}_A &= 2AA^* - I \\ \text{ref}_B &= 2BB^* - I \end{aligned} \tag{4}$$

We shall show how this looks for the Ergodic Markov chain in Fig [1]. Note that since A is an operator on $\Omega \otimes \Omega$ that for example the fourth column of A:

$$A^4 = \phi_4 = \frac{\sqrt{3}}{2} |4\rangle |3\rangle + \frac{1}{2} |4\rangle |5\rangle$$

is a vector of many components. This vector has 5^2 many slots. As a vector this is expressed as $3 \cdot 5$ many 0's followed by $(0, 0, \frac{\sqrt{3}}{2}, 0, \frac{1}{2})$ followed by 5 more zeroes, hence $A \in \mathbf{C}^{5^2 \times 5}$. We also see that $A^*A = I$ since $(A^*A)_{i,j} = \phi_i \cdot \phi_j$ and by definition of ϕ_i will be zero for all for all coordinates not corresponding to transitions from state i, thus $(A^*A)_{i,j} = \phi_i \cdot \phi_j = \delta_i^j \|\phi_i\| = \delta_i^j$. So A^*A is in fact the identity on $\Omega \otimes \Omega$. Similarly so is B^*B.

From this we get the important description of what ref_A is. We see that $(2AA^* - I)A = 2AA^*A - A = A$. Now let $v \in \mathcal{C}(A)^\perp$ (v is orthogonal to the column space of A), then $(2AA^* - I)v = 2AA^*v - v = 0 - v = -v$ since the column space A is equal to the row space of A and hence equal to the column space of A^* (and consequently v is orthogonal to the column space of A^*). So in total we get the following equations:

$$(2AA^* - I)A = A \tag{5}$$

$$(2AA^* - I)v = -v \qquad \text{for } v \in \mathcal{C}(A)^\perp \tag{6}$$

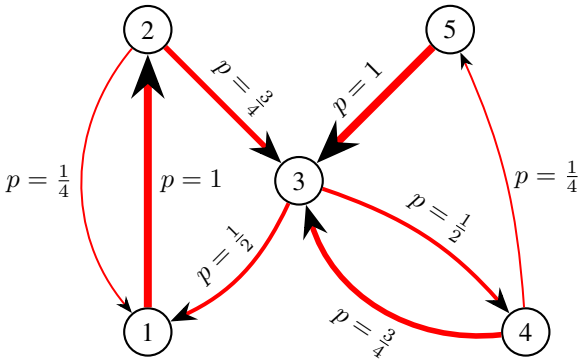


Fig. 1. An Ergodic Markov chain. Every node can be reached from any node, and for node 4 there is a path back to 4 of length 2 and 3, so this graph is aperiodic.

We can thus view ref_A and ref_B as *reflections* over the column spaces of A and B (hence the name "*ref*"). It then follows that as we can write $v \in \Omega \otimes \Omega$ as $v = c + o$ where $c \in \mathcal{C}(A)$ and $o \in \mathcal{C}(A)^\perp$, then $\text{ref}_A v = c - o$ and is thus unitary. Finally, we define $D(A, B) = A^*B$ to be the *discriminant* matrix of the quantized random walk operator W. It is easy to see that $D(A, B)_{x,y} = \phi_x \cdot \psi_y = \sqrt{p_{x,y}}\sqrt{q_{y,x}}$. In the case of a classical random walk with a symmetric transition matrix we will have that $q_{y,x} = p_{y,x} = p_{x,y}$, and so $D(A, B)$ will just be the transition matrix for this walk.

IV. Spectral Analysis of Szegedy Walk Operator

We now attempt to characterize the eigenvalues of W as this will be tied to many important properties of the random walk such as the rate at which it mixes (converges to its stationary distribution).

Given a vector $v \in \Omega \otimes \Omega$ then $D(A, B)$ can be viewed as a map from $\mathcal{C}(B) \longrightarrow \mathcal{C}(A)$ since $D(A, B)v = A^*(Bv)$, i.e. it takes a vector in $\mathcal{C}(B)$ of the form Bv and sends it to the column space of A^* which equals the column space of A. Now we have that $A \cdot D(A, B)$ is an orthogonal projector from the column space of B to A since AA^* is an orthogonal projector for column space of A. Mathematically, we see that $(\phi_i)_{i=1}^{|\Omega|}$ form an orthonormal basis for $\mathcal{C}(A) \subset \Omega \otimes \Omega$ and that:

$$\begin{aligned} (AA^*)v &= A\big(\langle \phi_1, v\rangle, \langle \phi_2, v\rangle, \ldots, \langle u_{|\Omega|}, v\rangle\big)^T \\ &= \Big(\sum_{i=1}^{|\Omega|} \phi_i(j)\langle \phi_i, v\rangle\Big)_{j=1}^{|\Omega|^2} \\ &= \sum_{i=1}^{|\Omega|} \langle \phi_i, v\rangle \phi_i \end{aligned}$$

Hence, AA^* projects v onto $\mathcal{C}(A)$. There $AD(A, B)$ is an orthogonal projection from $\mathcal{C}(B) \longrightarrow \mathcal{C}(A)$ and similarly for $D(A, B)^*$. We now let λ be a singular value of $D(A, B)$ with right singular unit-vector v and left singular unit-vector w. This gives us:

$$A^*Bv = \lambda w \tag{7}$$

$$B^*Aw = \lambda v \tag{8}$$

We now observe that $\|Bv\| = \|v\|$. To see why it is best illustrated via concrete example. Let $|\Omega| = 3$. Then take an ordering on the components of a vector in $\Omega \times \Omega$ as:

$$\begin{aligned} v = \big(&v_1 |11\rangle, v_2 |12\rangle, v_3 |13\rangle, \\ &v_4 |21\rangle, v_5 |22\rangle, v_6 |23\rangle, \\ &v_7 |31\rangle, v_8 |32\rangle, v_9 |33\rangle \big) \end{aligned}$$

Then $B = (\psi_1, \psi_2, \psi_3)$ where ψ_y are column vectors from (2). Note that ψ_y is still written in the order of $|x\rangle |y\rangle$, and so for example:

$$\begin{aligned} \psi_1 = \big(&\sqrt{q_{1,1}}, 0, 0, \\ &\sqrt{q_{1,2}}, 0, 0, \\ &\sqrt{q_{1,3}}, 0, 0\big) \end{aligned} \qquad \begin{aligned} \psi_2 = \big(&0, \sqrt{q_{2,1}}, 0, \\ &0, \sqrt{q_{2,2}}, 0, \\ &0, \sqrt{q_{2,3}}, 0\big) \end{aligned}$$

Then the matrix B acts on $v \in \Omega$ not $v \in \Omega \times \Omega$. We can think of v as being a probability distribution on Ω of all states $y \in \Omega$. Now we see that $(Bv)_{|x\rangle|y\rangle} = \sqrt{q_{y,x}} v_y$. So finally,

$$\begin{aligned} \|Bv\|^2 &= \sum_{x,y \in \Omega} q_{y,x} v_y^2 \\ &= \sum_{y \in \Omega} v_y^2 \sum_{x \in \Omega} q_{y,x} \\ &= \sum_{y \in \Omega} v_y^2 \cdot 1 \\ &= \|v\|^2 \end{aligned}$$

So we see that $\|Bv\| = \|v\|$ and similarly $\|Aw\| = \|w\|$. We also observe that projections do not increase length of vectors since at most the component of a vector w.r.t. a subspace can be at mos the vector itself (corresponding to that vector belonging to the subspace that we are projecting to). Thus, we get the following result by combining this with (6) and (7) and since v, w are both of unit norm:

Lemma 2. *All singular values of $D(A, B)$ are at most one.*

Now since singular values (for real-valued matrices) are always non-negative, we may write them all as $\cos(\theta) = \lambda$ where $0 \leq \theta \leq \frac{\pi}{2}$. The angle has important meaning because:

$$\begin{aligned} \langle Aw, Bv \rangle &= w^* A^* B v \\ &= w^* \lambda w \\ &= \lambda \|w\| \\ &= \cos(\theta) \end{aligned}$$

We now prove the main result of this section.

Theorem 3. *Let $\cos(\theta_1), \cos(\theta_2), \ldots, \cos(\theta_\ell)$ be the singular values of $D(A, B)$ that are in the open interval $(0, 1)$ with associated singular unit vector pairs (v_k, w_k) for $1 \leq k \leq \ell$. Then the eigenvalues of $W = ref_A ref_B$ with non-zero imaginary parts are:*

$$e^{-i2\theta_1}, e^{i2\theta_1}, e^{-i2\theta_2}, e^{i2\theta_2}, \ldots, e^{-i2\theta_\ell}, e^{i2\theta_\ell} \tag{9}$$

Proof. We wish to characterize the eigenvalues of W. We can do this by using Rank-Nullity applied to the subspace $\mathcal{C}(A) \cap \mathcal{C}(B)$. We can write a vector $v \in \mathcal{H}$ as:

$$v = v_{A,B} + v_{A^\perp,B} + v_{A,B^\perp} + v_{A^\perp,B^\perp} \tag{10}$$

We also observe that $AD(A, B)$ can be viewed in a way as the orthogonal projection of $\mathcal{C}(B) \longrightarrow \mathcal{C}(A)$ and vice-versa for $BD(A, B)^*$. We can gain insight then by looking at the singular values for $D(A, B)$ since this will help us understand how W acts on the overlap of $\mathcal{C}(A) \cap \mathcal{C}(B)^\perp$ and $\mathcal{C}(A)^\perp \cap \mathcal{C}(B)$. It is also understood that the singular values of $D(A, B)$ form a complete system, therefore $AD(A, B)$ gives us back all of $\mathcal{C}(A)$. Finally, putting everything together we will be able to see how W acts on a vector of $\mathcal{H}$ since we can always decompose v into its projected components in (9).

Let $\pi_A = AA^*$ and $\pi_B = BB^*$ be the orthogonal projectors. Then for our singular value vector pairs we have that:

$$\pi_A B v = \cos(\theta) A w, \quad \pi_B A w = \cos(\theta) B v \tag{11}$$

Therefore we see that:

$$\begin{aligned} W(Bv) &= (2\pi_B - I)(2\pi_A - I) Bv \\ &= (2\pi_B - I)(2\lambda Aw - Bv) \\ &= 4\lambda^2 Bv - 2Bv - 2\lambda Aw + Bv \\ &= (4\lambda^2 - 1) Bv - 2\lambda Aw \end{aligned} \tag{12}$$

Similarly,

$$\begin{aligned} W(Aw) &= (2\pi_B - I)(2\pi_A - I) Aw \\ &= (2\pi_B - I) Aw \qquad (13) \\ &= 2\lambda Bv - Aw \qquad (14) \end{aligned}$$

These two vectors belong to the subspace generated by $X = \text{span}\{Aw, Bv\}$, but since they are not multiples of each other we get that they generate the same subspace. Hence W acting on this subspace is invariant.

Now if X is two dimensional then W is reflecting this subspace along the two axes defined by Aw and Bv, and it is well known that this will result in a single reflection that is twice the angle between Aw and Bv. Then an eigenvalue for W that corresponds to a non-trivial reflection here would give us $e^{\pm i2\theta}$.

Putting this all together we know that for a vector $v \in \mathcal{C}(A)^\perp \cap \mathcal{C}(B)$ we get that:

$$\begin{aligned} Wv &= \text{ref}_B \text{ref}_A v \qquad (15) \\ &= \text{ref}_B (2AA^* - I) v \qquad (16) \\ &= (2BB^* - I)(-v) \qquad (17) \\ &= -v \qquad (18) \end{aligned}$$

Similarly, for $v \in \mathcal{C}(A) \cap \mathcal{C}(B)^\perp$ we have:

$$\begin{aligned} Wv &= \text{ref}_B \text{ref}_A v \qquad (19) \\ &= \text{ref}_B (2AA^* - I) v \qquad (20) \\ &= (2BB^* - I)(v) \qquad (21) \\ &= -v \qquad (22) \end{aligned}$$

Therefore to summarize we get:

- Vectors in $\mathcal{C}(A) \cap \mathcal{C}(B)$ have singular value 1 (W is the identity).
- Vectors in $\mathcal{C}(A)^\perp \cap \mathcal{C}(B)$ and $\mathcal{C}(A) \cap \mathcal{C}(B)^\perp$ correspond to singular value -1 since W acts as the negative identity here.
- Vectors in $\mathcal{C}(A)^\perp \cap \mathcal{C}(B)^\perp$ are singular value 1 (W is again the identity).
- Eigenvectors elsewhere may be decomposed by (9) and are thus non-trivial reflections corresponding to eigenvalues: $e^{\pm i 2\theta_\ell}$ where θ_ℓ is an angle corresponding to a singular value of $D(A, B)$.

□

Theorem 4. *Eigenvectors of W can be expressed as:*

$$Aw - e^{\pm i \arccos \lambda} Bv \tag{23}$$

Where λ is a singular value of $D(A, B)$.

Proof. Let $v = Aw - \mu Bv$ for some $\mu \in \mathbf{C}$. We will use μ to force v to be an eigenvector of W. Computing Wv we get:

$$\begin{aligned} Wv &= (2\pi_B - I)(2\pi_A - I)(v) \\ &= (2\pi_B - I)(Aw - \mu 2\lambda Aw + \mu Bv) \\ &= 2\lambda Bv - Aw - 4\mu\lambda^2 Bv + 2\mu\lambda Aw + \mu Bv \\ &= (2\mu\lambda - 1)Aw + (2\lambda - 4\mu\lambda^2 + \mu)Bv \end{aligned}$$

Therefore for this to be an eigenvector we will need:

$$(2\mu\lambda - 1)(-\mu) = (2\lambda - 4\mu\lambda^2 + \mu)$$

Hence we get:

$$\mu^2 - 2\lambda\mu + 1 = 0$$

Therefore we get $\mu = \lambda \pm i\sqrt{1 - \lambda^2} = e^{\pm i \arccos \lambda}$ (view this as a triangle with hypotenuse of length 1 and adjacent side of length λ) which is the coefficient we wanted.

Finally, we note that as the set of singular vectors forms a spanning set for $\Omega \otimes \Omega$ we get that every vector can be expressed as a multiple of $Aw - uBv$ where $u \in \mathbf{C}$, hence all eigenvectors must be of the form we attained. □

Theorem 5. *The phase gap of W scales as the square root of the spectral gap of $D(A, B)$ for small gaps.*

Proof. The previous theorem tells us that the phase gap of W will be given by $\arccos \lambda$ for when λ is close to 1 (the singular vector pair will be 1-dimensional). We also know that λ will be the second largest eigenvalue of $D(A, B)$. Hence, if we let Δ be the spectral gap of $D(A, B)$ we get that:

$$\arccos(\lambda) = \arccos(1 - \Delta)$$

Now it is easy to see that near zero (from the right) $\arccos(1 - x) \approx \frac{1}{\sqrt{2}}\sqrt{x}$ (see appendix [B]). Hence, we have that as the gap $\Delta \longrightarrow 0$ that the gap of W scales on the order of $\sqrt{\Delta}$. □

We care so much about the phase gap here because as we repetitively apply W we are able to separate the eigenvector corresponding to eigenvalue 1 from the eigenvalue-vector pair closest to it. The stationary distribution for which the classical markov chain must converge to is given by the eigenvector with eigenvalue 1, hence why we wish to distinguish it.

V. Metropolis-Hastings

A. Overview

We first introduce the classical Metropolis-Hastings algorithm. The insight for this algorithm can be seen by introducing it for an Ising model system as done in [4]. Given a graph of 100 particles with spin (each with two different configurations, $+1$ or -1) there are 2^{100} different possible arrangements, it is therefore intractable to perform global computations such as finding the mean. The solution provided by the algorithm is to instead create a Markov chain to sample from the distribution. We can just start with a random configuration of the 100 particles and then at each step of the walk randomly flip spins of certain chosen particles in our configuration. Then the Metropolis-Hastings approach would be then to calculate the energy of this new configuration and accept it as the next state of our Markov chain if the energy is lower. If the energy is instead higher than the current state then we accept the higher energy state as our new state with probability that decays exponentially as the energy gap increases.

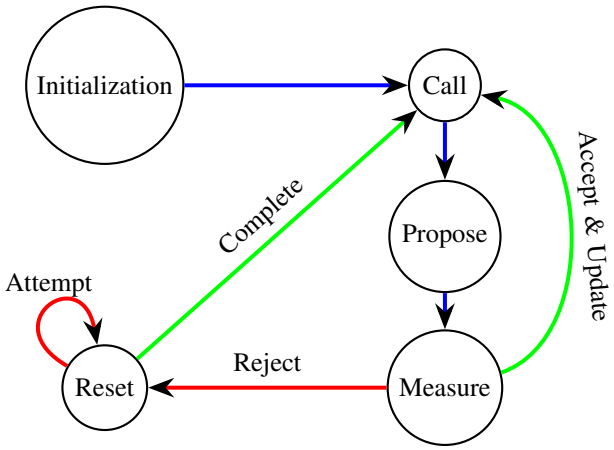


Fig. 2. State Diagram for Quantum Metropolis-Hastings

B. Quantization

We must adapt this algorithm to be a unitary operator that we may repetitively apply. This will be done in five steps as outlined in [4] which consist of an initialization phase and then four algorithmic phases. A key role is played in this algorithm by quantum phase estimation. This is an algorithm originally introduced by [5] which will estimate a given eigenvector's eigenvalue (to some fixed number r bits of precision) eigenvalue, i.e. if ϕ_i is an eigenvector with eigenvalue E_i then we have a means of implementing the following transformation:

$$|\phi_i\rangle |0\rangle \longrightarrow |\phi_i\rangle |E_i\rangle$$

Step 0: Initialization: We will need four quantum registers to make each step of the algorithm reversible. The first register will store our current state in the Markov chain, the second register will store the energy of this state up to some fixed precision ($r-$bits), the third register will store the energy of the proposed new state (also up to $r-$bits of precision) which will be achieved by the quantum phase estimate algorithm, and finally the fourth register will store whether we accept or reject the new state. We use [6] to find a random eigenvector, ϕ_i, for our Hamiltonian and measure its eigenvalue (to some r-bit precision) to achieve the state:

$$|\phi_i\rangle |E_i\rangle |0\rangle^r |0\rangle$$

which we pass to step (1).

Step 1: Function Call: We reset the upper two registers and proceed to step (2).

Step 2: Propose New State: We must have a set of unitary operations $\mathcal{C} = \{C_1, \ldots, C_n\}$ which correspond to changes to our current state. In the Ising spin system this would be a set of 100 different X gates. The only real requirements on this set of operations is that we need all possible configurations of the system to be achievable by the set and to be closed under the Hermitian. We may also allow selection from this set to be non-uniform in which case we also require that the probability of choosing $C \in \mathcal{C}$ is the same as choosing $C^\dagger$.

$C \in \mathcal{C}$ is then selected and applied to the first register which gives us the superposition of eigenstates:

$$C |\phi_i\rangle \longrightarrow \sum_k x_k |\phi_k\rangle$$

We then apply quantum phase estimation in the third register to achieve the total state:

$$\sum_k x_k |\phi_k\rangle |E_i\rangle |E_k\rangle |0\rangle$$

We now calculate the acceptance probability that we described in the overview. This will be:

$$a_k = \min(1, e^{-\beta(E_k - E_i)})$$

where β is the inverse temperature of the Boltzmann distribution. We then apply the following unitary transformation onto the fourth register (which takes as input the two energy registers):

$$W(E_i, E_k) = \begin{pmatrix} \sqrt{1-a_k} & \sqrt{a_k} \\ \sqrt{a_k} & -\sqrt{1-a_k} \end{pmatrix}$$

This gives us the following state:

$$\sum_k x_k \sqrt{1-a_k} |\phi_k\rangle |E_i\rangle |E_k\rangle |0\rangle + \sum_k x_k \sqrt{a_k} |\phi_k\rangle |E_i\rangle |E_k\rangle |1\rangle \quad (24)$$

The importance of this state will be seen in the following step.

Step 3: Measurement: We now measure the fourth register of (24) which is a single bit, and we associate measuring 1 as accepting the move. In the case of observing 1 in the fourth register we then measure the eigenvalue in the third register which collapses us to:

$$|\phi_k\rangle |E_i\rangle |E_k\rangle |1\rangle$$

The significance is that we moved to this exact state with probability $a_k \|x_k\|^2$ which is exactly the probability of transitioning to the eigenstate ϕ_k from ϕ_i. Thus, we feed the collapsed state back to step (1) which completes the iteration of the algorithm.

On the other hand things become more difficult if we measure the fourth register and find 0 then we are in a reject state and must try to undo the change. Normally, this would be simple since we have applied three unitary operators, but unfortunately we have just measured the fourth register which is not a reversible operation. Surprisingly, however, we still have a means of generating an eigenstate with the same energy as the starting state. Denote U as the unitary operator we get by applying the sequence of operators from step (2), i.e. the specific $C \in \mathcal{C}$ that we used, phase estimation, and then W. Given that we are in the reject state we now apply $U^\dagger$ and pass to step (4).

Step 4: Reset Upon Rejection: We are able prepare a state with the same energy as our initial eigenstate (which we will prove shortly). To do this we must first construct the following measurement projectors:

$$P_0 = \sum_i \sum_{E_\alpha \neq E_i} |\phi_\alpha\rangle \langle\phi_\alpha| \otimes |E_i\rangle \langle E_i| \otimes \mathbb{I} \otimes \mathbb{I}$$
$$P_1 = \sum_i \sum_{E_\alpha = E_i} |\phi_\alpha\rangle \langle\phi_\alpha| \otimes |E_i\rangle \langle E_i| \otimes \mathbb{I} \otimes \mathbb{I}$$

Here, P_1 acts on the first two registers and collapses us to a state: $|\phi_\alpha\rangle |E_\alpha\rangle$ where ϕ_α is an eigenstate with the same energy as our initial eigenstate. It is clear that $P_0 + P_1 = \mathbb{I}$.

We also have the projectors:

$$Q_0 = U^\dagger \mathbb{I} \otimes \mathbb{I} \otimes \mathbb{I} \otimes |0\rangle \langle 0| U$$
$$Q_1 = U^\dagger \mathbb{I} \otimes \mathbb{I} \otimes \mathbb{I} \otimes |1\rangle \langle 1| U$$

Q_1 measures whether the fourth register bit is 1. It is clear then that Q_1 splits the Hilbert space in half, and so the rank of Q_1 is half of the dimension of $\mathcal{H}$. We also denote $\text{rank}(P_1) = p$, which we may assume to be small (since we will like to bound the probability of obtaining P_1).

Lemma 6 (Jordan 1875)**.** *Let P_1 and Q_1 be projectors on a Hilbert space $\mathbb{C}^n$ such that rank$(P_1) = p$ and rank$(Q_1) = q$ which satisfy $p \leq q$ and $p + q \leq n$. There there is a basis for*

$\mathcal{H}$ such that P_1 and Q_1 can be written as block matrices:

$$P_1 = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$$

$$Q_1 = \begin{pmatrix} D_p & \sqrt{D_p(I_p - D_p)} & 0 & 0 \\ \sqrt{D_p(I_p - D_p)} & I_p - D_p & 0 & 0 \\ 0 & 0 & I_{q-p} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where D_p is a $p \times p$ diagonal matrix with real entries in $[0, 1]$. Note that since $q \geq p$ the $q \times q$ block in Q_1 can be written in terms of the first p-many dimensions of q as it is in our expression for Q_1.

Clearly by our assumptions stated before the lemma about the ranks of our projectors, we may apply this lemma to our measurement operators P_1 and Q_1. Since these are projectors whose pairs must sum to the identity we get that $P_0 = \mathbb{I} - P_1$, and $Q_0 = \mathbb{I} - Q_1$. We shall now combine all of this to show that the probability of failing to measure P_1 exponentially decays in the number of repetitive applications of Q and then P.

We now outline a recursive process, referred to as $\mathcal{M}$, that terminates upon measuring P_1.

- Measure P_1, and terminate if P_1 is observed.
- Apply Q_0 and Q_1 if P_1 isn't observed.
- Repeat.

We need to show that the probability to measure P_1 exponentially grows, to do this we shall calculate the probability of failing to observe P_1 directly after a sequence of applications of $\mathcal{M}$. Let n be the number of times we applied $\mathcal{M}$. Then we may have observed Q_1 m-many times where $0 \leq m \leq n$, hence all the various ways we could have failed to ever observe P_1 given n many repetitions of $\mathcal{M}$ is given by:

$$\sum_{m=0}^{n} \binom{n}{m} (P_0 Q_0 P_0)^{n-m} (P_0 Q_1 P_0)^m$$

Note that we can indeed group the occurrences of $P_0 Q_0 P_0$ and $P_0 Q_1 P_0$ in this manner because they commute with each other. Given that these are measurements, we now calculate the probability to fail given that we start with a state initially in step (2) that is $|\phi_i\rangle |E_i\rangle |0\rangle^r |0\rangle$. Then by Born's Rule:

$$p_i^{fail}(n) = tr\Big(|\phi_i\rangle |E_i\rangle |0\rangle^{2r+1} \sum_{m=0}^{n} \binom{n}{m} (P_0 Q_0 P_0)^{n-m} (P_0 Q_1 P_0)^m P_0 Q_0 \Big)$$

By the binomial theorem and the fact that our initial state is a pure state we may express this as:

$$p_i^{fail}(n) = \langle \phi_i | \langle E_i | \langle 0|^{2r+1} |Q_0 P_0 \big(P_0(Q_0 P_0 Q_0 + Q_1 P_0 Q_1)P_0\big)^n P_0 Q_0 |\phi_i\rangle |E_i\rangle |0\rangle^{2r+1} \tag{25}$$

We now employ Lemma (6) to obtain a basis change, B, that brings us to a basis where we may express P_1 and Q_1 in the form given by Lemma (6). Then our formula for the failure probability becomes:

$$p_i^{fail}(n) = \langle \phi_i | \langle E_i | \langle 0|^{2r+1} B^\dagger D_{fail}(n) B |\phi_i\rangle |E_i\rangle |0\rangle^{2r+1} \tag{26}$$

where $D_{fail}(n)$ is given below by (28):

The calculation for why $D_{fail}(n)$ is exactly this is given in Appendix [C].

We wish to see what happens to our initial vector when $D_{fail}(n)B$ acts on it. We know that our original state is in the P_1 subspace, and hence when expressed in the B-basis, $D_{fail}(n)$ acts on $B \langle \phi_i | \langle E_i | \langle 0|^{2r+1}$ via the upper left block only, and since this is an eigenvector of P_1 we get that D acting upon it results in a scaling by some $d \in \text{diag}(D)$ (note that $d \in [0, 1]$ by the lemma), and hence there is some $d^* \in \text{diag}(D)$ that corresponds to the largest entry of the upper left block of $D_{fail}(n)$. Thus,

$$p_i^{fail}(n) \leq d^*(1 - d^*)\big(d^{*^2} + d^*(1 - d^*)^2\big)^n \tag{27}$$

Clearly, due to the factor of n we have that this probability exponentially decays to 0 as $n \to \infty$. Hence we will eventually obtain a measurement of P_1, and have some state that has the same energy as our initial state, and so we can pass this to state to step (1).

$$D_{fail}(n) = \begin{pmatrix} D(\mathbb{I} - D)\big(D^2 + (\mathbb{I} - D)^2\big)^n & -\sqrt{D - (\mathbb{I} - D)}(D^2 + (\mathbb{I} - D)^2\big)^n & 0 & 0 \\ -\sqrt{D - (\mathbb{I} - D)}(D^2 + (\mathbb{I} - D)^2\big)^n & D^2\big(D^2 + (\mathbb{I} - D)^2\big)^n & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{28}$$

REFERENCES

[1] M. Szegedy, "Quantum speed-up of markov chain based algorithms," in *45th Annual IEEE Symposium on Foundations of Computer Science*, 2004, pp. 32–41.

[2] J. Lemieux, B. Heim, D. Poulin, K. Svore, and M. Troyer, "Efficient quantum walk circuits for metropolis-hastings algorithm," *Quantum*, vol. 4, p. 287, jun 2020. [Online]. Available: https://arxiv.org/abs/1910.01659

[3] S. Lalley, "Markov chains: Basic theory," 2018. [Online]. Available: http://galton.uchicago.edu/ lalley/Courses/383/MarkovChains.pdf

[4] K. Temme, T. J. Osborne, K. G. Vollbrecht, D. Poulin, and F. Verstraete, "Quantum metropolis sampling," *Nature*, vol. 471, no. 7336, pp. 87–90, mar 2011. [Online]. Available: https://doi.org/10.10382Fnature09770

[5] A. Y. Kitaev, "Quantum measurements and the abelian stabilizer problem," 1995.

[6] D. S. Abrams and S. Lloyd, "Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors," *Physical Review Letters*, vol. 83, no. 24, pp. 5162–5165, dec 1999. [Online]. Available: https://arxiv.org/abs/quant-ph/9807070

APPENDIX
VARIOUS CALCULATIONS

A. Positive Span of Co-Primes

Proof. Let a, b be positive integers that are coprimes. We wish to prove that for all $c > ab$ we have that there exists positive coefficients ℓ, k such that $c = \ell a + kb$. By Bezout's Theorem there exists integer coefficients (not necessarily positive) s_0, t_0 such that:

$$c = s_0 a + r_0 b$$

Then we will also have the following set of solutions:

$$s = s_0 + bt, \qquad r = r_0 - at, \qquad t \in \mathbb{Z}$$

We can then find a positive solution if we can find t such that:

$$-\frac{s_0}{b} < t < \frac{r_0}{a}$$

This inequality defines an interval, and its length is:

$$\frac{r_0}{a} - (-\frac{s_0}{b}) = \frac{as_0 + br}{ab} = \frac{c}{ab}$$

This ratio is larger than 1 by our assumption that $c > ab$. Hence there must be some integer within the interval defined by this length. Letting t be this integer, we are done. □

B. Calculation of $\lim_{x\to 0^+} \frac{\arccos(1-x)}{\sqrt{x}} = \sqrt{2}$

Proof.

$$\begin{aligned}
\lim_{x\to 0^+} \frac{\arccos(1-x)}{\sqrt{x}} &\overset{\text{Lhöpital's}}{=} \lim_{x\to 0^+} \frac{\frac{1}{\sqrt{1-(1-x)^2}}}{\frac{1}{2\sqrt{x}}} \\
&= \lim_{x\to 0^+} \frac{2\sqrt{x}}{\sqrt{1-(1-x)^2}} \\
&= \lim_{t\to 1^-} \frac{2\sqrt{1-t}}{\sqrt{1-t^2}} \\
&= \lim_{t\to 1^-} \frac{2\sqrt{1-t}}{\sqrt{(1-t)(1+t)}} \\
&= \lim_{t\to 1^-} \frac{2}{\sqrt{1+t}} \\
&= \frac{2}{\sqrt{2}} \\
&= \sqrt{2}
\end{aligned}$$

□

C. Calculation of $D_{fail}(n)$.

Proof. Let B be a basis change that brings P_1 and Q_1 to the form given by Lemma (6). Then we wish to express the operation:

$$Q_0 P_0 \big(P_0(Q_0 P_0 Q_0 + Q_1 P_0 Q_1) P_0\big)^n P_0 Q_0$$

as $B^\dagger D_{fail} B$ From the forms given by the Lemma we have:

$$Q_1 P_0 Q_1 = \begin{pmatrix} D(\mathbb{I}-D) & (\mathbb{I}-D)\sqrt{D(\mathbb{I}-D)} & 0 & 0 \\ (\mathbb{I}-D)\sqrt{D(\mathbb{I}-D)} & (\mathbb{I}-D)^2 & 0 & 0 \\ 0 & 0 & \mathbb{I} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$Q_0 P_0 Q_0 = \begin{pmatrix} D(\mathbb{I}-D) & -D\sqrt{D(\mathbb{I}-D)} & 0 & 0 \\ -D\sqrt{D(\mathbb{I}-D)} & D^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbb{I} \end{pmatrix}$$

Adding them together gives us and multiplying by both sides by P_0 gives us:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & D^2 + (I-D)^2 & 0 & 0 \\ 0 & 0 & \mathbb{I} & 0 \\ 0 & 0 & 0 & \mathbb{I} \end{pmatrix}$$

This is block diagonal so raising it to any exponent gives us the diagonals raised to that exponent. Then applying P_0 to both sides leaves it unaffected. Finally multiplying by both sides by Q_0 will give us exactly the form we asserted in $D_{fail}(n)$ since we are multiplying Q_0 by a block diagonal matrix.

□