Metropolis algorithm and its cousins

Paulo F. Bedaque*
Department of Physics
University of Maryland
College Park, MD 20742

We summarize importance sampling and the Metropolis algorithm.

CONTENTS

I. METROPOLIS-HASTINGS ALGORITHM

One can think of the Metropolis algorithm as a method to numerically compute integrals. It beats other methods in dimensions larger than 3 or 4 and it is the only viable method for very large number of dimensions (tens, hundreds or more).

A. Stupid Monte Carlo

Suppose we want to compute the area of an irregular shape, that is, to compute the two dimensional integral of a function that vanishes outside the circle and equals 1 inside it.. We can put the shape inside a square ok known area, say A and "throw" darts inside the square (with a computed with the help of pseudo-random number generators). By counting the proportion of "darts" falling inside the shape we can find the ratio between the area of the shape and the area of the square. This method is very inefficient if the integrand is sizable only on a small region of the square since most of the samplings (darts) fall outside the shape. This happens typically in higher dimensions. In fact, in two dimensions the ratio between the area of a circle inscribed within a square is $\pi r^2/(2r)^2 = \pi/4$. The same problem in three dimensions would give the ratio $\frac{4}{3}\pi r^3/(2r)^3 = \pi/6$. As the dimensions grow this ratio approaches 0.

* bedaque@umd.edu

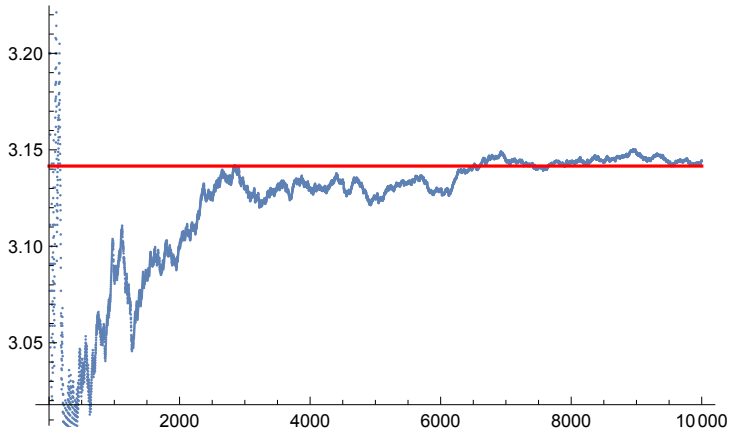

FIG. 1: Stupid Monte Carlo computing the area of a circle of radius 1 inside a square. The number of samplings is on the horizontal axis, the ration of the areas of a inscribed circle to the square (times 4) is on the vertical axis.

B. Importance sampling

It would be advantageous to sample the integrand more frequently where the integrand is large but without biasing the sampling. For instance, suppose we want to compute

$$\langle x^2 \rangle = \frac{\int_{-\infty}^{\infty} dx\ e^{-S(x)} x^2}{\int_{-\infty}^{\infty} dx\ e^{-S(x)}}. \tag{1.1}$$

(The method works better with ratios like that and that's what we need in Physics anyway). Since

$$\pi(x) = \frac{e^{-S(x)}}{\int_{-\infty}^{\infty} dx\ e^{-S(x)}} \tag{1.2}$$

can be thought as a probability density, $\langle x \rangle$ can be thought as an average of x^2 and can be estimated by

$$\langle x^2 \rangle \approx \frac{1}{\mathcal{N}} \sum_{a=1}^{\mathcal{N}} x_a^2, \tag{1.3}$$

where $(x_1, x_2, ..., x_{\mathcal{N}})$ are random numbers distributed according to the probability density $\pi(x) = e^{-S(x)} / \int_{-\infty}^{\infty} dx\ e^{-S(x)}$. Notice that $p(x)$ can be thought as a probability since it is positive and $\int_{-\infty}^{\infty} p(x) = 1$. The advantage here is that most samples x_a will be in the region where the $e^{-S(x)}$ is large. Another advantage is that, after having $(x_1, x_2, ..., x_{\mathcal{N}})$ in hands, it is easy to compute $\langle \mathcal{O}(x) \rangle$ for any function $\mathcal{O}(x)$:

$$\langle \mathcal{O}(x) \rangle \approx \frac{1}{\mathcal{N}} \sum_{a=1}^{\mathcal{N}} \mathcal{O}(x_a). \tag{1.4}$$

I won't show it here but, if the samples x_a are statistically independent of each other, the error made in the equation above scales as $\sim 1/\sqrt{\mathcal{N}}$.

The hard part is to collect the random samples $(x_1, x_2, ..., x_{\mathcal{N}})$ distributed according to $\pi(x)$. For that the standard method

is to use a Markov chain, that is, to have a method to generate x_{a+1} from x_a.[1] Since we are taking random samples, the method to get x_{a+1} from x_1 will be probabilistic. We denote by $T(x' \leftarrow x)$ the probability of getting x' as the next element of the chain if the previous one was x. The transition matrix $T(x' \leftarrow x)$ (think of x and x' as the indices for the rows and columns of a matrix [2]) satisfies:

$$\int dx'\ T(x' \leftarrow x) = 1,$$
$$T(x' \leftarrow x) \geq 0. \tag{1.5}$$

There are many choices for $T(x' \leftarrow x)$ that we will discuss below but for now just assume that we chose one satisfying *detailed balance*:

$$\pi(x)T(x' \leftarrow x) = \pi(x')T(x \leftarrow x'). \tag{1.6}$$

We will also assume that $T(x' \leftarrow x)$ is *ergodic*, that is, $T(x' \leftarrow x) > 0$ for any x', x[3].

Suppose $p_t(x)$ is the distribution probability of x obtained by applying $T(x' \leftarrow x)$ t times. Of course, we would like $p_t(x)$ to approach $\pi(x)$ as t grows One thing is easy to see: if $p_t(x) = \pi(x)$ for some t then $p_{t+1}(x) = \pi(x)$ and similarly for all subsequent $p_t(x)$. In fact,

$$p_{t+1}(x) = \int dy\ T(x \leftarrow y)p_t(y) = \int dy\ T(x \leftarrow y)\pi(y) = \int dy\ T(y \leftarrow x)\pi(x) = \pi(x) = p_t(x). \tag{1.7}$$

In other words, $p(x) = \pi(x)$ is a *fixed-point* of the transition $T(x' \leftarrow x)$.

One can also prove that even if $p_t(x) \neq \pi(x)$, $p_t(x)$ will approach $\pi(x)$ as t grows, in other words, $p(x) = \pi(x)$ is an *attractor*. For that we rely on the fact that the matrix $T(x' \leftarrow x)$. The statement

$$\int dx\ \pi(x)T(x' \leftarrow x) = \int dx\ \pi(x')T(x \leftarrow x') = \pi(x') \tag{1.8}$$

is the statement that the matrix $T(x' \leftarrow x)$ has $\pi(x)$ as an eigenvector with eigenvalue 1. Since $T(x' \leftarrow x) > 0$, Perron's theorem guarantees that all other eigenvalues λ_k are smaller than $\lambda_0 = 1$. So, if we expand an initial distribution $p_0(x)$ into eigenvectors $\phi_k(x)$ of $T(x' \leftarrow x)$

$$p_0(x) = \sum_k c_k\phi_k(x), \tag{1.9}$$

the repeated application of $T(x' \leftarrow x)$ will shrink the components in all directions but the one corresponding to the eigenvalue 1. Using matrix notation to simplify things we have:

$$p_t = \underbrace{T...T}_{t\ \text{times}}\ p_0 = (T)^t p_0 = (T)^t \sum_k c_k\phi_k = \sum_k c_k(T)^t\phi_k = \sum_k c_k\lambda_k^t\phi_k \underset{t\to\infty}{\to} c_0\phi_0 \sim \pi(x), \tag{1.10}$$

since $\lambda_k < 1$ for all $k \neq 0$.

C. Metropolis-Hastings

Usually the transition probability $T(x' \leftarrow x)$ is split into a proposal $\mathcal{P}(x' \leftarrow x)$ and an acceptance $\mathcal{A}(x' \leftarrow x)$ stages: $T(x' \leftarrow x) = \mathcal{A}(x' \leftarrow x)\mathcal{P}(x' \leftarrow x)$. If the acceptance is chosen as

---

[1] The work "Markov" refers to the fact that x_{a+1} depends only on x_a but not on the previous $x_{a-1}, x_{a-2}, ...$

[2] I know, this is strictly true for discrete x, not continuous ones. But just go along with me.

[3] This can be relaxed to the condition that any value of x can be reached after many transitions, instead of only one.

$$\mathcal{A}(x' \leftarrow x) = \min\left(1, \frac{\pi(x')}{\pi(x)} \frac{\mathcal{P}(x \leftarrow x')}{\mathcal{P}(x' \leftarrow x)}\right) \tag{1.11}$$

detailed balance is satisfied. In the Metropolis-Hastings algorithm, the proposal is chosen to be symmetric: $\mathcal{P}(x' \leftarrow x) = \mathcal{P}(x \leftarrow x')$. For instance, x' can be chosen to be $x' = x + \zeta$, where ζ is a random number uniformly distributed between $-\epsilon$ and ϵ.

In a nutshell, the Metropolis-Hastings algorithm is:

Metropolis

1. Start form any x_0.
2. propose $x_1 = x_0 + \zeta$.
3. accept x_1 with probability

$$\min\left(1, \frac{\pi(x')}{\pi(x)} \frac{\mathcal{P}(x \leftarrow x')}{\mathcal{P}(x' \leftarrow x)}\right). \tag{1.12}$$

   If x_1 is rejected, $x_1 \leftarrow x_0$, otherwise x_1 is the new proposed one.
4. Rinse and repeat.

This way a chain of $(x_0, x_1, ...)$ of values of x distributed according to $\pi 9x)$ is generated. They can then be used to compute Eq. 1.4.

As an example, take $S(x) = -x^2 + x^4$. Using the Metropolis algorithm, I sampled the $\pi(x)\ e^{-S(x)}$ distribution 10^4 times and obtained the histogram below.

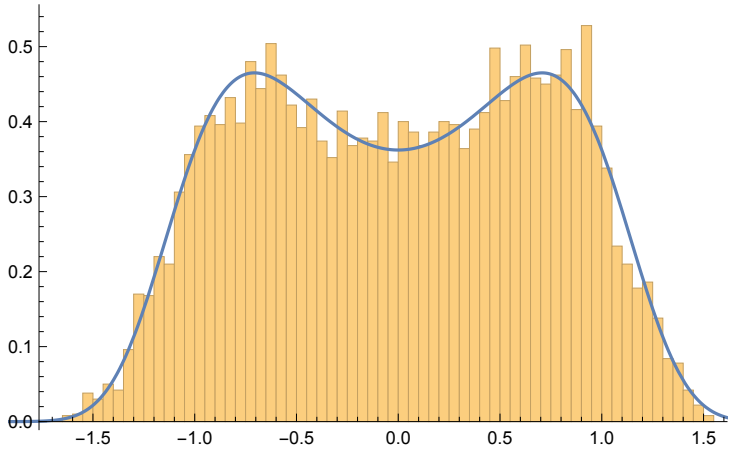

FIG. 2: Sampling of the distribution $\pi(x)\ e^{-S(x)}$ from 10^4 steps of the Metropolis algorithm with $\epsilon = 0.1$

II. JACKNIFE AND UNCERTAINTY ESTIMATION

The Monte Carlo method converges as $\mathcal{N} \to \infty$ but, at finite $\mathcal{N}$, there will be errors. Estimating this errors is as important as it is to find the average value. There is a general, easy to use, albeit computationally expensive, method to estimate the

errors called jacknife (I'm told that "jacknife" is a tool that can be used for many things, hence the name). It's one version of a class of "resampling" methods that include also the "bootstrap" method.

Suppose you repeat an experiment $\mathtt{n}$ time (maybe a numerical experiment) and, after a lot of processing, each one of these experiments predict a value θ_i, $i = 1, ..., \mathtt{n}$ for the quantity of interest. Define $\hat{\theta}_i$ and the average value obtained *excluding* the i^{th} measurement:

$$\hat{\theta}_i = \frac{1}{\mathtt{n}-1}\sum_{j\neq i}\theta_j. \tag{2.1}$$

Jacknife says that the best estimate of θ and its uncertainty are given by

$$\langle\theta\rangle = \frac{1}{\mathtt{n}}\sum_i \hat{\theta}_i, \tag{2.2}$$

$$\Delta\theta = \sqrt{\frac{\mathtt{n}-1}{\mathtt{n}}\sum_i(\hat{\theta}_i - \langle\theta\rangle)^2}. \tag{2.3}$$

$\langle\theta\rangle$ is just the average of θ_i, nothing new there. The point of the jacknife method is to estimate the uncertainty $\Delta\theta$. This method is particularly useful if the measurements θ_i are complicated functions of x_i. This error estimate assumes the measurements are statistically independent. A common method to guarantee that is to *block* the measurements. For instance, if we have $\mathtt{n} = 1000$ measurements we can separate them into 10 groups of 100 measurements and treat the average of each block $\theta_A, A = 1, \cdots, 10$ as a single measurement and then apply the jacknife method. The estimated error increases with the size of the block until it plateaus for a while (before reaching huge values when the size of the blocks is too large).

You may want to take a look at `https://www.youtube.com/watch?v=p9XPclE7NtA`.

III. VARIATIONAL MONTE CARLO

In the variational method one finds an approximation to the ground state energy/wavefunction by finding the wave function which minimizes the expected value of the energy

$$\mathcal{E} = \frac{\int dX\ \psi^\dagger(X)\hat{H}\psi(X)}{\int dX\ \psi^\dagger(X)\psi(X)} \tag{3.1}$$

(X stands for all the coordinates of the system) within a parametrized set of wavefunctions (the "ansatz"). In systems with many particles the computation of $\mathcal{E}$ is difficult so one can resort to Monte Carlo integration. In order to use MC integration we use the fact that the ground state wavefunction of bosonic systems is positive ane write $\mathcal{E}$ as an average statistical value:

$$\mathcal{E} = \frac{\int dX\ \psi^2(X)\psi^{-1}\hat{H}\psi(X)}{\int dX\ \psi^2(X)} = \int dX\ P(X)\psi^{-1}(X)\hat{H}\psi(X) = \langle\psi^{-1}\hat{H}\psi\rangle \tag{3.2}$$

where the average $\langle ...\rangle$ is defined by the probability distribution $P(X) = \psi^2(X)/\int dX\psi^2(X)$.

Gradient descent (or some improvement upon it) is used to find the best wavefunction within the ansatz family. This requires the computation of the gradient in parameter space. If we denote by θ the parameters of the variational family :

$$\frac{\partial\mathcal{E}}{\partial\theta} = \frac{\int dX\ [\partial\psi/\partial\theta\hat{H}\psi + \psi\hat{H}\partial\psi/\partial\theta\psi]}{\int dX\ \psi^2} - 2\frac{\int dX\ \psi\hat{H}\psi\int dX\ \psi\partial\psi/\partial\theta}{\left(\int dX\ \psi^2\right)^2} \tag{3.3}$$

$$= 2\langle\psi^{-1}\frac{\partial\log\psi}{\partial\theta}\hat{H}\psi\rangle - 2\langle\psi^{-1}\hat{H}\psi\rangle\langle\frac{\partial\log\psi}{\partial\theta}\rangle, \tag{3.4}$$

where we used the fact that $\hat{H}$ is hermitian. The expectation values appearing in $\partial\mathcal{E}/\partial\theta$ can be computed using Monte Carlo. Notice that a poor Monte Carlo is not a serious problem as it may delay the finding of the minimum of $\mathcal{E}$ but it will not change its location. In fact, a poor Monte Carlo may introduce some noise that actually *helps* the gradient descent to avoid

local minima where the minimization may get stuck. Also, the thermalization part of the MC chain can be reduced as the thermalized distribution at one step of the gradient descent is close to the distribution of the previous step.

The success of the variational method relies in having a good ansatz for the wavefunction. A very general ansatz is provided by a feed-forward neural net. In the case of two particles moving in one dimensions we can choose $\psi(x_1, x_2) = e^{u(x_1,x_2)}$ where $u(x_1, x_2)$ is the output of the neural net with inputs x_1, x_2. The role of the parameters θ are played by the biases and weights. Bose statistics can be enforced, in one dimension, by feeding the network *ordered* coordinates $x_1 < \cdots < x_N$. In higher dimensions it's a little harder to impose Bose symmetry but we have ideas ...

IV. HYBRID MONTE CARLO

Notice that

$$\langle \mathcal{O}| = \frac{\int D\phi\ \mathcal{O}e^{-S[\phi]}}{\int D\phi\ e^{-S[\phi]}} = \frac{\int D\pi D\phi\ \mathcal{O}e^{-(\frac{\pi^2}{2}+S[\phi])}}{\int D\pi D\phi\ e^{-(\frac{\pi^2}{2}+S[\phi])}}. \tag{4.1}$$

Averaging $\mathcal{O}$ over a trajectory in the (π, ϕ) phase space obtained by solving the classical equations of motion for the hamiltonian $\mathcal{H} = \pi^2/2 + S[\phi]$ would, by the ergodic hypothesis, give the average of the observable over an energy shell $\mathcal{H} = E$. The canonical ensemble average in Eq. 4.1 is obtained by a further averaging over energy shells with a weight determined by the kinetic energy (that measures the temperature). This suggests the algorithm:

Hybrid Monte Carlo

1. Pick a π' from a distribution $\sim e^{-\pi^2/2}$.
2. Evolve (π, ϕ) by Hamilton's equation ($\mathcal{H} = \pi^2/2 + S[\phi]$) by a certain trajectory time T to obtain (π', ϕ').
3. Accept the new (π', ϕ') with probability

$$\mathcal{A}[\pi', \phi' \leftarrow \pi, \phi] = \min\left(1, \frac{e^{-\mathcal{H}[\pi',\phi']}}{e^{-\mathcal{H}[\pi,\phi]}}\right). \tag{4.2}$$

4. Rinse and repeat.

Notice that, if the Hamilton equations are solved exactly, $\mathcal{H}[\pi', \phi'] = \mathcal{H}[\pi, \phi]$ and the acceptance is 1. Solving Hamilton's equations numerically there will always be some variation in the energy, hence the need for an accept/reject step. It is important that, even for non-zero time steps Δt we demand that $\mathcal{F}_T(\pi', \phi' \leftarrow \phi, \phi) = \delta(\phi' - \phi(T))\delta(\pi' - \pi(T))$ be time reversible:

$$\mathcal{F}_T(\pi', \phi' \leftarrow \phi, \phi) = \mathcal{F}_T(-\pi, \phi \leftarrow -\phi, \phi). \tag{4.3}$$

The formal proof of the correctness of HMC comes from proving the detailed balance condition. In fact,

$$
\begin{aligned}
e^{-S[\phi]}T(\pi',\phi'\leftarrow\pi,\phi)d\pi d\phi &= e^{-S[\phi]}d\pi d\phi\int d\pi d\pi' \underbrace{\mathcal{A}(\pi',\phi'\leftarrow\pi,\phi)}_{\begin{cases}\frac{e^{-\mathcal{H}'}}{e^{-\mathcal{H}}}, & \mathcal{H}'>\mathcal{H}\\ 1, & \mathcal{H}>\mathcal{H}'\end{cases}=\frac{e^{-\mathcal{H}'}}{e^{-\mathcal{H}}}\begin{cases}1, & \mathcal{H}'>\mathcal{H}\\ \frac{e^{-\mathcal{H}}}{e^{-\mathcal{H}'}},, & \mathcal{H}>\mathcal{H}'\end{cases}} \mathcal{F}_T(\pi',\phi'\leftarrow\pi,\phi)e^{-\pi^2/2}\\
&= e^{-S[\phi]}d\pi d\phi\int d\pi d\pi'\ \mathcal{A}(\pi,\phi\leftarrow\pi,'\phi')\frac{e^{-\mathcal{H}'}}{e^{-\mathcal{H}}}\underbrace{\mathcal{F}_T(\pi',\phi'\leftarrow\pi,\phi)}_{\mathcal{F}_T(-\pi,\phi\leftarrow-\pi',\phi')}e^{-\pi^2/2}\\
&= e^{-S[\phi']}d\pi' d\phi'\int d\pi d\pi'\ \mathcal{A}(\pi,\phi\leftarrow\pi,'\phi')\mathcal{F}_T(\pi',\phi'\leftarrow\pi,\phi)\\
&= e^{-S[\phi']}d\pi' d\phi' T(\pi,\phi\leftarrow\phi',\phi'),
\end{aligned}
\tag{4.4}
$$

where $\mathcal{H}=\mathcal{H}[\pi,\phi],\mathcal{H}'=\mathcal{H}[\pi',\phi']$ and we used $d\pi d\phi = d\pi' d\phi'$ that follows from the symplecticity of the hamiltonian evolution. This proof also shows that the hamiltonian evolution can be performed with a different, perhaps simpler, hamiltonian, while keeping the algorithm correct. The cost, of course, is in a reduce acceptance rate.

The practical implementation of this idea hinges on the existence of time reversible, symplectic integrators for the hamilton's equations. There is a whole literature on them. I'll give here a simple example, the leap-frog method:

$$
\begin{aligned}
\pi(\Delta t+\frac{\Delta t}{2}) &= \pi(t)-\frac{\partial S}{\partial\phi}|_{\phi(t)}\frac{\Delta t}{2},\\
\phi(\Delta t+\Delta t) &= \pi(t)+\pi(\Delta t+\Delta t)\Delta t,\\
\pi(\Delta t+\Delta t) &= \pi(\Delta t+\frac{\Delta t}{2})-\frac{\partial S}{\partial\phi}|_{\phi(t+\Delta t)}\frac{\Delta t}{2}.
\end{aligned}
\tag{4.5}
$$

Each one of these steps modifies only half of the variables (either π or ϕ) and their jacobian is a triangular matrix with diagonal elements equal to 1 so its determinant must be 1. Consequently, the whole leapfrog step has jacobian 1 and is symplectic. Later I'll show it is also time reversible.

V. NUMERICAL PATH INTEGRALS

Path integrals arising in statistical mechanics, quantum mechanics or field theory can be computed by Monte Carlo. In the case of quantum systems, the path integral

$$
\int Dx\ e^{iS[x]},
\tag{5.1}
$$

is ill defined (it requires some $i\epsilon$ prescriptions to make sense) and hopeless to be computed numerically. Fortunately, a version of it obtained by analytically continuing it to imaginary time (called a "Wick rotation" in field theory) is suitable for numerics. It is also the way they are formally defined. One way to arrive at this result is to observe that the density matrix of the canonical ensemble $e^{-\beta H}$ (H is the hamiltonian) is the time-evolution operator $U(t=i\beta)=e^{i(i\beta)H}$ and, therefore, has a similar path integral representation (one can go through exactly the same steps as the usual derivation of the path integral representation with $t\to i\beta$). The result is that the partition function is given by

$$
Z=\operatorname{tr}e^{-\beta H}=\int Dx(\tau)\ e^{-S[x]},
\tag{5.2}
$$

$S[x]$ given by (in quantum mechanics, for instance)

$$S[x] = \int_0^\beta d\tau \ \left[\frac{m}{2} \left(\frac{dx}{d\tau} \right)^2 + V(x) \right]. \tag{5.3}$$

The sum over imaginary time trajectories should be restricted to periodic (with $x(0) = x(\beta)$ and) functions, a condition arising from the trace $\mathrm{tr}\, A = \int dx \, \langle x| \, A \, |x\rangle$. Any field theory book will have more details on this and its generalizations to field theories and fermions.

Notice that Eq. 5.2 is both the *quantum* partition function of a particle moving in 1D or the *classical* partition function of a string with hamiltonian $S[x]$. This observation extends to field theory. The quantum theory of a field in $d+1$ dimensions (d spatial + one time) is the same as teh classical statistical mechanics of a theory in $d+1$ spatial dimensions.

Unfortunately, the Monte Carlo method does not allow us to compute Z directly. It can be used to compute thermal expectation values though:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \, \mathrm{tr}\, \mathcal{O} e^{-\beta H} = \frac{\int Dx \ \mathcal{O} e^{-S[x]}}{\int Dx \ e^{-S[x]}} \approx \frac{1}{\mathcal{N}} \sum_{a=1}^{\mathcal{N}} \mathcal{O}(x_a), \tag{5.4}$$

where $\mathcal{O}$ is any function of the operator x.

Also, individual energy levels can be extracted from two-point functions (correlators)

$$\begin{aligned}
\langle x(\tau) x(0) \rangle &= \frac{1}{Z} \sum_n e^{-\beta E_n} \, \langle n| \, x(\tau) x(0) \, |n\rangle \\
&= \frac{1}{Z} \sum_{n,m} e^{-\beta E_n} \, \langle n| \, e^{\tau H} x e^{-tH} \, |m\rangle \, \langle m| \, x \, |n\rangle \\
&= \frac{1}{Z} \sum_{n,m} e^{-\beta E_n + \tau (E_n - E_m)} \, \langle n| \, x \, |m\rangle \, \langle m| \, x \, |n\rangle \\
&\xrightarrow[\beta \to \infty]{} \frac{1}{Z} \sum_n e^{-\beta E_0 + \tau (E_0 - E_n)} | \, \langle 0| \, x \, |n\rangle \, |^2 \\
&\xrightarrow[\tau \to \infty]{} \frac{1}{Z} \sum_n e^{-\beta E_0} e^{-\tau (E_1 - E_0)} | \, \langle 1| \, x \, |0\rangle \, |^2
\end{aligned} \tag{5.5}$$

where $x(\tau) = e^{\tau H} x e^{-\tau H}$ are the imaginary-time Heisenberg picture operators. The eigenenergy E_1 is the smallest among those states for which the matrix element $\langle n| \, x \, |n\rangle$ does not vanish. By plotting the correlator as a function of τ and fitting an exponential one can extract the energy gap $E_1 - E_0$. By looking at other correlators or subleading exponentials in the $\tau \to \infty$ limit, higher excited states can be extracted.

ACKNOWLEDGMENTS

We thank coffee and corn bread with fennel seeds.