SLAC-PUB-14373

Reducing Memory Cost of Exact Diagonalization using Singular Value Decomposition*

Marvin Weinstein[1], Assa Auerbach[2,3], and V. Ravi Chandra[3]
[1]*SLAC National Accelerator Laboratory, Stanford, 2575 Sand Hill Road, CA 94025, USA.*
[2]*Department of Physics, Stanford University, Stanford CA 94306, USA. and*
[3]*Physics Department, Technion, Haifa, 32000, Israel*
(Dated: April 29, 2011)

We present a modified Lanczos algorithm to diagonalize lattice Hamiltonians with dramatically reduced memory requirements. The lattice of size N is partitioned into two subclusters. At each iteration the Lanczos vector is projected into a set of $n_{\rm svd}$ smaller subcluster vectors using singular value decomposition. For low entanglement entropy S_{ee}, (satisfied by short range Hamiltonians), we expect the truncation error to vanish as $\exp(-n_{\rm svd}^{1/S_{ee}})$. Convergence is tested for the Heisenberg model on Kagomé clusters of up to 36 sites, with no symmetries exploited, using less than 15GB of memory. Generalization to multiple partitioning is discussed.

PACS numbers: 05.30.-d, 02.70.-c, 03.67.Mn, 05.50.+q

Numerical ("Exact") diagonalization (ED) of quantum many-body Hamiltonians on finite clusters is often used to advance our understanding of larger lattices. For example, Contractor Renormalization [1–3] uses ED to compute the short range interactions of the effective hamiltonian. Approximation schemes, mean field theories and variational wavefunctions, are routinely tested on small clusters by comparisons to ED. ED also yields short wavelength dynamical response functions[4] and Chern numbers for computing Hall conductivity [5].

Lanczos algorithms [6, 7] are particularly efficient for sparse Hamiltonians, such as the Hubbard and Heisenberg models. Much effort has been devoted to speeding up Lanczos convergence, and avoiding error accumulation. However, for a lattice of size N, with m states per site, the dimension of the Lanczos vectors (which are stored in the dynamical memory) goes as m^N. Therefore, increasing ED to larger lattice sizes is prevented primarily by memory limitations, rather than processor speed.

The central idea of this paper is to significantly reduce the dynamical memory cost of ED. The lattice is divided into two subclusters. We use singular value decomposition (SVD) to reduce "large" vectors to a set of $n_{\rm svd}$ "small" subcluster vectors. High accuracy can be guaranteed if the truncation error is a rapidly decreasing function of $n_{\rm svd}$. Significant memory compression is achieved when $n_{\rm svd} << m^{N/2}$, which in turn depends on having low *entanglement entropy* S_{ee} [8] of the target eigenstates. Indeed, many of the interesting short range models in one and two dimensions [9–12] are known to have low entanglement entropy. This explains the remarkable success of density matrix renormalization group [13], matrix product states [9, 14] , and multiscale entanglement renormalization approaches [15] for these models. In higher dimensions d, the area law scaling $S_{ee} \sim cN^{(d-1)/d}$ is expected [8]. This allows us to choose $S_{ee} << \log(n_{\rm svd}) << N/2\log(m)$ and still obtain a small truncation error, as we shall demonstrate below.

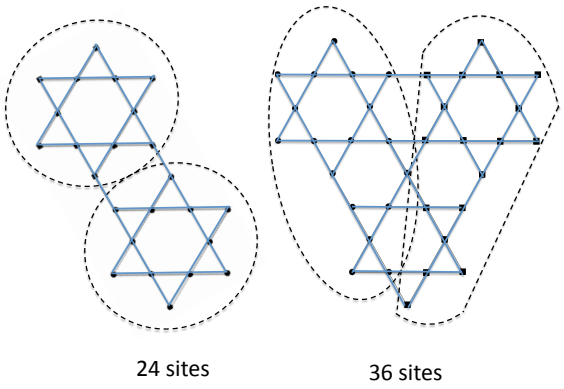

FIG. 1. Partitioning Kagomé clusters for application of the Lanczos-SVD algorithm

The paper is organized as follows. The SVD projection of a large vector into small vectors is defined. The convergence of the truncation error as a function of $n_{\rm svd}$ and S_{ee} is obtained for a generic model of entanglement spectrum. We describe a Lanczos-SVD iteration step with its orthonormalizations and diagonalizations. We test the Lanczos-SVD algorithm for the spin half Heisenberg antiferromaget on the Kagomé clusters of 24 and 36 sites, which are depicted in Fig. 1. The results agree with our

*This work was supported by the U. S. DOE, Contract No. DE-AC02-76SF00515.

estimate of the truncation error. The ground state energy of 36 sites converges to a relative error of $\approx 7\times10^{-5}$ as compared to the exact result [16], where we use a desktop computer with less than 15GB of memory and no lattice symmetries. Finally, we discuss a possible extension of this approach to multi-partitioning, and estimate the optimal reduction in memory cost that could be achieved.

The SVD projection. Any state $|\psi\rangle$ of the full cluster can be represented in a unique SVD form as

$$|\psi\rangle = \sum_\alpha \lambda_\alpha |\alpha\rangle_1 |\alpha\rangle_2,$$
$$\sum_\alpha \lambda_\alpha^2 = 1, \quad \langle\alpha|\alpha'\rangle_i = \delta_{\alpha\alpha'}, \tag{1}$$

where the λ_α are positive, and $|\alpha\rangle_i$ are "small" basis vectors of subclusters $i=1,2$. Truncating the sum into the largest $n_{\rm svd}$ terms defines the SVD projector,

$$P_{\rm svd}|\psi\rangle = \sum_{\alpha=1}^{n_{\rm svd}} \lambda_\alpha |\alpha\rangle_1 |\alpha\rangle_2, \tag{2}$$

which introduces a wavefunction error $\epsilon = ||(P_{\rm svd}-1)|\psi\rangle||^2$.

To get an idea of how $\epsilon(n_{\rm svd})$ converges, we must know the "entanglement spectrum" $\{s_\alpha\}$ defined by $\lambda_\alpha^2 \equiv e^{-s_\alpha}$. A generic density of states can be modelled by a power law form

$$\rho_p(s) = \sum_\alpha \delta(s-s_\alpha) = \frac{s^p}{\Gamma(p+1)}, \quad p>-1, \tag{3}$$

which describes the many-body density of states of a *classical* gas with constant (Dulong-Petit) specific heat [20]. p counts with the number of entangled degrees of freedom. The corresponding entanglement entropy is easy to evaluate,

$$S_{ee} = -\sum_\alpha \lambda^2 \log(\lambda^2) = \int_0^\infty ds s\rho_p(s) e^{-s} = p+1. \tag{4}$$

Choosing a cut-off exponent s_c such that $n_{\rm svd} = \int_0^{s_c} ds\rho_p(s)$, we arrive at the error estimate at large $n_{\rm svd}$,

$$\epsilon \sim e^{-n_{\rm svd}^{1/S_{ee}}}. \tag{5}$$

Hence a choice of $n_{svd} > e^{S_{ee}}$ will ensure an exponentially small truncation error.

Lanczos-SVD step. Lanczos-SVD economizes on the storage space by applying an SVD projection after each application of the Hamiltonian on the Lanczos vector,

$$|\psi\rangle' = P_{\rm svd} H|\psi\rangle. \tag{6}$$

The projection entails the following computational steps. H can be written as a sum of products of the two subcluster operators,

$$H = H_1^0 \otimes I_2 + I_1 \otimes H_2^0 + \sum_{\mu=3}^{M} H_1^\mu \otimes H_2^\mu. \tag{7}$$

For example, a nearest neighbor Heisenberg model ($\sum_{ij} \mathbf{S}_i\cdot\mathbf{S}_j$) with K bonds connecting the two subclusters has $M = 2+3K$ terms.

Acting with H on $|\psi\rangle$ produces a new state,

$$H|\psi\rangle = \sum_{\nu=1}^{n_{\rm svd}M} |\nu)_1 |\nu)_2, \tag{8}$$

where the new (non orthonormal) small vectors are labelled by $|\nu)_i = H_i^\mu |\alpha\rangle_i$. To project the new state by $P_{\rm svd}$, we must first orthonormalize $|\nu)_i$. We diagonalize the two Hermitian overlap matrices (of row dimensions $n_{\rm svd}M$)

$$\langle\nu'|\nu\rangle_i = \left(V_i^\dagger D_i V_i\right)_{\nu\nu'} \qquad i=1,2$$
$$|\beta\rangle_i = \sum_\nu \left(D_i^{-\frac{1}{2}} V_i\right)_{\beta\nu} |\nu)_i, \tag{9}$$

where D_i are diagonal and positive semidefinite, and V_i are unitary. $|\beta\rangle_i, i=1,2$ are orthonormal sets in their respective spaces. Thus, the new vector is given by

$$H|\psi\rangle = \sum_{\beta\beta'}^{n_\beta n_{\beta'}} C_{\beta\beta'} |\beta\rangle_1 |\beta'\rangle_2$$
$$C = \sqrt{D_1} V_1^* V_2^\dagger \sqrt{D_2}. \tag{10}$$

Now we perform an SVD on the matrix C,

$$C = \nu_c^2 U_1^t \Lambda' U_2, \tag{11}$$

where ν_c is the normalization. $U_1^t, U_2^\dagger$ are unitary matrices which diagonalize the Hermitian products $CC^\dagger$ and $C^\dagger C$ respectively. After computing Λ, U_1, U_2 we obtain the SVD form of the new state. Λ is diagonal and normalized to ${\rm Tr}\Lambda^2 = 1$ with positive eigenvalues λ'_α. We keep only the $n_{\rm svd}$ largest λ'_α and obtain

$$P_{\rm svd} H|\psi\rangle = \sum_{\alpha=1}^{n_{\rm svd}} \lambda'_\alpha |\alpha\rangle'_1 |\alpha\rangle'_2, \tag{12}$$

where the new small vectors of $i=1,2$ are,

$$|\alpha\rangle'_i = \sum_{\nu=1}^{n_{\rm svd}M} \left(U_i D^{-\frac{1}{2}} V_i\right)_{\alpha\nu} |\nu)_i. \tag{13}$$

Convergence to eigenstates The Lanczos algorithm rotates a set of basis states $|\psi_n\rangle$ into the lowest energy eigenstates with which the basis has a finite overlap. If we choose ϵ to be much smaller than the lowest relative energy gap, the Lanczos-SVD vectors converge to a states which are within ϵ distance from the SVD projection of the corresponding exact eigenstate.

ED of Kagomé clusters. We test the convergence of the Lanczos-SVD algorithm for the spin-half Heisenberg antiferromagnet

$$H = \sum_{\langle ij\rangle} \mathbf{S}_i \cdot \mathbf{S}_j \tag{14}$$

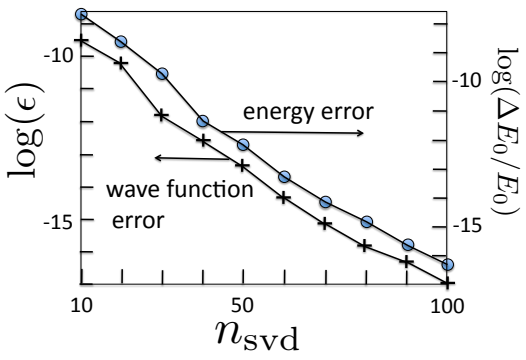

FIG. 2. Lanczos-SVD truncation errors, for the 24 sites Kagomé cluster. $n_{\rm svd}$ is the number of retained products in the SVD representation of the ground state (1). ϵ is the wavefunction error Eq. (2) and following text). $\Delta E_0/E_0$ is the relative error in the the Lanczos-SVD ground state energy as compared to the exact result. The low entanglement entropy of the subclusters ($S_{ee} = 1.510$) is the reason for the rapid decay of the errors in the regime $n_{\rm svd} << 2^{12}$.

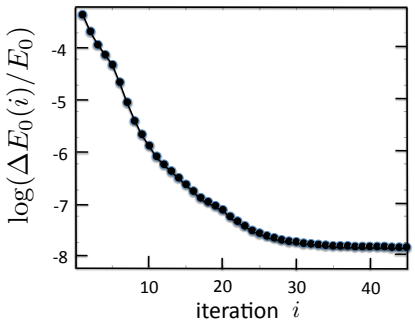

FIG. 3. Lanczos-SVD convergence of the ground state energy of the 36 site Kagomé cluster. $\Delta E_0(i)$ is difference between the energy at iteration i and the ED result of Ref. [16]. We used $n_{\rm svd} = 100$, and $L = 4$ Lanczos vectors.

on Kagomé clusters of N=24,36 sites, as depicted in Figs.1 . The Lanczos-SVD routine proceeds as follows: We initialize $|\psi\rangle^{(0)}$ as a direct product of the two subcluster states. We compute $(\mathcal{P}_{\rm svd}H)^n|\psi^{(0)}\rangle = |\psi^{(n)})$ as described above. Since our method is economical in memory, we can afford to retain L sequential Lanczos vectors $|\psi^{(n)}), |\psi^{(n+1)}), \ldots |\psi^{(n+L)})$, which speeds up the convergence with iteration number considerably. (If memory is scarce, one could use the slower method of keeping only two Lanczos vectors).

Now, we compute the overlap matrix and orthonormalize this set of Lanczos vectors. This produces a "rotating basis" of dimension L

$$|\varphi^{(i)}\rangle = \sum_{n'=n}^{n+L} A_{in'}|\psi^{(n')}\rangle, \qquad \langle\varphi^{(i)}|\varphi^{(j)}\rangle = \delta_{ij}, \qquad (15)$$

where A are the coefficients determined by diagonalizing the overlap matrix (see e.g. Eq. (9)).

Subsequently, we compute the matrix elements of the reduced Hamiltonian,

$$H_{ij} = \langle\varphi^{(i)}|H|\varphi^{(j)}\rangle, \qquad i,j = 1,\ldots L. \qquad (16)$$

The reduced Hamiltonian matrix is diagonalized. Its lowest eigenvalue and eigenvector yield the best variational approximations to the ground state energy and wavefunction at this level of iteration [17].

Finally, we bring the resulting wavefunction to the SVD form truncated into $n_{\rm svd}$ terms. This provides the new initial state $|\psi^{(n+L+1)}\rangle$ for further Lanczos steps.

Results. For 24 sites, we compare standard Lanczos to Lanczos-SVD in order to investigate the dependence of truncation error on $n_{\rm svd}$. The ground state entanglement entropy is $S_{ee} = 1.51$. In Fig.2 we plot the wave function and ground state energy errors versus $n_{\rm svd}$, and find that they decrease exponentially as expected by Eq. (5). We also verified that the SVD projection $\mathcal{P}_{\rm svd}$ does not slow down the energy convergence with iteration number.

The 36 site Kagomé cluster was split into two 18 site clusters as shown in Fig. 1. The Hamiltonian has 7 interconnecting bonds, which implies $M = 23$. The number of retained small vectors before orthonormalization is therefore $2Mn_{\rm svd}$. In Fig 3 we plot ground state energy of Lanczos-SVD versus iteration, using $n_{\rm svd} = 100$, and $L = 4$, and compare it to the exact ground state energy as determined by standard Lanczos $E_0(36) = -14.859397$ [16]. The entanglement entropy was measured to be $S_{ee} \approx 2.5$. The calculation converges to relative energy accuracy of 10^{-4}, and an SVD truncation error of similar magnitude.

Comparing ϵ and S_{ee} of 36 sites and 24 sites for the same $n_{\rm svd} = 100$ is consistent with the errors being 3.8×10^{-8} and 6.6×10^{-5} respectively, as expected by Eq. (5).

Extension to larger lattices. The Lanczos-SVD compresses the memory requirement by a single division of the cluster into two subclusters $i = 1, 2$. This idea could be extended to recursive partitioning. Consider that each small vector (e.g. $|\alpha\rangle_i$ in Eq. (1) gets further decomposed into $n_{\rm svd}$ products of even smaller subcluster vectors. If the SVD is thus iterated p times, one obtains a representation in terms of small vectors of $P = 2^p$ subclusters. $|\psi\rangle$ is thus stored in terms of a set of the smallest vectors. Let us estimate the memory cost of storing such a decomposition and applying to it the Hamiltonian.

For concreteness, consider a two dimensional disk of radius $R >> 1$, containing $N \simeq \pi R^2$ sites of spin half, divided into P equal sections as shown in Fig.4. The sections are labelled by a binary number $\mathbf{i} = (i_1, i_2, \ldots i_p)$, $i_k = 0, 1$. The recursive SVD decomposition yields the

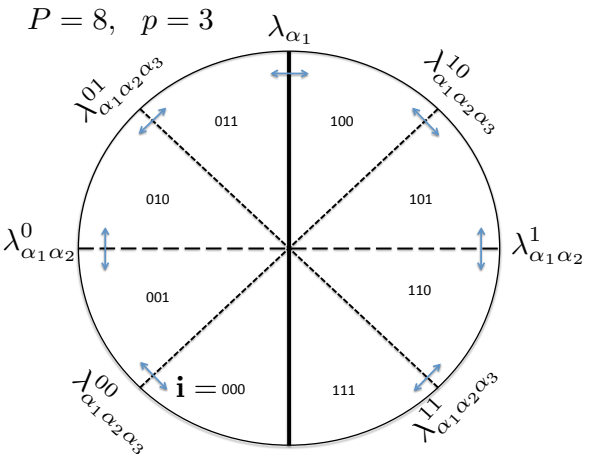

FIG. 4. Multiple subclusters

expression

$$|\psi\rangle = \sum_{\alpha_1,\alpha_2\ldots\alpha_p} \lambda_{\alpha_1}\lambda^{i_1}_{\alpha_1,\alpha_2}\cdot\cdot\lambda^{i_1,\ldots i_{p-1}}_{\alpha_1,\ldots\alpha_p}\prod_{\mathbf{i}}|\alpha_1,\ldots\alpha_p\rangle_{\mathbf{i}}. \tag{17}$$

The SVD weights $\lambda^{\mathbf{i}}$ are labelled according to the boundaries they describe, as shown in Fig. (4). Each α_i runs over $n_{\rm svd}$ numbers, which means that each section $\mathbf{i}$ is represented by $n^p_{\rm svd}$ vectors of dimension $2^{N/P}$. By the "area law" $S_{ee} \propto R$ on each boundary. As shown before, we must retain $n_{\rm svd} \sim e^{cR}$ terms in each SVD, where $c(\epsilon) > 1$ in order to achieve a desired truncation error ϵ.

After applying H to $|\psi\rangle$, we generate a factor of $M \approx 6R$ more small vectors. Thus we should store $6PRn^{p)}_{\rm svd}$ small vectors in the memory. Thus the memory cost is

$$\mathcal{M}_c \approx 6PR\exp\left(cR\log(P) + \frac{\pi R^2\log(2)}{P}\right). \tag{18}$$

Minimizing $\mathcal{M}_c(P)$ one finds the optimal partitioning P^{opt}, and the optimal memory cost $\mathcal{M}^{opt}$ at large N to scale as

$$P^{opt} \approx \frac{\pi\log(2)R}{c}, \quad \mathcal{M}^{opt}_c \sim Ne^{\frac{c}{2}\left(\sqrt{N/\pi}\log(N/\pi)-2\right)}. \tag{19}$$

This would amount to a significant compression of memory as compared to standard Lanczos $\mathcal{M} \sim 2^N$. The remaining challenge is to speed up the significantly larger computational time needed to orthonormalize and SVD large sets of small vectors.

Acknowledgements We thank Dan Arovas, Yosi Avron, Sylvain Capponi, Steve Kivelson, Israel Klich and Netanel Lindner for conversations. AA acknowledges support from israel Science Foundation, and US Israel Binational Science Foundation, and the hospitailty of Aspen Center For Physics.

[1] C. J. Morningstar, M. Weinstein Phys.Rev. D **54**, 4131 (1996); Marvin Weinstein *Phys.Rev.* **B63**, 174421 (2001); M.Stewart Siu, Marvin Weinstein Phys. Rev. B **77**, 155116 (2008).
[2] E. Altman and A. Auerbach, Phys. Rev. B **65**, 104508 (2002); E. Berg, E. Altman, and A. Auerbach, Phys. Rev. Lett. **90**, 147204 (2003); R. Budnik and A. Auerbach Phys. Rev. Lett. **93**, 187205 (2004).
[3] S. Capponi, A. Läuchli, and M. Mambrini Phys. Rev. B **70**, 104424 (2004).
[4] N. H. Lindner and A. Auerbach, Phys. Rev. B **81**, 054512 (2010).
[5] J. E. Avron and R. Seiler, Phys. Rev. Lett. **54**, 259 (1985); N. Lindner, A. Auerbach, D. P. Arovas, Phys. Rev. B **82** 134510 (2010).
[6] C. Lanczos. *J. Res. Nat. Bur. Stand* **45**, 255 (1950)
[7] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Birkhäuser (1985)
[8] L. Amico, R. Fazio, A. Osterloh, and V. Vedral, Rev. Mod. Phys. **80**, 517 (2008) (and references therein).
[9] F. Verstraete, M. A. Martin-Delgado, and J. Cirac, Phys. Rev. Lett. **92**, 087201 (2004).
[10] M. B. Hastings,Phys. Rev. B 69, 104431 (2004);
[11] E. Fradkin, and J. E. Moore, Phys. Rev. Lett. **97**, 050404 (2006).
[12] I. Klich, G. Refael, and A. Silva, Phys. Rev. A **74**, 032306 (2006).
[13] S. R. White, Phys. Rev. Lett. **69**, 2863 (1992); Phys. Rev. B **48**, 10345 (1993).
[14] S. Ostlund and S. Rommer, Phys. Rev. Lett. **75**, 3537 (1995).
[15] G. Vidal, *Phys. Rev. Lett.* **99**, 220405 (2007); G. Evenbly and G. Vidal, *Phys. Rev. Lett.* **104**, 187203 (2010).
[16] S. Capponi, *Private communication*
[17] M. Weinstein, *Phys. Rev. D* **47**, 5499 (1993).
[18] M. Weinstein, A. Auerbach, V. Ravi Chandra (unpublished)
[19] A.M. Läuchli, J. Sudan, E.S. Sorensen, arXiv:1103.1159.
[20] A power law can fit the numerical entanglement spectra asymptotics of two dimensional systems computed by e.g. M-C. Chung and I. Peschel , Phys. Rev. B **64**, 064412 (2001). H. Li and F. D. M. Haldane, Phys. Rev. Lett. **101**, 010504 (2008).