Projector Quantum Monte Carlo Method for Nonlinear Wave Functions

Lauretta R. Schwarz,[1,*] A. Alavi,[2,†] and George H. Booth[3,‡]
[1]University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
[2]Max Planck Institute for Solid State Research, Heisenbergstraße 1, 70569 Stuttgart, Germany and University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
[3]Department of Physics, King's College London, Strand, London WC2R 2LS, United Kingdom
(Received 28 October 2016; published 25 April 2017)

We reformulate the projected imaginary-time evolution of the full configuration interaction quantum Monte Carlo method in terms of a Lagrangian minimization. This naturally leads to the admission of polynomial complex wave function parametrizations, circumventing the exponential scaling of the approach. While previously these functions have traditionally inhabited the domain of variational Monte Carlo approaches, we consider recent developments for the identification of deep-learning neural networks to optimize this Lagrangian, which can be written as a modification of the propagator for the wave function dynamics. We demonstrate this approach with a form of tensor network state, and use it to find solutions to the strongly correlated Hubbard model, as well as its application to a fully periodic *ab initio* graphene sheet. The number of variables which can be simultaneously optimized greatly exceeds alternative formulations of variational Monte Carlo methods, allowing for systematic improvability of the wave function flexibility towards exactness for a number of different forms, while blurring the line between traditional variational and projector quantum Monte Carlo approaches.

DOI: 10.1103/PhysRevLett.118.176403

The description of quantum many-body states in strongly correlated systems is central to understanding a wealth of complex emergent phenomena in condensed matter physics and quantum chemistry. The problem is well defined; the Hamiltonian is known, and the solution is a linear superposition of all possible classical configurations of particles. However, this conceals exponential complexity in the wave function, which in general prohibits both storage and manipulation of these linear coefficients.

To deal with this exponentially large Hilbert space, one approach is to sample the space stochastically. For studies of the ground state of quantum systems, this is broadly split into two separate categories, *projector* (PMC) and *variational* Monte Carlo (VMC) methods [1,2]. In PMC, a decaying function of the Hamiltonian is continually applied to a stochastic representation of the full wave function. This projects out the higher-energy components, leaving a stochastic sampling of the dominant (generally ground-state) eigenfunction. By contrast, in VMC a polynomial-complex approximate wave function *Ansatz* is imposed, generally with a small number of variational parameters. State-of-the-art methods to optimize this wave function then involve sampling the gradient and Hessian of the energy with respect to the parameters in the tangent space of the current wave function. This is done by projecting into and sampling from the exponential configurational space. Once a stochastic representation of these quantities is obtained, updates to the wave function parameters are found by a variety of iterative techniques until convergence of this nonlinear parametrization is achieved.

One promising emerging technique is the full configuration interaction quantum Monte Carlo (FCIQMC) approach, a projector quantum Monte Carlo method that stochastically samples both the wave function and the propagator in Fock space [3,4]. By exploiting sparsity inherent in the wave function of many representations of quantum systems, essentially exact results can be obtained with only small fractions of the Hilbert space simultaneously occupied. However, despite often admitting highly accurate solutions for systems far out of reach of many alternative approaches, the method is formally exponentially scaling with system size, albeit often weakly. In order to advance to larger and condensed phase systems, one approach is to exploit the fact that electron correlation is, in general, inherently local. Two-point correlation functions (away from criticality) will decay exponentially with distance, while the screening of the Coulomb interaction in bulk systems will result in local entanglement of nearby electrons, with distant electrons behaving increasingly independently [5].

Following the success of the FCIQMC approach for finite systems, we aim to exploit this locality, to formally contain the scaling to polynomial cost. This is done by imposing a nonlinear, yet systematically improvable, *Ansatz* of the form of a correlator product state (CPS), which explicitly correlates plaquettes of locally neighboring degrees of freedom [6,7]. Related wave functions have also been called entangled plaquette states or complete graph tensor networks

to stress their connection to tensor network states [8–11]. In formulating this, we develop connections between projector and variational quantum Monte Carlo approaches, and propose new methodology for the optimization of arbitrary nonlinear wave function parametrizations. This approach is shown to confer a number of benefits compared to state-of-the-art wave function optimization [12–16]. The number of parameters that can be handled even brings into scope more sophisticated wave functions, including other tensor network parametrizations [17,18]. We apply this approach to a number of model and *ab initio* systems, showing that systematic improvability and exceedingly large parameter spaces can be handled within these complex optimization problems.

The CPS wave function defines "correlators" as diagonal operators (to optimize) which directly encode the entanglement within sets of single-particle states (which in this work are exclusively neighboring), as $\hat{C}_\lambda = \sum_{\mathbf{n}_\lambda} C_{\mathbf{n}_\lambda} \hat{P}_{\mathbf{n}_\lambda}$, where $\hat{P}_{\mathbf{n}_\lambda} = |\mathbf{n}_\lambda\rangle\langle\mathbf{n}_\lambda|$ is the projection operator for the set of *all* many-body Fock states $\mathbf{n}_\lambda$ in the correlator λ, with adjustable amplitudes $C_{\mathbf{n}_\lambda}$. The CPS is then written as a multilinear product of correlators acting on a chosen reference state $|\Phi\rangle$. In this work, this reference state is a single Slater determinant (which can also be variationally optimized), but other reference states are possible [19,20]. The final CPS wave function is, therefore, represented as $|\Psi_{\mathrm{CPS}}\rangle = \prod_\lambda \hat{C}_\lambda |\Phi\rangle$. It can be shown that a number of different phases and wave functions can be expressed in this form, including resonanting valence bond (RVB) and Laughlin wave functions [6]. As the number of degrees of freedom in the system grows, the complexity of the wave function grows only linearly. Additionally, this choice of low-rank factorization of the wave function is systematically improvable with increasing correlator size as it recovers longer-ranged entanglement effects, but this admits many variables to optimize. VMC techniques have been used previously for similar tensor network forms, but the growth of parameters has led to limited success in recovering long-range entanglement or thermodynamic limit results [17,18]. We now consider a new, efficient approach to handle these many parameters, derived in part from the FCIQMC approach, which can be considered as the limit of a single large correlator.

Combining PMC and VMC approaches.—The FCIQMC (and some other PMC [21]) methods are simulated through stochastic dynamics given by

$$|\Psi_0\rangle = \lim_{k\to\infty} [1 - \tau(\hat{H} - \hat{I}E_0)]^k |\psi^{(0)}\rangle, \tag{1}$$

with τ chosen to be sufficiently small, where Ψ_0 is the ground state of the system and E_0 is the self-consistently obtained ground-state energy [3]. This can be considered both as a first-order approximation to imaginary time dynamics as $e^{-\beta\hat{H}}|\psi^{(0)}\rangle$ and as a power method to project out the dominant, lowest-energy eigenvector of $\hat{H}$ [22]. Alternatively, a VMC perspective considers finding the variational minimum of the Ritz functional $\langle\Psi|\hat{H}|\Psi\rangle / \langle\Psi|\Psi\rangle$ through optimization of the wave function parameters.

These approaches can be shown to be analogous by considering the minimization of a positive-definite Lagrangian,

$$\mathcal{L}[\Psi(Z_\sigma)] = \langle\Psi|\hat{H}|\Psi\rangle - E_0(\langle\Psi|\hat{I}|\Psi\rangle - A), \tag{2}$$

where normalization (A) is enforced by a Lagrange multiplier, which at convergence is given by E_0. It is simple to show that the minimum of this functional is the same as that given by the Ritz functional. We can consider a simple gradient descent minimization of all variational parameters, $\{Z_\sigma\}$ in Eq. (2), with step size τ_k, as

$$Z_\sigma^{(k+1)} = Z_\sigma^{(k)} - \tau_k \frac{\partial\mathcal{L}[\Psi^{(k)}]}{\partial Z_\sigma}. \tag{3}$$

Projecting the equations into the full Hilbert space of configurations $\{|\mathbf{m}\rangle\}$, we obtain

$$Z_\sigma^{(k+1)} = Z_\sigma^{(k)} - \tau_k \sum_{\mathbf{nm}} \left\langle \frac{\partial\Psi^{(k)}}{\partial Z_\sigma} \middle| \mathbf{m} \right\rangle \times (H_{\mathbf{mn}} - E^{(k)}\delta_{\mathbf{mn}})\langle\mathbf{n}|\Psi^{(k)}\rangle. \tag{4}$$

If the chosen wave function is an expansion of linearly independent configurations, then this will return exactly the "imaginary-time" dynamics of Eq. (1) and the FCIQMC master equations, demonstrating the deep connection between imaginary-time propagation, gradient descent, and the power method [23].

However, here we aim to go beyond this. In keeping with the FCIQMC approach, the summations are replaced by random samples of both the wave function and Hamiltonian connections. The sum over $\{\mathbf{n}\}$ is stochastically sampled via a Metropolis Markov chain, to evaluate a stochastic representation of the wave function [22,24–27]. Each iteration consists of 100 000–200 000 random samples of the wave function. For each, a small selection of configurations $\{\mathbf{m}\}$ are sampled from the set of nonzero connections via $H_{\mathbf{mn}}$ in the manner of the FCIQMC approach, while appropriately unbiasing for the probability of this selection [28,29]. Furthermore, the derivatives $\langle(\partial\Psi^{(k)}/\partial Z_\sigma)|\mathbf{m}\rangle$ can be efficiently evaluated from the respective wave function amplitudes $\langle\Psi^{(k)}|\mathbf{m}\rangle$. Technical details on the sampling of this gradient can be found in the Supplemental Material [30].

This stochastic gradient descent (SGD) of the Lagrangian results in an iteration cost that is independent of the size of the Hilbert space and thus renders this methods inherently suitable for large scale systems. It also

admits a number of advantages over state-of-the-art VMC optimization [12–14], such as the avoidance of the construction of matrices in the tangent space, whose sampling and manipulation becomes a bottleneck for large numbers of parameters. While Krylov subspace techniques have been proposed to circumvent this by projecting down to more manageable spaces [15], ill conditioning can limit the efficiency of this approach [16]. Furthermore, diagonalization of the randomly sampled matrices required in some optimizations can lead to biases in the final parameters [31,32]. Our approach also bears similarities with stochastic reconfiguration [13,14], which can also be considered an imaginary time propagation that differs from SGD in its definition of the metric for the updates [33]. Because of this, stochastic reconfiguration also requires projection of the equations into the tangent space of the current wave function and stabilization of the resultant matrix equations [14]. However, the proposed matrix-free stochastic application of Eq. (3) describes a quasicontinuous optimization, where the error bar at convergence represents both the stochastic error in the sampling and fluctuations in the wave function. In addition, the dynamic also provides a straightforward route to unbiased computation of the two-body reduced density matrix [34,35], $\Gamma_{pq,rs} = \langle \Psi | a_p^\dagger a_q^\dagger a_s a_r | \Psi \rangle$. By evaluating $\langle Q \rangle = \mathrm{Tr}[\Gamma \hat{Q}]$, arbitrary one- and two-body static properties can be found. This includes the energy, spin, and magnetic properties which here are computed from the density matrix, rather than from the local energy as is common in VMC calculations.

However, similar SGD approaches have been considered before with little success for large numbers of variables, due to the slow convergence of the parameters as $\mathcal{O}[(1/k) + (\sigma/\sqrt{k})]$, where σ is the variance in the gradient [36,37]. Improving on this involves advances in SGD methods, used in the field of deep-learning algorithms of neural networks [38,39]. Analogously, these networks represent a flexible nonlinear function with parameters to be optimized via minimization of a cost function, often achieved via SGD schemes similar to the one in Eq. (3) [40,41].

The convergence can be accelerated via the addition of a "momentum," whereby the update retains a memory of the previous updates. Propagation then results in the accumulation of velocity in the direction of persistent decrease in energy, thereby accelerating the update in directions of low curvature over multiple iterations [42], formally accelerating the convergence rate to a second-order $\mathcal{O}[(1/k^2) + (\sigma/\sqrt{k})]$. Mathematically, the stochastic projection is given by a monic polynomial of the propagator, such that $\mathbf{\Psi}^{(k)} = p_A^k(\mathbf{A})\mathbf{\Psi}^{(0)}$. In the SGD scheme of Eq. (1), this is akin to the power method. However, the optimal projection will be a polynomial approximation to a function whose value at the desired eigenvalue of the propagator is one, and whose maximum absolute value in the range of the rest of the spectrum is minimized. This is best represented by using a shifted and scaled Chebyshev polynomial approximation to the projection. The success of the Lanczos approach as a second-order optimization, as well as other deterministic projections, can also be rationalized in this fashion [43,44].

An optimal version of this projector can be formulated as Nesterov's accelerated approach [45], whereby the sequence $\lambda_0 = 0, \quad \lambda_k = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\lambda_{k-1}^2}, \quad \gamma_k = (1 - \lambda_k)/\lambda_{k+1}$ is defined, and starting at an initial point $Z_\sigma^{(1)} = Y_\sigma^{(1)}$, the algorithm stochastically iterates the equations [46]

$$Y_\sigma^{(k+1)} = Z_\sigma^{(k)} - \tau_k \frac{\partial \mathcal{L}[\Psi^{(k)}]}{\partial Z_\sigma}, \tag{5}$$

$$Z_\sigma^{(k+1)} = (1 - \gamma_k) Y_\sigma^{(k+1)} + \gamma_k Y_\sigma^{(k)}, \tag{6}$$

for $k \geq 1$. While an optimal projection overall, this is no longer a gradient descent scheme, and as such there is no requirement that each iteration will decrease the energy, and instabilities can be observed [47,48]. To mitigate this behavior, we have found it beneficial to include a damping for the momentum d as $\gamma_k \to \gamma_k e^{-(1/d)(k-1)}$ [47,49]. With a suitably chosen damping parameter the rate of convergence of the optimization should not be hindered, since this is dominated in the latter stages by the $\sigma/\sqrt{k}$ term for both accelerated and conventional gradient descent [50].

The remaining arbitrariness concerns the step size (or learning rate) τ_k. While decreasing the step size generally improves robustness, it slows convergence and increases autocorrelation time [40,41]. We found optimal convergence and accuracy achieved with a deep-learning technique denoted RMSprop [51], an adaptive step size method which dynamically estimates an independent $\tau_{Z_\sigma}^{(k)}$ for each parameter. This gives $\tau_{Z_\sigma}^{(k)} = \eta(\mathrm{RMS}[g_{Z_\sigma}]^{(k)})^{-1}$, where η is a global parameter for all variables and $\mathrm{RMS}[g_{Z_\sigma}]^{(k)}$ represents the root mean square of previous gradients for the variable up to the current iteration, $\mathrm{RMS}[g_{Z_\sigma}]^{(k)} = \sqrt{E[g_{Z_\sigma}^2] + \epsilon}$, evaluated by accumulating an exponentially decaying average of the squared gradients of the Lagrangian g: $E[g_{Z_\sigma}^2]^{(k)} = \rho E[g_{Z_\sigma}^2]^{(k-1)} + (1 - \rho) g_{Z_\sigma}^2$. The small constant ϵ is added to better condition the denominator and ρ is the decay constant. This dynamically adaptive, parameter-specific step size acts much like a preconditioner for the system, and allows the optimization to take larger steps for those parameters with small and consistent gradients, and vice versa. This ensures robustness of the algorithm to large changes in gradients due to the stochastic nature of the gradient evaluation.

Results.—The demonstration of the ability of the algorithm to converge wave functions with many parameters is shown in Fig. 1, which considers a 98-site 2D Hubbard model at half filling, with $U/t = 8$. In this study, independent, overlapping five-site correlators centered on every site in the lattice were chosen to correlate with nearest neighbors, allowing up to ten-electron short-ranged

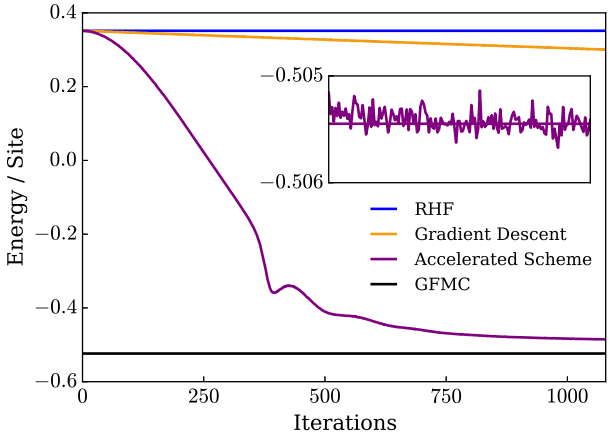

FIG. 1. Convergence of CPS with $\mathcal{O}[10^5]$ parameters for SGD and accelerated scheme with RMSProp algorithm for the 98-site (tilted) 2D Hubbard model at $U = 8t$. Green's-function Monte Carlo (GFMC) energy is taken from Ref. [52], while the wavefunction is initialized at the restricted Hartree-Fock (RHF) solution, as given above. Inset shows fluctuations both in the statistical sampling of expectation values and in the variation of the parameters.

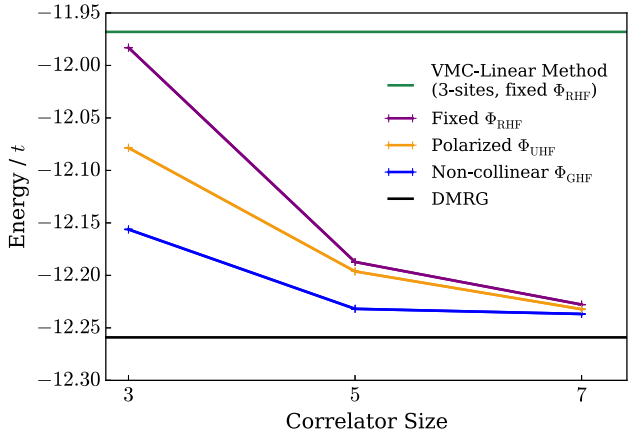

FIG. 2. Convergence of energy for a range of Ψ_{CPS} for 1×22 Hubbard model. VMC linear method and DMRG energies are taken from Ref. [53]. Error bars are too small to be visible.

correlation to be directly captured, as well as long-range correlation and symmetry breaking through coupling between the overlapping correlators and the optimization of the Slater determinant. The lattice and tiling of these correlator plaquettes is depicted in the Supplemental Material [30]. Accurate results for this system are given by the Green's-function Monte Carlo technique [52]. Our CPS captures 97.9% of this correlation energy, with the remaining likely to be due to the lack of direct long-range two-body correlation. However, this parametrization still requires the simultaneous optimization of over 10^5 parameters, beyond the capabilities of most VMC implementations, and demonstrates a striking advance in the rate of convergence afforded by the accelerated algorithm.

To consider the systematic improvability of the CPS *Ansätze*, we consider the 1D, 22-site Hubbard model, such that benchmark data can be found from the density matrix renormalization group (DMRG), which can be made numerically exact for this 1D system [53]. Results at half filling and $U = 4t$ are shown in Fig. 2. For a wave function of three-site overlapping correlators and a fixed, noninteracting reference, we find a variationally lower result than previously published for an identical parametrization via linear method optimization [12,53]. This could be due to the bias from the nonlinear operations (diagonalization) of random variables present in these alternate algorithms [31,32]. We also investigate how increasing the size of the correlators in order to *directly* capture longer-ranged many-body correlation, as well as optimizing "unrestricted" spin-polarized (Φ_{UHF}) or "generalized" noncollinear (Φ_{GHF}) Slater determinants rather than a paramagnetic orbital component (Φ_{RHF}), affects the quality of the wave function. The increased flexibility of this democratic wave function gives rise to systematic convergence towards DMRG with very small error bars, despite requiring over 250 000 variables.

Ab initio systems can also be well treated in the same vein, stochastically sampling from both the configuration space of the wave function and from its $\mathcal{O}[N^4]$ connected configurations in Eq. (4), which are now far larger than found in the Hubbard model due to long-range interactions. In Fig. 3, we consider the symmetric dissociation of H_{50}, a molecular model for strongly correlated systems and a nontrivial benchmark system [54]. This system has been treated not only with conventional quantum chemistry methods such as coupled cluster (which fail to converge at stretched bond

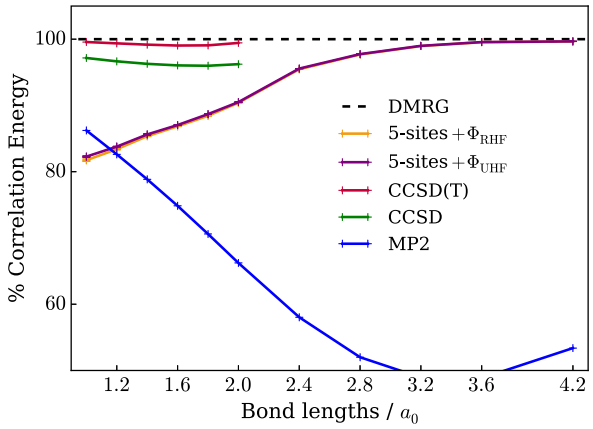

FIG. 3. Percentage of DMRG correlation energy captured by Ψ_{CPS} for the symmetric dissociation of a linear chain of 50 hydrogen atoms in a STO-6G basis. Numerically exact DMRG, as well as high-level correlated quantum chemical methods of Møller-Plesset perturbation theory (MP2), coupled cluster up to double excitations (CCSD) and with perturbative triple excitations [CCSD(T)] are included, with values taken from Ref. [55]. The largest deviation in the total energy compared to DMRG across all bond lengths shown is 1.1 kcal/mol per atom.

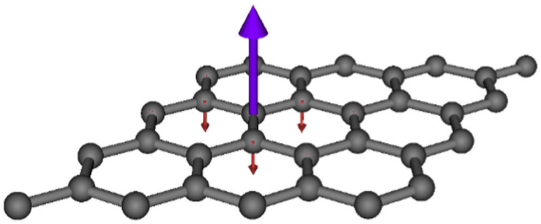

FIG. 4. Spin correlation function $\langle \Psi_{\mathrm{CPS}} | \mathbf{S}_i \cdot \mathbf{S}_j | \Psi_{\mathrm{CPS}} \rangle$ of graphene in the p_z space with a six-site CPS (with i as the atomic site with maximal spin) [61].

lengths beyond $2.0a_0$)[55], but also strongly correlated approaches including dynamical mean-field theory and other embedding methods [56–58], due to the availability of numerically exact DMRG values for comparison [55]. We parametrize our CPS with five-atom overlapping correlators, and both a fixed unpolarized reference and a stochastically optimized unrestricted reference determinant. At stretched bond lengths, nearly all of the DMRG correlation energy is captured, as the correlation length spans few atoms, and on-site repulsion dominates. However, as the bond length decreases, a successively smaller percentage of the DMRG correlation energy is captured, as the entanglement of the electrons spans larger numbers of atoms, as can also be seen in the larger bond dimension required of DMRG at these geometries [55]. Despite this, the correlation energy is so small at these lengths that the maximum error in the total energy is only 1.1 kcal/mol per atom, achieving chemical accuracy for the stretching of this system.

Fully periodic localized orbitals can also be used to construct a Fock space in which to form a CPS, and here we consider an infinitely periodic graphene sheet with 4×4 k-point sampling [59]. From a double-zeta periodic Gaussian basis, we choose one localized, translationally invariant $2p_z$ orbital centered on each carbon atom. Overlapping correlators consisting of the atoms on each hexagonal six-membered ring can then be constructed and the full Hamiltonian projected into this low-energy space, including a potential from the core electrons at the Hartree-Fock level [60]. A generalized reference determinant is then stochastically optimized along with the correlators, giving a wave function parametrization of 67,584 parameters—we believe the largest number of nonlinear parameters for an *ab initio* system to date. This is equivalent to a quantum chemical calculation of a complete active space of 32 orbitals, which is beyond that which could be treated by conventional techniques. This spans the dominant strong correlation effects, but precludes high-energy many-body dynamic correlation and screening.

From the sampled density matrix, we can construct the spin correlation function to analyze the extent to which spin fluctuations among the π/π^* bands around the Fermi level affect the magnetic order of the system. The spin correlation functions are constructed from two-point functions, rather than from symmetry breaking in the wave function, and show a rapid decay of antiferromagnetic correlations which only substantially affect nearest neighbors (Fig. 4).

Conclusions.—In this work we have presented a novel approach to sample and optimize arbitrary nonlinear wave functions of many-body quantum systems. The optimization is written as an accelerated propagator inspired by ideas from developments in deep-learning algorithms and the FCIQMC approach. This allows for large numbers of parameters to be handled, and systematically improvable Fock-space wave functions to be used in both lattice and *ab initio* systems.

The calculations made extensive use of computing facilities of the Rechenzentrum Garching of the Max Planck Society with calculation data available from Ref. [62].

G. H. B. gratefully acknowledges funding from the Royal Society via a University Research Fellowship, as well as support from the Air Force Office of Scientific Research via Grant No. FA9550-16-1-0256. A. A. acknowledges support from the EPSRC, Grant No. EP/J003867/1. L. R. S. is supported by an EPSRC studentship.

*lrs37@cam.ac.uk
†A.Alavi@fkf.mpg.de
‡george.booth@kcl.ac.uk

[1] *Quantum Monte Carlo Methods in Physics and Chemistry*, NATO ASI Series C, Vol. 525, edited by M. P. Nightingale and C. J. Umrigar (Kluwer, Dordrecht, 1999).
[2] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).
[3] G. H. Booth, A. J. W. Thom, and A. Alavi, J. Chem. Phys. **131**, 054106 (2009).
[4] G. H. Booth, A. Gruneis, G. Kresse, and A. Alavi, Nature (London) **493**, 365 (2013).
[5] J. Eisert, M. Cramer, and M. B. Plenio, Rev. Mod. Phys. **82**, 277 (2010).
[6] H. J. Changlani, J. M. Kinder, C. J. Umrigar, and G. K.-L. Chan, Phys. Rev. B **80**, 245116 (2009).
[7] E. Neuscamman and G. K.-L. Chan, Phys. Rev. B **86**, 064402 (2012).
[8] F. Mezzacapo, N. Schuch, M. Boninsegni, and J. I. Cirac, New J. Phys. **11**, 083026 (2009).
[9] F. Mezzacapo and J. I. Cirac, New J. Phys. **12**, 103039 (2010).
[10] K. H. Marti, B. Bauer, M. Reiher, M. Troyer, and F. Verstraete, New J. Phys. **12**, 103008 (2010).
[11] K. H. Marti and M. Reiher, Phys. Chem. Chem. Phys. **13**, 6750 (2011).
[12] J. Toulouse and C. J. Umrigar, J. Chem. Phys. **126**, 084102 (2007).
[13] S. Sorella, Phys. Rev. B **71**, 241103 (2005).
[14] S. Sorella, M. Casula, and D. Rocca, J. Chem. Phys. **127**, 014105 (2007).
[15] E. Neuscamman, C. J. Umrigar, and G. K.-L. Chan, Phys. Rev. B **85**, 045103 (2012).
[16] Z. Luning and E. Neuscamman, arXiv:1702.01481.

[17] A. W. Sandvik and G. Vidal, Phys. Rev. Lett. **99**, 220602 (2007).
[18] O. Sikora, H.-W. Chang, C.-P. Chou, F. Pollmann, and Y.-J. Kao, Phys. Rev. B **91**, 165113 (2015).
[19] E. Neuscamman, J. Chem. Phys. **139**, 194105 (2013).
[20] M. Casula and S. Sorella, J. Chem Phys. **119**, 6500 (2003).
[21] M. Casula, C. Filippi, and S. Sorella, Phys. Rev. Lett. **95**, 100201 (2005).
[22] C. J. Umrigar, J. Chem. Phys. **143**, 164105 (2015).
[23] S. Bubeck, arXiv:1405.4980.
[24] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[25] W. K. Hastings, Biometrika **57**, 97 (1970).
[26] R. M. Lee, G. J. Conduit, N. Nemec, P. López Ríos, and N. D. Drummond, Phys. Rev. E **83**, 066706 (2011).
[27] J. R. Trail and R. Maezono, J. Chem. Phys. **133**, 174120 (2010).
[28] G. H. Booth, S. D. Smart, and A. Alavi, Mol. Phys. **112**, 1855 (2014).
[29] A. A. Holmes, H. J. Changlani, and C. J. Umrigar, J. Chem. Theory Comput. **12**, 1561 (2016).
[30] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.118.176403 for a description on the sampling of the lagrangian gradient.
[31] N. S. Blunt, A. Alavi, and G. H. Booth, Phys. Rev. Lett. **115**, 050603 (2015).
[32] L. Zhao and E. Neuscamman, J. Chem. Theory Comput. **12**, 3719 (2016).
[33] M. Casula, C. Attaccalite, and S. Sorella, J. Chem. Phys. **121**, 7110 (2004).
[34] L. K. Wagner, J. Chem. Phys. **138**, 094106 (2013).
[35] C. Overy, G. H. Booth, N. S. Blunt, J. J. Shepherd, D. Cleland, and A. Alavi, J. Chem. Phys. **141**, 244117 (2014).
[36] A. Harju, B. Barbiellini, S. Siljamäki, R. M. Nieminen, and G. Ortiz, Phys. Rev. Lett. **79**, 1173 (1997).
[37] H. Robbins and S. Monro, Ann. Math. Stat. **22**, 400 (1951).
[38] M. A. Nielsen, *Neural Networks and Deep Learning* (Determination Press, 2015).
[39] V. Dunjko, J. M. Taylor, and H. J. Briegel, Phys. Rev. Lett. **117**, 130501 (2016).
[40] D. R. Wilson and T. R. Martinez, in *Proceedings of the 2001 International Joint Conference on Neural Networks (IJCNN '01), Washington, DC* (Neural Networks, 2001), Vol. 1, pp. 115–119 .
[41] R. A. Jacobs, Neural Netw. **1**, 295 (1988).
[42] N. Qian, Neural Netw. **12**, 145 (1999).
[43] J. Cullum and R. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations* (Birkhäuser, Boston, 1985), Vol. 2.
[44] T. Zhang and F. A. Evangelista, J. Chem. Theory Comput. **12**, 4326 (2016).
[45] Y. Nesterov, Sov. Math. Dokl. **27**, 372 (1983).
[46] A. Beck and M. Teboulle, SIAM J. Imag. Sci. **2**, 183 (2009).
[47] W. Su, S. Boyd, and E. Candes, in *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014), Vol. 27, pp. 2510–2518.
[48] A. Beck and M. Teboulle, IEEE Trans. Image Process. **18**, 2419 (2009).
[49] B. O'Donoghue and E. Candès, Found. Comput. Math. **15**, 715 (2015).
[50] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, edited by S. Dasgupta and D. Mcallester (JMLR.org, Atlanta, GA, 2013), Vol. 28, pp. 1139–1147.
[51] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, arXiv:1502.04390.
[52] J. P. F. LeBlanc *et al.* (Simons Collaboration on the Many-Electron Problem), Phys. Rev. X **5**, 041041 (2015).
[53] E. Neuscamman, H. Changlani, J. Kinder, and G. K.-L. Chan, Phys. Rev. B **84**, 205132 (2011).
[54] W. J. Hehre, R. F. Stewart, and J. A. Pople, J. Chem. Phys. **51**, 2657 (1969).
[55] J. Hachmann, W. Cardoen, and G. K.-L. Chan, J. Chem. Phys. **125**, 144101 (2006).
[56] T. Tsuchimochi and G. E. Scuseria, J. Chem. Phys. **131**, 121102 (2009).
[57] K. Boguslawski, P. Tecmer, P. W. Ayers, P. Bultinck, S. De Baerdemacker, and D. Van Neck, Phys. Rev. B **89**, 201106 (2014).
[58] N. Lin, C. A. Marianetti, A. J. Millis, and D. R. Reichman, Phys. Rev. Lett. **106**, 096402 (2011).
[59] G. H. Booth, T. Tsatsoulis, G. K.-L. Chan, and A. Grüneis, J. Chem. Phys. **145**, 084111 (2016).
[60] R. E. Thomas, Q. Sun, A. Alavi, and G. H. Booth, J. Chem. Theory Comput. **11**, 5316 (2015).
[61] H. Childs, E. Brugger, B. Whitlock, J. Meredith, S. Ahern, D. Pugmire, K. Biagas, M. Miller, C. Harrison, G. H. Weber, H. Krishnan, T. Fogal, A. Sanderson, C. Garth, E. W. Bethel, D. Camp, O. Rübel, M. Durant, J. M. Favre, and P. Navrátil, *High Performance Visualization-Enabling Extreme-Scale Scientific Insight* (2012), pp. 357–372.
[62] See DOI:10.17863/CAM.8161.