PHYSICAL REVIEW
LETTERS

VOLUME 79 | 18 AUGUST 1997 | NUMBER 7

Stochastic Gradient Approximation: An Efficient Method to Optimize Many-Body Wave Functions

A. Harju,* B. Barbiellini, S. Siljamäki, and R. M. Nieminen
Laboratory of Physics, Helsinki University of Technology, FIN-02150 Espoo, Finland

G. Ortiz†
Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801
(Received 13 September 1996)

A novel, efficient optimization method for physical problems is presented. The method utilizes the noise inherent in stochastic functions. As an application, an algorithm for the variational optimization of quantum many-body wave functions is derived. The numerical results indicate superior performance when compared to traditional techniques. [S0031-9007(97)03868-4]

PACS numbers: 02.60.Pn, 31.25.–v, 71.10.–w, 71.60.+z

Optimization in the presence of noise is a difficult task. Most of the optimization methods available are totally deterministic in nature, and, when applied to problems affected by noise, they are either unable to reach an optimum or they may reach a false one. In this Letter we present a novel optimization scheme, called the stochastic gradient approximation (SGA). The method has its roots in the mathematics of automatic control theory [1,2], but to the authors' knowledge it has not been applied for optimization problems in physics before.

The algorithm belongs to the class of probabilistic iterative methods with variable step size. We consider one important application, namely, the optimization of many-body wave functions using the variational Monte Carlo (VMC) method. The results obtained show conclusively that the SGA constitutes a method tailor-made for quantum Monte Carlo (QMC) techniques. The excellent performance obtained makes the SGA an attractive tool to other difficult optimization problems as well.

Quantum Monte Carlo methods are powerful tools for studying interacting many-particle systems. For fermion systems, the fixed-node diffusion QMC (DMC) can be thought of as a supervariational approach with an energy which is guaranteed to be closer to the exact answer than the starting VMC parent wave function [3]. For a given nodal surface the DMC provides the lowest energy compatible with such a constraint. In atomic and molecular systems, the energies computed with the DMC are comparable in accuracy to the ones obtained using traditional configuration-interaction approaches [3]. This is remarkable when one realizes that in DMC very simple and compact wave functions are used. A popular choice is the Slater-Jastrow form with molecular orbitals from a mean-field calculation, and a parametrized bosonic correlation factor. In such a case, the nodal structure is determined solely by the one-body molecular orbitals. The study of Si_n clusters by Grossman and Mitáš [4] clearly indicates the importance of the optimization of such orbitals, as done in, for example, [5]. Optimization of the full many-body wave function is crucial for the success not only of the VMC method but of the DMC itself. However, particularly for complex molecules, this is a very time-consuming process. It is clear that an efficient optimization scheme is a very important ingredient for the ultimate success of the QMC methods.

Suppose that the quantity one is interested in optimizing is given by

$$\mathcal{F}[\boldsymbol{\alpha}] = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} \mathcal{Q}(\mathbf{R}_j; \boldsymbol{\alpha}), \qquad (1)$$

where $\mathbf{R}_j$ is given by a random sequence, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ represents the vector of n parameters to be optimized, and $\mathcal{Q}$ is the cost function. The standard way is to take k large and use the numerical approximation of $\mathcal{F}$

as a deterministic function. The approach of SGA is different. The basic idea is to define a sequence of vectors $\boldsymbol{\alpha}_i$ by the recursive algorithm

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i-1} - \gamma_i \nabla_{\boldsymbol{\alpha}} \mathcal{Q}(\mathbf{R}_j; \boldsymbol{\alpha}_{i-1}), \tag{2}$$

where γ_i is a weighting factor. Notice that the change in the parameters is given by the unbiased *stochastic approximation of the gradient* and not the gradient of the functional $\mathcal{F}$, which remains unknown. Even for the optimal parameters $\boldsymbol{\alpha}^*$ we have $\nabla_{\boldsymbol{\alpha}} \mathcal{Q}(\mathbf{R}; \boldsymbol{\alpha}^*) \not\equiv 0$. The weighting factors γ_i should be such that the recursion converges to the optimum parameters in a stochastic sense. In fact, γ_i should satisfy the conditions [1]

$$\sum_{i=1}^{\infty} \gamma_i^2 < \sum_{i=1}^{\infty} \gamma_i = \infty. \tag{3}$$

There is a simple interpretation for these conditions: The sum of γ^2 should be finite to dissipate the cumulative error given by the noise in the approximative gradient and the sum of γ should diverge, because otherwise the maximum distance from the initial parameters would be limited. The simplest choice is to use $\gamma_i = 1/i$. With this choice one should notice the resemblance to the recursive calculation of a mean: $\bar{x}_i = \bar{x}_{i-1} - \frac{1}{i}(\bar{x}_{i-1} - x_i)$.

For the sake of clarity, and because of the envisaged applications, let us sketch the steps to follow in a VMC calculation. Suppose that the function one wants to optimize is the mean value of a local function $\mathcal{O}_{\mathrm{L}}[\mathbf{R}; \boldsymbol{\alpha}] = \Psi^{-1} \mathcal{O} \Psi$, where $\mathcal{O}$ represents a physical observable. This can be, for example, the local energy $\mathcal{E}_{\mathrm{L}} = \Psi^{-1} H \Psi$, where H is the Hamiltonian and Ψ is the N-body wave function. Other examples would be the variance of the local energy or a linear combination of the local energy and its variance. The expectation value of the operator $\mathcal{O}$ is calculated as a mean of the local function at configurations distributed according to Ψ^2. Then the SGA method can be summarized by the following recursive algorithm:

(1) Metropolis sampling is done with parameters $\boldsymbol{\alpha}_{i-1}$ starting from a sample of m configurations $\{\mathbf{R}_j\}_{i-1}$ to obtain a new sample $\{\mathbf{R}_j\}_i$ of the same size. We use i for characterizing the iteration number and j for different configurations in a sample.

(2) Update variational parameters according to

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i-1} - \gamma_i \nabla_{\boldsymbol{\alpha}} \left\{ \frac{1}{m} \sum_{j=1}^{m} \mathcal{O}_{\mathrm{L}}[\{\mathbf{R}_j\}_i; \boldsymbol{\alpha}_{i-1}] \right\}, \tag{4}$$

where γ_i is as previously.

While iterating, the sample of configurations follows the change in the parameters. In this way there is no bias due to the use of a fixed set. One important technical remark is that if one uses finite differences for the calculation of the gradient, one should note that the configurations are distributed according to $|\Psi(\boldsymbol{\alpha})|^2$ and *not* according to $|\Psi(\boldsymbol{\alpha} \pm \Delta)|^2$, where Δ represents a small change. Thus, the expectation value of the operator $\mathcal{O}$ is calculated as

$$\frac{1}{\tilde{m}} \sum_{j=1}^{m} w_j \mathcal{O}_{\mathrm{L}}[\mathbf{R}_j; \boldsymbol{\alpha} \pm \Delta], \tag{5}$$

where the points $\mathbf{R}_j$ are distributed according to $|\Psi(\boldsymbol{\alpha})|^2$, $w_j = |\Psi(\mathbf{R}_j; \boldsymbol{\alpha} \pm \Delta)/\Psi(\mathbf{R}_j; \boldsymbol{\alpha})|^2$, and $\tilde{m} = \sum w_j$. The "weights" w_j of the local functions are very close to unity, because Δ is only a small change.

An important feature of the SGA is that it is less sensitive to the local optima than the traditional steepest-descent methods. The noise in the gradient helps the fictitious dynamics of the parameters not to get stuck in local minima. In fact, the SGA has some similarity to the simulated annealing technique. As we will see, another remarkable property of the SGA method is its scalability; i.e., the size of the sample m and the number of iterations it takes to converge to the optimum are almost independent of the physical size (N) of the system. The number of configurations m controls the amount of noise. As m gets smaller the noise in the approximate gradient increases. In the examples below, we found that the method is more efficient when m is of the order of one. The reason is that the SGA takes advantage of the stochastic noise to perform the global minimization while other methods try to get rid of it. In the following, we have used $m \approx 5$.

To illustrate the method, we will first consider the ground state of a He atom. In this whole example, we will use as cost function the local energy $\mathcal{E}_{\mathrm{L}}$. Umrigar *et al.* [6] have shown that variance minimization is generally more efficient, but in some cases energy minimization is required. The simplest variational wave function is a product of two hydrogenlike $1s$ orbitals with an effective charge which plays the role of variational parameter. The optimal value of α is known to be $\alpha^* = 27/16$, with a corresponding energy of -2.8477 (a.u.). To show the importance of the choice of γ to ensure convergence, we present in Fig. 1 the first 250 SGA iterations using different dampings. The number of configurations is chosen to be $m = 5$. We can see that the SGA method with appropriate damping converges to the correct value of α. The "wrong" choices do not converge or do it very slowly. One needs less than $N_i = 10\,000$ iterations to obtain the correct value of the parameter and reach an accuracy in energy of 0.0001 (a.u.). The convergence properties of the algorithm strongly depend upon the cost function, although its asymptotics is at least linear with $1/i$. Using the traditional optimization method for the energy [6], we find that even with $m = 25\,000$ fixed configurations, the optimum obtained contains a large uncertainty of 0.006 (a.u.) in energy. This is due to the use of a finite number of configurations. It is important to stress that the SGA does *not* have such a source of error, as the limitation to a finite number of configurations is only intrinsic to

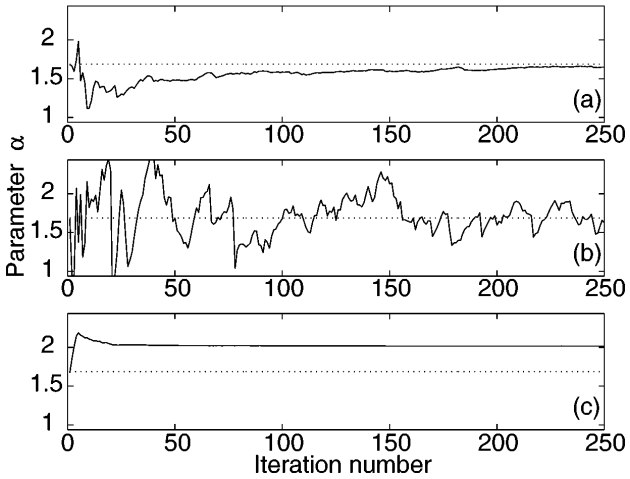

FIG. 1. The first 250 iterations of SGA with different choices of the damping rate γ_i. In (a) we have used $1/i$, in (b) $1/\sqrt{i}$, and in (c) $1/i^2$. We can see that (a) converges stochastically to the optimal value α^* given by the "dotted" line. In addition, one can see how the simulation (b) is oscillating, and the simulation in (c) fails because the damping is too large.

the deterministic approaches. In the most favorable case, the traditional method needs only a few iterations to converge to the minimum of the chosen configurations, but that minimum has a large uncertainty and several updates of the sample are needed to get the closest realization of the true minimum.

To approach the Hartree-Fock (HF) result for the He atom, we have added to the previous one-body wave function an additional $1s$ orbital. This increases the number of variational parameters to three. We have optimized those parameters, again using $m = 5$. The desired accuracy [0.001 (a.u.)] is found in less than 10 000 iterations, and the corresponding energy is -2.861 (a.u.) which is in good agreement with the HF energy. Using the traditional method, the optimal parameters are found with $m = 10\,000$ configurations and $N_i \approx 30$ iterations.

Next, we multiply the previous two-body wave function by a simple Jastrow correlation factor $\exp[r_{12}/(2 + br_{12})]$, where r_{12} is the distance between the electrons and b is a new parameter. The four optimum parameters are again found in less than 10 000 iterations ($m = 5$), and the energy we get is -2.8995 ± 0.0003 (a.u.). To get a similar accuracy using the traditional approach we found $m \approx 6000$ and $N_i \approx 30$. For a similar wave function Umrigar *et al.* [6] have found the energy to be -2.8996 ± 0.0003 (a.u.), using the traditional method in the variance minimization version. In such a case, the efficiency of both approaches is comparable (product $mN_i \approx 10^3$).

In this simple example, we can recognize two important features of the SGA. First, a small number of configurations is enough for an accurate optimization. The desired accuracy is reached by iterating long enough even if the number of configurations is as small as one. On the other hand, in the traditional approach, both the number of configurations and iterations determine the accuracy of the solution. *For a fixed and finite value of m, the traditional method carries an intrinsic bias.* This bias can be avoided to some extent by a sequence of optimizations and averaging processes which, of course, degrade its efficiency. Second, the SGA is not sensitive to the quality of the wave function. The previous examples show how the performance of the SGA is similar for all the variational wave functions.

To investigate the impact that the dimension of the vector of variational parameters $\boldsymbol{\alpha}$ has on the efficiency and convergence properties of the SGA algorithm, we have considered the case of small positronic atoms. Using a wave function with 8 variational parameters we found [7,8] similar performance to that in our previous examples. For instance, using the variance of the local energy as the cost function we found for hydrogen positronium (HPs) $m = 5$, $N_i < 100$. Besides, preliminary SGA tests for molecules, with even larger number of variational parameters, confirm this assertion.

Recently, Williamson *et al.* [9] showed that the efficient optimization of the many-body wave function of *extended* systems [for example, the homogeneous electron gas (HEG)] raises new challenges for QMC. They considered forms for the Jastrow factor, proposed by Ortiz and Ballone [10,11], which are convenient to optimize. However, Williamson *et al.* found their minimization scheme to be unstable as the number of particles increases. The problem arises from the reweighting factors which allow one to evaluate properties of a modified wave function. To cure this problem they set the reweighting factors equal to unity and allowed the parameters to change at each iteration step only slightly. Their modified scheme uses $m = 10\,000$ statistically independent configurations. In our previous work [12] we have noticed a similar instability. However, this is not the case when we use the SGA technique: 7 configurations and a few hundred iterations are sufficient to reach convergence with variance minimization as the optimization criterion.

A more stringent test of the SGA method is the problem of one positron embedded in HEG. This problem constitutes a benchmark for the interpretation of positron annihilation spectroscopy in solids [13]. Moreover, since the added positron represents a tiny fraction of the total number of particles in the system, the optimization of the electron-positron Jastrow function is computationally challenging. Here, we consider 226 electrons and a positron in an fcc unit cell with periodic boundary conditions and $r_s = 2$ [$r_s = (\frac{3}{4\pi n})^{1/3}$, where n is the electron density]. The variational wave function is of the form

$$\Psi = D_\downarrow D_\uparrow J \varphi_+ , \tag{6}$$

where $D_{\uparrow(\downarrow)}$ is a Slater determinant for spin up (down) electrons, φ_+ is the positron wave function (constant), and J is the Jastrow factor proposed by Ortiz and Ballone [10,11],

$$J = \exp\left[\sum_{i<j} \frac{a r_{ij}}{1 + b r_{ij}} + A e^{-\alpha r_{ij}^2} + s\right] \quad (7)$$

for $r_{ij} < R$, otherwise its contribution is truncated smoothly to zero. Here the indices i and j span all the particle pairs, including the positron. In the previous work [10], it was found that for the parameter R one can use the scaling law $R \cong 0.46L$, where L is the radius of the sphere inscribed within and tangent to the simulation cell. The variational parameters A, b, α, and s take different values depending on the type of particle pair (positron versus electron, spin-up electron versus spin-down electron, etc.), and a is fixed by the cusp condition. The total number of parameters is 10. Table I displays the optimal parameters. During the optimization, we have attempted the displacement of the positron much more often than for the electrons (up to 50 times more frequently) in order to get a significant signal for positron-related observables. With this procedure, 7 configurations and a few hundred iterations are sufficient to reach convergence, as for the case of HEG without the positron. The total energy per electron corresponding to the parameters in Table I is $0.0032 \pm 3 \times 10^{-4}$ (a.u.) and the standard deviation 0.0050 (a.u.). The run consists of about 15 000 steps per particle. We have also checked that the parameters were at their optimal values by varying them slightly and confirming the minimum variance condition.

To test the wave function, we have also calculated the electron-positron correlation energy E_{ep}, defined as the change in energy of the HEG after introducing the positron. Strong size dependent effects may occur [11] when E_{ep} is computed as the total energy difference of the systems with and without the positron. To avoid this problem, we have calculated E_{ep} by fitting over the range R the electron-positron pair distribution function from our simulation with an analytic form fulfilling the cusp and Friedel sum rule conditions [8],

$$g(r, Z) = 1 + cZ \exp\left(-\frac{(Zc + 1)r}{c}\right) \cos\left(\frac{(Zc + 1)r}{c\sqrt{3}}\right) + \frac{Z^4 r_s^3}{6} \exp(-Zr), \quad (8)$$

TABLE I. The Jastrow parameters for the optimized wave function for HEG $r_s = 2$ with a positron. The different pairs are electrons with parallel (antiparallel) spins ↑↑ (↑↓) and an electron-positron pair $e^+ e^-$.

Pair	b	s	A	α
↑↑	0.31	0.58		
↑↓	0.44	0.90	0.00	0.00
$e^+ e^-$	0.39	−1.02	0.18	0.30

where Z is a coupling constant ($0 < Z < 1$) characterizing the charge of the positron, and c a fitting parameter [14]. This form is motivated by the fact that the shape of the screening cloud resembles that of a positronium, namely, $\exp(-Zr)$ [15]. The fit of the screening charge shows an overall agreement with both the Arponen-Pajanne (AP) theory [15] and data. However, in the cusp region both the fitted and AP curves are above the simulated data. Using the coupling constant integration technique [16] one obtains $E_{ep} = -0.30 \pm 0.03$, where the uncertainty is due to discrepancy between the data and the fit. More variational freedom, such as three-body correlations, may be necessary to recover all the electron-positron correlations.

In conclusion, we have presented a method for the optimization of noisy functions. We have tailored the scheme for the optimization of Monte Carlo many-body wave functions. The main advantages of the method are robustness, simultaneous updating of parameters and wave function configurations, scalability with the number of degrees of freedom, and a lack of sensitivity to the quality of the wave function. We have been able to use a small number of configurations even for a complex system such as a positron in an HEG. In fact, the same number of configurations can be used from a two-particle system up to a 227-particle system. In particular, the good performance in the HEG illustrates the potential of SGA in extended systems, where the optimization of the many-body wave function is a difficult task. Overall, the SGA method opens up new possibilities for the variational freedom of the wave functions used in stochastic simulations for a large variety of systems.

B. B. and G. O. are supported by the Swiss NSF Grant No. 8220-037167 and the U.S. NSF Grant No. DMR-91-17822, respectively.

*Author to whom correspondence should be addressed. Electronic address: Ari.Harju@hut.fi

†Present address: Theoretical Division, T-11, Los Alamos National Laboratory, Los Alamos, NM 87545.

[1] P. Young, in *Optimization in Action*, edited by L. C. W. Dixon (Academic Press, London, 1976).

[2] H. Robbins and S. Monro, Ann. Math. Stat. **22**, 400 (1951).

[3] P. J. Reynolds, D. Ceperley, B. Alder, and W. A. Lester, J. Chem. Phys. **77**, 5593 (1982).

[4] J. C. Grossman and L. Mitáš, Phys. Rev. Lett. **74**, 1323 (1995).

[5] C. Filippi and C. J. Umrigar, J. Chem. Phys. **105**, 213 (1996).

[6] C. J. Umrigar, K. G. Wilson, and J. W. Wilkins, Phys. Rev. Lett. **60**, 1719 (1988).

[7] A. Harju, B. Barbiellini, and R. M. Nieminen, Phys. Rev. A **54**, 4849 (1996).

[8] A. Harju, B. Barbiellini, S. Siljamäki, R. M. Nieminen, and G. Ortiz, J. Radioanal. Nucl. Chem. **211**, 193–202 (1996).

[9] A. J. Williamson *et al.*, Phys. Rev. B **53**, 9640 (1996).
[10] G. Ortiz and P. Ballone, Europhys. Lett. **23**, 7 (1993); Phys. Rev. B **50**, 1391 (1994).
[11] G. Ortiz, Ph. D. thesis, Ecole Polytechnique Fédérale de Lausanne, 1992.
[12] S. Siljamäki, B. Barbiellini, A. Harju, and R. M. Nieminen (unpublished).
[13] M. J. Puska and R. M. Nieminen, Rev. Mod. Phys. **66**, 841 (1994).
[14] The fit yields a contact term $g(r = 0, Z = 1) = 4.4$ and the simulation with 178 electrons gives the same value within 1% [8].
[15] J. Arponen and E. Pajanne, Ann. Phys. (N.Y.) **121**, 343 (1979); in *Positron Annihilation,* Proceedings of the ICPA7, edited by P. C. Jain *et al.* (World Scientific, Singapore, 1985), parametrized by E. Boroński and R. M. Nieminen, Phys. Rev. B **34**, 3820 (1986).
[16] C. H. Hodges and M. J. Stott, Phys. Rev. B **7**, 73 (1973).