# Molecular potential energy surfaces by interpolation: Strategies for faster convergence

Gloria E. Moyano and Michael A. Collins

Citation: The Journal of Chemical Physics **121**, 9769 (2004); doi: 10.1063/1.1809579
View online: http://dx.doi.org/10.1063/1.1809579
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/121/20?ver=pdfcov
Published by the AIP Publishing

## ARTICLES

# Molecular potential energy surfaces by interpolation: Strategies for faster convergence

Gloria E. Moyano and Michael A. Collins
*Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia*

(Received 17 August 2004; accepted 3 September 2004)

A method for interpolating molecular potential energy surfaces introduced [Ischtwan and Collins, J. Chem. Phys. **100**, 8080 (1994)] and developed as an iterative scheme has been improved by different criteria for the selection of the data points. Refinements in the selection procedure are based on the variance of the interpolation and the direct exploration of the interpolation error, and produce more accurate surfaces than the previously established scheme for the same number of data points. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1809579]

## I. INTRODUCTION

Molecular potential energy surfaces (PESs) are necessary for the computation of reaction dynamics. In the Born-Oppenheimer approximation, the molecular potential energy is the total electronic energy, which can be evaluated using the methods of *ab initio* quantum chemistry. Current approaches to quantum reaction dynamics require calculation of this energy at each node in a very large grid of molecular configurations. The numerical implementation of classical reaction dynamics requires the gradient of the energy with respect to the nuclear positions for a very large number of molecular configurations. The direct determination of these energies and/or energy gradients by *ab initio* calculations is an extremely expensive task, and has only been applied to small molecules or with relatively low level *ab initio* methods. In recent years, methods have been developed to approximate the PES by interpolation over a number of *ab initio* calculations which is orders of magnitude smaller than that required for the direct approach to dynamics.[1,2]

A systematic method of generating accurate PES by interpolation has been proposed[1,3–7] and applied to the treatment of several reactions using both classical and quantum dynamics. The method has been presented in detail elsewhere[1,3–7] so it will be described only briefly in Sec. II of this paper. The interpolation of the PES is based on ''local'' approximations to the surface by Taylor expansions, but is global in character since these local approximations in different regions are combined in a weighted average. The method involves an iterative procedure in which molecular configurations, ''data points,'' and the associated *ab initio* calculations are accumulated. The PES defined by this data converges to the exact PES (for the *ab initio* level employed) as the number of data points increases. The number of *ab initio* data points needed to accurately describe the relevant region of a PES varies from system to system. However, given the computational cost of the high level *ab initio* methods which may be necessary, there is always substantial computational saving to be made by reducing the number of data points required. Refinements of the established method which improve the rate of convergence are presented in this paper. These refinements modify the way in which molecular configurations are chosen for the location of each additional data point in the iterative procedure. These methods are presented in Sec. III. Section IV contains a description of the tests carried out to verify the utility of these methods. Concluding remarks are presented in Sec. V.

## II. THE CURRENT METHOD

For a molecular system with $N$ atoms in a specific electronic state the interpolated PES is constructed using inverse interatomic distances $\mathbf{Z}=\{Z_1,\ldots,Z_k,\ldots,Z_{N(N-1)/2}\}$, where $Z_k=1/R_k$, as the basis for coordinates to describe molecular configurations. Near any data point $\mathbf{Z}(i)$ a set of $3N-6$ locally independent internal coordinates $\{\xi_k(\mathbf{Z})\}$ can be defined in terms of $\mathbf{Z}$, and used to construct a Taylor series expansion (to second order) $T_i(\mathbf{Z})$ of the PES. The energy and derivatives required for each of the Taylor expansions have to be evaluated by (usually *ab initio*) electronic structure methods. With no additional *ab initio* calculation, Taylor series can be constructed about all symmetry related configurations. These are the configurations which differ from the original by permutation of the positions of indistinguishable nuclei. These permutations form the complete nuclear permutation (CNP) symmetry group of the molecule. The electronic energy is the same at all permuted geometries, the elements of the Cartesian energy gradient vector are simply permuted at a permuted configuration, and the rows and columns of the Cartesian Hessian are similarly permuted. Hence, no additional *ab initio* calculations are required to produce Taylor series about all permuted versions of some data point configuration $\mathbf{Z}(i)$. The CNP symmetry group is denoted as $G$, and a permutation operation (for $g \in G$) on a data point configuration is denoted as $go\mathbf{Z}(i)$, or simply as $goi$. Hence, from one set of *ab initio* calculations, we can actually add as many points to the interpolation data set as there are elements in $G$. This not only improves the accuracy

 

of the interpolation, but ensures that the resultant PES has the correct symmetry (invariant to the permutation of indistinguishable nuclei). The form of the interpolated PES is then

$$E(\mathbf{Z})=\sum_{g\in G}\sum_{i=1}^{N_{\text{data}}} w_{goi}(\mathbf{Z})T_{goi}(\mathbf{Z}). \tag{2.1}$$

This expression corresponds to a particular case of modified Shepard interpolation,[8] where a number $N_{\text{data}}$ of local Taylor expansions (and their symmetry equivalents) are combined as a weighted average to give an explicit global interpolation formula. The weights $w_i$ in Eq. (2.1) are normalized to unity according to

$$w_{goi}=\frac{v_{goi}}{\sum_{g\in G}\sum_{j=1}^{N_{\text{data}}} v_{goj}}. \tag{2.2}$$

The $v_i$ or ''primitive weights'' in Eq. (2.2) are functions of the distance coordinates $\mathbf{Z}$ with respect to the data point configuration $\mathbf{Z}(i)$. In a simple implementation of the Shepard interpolation formula,[8] the primitive weights are given by

$$v_i=\|\mathbf{Z}-\mathbf{Z}(i)\|^{-p}, \quad p>3N-3. \tag{2.3a}$$

However, a more suitable form is given by

$$v_i=\left(\left\{\sum_{k=1}^{N(N-1)/2}\left[\frac{Z_k-Z_k(i)}{d_k(i)}\right]^2\right\}^q + \left\{\sum_{k=1}^{N(N-1)/2}\left[\frac{Z_k-Z_k(i)}{d_k(i)}\right]^2\right\}^p\right)^{-1}, \tag{2.3b}$$

where the parameters $q$ and $p$ are usually given values of 2 and 12, respectively. The $d_k(i)$ denote confidence lengths about the $i$th data point for each direction $k$ in space, and have been derived from a Bayesian analysis of the errors in the energy gradients.[3]

The *data set* is a set of $N_{\text{data}}$ molecular configurations about which the energy is approximated by the local Taylor expansions. This is a sparse set of data points, which the method seeks to concentrate in the domain of molecular configurations relevant for the dynamics under study.[1] An initial set of data points is usually taken along the minimum energy path (MEP) of the reaction. Additional data points are accumulated in an iterative fashion from a collection or batch of molecular configurations encountered in classical trajectories for the reaction being studied [using the existing data set in Eq. (2.1)]. The aim of adding a new data point to the data set is to improve the accuracy of the interpolated PES as much as possible. The current methods for selecting new data points from trajectory configurations are based on two arguments. One suggests that the quantity

$$h(k)=\frac{\sum_{\substack{m=1\\ m\neq k}}^{N_{\text{traj}}} v_m[\mathbf{Z}(k)]}{\sum_{g\in G}\sum_{i=1}^{N_{\text{data}}} v_{goi}[\mathbf{Z}(k)]}, \tag{2.4}$$

evaluated at each of the $N_{\text{traj}}$ trajectory points in the batch, gives the largest values to points, $\mathbf{Z}(k)$, in the regions most frequently ''visited'' by the classical trajectories, so long as they are distant from points already in the data.[7] Note that the numerator of Eq. (2.4) involves evaluating the primitive weight $v$ between trajectory configurations. Since the confidence lengths in Eq. (2.3b) are not known at a trajectory configuration (only at data points), we use the simple weight function of Eq. (2.3a) everywhere in Eq. (2.4). The trajectory point with largest $h$ value is chosen to be a new data point. This assumes that if a trajectory point is far from the data set, additional data are required in its neighborhood. The second argument considers the quantity known as the variance

$$\sigma^2(k)=\sum_{g\in G}\sum_{i=1}^{N_{\text{data}}} w_{goi}[\mathbf{Z}(k)]\{T_{goi}[\mathbf{Z}(k)]-E[(\mathbf{Z}(k)]\}^2, \tag{2.5}$$

where $E[\mathbf{Z}(k)]$ is the interpolated energy as expressed in Eq. (2.1). If all of the Taylor polynomials with significant weight in the expansion (2.1) agree on the value of the interpolated energy at the point $k$, then $T_i\approx E[\mathbf{Z}(k)]$ and $\sigma^2\approx 0$ at $k$. The variance is a measure of the uncertainty of the weighted average or the disagreement between the values of the local Taylor polynomials at $\mathbf{Z}(k)$. On the assumption that the molecular configurations where the interpolated energy is inaccurate are associated with large variances, the trajectory points with largest variances are chosen as new elements of the data set.[5]

The accuracy of the interpolated PES is evaluated by determination of observables like the reaction cross section or the thermal rate coefficient at different sizes of the data set. When the values of these observables do not change, beyond a certain tolerance, with increasing $N_{\text{data}}$, the PES is considered ''converged.''[1,5,6] Complementary indicators of convergence are the average and maximum interpolation errors [the interpolation error is defined as the absolute difference between the *ab initio* and interpolated energies at a point $\mathbf{Z}(k)$] for a sample of trajectory points, also calculated at different sizes of the data set.[1,5]

This method for constructing PES is implemented in a program package called GROW and will be denoted by that name below.[1]

## III. REFINEMENTS FOR FASTER CONVERGENCE

The method as described in Sec. II is in a state of refinement. Different aspects of the method susceptible to change have been pointed out (see, e.g., Refs. 3 and 5) and examined in order to improve the accuracy and rate of convergence. In this paper we concentrate on the sampling of trajectory points to select new elements for the data set.

### A. Reference PES

In order to illustrate some aspects of a PES which describes a chemical reaction, we will make use of the analytic PES proposed by Schatz and Elgersma[9] for the reaction of

$$OH+H_2\rightarrow H_2O+H \tag{3.1}$$

(hereafter denoted as the SE surface). We also use this PES as the basis for testing the accuracy of various ways of constructing the interpolation data set. The interpolated PES of

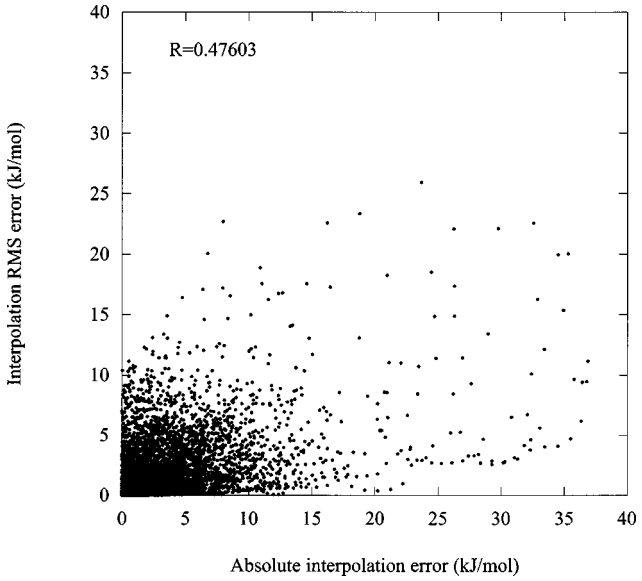


FIG. 1. Scatter plot of the interpolation RMS error vs the absolute interpolation error, for 7882 configurations from trajectories for the reaction between OH and $H_2$. The value of $R$ corresponds to the Pearson correlation coefficient between the variables with respect to this set of configurations.

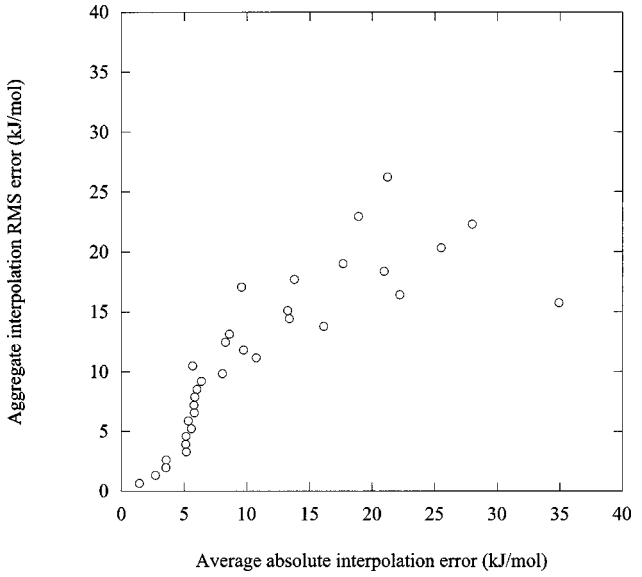


FIG. 2. The 7882 configurations of Fig. 1 were aggregated according to their rms error, using a bin size of 0.656 kJ/mol. This figure shows the aggregate (bin center) interpolation rms error vs the average absolute interpolation error for the configurations in each bin.

Eq. (2.1) is constructed with energies, energy gradients, and second derivatives evaluated from the SE surface, so that the interpolated PES is an approximation to the SE surface. It is an easy matter to examine the accuracy of the interpolations by comparison with the SE surface.

#### *1. Computational details*

In this and later sections, we consider the dynamics of reaction (3.1) with both reactants given zero rotational angular momentum, a relative kinetic energy of 78.75 kJ/mol, and vibrational energies of 26.25 kJ/mol for OH and 78.75 kJ/mol for $H_2$, at the start of the classical collisions. The efficient microcanonical sampling method of Schranz, Nordholm, and Nyman[10] was used to generate the initial atomic positions and velocities for the reactants. The initial center of mass separation for the fragments was set at 4.5 Å, and the trajectories were terminated when the separating fragments reached this distance. An impact parameter $b=0$ was assumed for the collisions used in the iterative construction of the interpolated surfaces. The trajectory integration was carried out with a velocity-Verlet algorithm[11] using a step size of $1.0\times10^{-17}$ s. For construction of interpolated approximations to the SE surface, an initial data set of 30 points along the MEP of reaction (3.1) was used.[6] The interpolated PESs data sets were accumulated in the standard iterative scheme, with new data points chosen from configurations sampled from batches of ten classical trajectories started from the $OH+H_2$ reactants.

### B. Maximization of the variance

The variance of Eq. (2.5) is a function of the molecular configuration $\sigma^2(\mathbf{Z})$. This quantity is a measure of the uncertainty in the estimate of the energy given by the interpolation formula. Does this uncertainty correlate with the actual interpolation error? As an example, Fig. 1 presents a comparison between the square root of the variance, also known as the root-mean-square (RMS) error, and the actual absolute interpolation error for a set of molecular configurations. These are 7882 configurations sampled from classical trajectories for reaction (3.1), on an interpolated approximation to the SE surface, with 80 data points chosen using the standard GROW algorithm under the conditions described above.

The correlation coefficient between the square root of the variance and the interpolation error, $R=0.476\,03$, is far from unity. However, if we "bin" the RMS errors by magnitude and consider the average absolute error for configurations in each bin, we obtain Fig. 2. Clearly there is, on average, a tendency for configurations which have large values of the RMS error to have large interpolation errors. Hence, to estimate where the PES is most in error, we might find the configuration where $\sigma^2(\mathbf{Z})$ is largest. In the standard approach, this is achieved by choosing the geometry of largest $\sigma$ from a sample of configurations generated from trajectories. However, the geometry of largest $\sigma$ could be found by using a steepest descent algorithm to minimize $-\sigma^2(\mathbf{Z})$. If $\mathbf{x}$ denotes the $3N$ dimensional vector of atomic Cartesian coordinates, then beginning at some configuration, such a steepest descent path would be given by

$$\frac{dx_i}{dt}=\frac{\partial\sigma^2}{\partial x_i}, \quad i=1,...,3N. \tag{3.2}$$

However, $\sigma^2(\mathbf{Z})$ is also correlated with the energy, $E(\mathbf{Z})$, as indicated by Fig. 3. Configurations of high energy tend to have large values of $\sigma^2(\mathbf{Z})$ and large absolute interpolation errors, because the number of data points is generally low at high energy. Of course, if we consider energies far above that of the highest energy data point, then there would be no data points near that energy, and most probably, both the variance and interpolation error at such a configuration would be very high. Hence, if we simply attempted to

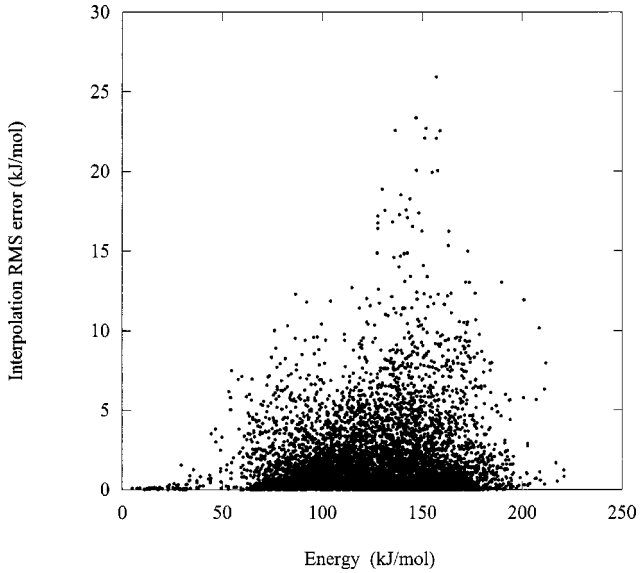


FIG. 3. Scatter plot of the interpolation rms error as a function of the energy for the configurations of Fig. 1. The reference energy level for this plot corresponds to $H_2O+H$.

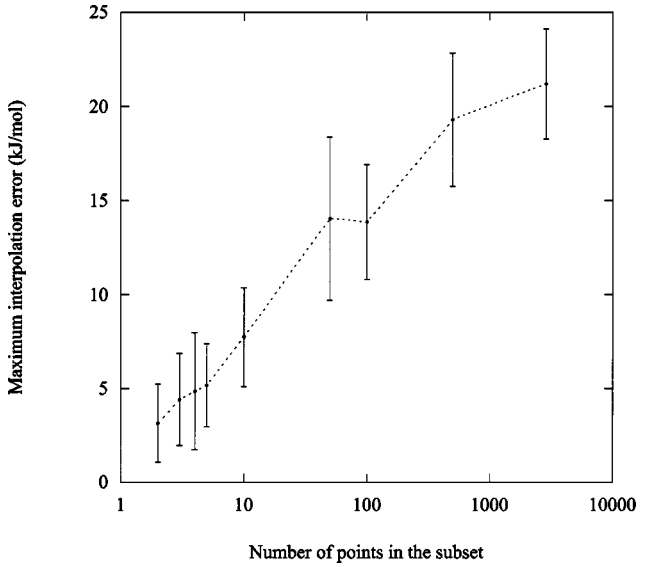


FIG. 4. The maximum interpolation error in a subset of a large collection of trajectory configurations is shown as a function of the number $n_s$ of points in the subset. The value shown is the average of the maximum error in independent samples of $n_s$ configurations from ten large collections of trajectory configurations. The error bars indicate the standard deviation of the ten trials. The connecting line is included merely as a visual aid.

solve Eq. (3.2), the path traced out would almost inevitably lead to a configuration of unphysically high energy. To prevent this, we constrain the steepest descent path to conserve the energy of the initial configuration $\mathbf{Z}(t=0)$, by introducing a Lagrange multiplier:

$$\frac{dx_i}{dt}=\frac{\partial\sigma^2}{\partial x_i}-\lambda\frac{\partial E}{\partial x_i}. \tag{3.3}$$

By requiring that $E(\mathbf{Z})-E[\mathbf{Z}(t=0)]=0$, $\lambda$ is determined:

$$\frac{dx_i}{dt}=\frac{\partial\sigma^2}{\partial x_i}-\frac{\partial E}{\partial x_i}\left(\frac{\frac{\partial E}{\partial \mathbf{x}}\cdot\frac{\partial\sigma^2}{\partial \mathbf{x}}}{\left\|\frac{\partial E}{\partial x}\right\|^2}\right). \tag{3.4}$$

In practice, the configuration of highest variance from a trajectory sample is chosen as the initial configuration, $\mathbf{Z}(t=0)$. Equation (3.4) is then solved to estimate the configuration which has the highest variance at the same energy $E[\mathbf{Z}(t=0)]$. Since there is no reason to assume that the path followed by Eq. (3.4) leads to the absolute minimum for the constrained optimization, the final configuration is likely to be simply a local (constrained) maximum in the variance. To indicate the utility of this method for finding configurations where the interpolation error is large, we accumulated samples of configurations from ten sets of ten trajectories each, evaluated on the 80 data point interpolated PES. For every set, the average interpolation error and the average of the largest 10% of errors were calculated. The configuration of maximum variance was found using the method described above and the interpolation error at this configuration was determined. For the samples analyzed, the average absolute interpolation error was 2.3 kJ/mol, the average of the largest 10% of absolute errors was 8.7 kJ/mol, and the average absolute interpolation error at the configurations of maximum variance was 6.0 kJ/mol. This indicates that the criterion of maximum variance is a useful method for locating configurations where the interpolation error is large.

In Sec. IV, interpolated approximations to the SE surface are reported where the data point was chosen at each iteration as the configuration of maximum variance. This method is denoted as Maxvar-GROW.

### C. Direct test of the interpolation error

If the ansatz for choosing a new data point is (say) to choose a configuration with large interpolation error, it is useful to examine the distribution of interpolation errors in our test case. The interpolation error for a PES with the form (2.1) is a function of the molecular geometry which varies over the domain of the PES. To estimate this error for any PES, we can generate a sample of trajectory points scattered on that domain, and calculate the average and the maximum interpolation errors for that sample. For several samples on the same domain, typically with a few thousand points each, the average interpolation error naturally tends to be more reproducible than the maximum interpolation error. This can be illustrated with the ten samples of trajectory points referred to in Sec. III B. These samples have an average size of 2882 points, and an average interpolation error of 2.3 kJ/mol with a standard deviation of 0.2 kJ/mol. In comparison, the mean for the maximum interpolation error in each sample is 21.2 kJ/mol with a standard deviation of 2.9 kJ/mol.

The maximum interpolation error in subsets of points chosen at random from the samples behaves as shown in Fig. 4. The maximum interpolation error increases with the number of points in the subset, although the error bars in Fig. 4 indicate the variability expected for relatively small samples. The trend indicated in Fig. 4 is reproducible when performing this analysis using a variety of interpolated PES (with

different data sets) and trajectory point samples for the same system. Note that the number of points in the subset is shown on a logarithmic scale in Fig. 4. So, if we choose only one configuration in each subset, the maximum interpolation error is the average error, here 2.3 kJ/mol; choosing three to seven configurations in each subset gives a maximum interpolation error of about 5 kJ/mol; while we would need to choose about 1000 configurations in each subset to obtain a maximum interpolation error of about 20 kJ/mol.

Thus, if we evaluate the interpolation error for several configurations chosen at random, we are likely to observe a maximum interpolation error of about twice the average error; but to observe much larger errors would require very much larger samples. This initial rapid increase of the maximum interpolation error with the size of the subset is the basis for a strategy to directly detect regions of large interpolation error, and to choose new data points during the construction of the PES.

When we construct an interpolated PES using *ab initio* data, the computational cost is the product of the number of data points with the computational cost of calculating the energy, energy gradient, and second derivatives. If the first and second derivatives are evaluated by so-called analytic methods, then the computational cost for each data point is typically some small multiple of the cost of calculating the energy. If only analytic gradients are available for the *ab initio* method, then the second derivatives require $[2(3N-6)+1]$ gradient evaluations for a central finite difference evaluation of the second derivatives, so the computational cost is some small multiple times $[2(3N-6)+1]$ times the cost of an energy calculation. If all derivatives have to be evaluated by central differences of the energy, then the cost is $[(3N-6)(3N-5)+1]$ times the cost of an energy calculation. In this case, the cost of three to seven additional energy calculations is a relatively small addition to the cost of calculating the required data. Hence, the direct estimate of a configuration of large interpolation error may be an efficient means of choosing a new data point, if this choice results in more rapid convergence of the PES with respect to data set size.

We have implemented this strategy in the following way. At each iteration of the PES growing procedure, a small number $n_s$ of trajectory points were chosen at random from the sample of trajectory configurations. The exact energy of these configurations was calculated (from the SE surface). These configurations were added to a set of all such configurations accumulated from previous iterations. The current data set was used to interpolate the energy at all these configurations, and the configuration with maximum interpolation error was determined. This configuration was chosen to be the new data point for this iteration. In the following section we examine the accuracy of PES constructed with this direct error detection procedure, denoting the methods as $E2$-GROW, $E3$-GROW, and $E5$-GROW, for $n_s=2$, 3, and 5, respectively.

A modified version of the direct error detection method has also been implemented as follows. This method differs from $E3$-GROW only in that while one trajectory configuration is chosen at random, one is chosen as the trajectory configuration of highest $h$ value, and the third point is chosen as the trajectory configuration of highest variance. This method is denoted as $E$-GROW as it combines direct error evaluation with the two standard methods for choosing new data points.

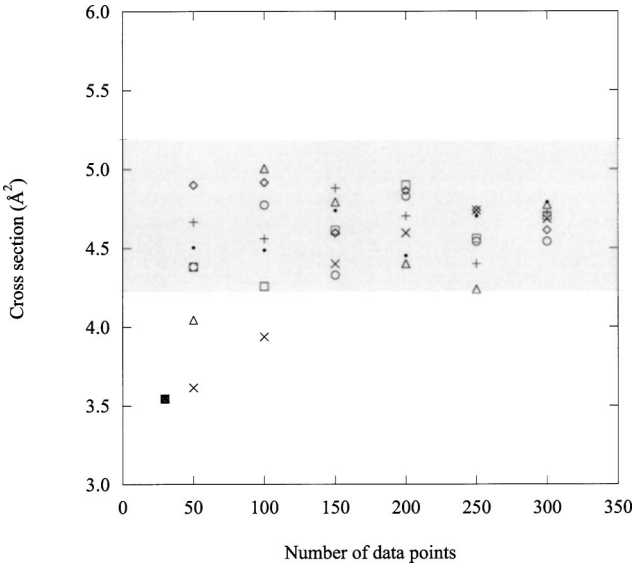


FIG. 5. The cross sections for reaction (3.1) calculated on the seven interpolated PESs are shown as a function of the PES data set size. The shaded area indicates the cross section determined using the SE surface and two expected standard deviations for statistics based on 1000 trajectories. The values shown are for the following: Standard GROW, ○; $E2$-GROW, □; $E3$-GROW, ◇; $E5$-GROW, ×; $E$-GROW, +; Random-GROW, △; and Maxvar-GROW, •.

## IV. PERFORMANCE COMPARISONS

In this section we compare the accuracy of PES constructed with data chosen in seven different ways. Surfaces have been constructed with the $E$-GROW, $E2$-GROW, $E3$-GROW, and $E5$-GROW methods; with data points chosen by the maximum variance procedure of Sec. III B, Maxvar-GROW; with data points chosen at random from the trajectory configurations, denoted as Random-GROW; and finally with data chosen by the standard GROW procedure.

An initial data set of 30 points along the MEP of reaction (3.1) was used for the construction of an interpolated approximation to the SE surface for all seven procedures. From the initial data set, interpolated PESs with a total of 300 points were produced for all the versions of the method. For the Maxvar-GROW method, it was found that stable determination of the configuration of maximum variance was difficult for small data sets. Hence, in this case the standard GROW method was used to construct a data set of 100 points, after which further data point selection followed the Maxvar-GROW scheme.

To evaluate the convergence of the reaction cross section, batches of 1000 trajectories were run on all the interpolated surfaces for data set sizes of 30, 50, 100, 150, 200, 250, and 300 points. A batch of 10 000 trajectories was also run on the SE surface. The initial conditions for all those trajectories were as specified above except that the impact parameters $b$ were sampled randomly from a distribution limited by

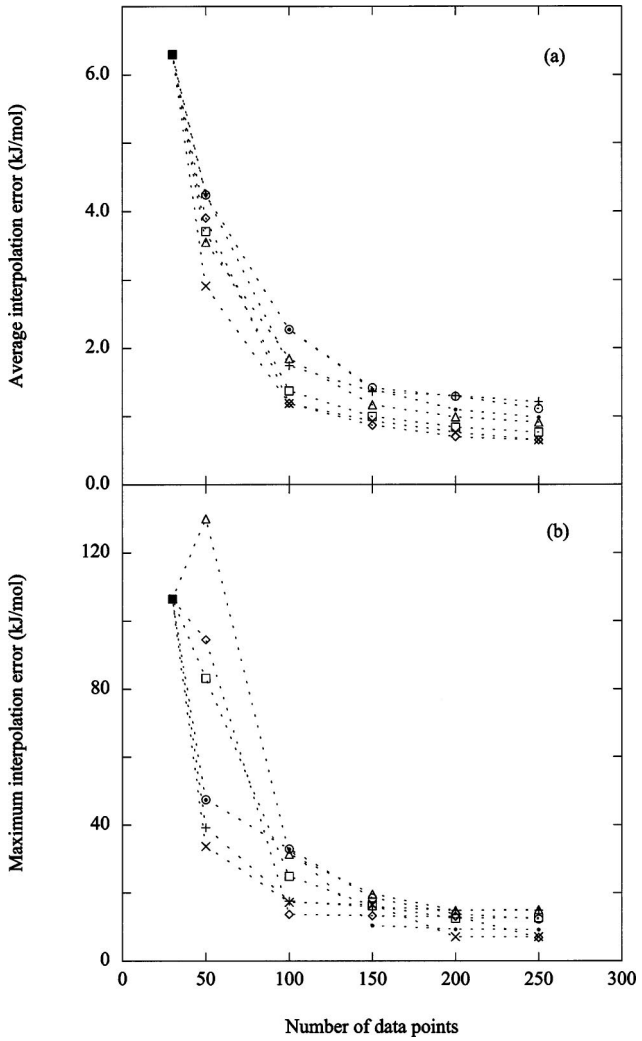


FIG. 6. (a) The average absolute interpolation error, and (b) the maximum interpolation error are shown as a function of the data set size for the seven interpolated PES. The errors were determined with reference to a set of 4757 points, sampled at random, along trajectories for the reaction between OH and $H_2$. The values shown are for the following: Standard GROW, ○; $E2$-GROW, □; $E3$-GROW, ◇; $E5$-GROW, ×; $E$-GROW, +; Random-GROW, △; and Maxvar-GROW, •. Connecting lines are included merely as a visual aid.

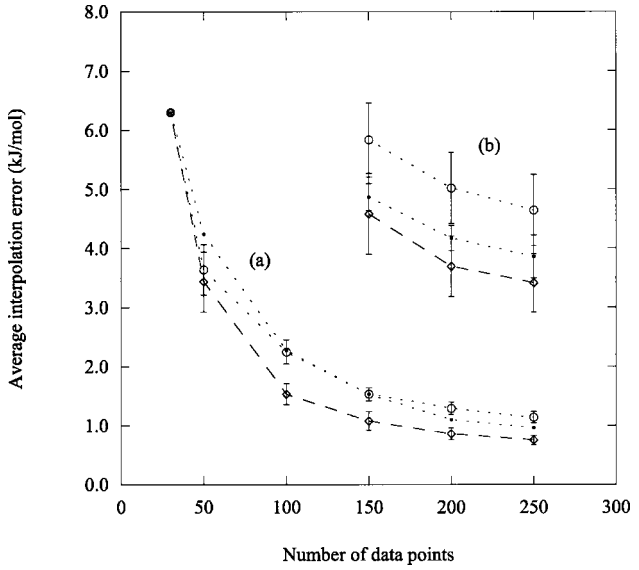


FIG. 7. (a) The average absolute interpolation error and (b) the average of the 10% of largest interpolation errors for ten replicas of the interpolated PESs are shown as a function of the data set size. The errors were determined with reference to the same set of points as used in Fig. 6. The values shown are for the following: Standard GROW, ○; $E3$-GROW, ◇; and Maxvar-GROW, •. Error bars represent one standard deviation for the ten replicas.

a maximum exceeding the largest value at which reaction was observed. The distributions of $b$ values was such that the probability of a trajectory having an impact parameter between $b$ and $b+db$ was proportional to $b$.

The cross section for reaction (3.1) on the SE analytic surface was found to be $4.71\pm0.08$ Å$^2$, based on 10 000 trajectories. The corresponding cross sections on the seven interpolated PES are compared with this value in Fig. 5. The smaller samples employed for the seven interpolated PESs imply an expected standard deviation of about 0.24 Å$^2$. To avoid visual congestion in Fig. 5, the shaded area indicates two standard deviations from the cross section on the analytic surface. For data sets containing 300 points all the interpolated PESs could be considered converged with reference to the reaction cross section, with the possible exception of the Random-GROW PES. Clearly, all the methods examined for choosing new data points from trajectory samples are successful in producing suitable PESs for reactive scattering calculations. Selective sampling does appear to produce more rapid convergence than simple random sampling.

In order to more closely examine the relative accuracy of the PES produced by these seven methods for selection of new data points, the absolute interpolation errors were determined for a set of 4757 configurations randomly sampled from trajectories. The variation of the average and maximum interpolation errors was monitored at different data set sizes, as shown in Fig. 6.

The average interpolation error for the seven surfaces decreases as the number of data points increases. There are no noticeable deviations of this trend for any of the versions of the method tested. Figure 6(a) may indicate that the $E2$-, $E3$-, and $E5$-GROW methods produce a lower average error than the other methods. Figure 6(b) may indicate that the Maxvar-GROW and the $E3$- and $E5$-GROW methods may produce lower maximum interpolation errors. However, all seven methods contain a statistical variation that arises from the Monte Carlo sampling of initial conditions for the small sets of trajectories used to generate samples of configurations at each iteration of the construction scheme. Hence, multiple PESs constructed using a single method would display some variation in accuracy. It is possible that the variation in accuracy of the several methods, indicated in Figs. 6(a) and 6(b), is simply due to this fact. Moreover, the observed maximum error is subject to large variation for all methods for a finite sample of configurations.

This variability was examined by the repeated construction of PESs for three of the methods. Ten replicas of each of the standard, Maxvar and $E3$-GROW PESs were constructed. The average absolute interpolation error for those PESs were determined for the set of 4757 trajectory points. The average of the largest 10% of errors was also determined for all 30

PESs. The results are shown in Fig. 7 for different sizes of the data sets. The error bars in these figures indicate the standard deviation (over the ten trials) for each method. Clearly, the $E3$-GROW method achieves a smaller average interpolation error than the Maxvar method which is again lower than the standard GROW approach. Figure 7(b) indicates that the largest interpolation errors are also reduced in a similar way.

There is no clear evidence that the $E3$-GROW method leads to lower interpolation errors than the $E2$- or $E5$-GROW methods. This may reflect the utility of the strategy of accumulating a database of actual interpolation errors with each iteration of the PES construction scheme. As Fig. 4 suggests, the probability of observing a large interpolation error increases only slowly with the size of a random sample of configurations. Hence, all three schemes, $E2$-, $E3$-, and $E5$-GROW might be very nearly equally effective in locating large interpolation errors and in reducing the overall interpolation errors for the PES.

The random addition of trajectory points to the data set resulted in a PES with the largest of the maximum interpolation errors, and that behavior was consistent at the different data set sizes considered. Moreover, the reaction cross section for the Random-GROW PES may not have converged at a data set size of 300 points.

## V. CONCLUSIONS

The accuracy and rate of convergence of the standard GROW methodology to construct PES by interpolation can be improved by using two new data point selection procedures.

One procedure selects new data points as configurations which are local maxima in the variance or uncertainty of the interpolated energy. This new procedure reduces the average and largest interpolation errors with the addition of data points at a faster rate than the standard GROW method. This method involves no additional *ab initio* calculations and therefore results in more accurate PESs at no extra computational cost.

The second procedure selects new data points as the configurations of largest interpolation error, determined by directly testing the error in a small (but iteratively increasing) sample of trajectory configurations. This approach requires the evaluation of a small number (say three) of additional *ab initio* energies at each iteration of the construction scheme. This additional computational cost would exceed the benefit of the consequent improvement in accuracy if the PES were constructed with an *ab initio* method for which efficient analytic second derivatives are available. However, if only analytic gradients are available, and certainly if only energy calculations are available, this method would be very computationally efficient. The PESs for $OH+H_2$ generated in this way were more accurate and converged at a faster rate than the standard GROW PES.

The performance evaluations and conclusions from this work are only based on calculations for the SE surface for the $OH+H_2$ reaction. The general utility of the procedures proposed here will only be established by the construction of *ab initio* PES for a variety of reactions.

It is worthwhile to note that we have only considered the selection of new data points from samples of configurations generated by classical trajectories. The use of classical mechanics to explore the configuration space involved in a chemical reaction is intuitively reasonable at high energy, but may be inappropriate at very low energy. Recently, interpolated PES have been constructed to describe the ground vibrational states of molecules, where the sampling of configurations was achieved with quantum diffusion Monte Carlo simulations.[12,13] It may be that in order to construct very accurate interpolated PES for very low energy reactions, a quantum approach to the exploration of the relevant configuration space is necessary.

## ACKNOWLEDGMENT

The authors wish to thank the Australian Partnership for Advanced Computing National Facility for an allocation of computer time.

[1] M. A. Collins, Theor. Chem. Acc. **108**, 313 (2002).
[2] T. Hollebeck, T.-S. Ho, and R. Herschel, Annu. Rev. Phys. Chem. **50**, 537 (1999).
[3] R. P. A. Bettens and M. A. Collins, J. Chem. Phys. **111**, 816 (1999).
[4] K. C. Thompson, M. J. T. Jordan, and M. A. Collins, J. Chem. Phys. **108**, 8302 (1998).
[5] K. C. Thompson and M. A. Collins, J. Chem. Soc., Faraday Trans. **93**, 871 (1997).
[6] M. J. T. Jordan, K. C. Thompson, and M. A. Collins, J. Chem. Phys. **102**, 5647 (1995).
[7] J. Ischtwan and M. A. Collins, J. Chem. Phys. **100**, 8080 (1994).
[8] R. Farwig, Math. Comput. **46**, 577 (1986); R. Farwig, in *Algorithms for Approximation*, edited by J. C. Mason and M. G. Cox (Clarendon, Oxford, 1987), p. 194.
[9] G. C. Schatz and H. Elgersma, Chem. Phys. Lett. **73**, 21 (1980).
[10] H. W. Schranz, S. Nordholm, and G. Nyman, J. Chem. Phys. **94**, 1487 (1991).
[11] M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids* (Clarendon, Oxford, 1987).
[12] R. P. A. Bettens, J. Am. Chem. Soc. **125**, 584 (2003).
[13] K. C. Thompson, D. L. Crittenden, and M. J. T. Jordan, J. Am. Chem. Soc. (to be published).