Advanced computational methods for simulating chemical reactions

D L Thompson¹, A F Wagner² and M Minkoff³

¹Department of Chemistry, University of Missouri, Columbia, MO 65211
²Chemistry Division, Argonne National Laboratory, Argonne, IL 60439
³Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

thomspndon@missouri.edu

Abstract. The Interpolative Moving Least Squares (IMLS) fitting scheme is being developed for the purpose of fitting potential energy surfaces used in chemistry. IMLS allows for automatic surface generation in which the fitting method selects the positions at which expensive electronic structure calculations determine specific values on the surface. The resulting surfaces are necessary for accurate kinetics and dynamics.

1. Introduction

The calculation of spectroscopic and dynamic properties of molecules and their reactivity with each other requires a potential energy surface. This hypersurface represents the forces produced by the interactions of all the electrons in the system under study. Those forces change as the atoms in the molecules change with respect to one another. Consequently, the potential energy surface is a function of all the molecular coordinates and thus can be highly multidimensional. (For example, the octane in gasoline has over 70 molecular coordinates to describe its exact configuration.) Because electrons are very light and very fast, their behavior can only be accurately followed quantum mechanically. As a consequence, a given point on the potential energy surface requires very expensive calculations. Such calculations can be run on powerful parallel computers but even under such circumstances, they take too long to map out the relevant portions of the potential energy surface. Hence, a pressing problem is to develop efficient and accurate ways to interpolate discrete quantum mechanical electronic structure calculations into a continuous potential energy surface. The necessary fitting methods hopefully will not require high performance computing but they will direct where high performance computing of discrete points should be carried out.

Our collaboration is exploring higher order Interpolative Moving Least Squares (IMLS) [1],[2],[3],[4],[5] as the solution to this major fitting problem. The IMLS approach involves solving a weighted linear least squares problem every time an evaluation of the surface is required. It has several advantages:

• Special weights confer non-linear flexibility to the fit. As a consequence IMLS can accurately interpolate broad regions of the surface with a quite sparse set of explicit calculations of points on the surface.
IMLS can by itself estimate where it is most inaccurate and thus automatically direct new electronic structure calculations for specific points on the surface to improve its accuracy. This allows IMLS to develop surface fits to a desired accuracy without human assistance. IMLS is a black box fitting method where the user need only specify the fitting basis set. Of course, superior basis sets will be more efficient than poorly chosen ones.

The major disadvantage of IMLS is that a least squares solution must be carried out every time a value on the fitted surface is required. In the past year, we have worked on developing automatic surface generation capabilities, on reducing the cost of the least squares procedure, and on developing a distributable “black box” IMLS code.

2. Automatic Surface Generation
The weights applied to each computed point being fit by IMLS change with where the surface is being evaluated. In particular, the weights decrease very rapidly with the distance of the evaluation point from a computed point. One consequence of this is that any IMLS fit, no matter what the fitting basis set, very accurately reproduces each computed points because the weight for that point becomes near singular. Of course, in between computed points different basis sets produce IMLS fits that do differ. Where they differ the most is an indication of the least certain region of the hypersurface. Consequently, after there are enough seed computed points for an attempted IMLS surface, finding the point of maximum difference between IMLS surfaces built with different basis sets identifies where the next computed point should be. As new points are systematically computed in this way, the maximum and the global rms differences between IMLS surfaces built with different basis sets decrease. If the basis sets are at all reasonable, this rms difference should track the rms error of either IMLS fit with the true surface. Thus new points can be systematically computed until the rms difference between the IMLS fits drops below an input value. The result is an automatically generated surface of predetermined accuracy.

We have most recently tested this approach in developing a six dimensional IMLS potential energy surface for HOOH accurate up to 100 kcal/mol.[5] Rather than use electronic structure calculations for this test, we used a pre-existing HOOH surface [6] that allows us to monitor the accuracy of the IMLS surface as it is being generated. We used 89 points generally selected in the vicinity of the HOOH minimum augmented by points selected on eight global grids of 100x2^n points where n = 0,..,7. By design, all grid points have energy values below 100 kcal/mol. The grids for n = 0,..,4 were considered as seed points from which to automatically grow the surface by contrasting a second order direct product basis with a third order direct product basis in the six distance coordinates. All fits were monitored for accuracy by computing the rms error on tens of thousands of randomly selected points below 100 kcal/mol on the surface. The results are show in Fig. 1 Rms error in the energy for various IMLS fully dimensional HOOH surfaces as a function of the number of computed points. The open circle and dash-dotted line are results for eight different global grids. The dotted lines are for automatically generated surfaces (see text for details). The thin solid line is the 1.0 kcal/mol standard for chemical accuracy.
Fig. 1. The open circles show the rms error for the eight grids. The rms error declines with each larger grid and reaches chemical accuracy of 1.0 kcal/mol at about 6500 points. The results with automatic surface generation all show a faster approach to chemical accuracy with increased numbers of computed points. No matter what the number of seed points are, the rms error for IMLS surfaces being automatically generated tends to merge into a common line that reaches 1.0 kcal/mol accuracy at ~1400 computed points. Thus automatic surface generation reduces the number of required computer points by a factor of 4 to 5. 1400 points is equivalent to a six-dimensional grid with only 3.3 points per dimension. However, rather than locating points on a grid, the automatically generated surface has points in the most optimal locations.

3. Reducing IMLS Costs

The flexibility of IMLS allows a relatively accurate representation of a hypersurface based on few computed points on the surface. However, that flexibility comes at the price of doing a least squares fit every time the surface must be evaluated. The cost of the least squares procedure goes as \( Nm^2 \) where \( N \) is the number of computed points to be fit by the least squares procedure and \( m \) is the number of basis functions in the fit. Consequently, there are only three approaches for reducing IMLS evaluation costs: reduce \( N \) by fitting only a local subset of points, reduce \( m \) by developing unusually flexible (and probably non-linear) basis functions, and reuse least squares solutions from prior “nearby” IMLS evaluation. Only our work reducing \( N \) will be described here.

Rigorously \( N \) should be the total number of computed points on the hypersurface being fit. However, the weights used in the IMLS method generally decrease by some high power in the distance of the evaluation point from the location of a computed point. Thus while formally global, in practice the IMLS fit is highly local. The weights then can be exploited as a screen to reduce \( N \) to the \( N_{\text{eff}} \) local computed points that really affect the least squares procedure. This screening must be done in a way that preserves a continuous representation of the hypersurface in value and derivative. We have developed two ways of doing this based on using a cutoff radius \( r_{\text{cut}} \) such that only computed points contained within the hypersphere of radius \( r_{\text{cut}} \) centered on the evaluation point are included in the least squares procedure. To ensure continuity in value and gradient, the normal weights are reduced to exactly zero at the surface of the hypersphere by using a multiplicative switch that is a function of only the radial distance and whose value and lowest few derivatives go to zero at the hypersphere surface.

The first and simplest method is called the fixed radius cutoff (FRC) and uses a fixed input value of \( r_{\text{cut}} \). With the appropriate switch for the weights, the FRC produces a reliable representation of the hypersurface with very low overhead costs (a single pass over all \( N \) points to determine their radial distances). However, in practice there are usually regions of the hypersurface with relatively sparse numbers of local computed points. In these regions, an aggressive value of \( r_{\text{cut}} \) can encircle too few
computed points, leading to significant errors in the IMLS evaluation relative to including all the points. In other words, to obtain a global reliability, the fixed value of \( r_{\text{cut}} \) must to set for the few regions of the hypersphere most sparsely populated by computed points, implying the FRC \( N_{\text{eff}} \) is excessive in all the other more densely sampled regions.

The second method is called the variable number cutoff (VNC) and dynamically derives an \( r_{\text{cut}} \) value appropriate to the density of computed points in the vicinity of the evaluation point. The same switch used to turn off the weights at the surface of the hypersphere is used to assign a fractional value to each computed point. The value derived for \( r_{\text{cut}} \) is that required for the accumulated fractional values of computed points within the hypersphere equal a fixed input value \( \zeta \). This ensures \( N_{\text{eff}} \geq \zeta \) and also allows an equally aggressive approach to \( r_{\text{cut}} \) in regions of sparse or dense populations of computed points. The overhead associated with this approach is somewhat higher than that for the FRC because \( r_{\text{cut}} \) requires an iterative solution.

In Fig. 2, both the FRC and the VNC approaches are applied to a six-dimensional HOOH IMLS fit with \( \sim3000 \) computed points. The results in the figure report rms errors on tens of thousands of randomly selected points. Both rms errors in the value as well as the derivatives are reported. (Note classical trajectories computed on the surface primarily use derivatives.) The rms errors are a function of the input value of \( r_{\text{cut}} \) (for FRC) or \( \zeta \) (for VNC). The smaller these values, the smaller \( N_{\text{eff}} \) and the faster the computation time relative to no cutoff. Rather than plot rms error versus input value, Fig. 2 shows the rms error versus the corresponding relative computation time. The results in the figure show that 90% to 95% reductions in computation time are possible at the cost of less than a 10% increase in the rms error for either value or derivative. The results also show that the VNC is more efficient than the FRC despite its higher overhead.

4. Code Developments
The “black box” nature of IMLS means that its associated software can be very general. Our current IMLS software is available at chem.missouri.edu/thompson/SciDAC.html. The user must supply a weight function, a cutoff method (if so desired), and a fitting basis set. The software will then produce for any set of explicitly calculated points a fit in value and derivative. The derivatives are calculated directly with no finite differences. The web site contains a selection of basis sets, weights, and cutoff routines the user can select from. Test cases and archived output ensure that any downloaded code is working properly.

5. Conclusion
The results in this and earlier papers show that the IMLS method can accurately represent hypersurfaces of interest to chemistry with very few computed points and without gradients or hessians. We have also shown that the promise of automatic surface generation can be realized in practice. However, all our studies to date have been on hypersurfaces of six dimensions or fewer. In the future, we plan to examine higher dimensional surfaces and develop several hopefully simple guidelines for how the minimum number of computed points scales with dimensions.

The IMLS method achieves its accuracy with so few points because every evaluation formally requires a least squares solution that emphasizes the hypersurface region around the evaluation point. Consequently, compared to other methods, IMLS evaluations are expensive. Even if this expense cannot be reduced, the IMLS evaluation costs is very much less than the cost of computing a point on hypersurfaces of interest to chemistry. At worst, an expensive IMLS fit could provide the much denser set of “computed” points required by other fitting methods that are less expensive to evaluate. However, the IMLS expense can be reduced in three ways, one of which, the screening out of uninteresting computed points, was discussed in this article. The other two approaches we are also currently at work on. In one of these two approaches, we are developing compact non-linear basis functions that reduce the number of coefficients the least squares procedure must determine. For example, the Morse oscillator representation of bond stretching motion on potential energy surfaces is quadratic in a basis of \((e^{-\alpha x})^n\) for \( n = 0, 1, \ldots \). Such a basis involves the selection of the parameter \( \alpha \) but offers great flexibility. The second approach is to reuse the least squares solution. Many applications
of hypersurface fitting involve propagating solutions (e.g., trajectories, steepest descent paths, etc.) on the hypersurface. In such cases, an evaluation of the surface is nearby some prior evaluation of the surface. The IMLS value changes due to both changes in the basis and changes in the least squares coefficients. The lower order derivatives of the coefficients can be computed at any one point at an expense generally less than solving the least squares from scratch. In effect one can establish a trust radius around each evaluation point where the least squares solution can be reused without compromising the accuracy of the representation. How effective this approach would be should depend on the basis. For example, if a bond stretch on a potential energy surface were exactly a Morse oscillator in form, then the least squares solution could be reused over the entire range of the stretching motion if a \((e^{\alpha x})^n\) for \(n = 0, 1,\ldots\) basis were used with the proper value of \(\alpha\). Thus the two approaches to cost reduction - nonlinear basis functions and least squares solution reuse – are connected.

While our focus has been on potential energy surfaces, IMLS is a general method for fitting hypersurfaces and it is especially well adapted to hypersurfaces for which the gradients and Hessians of computed points are difficult to obtain. In many simulation problems, the final conditions generated by propagating differential equations from input conditions for a specific length of time produce response hypersurfaces whose dimensions are the number of input variables. Optimizing performance by finding extrema on the surface or obtaining final conditions for input into a further multiscale simulation all require a representation of the response hypersurface where computed points on the hypersurface generally do not come with gradients or Hessians. IMLS can be an effective approach for fitting such surfaces.

Whether representing hypersurfaces in chemistry or more general hypersurfaces, the role of IMLS in advanced computing is the minimization of the number of high performance simulations needed to compute points on the surface. The automatic surface generation capabilities of IMLS make it the director of high performance simulations at points on the surface it selects without human assistance. We are at work in our chemistry applications in integrating IMLS and electronic structure codes to achieve the generation of complete hypersurfaces of specified accuracy upon command.

References
[1] Maisuradze G G and Thompson D L 2003 J. Phys. Chem. A 107 7118
[2] Maisuradze G G, Thompson D L, Wagner A F and Minkoff M 2003 J. Chem. Phys. 119 10002
[3] Kawano A, Guo Y, Thompson D L, Wagner A F and Minkoff M 2004 J. Chem. Phys. 120 6414
[4] Guo Y, Kawano A, Thompson D L, Wagner A F and Minkoff M 2004 J. Chem. Phys. 121 5091
[5] Maisuradze G G, Kawano A, Thompson D L, Wagner A F and Minkoff M 2004 J. Chem. Phys. 121 10329
[6] Kuhn B, Rizzo T R, Luckhaus D, Quack M and Suhm 1999 J. Chem. Phys. 111 2565