PS 10: Final Project – Paper Outline
Automated Computational Methods for Solving the Molecular Schrödinger Equation

Alexander Lin
Supervised by Dr. Michael Mavros

November 18, 2018

1 Main Concept

In the world of chemistry, we are greatly interested in the ability to accurately solve the Schrödinger equation. For molecular structures, solutions to the Schrödinger equation allow us to characterize many interesting properties, such as atomization energies, bond lengths, and oscillation frequencies, to name a few.

However, since there do not exist exact, analytical solutions to many-body electron systems, the scientific community has resorted to computational approaches to make numerical approximations. One of the most famous Schrödinger solvers that has been used throughout history is the Hartree-Fock algorithm [8]. By using a series of approximations – such as Born-Oppenheimer, the single Slater Determinant, and the variational method – Hartree-Fock allows us to compute an *upperbound* to the true ground state energy E of any given molecule, using nothing but the Cartesian coordinates of its constituent atoms.

In this project, we plan to automate the implementation of Hartree-Fock for a large dataset of molecules. Specifically, we wish to investigate the efficacy of Hartree-Fock in solving the Schrödinger equation for the GDB-13 dataset, an exhaustive enumeration of over 970 million small, organic, and druglike molecules containing up to 13 atoms of cabon, nitrogen, oxygen, sulfur, and chlorine that are saturated with hydrogens [1]. We will focus on a subset of this dataset, namely QM7b, which was created to reduce the original dataset to a more manageable 7211 structures, while attempting to maintain the rich diversity of GDB-13 [4].

Specifically, we wish to explore questions surrounding the efficiency of Hartree-Fock on this dataset. These include characterizing how long it takes for the algorithm to run, how this quantity varies across different molecules in the dataset, if there exists a relationship between the algorithm's efficiency and the molecule's length or its specific constituent atoms, etc. QM7b also provides atomization energies values that are essentially exact [4], so it would be nice to use these in benchmarking the accuracy of Hartree-Fock's approximation. Given more time, we could also benchmark its performance on other properties that are provided

in QM7b, such as polarizability, HOMO and LUMO eigenvalues, and excitation energies. All of these results will be displayed in the form of figures for the final paper.

Finally, there has been a recent interest in using machine learning methods as an alternative for more traditional, physics-based approaches in approximating solutions to the Schrödinger equation. It would be interesting to compare accuracy and efficiency of Hartree-Fock against existing machine learning algorithms that have been used for this dataset. These include kernel ridge regression [5] and the multilayer perceptron [3, 4]. The code for these algorithms are freely available http://quantum-machine.org/. Given more time, we could also develop our own machine learning algorithms for additional comparison. In summary, it would be interesting to observe if machine learning can perform better than classical methods, or if it is really not necessary for these types of problems.

2 Implementation of Hartree-Fock

We will implement Hartree-Fock using the Python programming language to automatically calculate molecular ground-state energies and other quantities of interest, such as atomization energies. Notation and methods are mainly adapted from [8] and [7], the latter of which has an excellent set of resources at http://vergil.chemistry.gatech.edu/notes/.

Let a molecule be comprised of m atoms and have n electrons. Given an m-by-3 matrix of atomic coordinates $\boldsymbol{R}$, Hartree-Fock approximately solves the multi-electron Schrödinger equation

$$\hat{H}_{el}(\boldsymbol{r}; \boldsymbol{R})\Psi(\boldsymbol{r}; \boldsymbol{R}) = E_{el}(\boldsymbol{R})\Psi(\boldsymbol{r}; \boldsymbol{R})$$

for the wavefunction $\Psi(\boldsymbol{r})$ over n electron coordinates $\boldsymbol{r}$ and the energy E_{el}. Taking the Born-Oppenheimer approximation, we can write the Hamiltonian operator as

$$\hat{H}_{el} = \sum_{i=1}^{n} \hat{h}(i) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \hat{v}(i,j) + \hat{V}_{NN}$$

where $\hat{h}(i) = -1/2 \cdot \nabla_i^2 - \sum_{a=1}^{m} Z_a/r_{ia}$ characterizes an electron's kinetic energy and its attraction to all nuclei, $\hat{v}(i,j) = 1/r_{ij}$ characterizes repulsion between two electrons, and $\hat{V}_{NN}$ characterizes all nucleus-nucleus repulsions. In general, we use the notation $r_{\alpha\beta}$ to denote the Euclidean distance between two entities – either electrons or nuclei – indexed by α and β, respectively.

We additionally assume that Ψ is an anti-symmetric product-sum of single-electron, molecular wavefunctions $\chi_1, \ldots, \chi_n$ over coordinates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ corresponding to the rows of matrix $\boldsymbol{r}$. This is known as the Slater Determinant, or

$$\Psi = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(\boldsymbol{x}_1) & \chi_2(\boldsymbol{x}_1) & \cdots & \chi_n(\boldsymbol{x}_1) \\ \chi_1(\boldsymbol{x}_2) & \chi_2(\boldsymbol{x}_2) & \cdots & \chi_n(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\boldsymbol{x}_n) & \chi_2(\boldsymbol{x}_n) & \cdots & \chi_n(\boldsymbol{x}_n) \end{vmatrix}$$

In finding the ground-state energy of a molecule, the goal of Hartree-Fock is to solve the following optimization problem,

$$E_{HF} = \min_{\Psi} E_{el} = \min_{\Psi}\langle\Psi|\hat{H}_{el}|\Psi\rangle = \min_{\chi_1,\dots,\chi_n} \sum_{i=1}^{n}\langle i|\hat{h}|i\rangle + \sum_{i=1}^{n}\sum_{j=i+1}^{n}[ii|\hat{v}|jj] - [ij|\hat{v}|ji]$$

where

$$\langle i|\hat{h}|i\rangle = \int_{\mathbb{R}} \chi_i^*(\boldsymbol{x})\hat{h}(i)\chi_i(\boldsymbol{x})d\boldsymbol{x}$$

$$[ii|\hat{v}|jj] = \int_{\mathbb{R}^2} \chi_i^*(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1)\hat{v}\chi_j^*(\boldsymbol{x}_2)\chi_j(\boldsymbol{x}_2)d\boldsymbol{x}_1\boldsymbol{x}_2$$

$$[ij|\hat{v}|ji] = \int_{\mathbb{R}^2} \chi_i^*(\boldsymbol{x}_1)\chi_j(\boldsymbol{x}_1)\hat{v}\chi_j^*(\boldsymbol{x}_2)\chi_i(\boldsymbol{x}_2)d\boldsymbol{x}_1\boldsymbol{x}_2$$

These integrals are analytically intractable, so we use the variational method. Specific, we use linear combinations of Gaussian basis sets to approximate atomic orbitals that themselves form linear combinations of these molecular orbitals. Originally developed by [2] and compiled by [6], we consult https://bse.pnl.gov/bse/portal for a list of basis approximations. Eventually, the algorithm boils down to the Hartree-Fock-Roothan equations [8], which iteratively solves pseudo-eigenvalue problems to determine the ground-state energy of a molecule.

I plan on heavily utilizing the numerical computing packages NumPy and SciPy in my implementation of Hartree-Fock.

3   Paper Outline

Here is a brief outline of the paper.

1. MOTIVATION - explain why Hartree-Fock is important, what quantities we can calculate by solving the Schrödinger equation

2. THEORY - explain the mathematics behind the algorithm

   - Approximations taken by Hartree-Fock - i.e. Born-Oppenheimer, Slater, variational method
   - Spin integration in exchange integrals
   - Hartree-Fock-Roothan equations and self-consistent field procedure

3. IMPLEMENTATION - explain the process of coding the algorithm

   - Packages utilized for numerical computations
   - Specific basis sets employed by our implementation
   - Hyperparameters of the algorithm - e.g. tolerance of self-consistent field procedure, number of iterations, etc.

4. RESULTS - show the utility of the algorithm

- Efficiency results - how long does it take to find the ground-state energy of molecules; are there certain molecules that take longer than others
- Accuracy results - how accurate is the algorithm at computing atomization energies; are there certain molecules that are more accurate than others
- (*Optional.*) Comparison with machine learning approaches

References

[1] Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25):8732–8733, 2009.

[2] David Feller. The role of databases in support of computational chemistry calculations. *Journal of computational chemistry*, 17(13):1571–1586, 1996.

[3] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, pages 440–448, 2012.

[4] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.

[5] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[6] Karen L Schuchardt, Brett T Didier, Todd Elsethagen, Lisong Sun, Vidhya Gurumoorthi, Jared Chase, Jun Li, and Theresa L Windus. Basis set exchange: a community database for computational sciences. *Journal of chemical information and modeling*, 47(3):1045–1052, 2007.

[7] C David Sherrill. A brief review of elementary quantum chemistry. *Georgia Institute of Technology, School of Chemistry and Biochemistry*, 2001.

[8] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.