PS 10: Final Project – Final Paper

Automated Computation Of Molecular Properties From First Principles

Alexander Lin
Supervised by Dr. Michael Mavros

December 16, 2018

1 Introduction

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 total energies, ionization potentials, and interatomic bond lengths, 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 that make numerical approximations. One of the most famous Schrödinger solvers that has been used throughout history is the Hartree-Fock algorithm [6]. By using a series of approximating methods – 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 and nuclear charges of its constituent atoms.

In this project, we automate the implementation of Hartree-Fock for small molecules and derive interesting, experimentally-verified properties from first principles. We begin by testing Hartree-Fock's ability to recover bond lengths of very simple inorganic molecules such as hydrogen fluoride and nitrogen gas. Next, we evaluate how well the algorithm is able to calculate the total energy of various multi-electron atoms.

Finally, we investigate the efficacy of Hartree-Fock in solving the Schrödinger equation for the GDB-13 dataset, an exhaustive enumeration of over 970 million organic and druglike molecules containing up to 13 atoms of cabon, nitrogen, oxygen, sulfur, and chlorine that are saturated with hydrogens [2]. We focus on a subset of this dataset – namely 59 small molecules with up to four non-hydrogen atoms from QM7b[1], which was created to reduce the original dataset to a more manageable number of structures while maintaining the rich diversity of GDB-13 [3]. For each organic molecule, we compare a Hartree-Fock calculation of its total energy and first ionization potential to ground-truth values. We also analyze trends

[1] Freely available at http://quantum-machine.org/datasets/.

within specific organic families such as alkanes and alkynes to show that Hartree-Fock can successfully recover these trends. We conclude the analysis with experimental evidence that the asymptotic running time of Hartree-Fock is $\mathcal{O}(n^4)$, where n is the number of orbitals considered by the algorithm for a given molecule.

The rest of this paper is organized as follows: Section 2 explains the mathematical theory behind the Hartree-Fock algorithm. Section 3 details the process of implementing the algorithm, along with some technical specifications. Section 4 presents the main results, thereby providing some evidence of the algorithm's utility. And finally, Section 5 concludes the paper and touches on some potential future work.

2 Theory

In this section, we heavily utilize the notation of [5] and [6]. We highly recommend interested readers to consult either of these comprehensive sources for additional information about Hartree-Fock theory.

2.1 Initial Approximations

Let us start with the Hamiltonian $\hat{H}$ and corresponding ground-state energy E_{tot} for a multi-electron system. The Hamiltonian can be characterized by five main components,

$$\hat{H} = \hat{T}_N(\boldsymbol{R}) + \hat{T}_e(\boldsymbol{r}) + \hat{V}_{NN}(\boldsymbol{R}, \boldsymbol{Z}) + \hat{V}_{eN}(\boldsymbol{r}, \boldsymbol{R}, \boldsymbol{Z}) + \hat{V}_{ee}(\boldsymbol{r}), \tag{1}$$

where $\hat{T}_N, \hat{T}_e$ respectively describe the kinetic energies of the nuclei and electrons; and $\hat{V}_{NN}, \hat{V}_{eN}, \hat{V}_{ee}$ respectively describe the Coulombic potential energies of nucleus-nucleus repulsion, electron-nucleus attraction, and electron-electron attraction. Here, $\boldsymbol{R} = \{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_M\}$ and $\boldsymbol{r} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$ are matrices that hold three-dimensional coordinates for the M nuclei and N electrons of the molecule in question. The vector $\boldsymbol{Z} = \{Z_1, \ldots, Z_M\}$ denotes the charges for the M nuclei. Note that $\boldsymbol{R}, \boldsymbol{Z}$ are inputs to the algorithm, whereas $\boldsymbol{r}$ is characterized by the wavefunction. In general, we will use $\{i, j, k\}$ to index electrons and $\{A, B, C\}$ to index nuclei.

The first approximation taken by Hartree-Fock is Born-Oppenheimer, which drops $\hat{T}_N$ from the Hamiltonian. The other four terms can be expanded as,

$$\hat{T}_e(\boldsymbol{r}) = -\frac{1}{2}\sum_{i=1}^{N} \nabla_i^2, \tag{2}$$

$$\hat{V}_{NN}(\boldsymbol{R}, \boldsymbol{Z}) = \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}}, \tag{3}$$

$$\hat{V}_{eN}(\boldsymbol{r}, \boldsymbol{R}, \boldsymbol{Z}) = -\sum_{A=1}^{M}\sum_{i=1}^{N} \frac{Z_A}{r_{Ai}}, \tag{4}$$

$$\hat{V}_{ee}(\boldsymbol{r}) = \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}}, \tag{5}$$

where $R_{AB} = \|\boldsymbol{R}_B - \boldsymbol{R}_A\|_2$ is an internuclear distance, $r_{ij} = \|\boldsymbol{r}_j - \boldsymbol{r}_i\|_2$ is an electron-electron distance, and $r_{Ai} = \|\boldsymbol{r}_i - \boldsymbol{r}_A\|_2$ is an nucleus-electron distance. One immediate observation from Equation 3 is that the operator $\hat{V}_{NN}$ has no dependence on electron coordinates $\boldsymbol{r}$; therefore, we can simply calculate this quantity at the beginning of the algorithm and leave it to the side. It follows that the electronic Schrödinger equation may be simplified as

$$\hat{H}_{ele}\Psi(\boldsymbol{r};\boldsymbol{R},\boldsymbol{Z}) = \left[\hat{T}_e(\boldsymbol{r}) + \hat{V}_{eN}(\boldsymbol{r},\boldsymbol{R},\boldsymbol{Z}) + \hat{V}_{ee}(\boldsymbol{r})\right]\Psi(\boldsymbol{r};\boldsymbol{R},\boldsymbol{Z}) = E_{ele}\Psi(\boldsymbol{r};\boldsymbol{R},\boldsymbol{Z}), \tag{6}$$

where the total energy of the multi-electron system $E_{tot} = E_{ele} + V_{NN}$ is the sum of the electronic and nuclear energies.

The antisymmetry principle states that for a system of fermions, the wavefunction must be antisymmetric with respect to changes in position *and* spin of any two fermions [5]. To satisfy this principle, we must introduce a new variable (i.e. the spin coordinate ω) for each electron and define the wavefunction Ψ in terms of $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, where each $\boldsymbol{x}_i = \{\boldsymbol{r}_i, \omega_i\}$. The following equation describes antisymmetry for any two electrons i, j,

$$\Psi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_N) = -\Psi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_N). \tag{7}$$

We would like to express Ψ as some aggregated function of single-electron, molecular wavefunctions $\chi_1, \ldots, \chi_n$ to make calculations easier. Perhaps the most straightforward way to do this while satisfying Equation 7 – and its immediate corollary, the Pauli exclusion principle – is to let Ψ be a Slater determinant,

$$\Psi = \frac{1}{\sqrt{n!}}\begin{vmatrix} \chi_1(\boldsymbol{x}_1) & \chi_2(\boldsymbol{x}_1) & \cdots & \chi_N(\boldsymbol{x}_1) \\ \chi_1(\boldsymbol{x}_2) & \chi_2(\boldsymbol{x}_2) & \cdots & \chi_N(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\boldsymbol{x}_N) & \chi_2(\boldsymbol{x}_N) & \cdots & \chi_N(\boldsymbol{x}_N) \end{vmatrix}. \tag{8}$$

That is, we assume Ψ to be an anti-symmetric product-sum. It is this assumption that makes Hartree-Fock a *mean field approximation*, which means that each electron feels the average repulsive cloud of other electrons, but not the individual effects. For this reason, the Hartree-Fock model fails to capture certain real-world phenomena, such as London dispersion, that occur between specific sets of electrons. There exists an entire body of literature on post-Hartree-Fock methods for improving this approximation [1].

2.2 Hartree-Fock Energy

Following Sherrill's notation [5], we can compactly re-express Equations 2 and 4 as a one-electron operator $\hat{h}$,

$$\hat{T}_e(\boldsymbol{r}) + \hat{V}_{eN}(\boldsymbol{r},\boldsymbol{R},\boldsymbol{Z}) = \sum_i \left(-\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{r_{iA}}\right) = \sum_i \hat{h}(i), \tag{9}$$

and Equation 5 as a two-electron operator $\hat{v}$,

$$\hat{V}_{ee}(\boldsymbol{r}) = \sum_{i<j} \hat{v}(i,j) = \sum_{i<j} \frac{1}{r_{ij}}. \tag{10}$$

It follows that we can simply re-write the electronic Hamiltonian of Equation 6 as

$$\hat{H}_{ele} = \sum_i \hat{h}(i) + \sum_{i<j} \hat{v}(i,j). \tag{11}$$

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

$$E_{HF} = \min_{\Psi} E_{ele} = \min_{\Psi} \langle \Psi | \hat{H}_{el} | \Psi \rangle = \min_{\chi_1,\ldots,\chi_N} \sum_{i=1}^{n} \langle i|\hat{h}|i\rangle + \sum_{i=1}^{n} \sum_{j=i+1}^{n} [ii|jj] - [ij|ji], \tag{12}$$

where

$$\langle i|\hat{h}|i\rangle = \int_{\mathbb{R}} \chi_i^*(\boldsymbol{x}) \hat{h}(i) \chi_i(\boldsymbol{x}) d\boldsymbol{x}, \tag{13}$$

$$[ii|jj] = \int_{\mathbb{R}^2} \chi_i^*(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1)\hat{v}(i,j)\chi_j^*(\boldsymbol{x}_2)\chi_j(\boldsymbol{x}_2) d\boldsymbol{x}_1\boldsymbol{x}_2, \tag{14}$$

$$[ij|ji] = \int_{\mathbb{R}^2} \chi_i^*(\boldsymbol{x}_1)\chi_j(\boldsymbol{x}_1)\hat{v}(i,j)\chi_j^*(\boldsymbol{x}_2)\chi_i(\boldsymbol{x}_2) d\boldsymbol{x}_1\boldsymbol{x}_2. \tag{15}$$

Working through Lagrange's method of undetermined multipliers for this optimization, as detailed in [4], we arrive at the eigenvalue problem,

$$f(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1) = \epsilon_i \chi_i(\boldsymbol{x}_1), \tag{16}$$

where the Fock operator f is defined by

$$\begin{aligned} f(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1) = h(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1) &+ \sum_{j\neq i} \left[\int |\chi_j(\boldsymbol{x}_2)|^2 \frac{1}{r_{12}} d\boldsymbol{x}_2 \right] \chi_i(\boldsymbol{x}_1) \\ &- \sum_{j\neq i} \left[\int \chi_j^*(\boldsymbol{x}_2)\chi_i(\boldsymbol{x}_2) \frac{1}{r_{12}} d\boldsymbol{x}_2 \right] \chi_j(\boldsymbol{x}_1), \end{aligned} \tag{17}$$

and ϵ_i is the energy of molecular orbital i. The operator involves the integration of complicated expressions, so to make things analytically tractable, we introduce a basis set of easy-to-integrate atomic orbitals $\tilde{\chi}_1, \ldots, \tilde{\chi}_K$. Typically, $\tilde{\chi}_\mu$ is a linear combination of Gaussians whose coefficients have been optimized to fit Slater-type orbitals; we elaborate more on this in Section 3. In doing so, we employ the variational principle – another source of approximation.

We now have that each molecular orbital i is a linear combination of atomic orbitals (i.e. MO-LCAO method) with coefficients $C_{1i}, \ldots, C_{Ki}$,

$$\chi_i = \sum_{\mu=1}^{K} C_{\mu i} \tilde{\chi}_\mu. \tag{18}$$

From this, we can rewrite Equation 16 as the Hartree-Fock-Roothan equations,

$$\sum_\nu F_{\mu\nu} C_{\nu i} = \epsilon_i \sum_\nu S_{\mu\nu} C_{\nu i}, \tag{19}$$

where we have the more tractable integrals,

$$S_{\mu\nu} = \int \tilde{\chi}^*_\mu(\boldsymbol{x}_1)\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{x}_1, \tag{20}$$

$$F_{\mu\nu} = \int \tilde{\chi}^*_\mu(\boldsymbol{x}_1)f(\boldsymbol{x}_1)\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{x}_1. \tag{21}$$

In matrix form, this can be written as

$$\boldsymbol{FC} = \boldsymbol{SC}\boldsymbol{\epsilon} \tag{22}$$

Equation 22 is a peculiar eigenvalue equation, because $\boldsymbol{F}$ depends on $\boldsymbol{C}$ and vice-versa. This means that both cannot be optimized simultaneously; instead, the Hartree-Fock algorithm must alternately update these two matrices until convergence.

2.3 Hartree-Fock Algorithm

The Hartree-Fock algorithm [6] can be divided into two main parts – (1) integration and (2) iteration. The integration part tends to dominate in terms of computation time.

During the integration part, there are four main integrals of interest that need to be pre-computed. The first is the overlap integral $S_{\mu\nu}$ for every pair of atomic orbitals μ, ν, as described by Equation 20. The other three – kinetic energy $T_{\mu\nu}$, nuclear-electron attraction $V^{\text{nucl}}_{\mu\nu}$, and electron-electron repulsion $[\mu\nu|\lambda\sigma]$ – are involved in the Fock integral of Equation 21. Their expressions come straight from the operators – as defined by Equations 2, 4, and 5 – applied to the atomic basis functions,

$$T_{\mu\nu} = \int \tilde{\chi}^*_\mu(\boldsymbol{x}_1)\left[-\frac{1}{2}\nabla^2_1\right]\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{r}_1 \tag{23}$$

$$V^{\text{nucl}}_{\mu\nu} = \int \tilde{\chi}^*_\mu(\boldsymbol{x}_1)\left[-\sum_A \frac{Z_A}{r_{A1}}\right]\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{r}_1 \tag{24}$$

$$[\mu\nu|\lambda\sigma] = \int\int \tilde{\chi}^*_\mu(\boldsymbol{x}_1)\tilde{\chi}_\nu(\boldsymbol{x}_1)\frac{1}{\boldsymbol{r}_{12}}\tilde{\chi}^*_\lambda(\boldsymbol{x}_2)\tilde{\chi}_\sigma(\boldsymbol{x}_2)d\boldsymbol{r}_1\boldsymbol{r}_2 \tag{25}$$

After these integrals are computed, the algorithm proceeds by alternately changing $\boldsymbol{F}$ and $\boldsymbol{C}$. A full description is given by [6] and summarized in Algorithm 1. Note that this is the restricted Hartree-Fock procedure, which treats electrons as paired fermions and works for an even number of electrons.

3 Implementation

Our implementation of Hartree-Fock follows Algorithm 1.

References

[1] Rodney J Bartlett and John F Stanton. Applications of post-hartreefock methods: A tutorial. *Reviews in computational chemistry*, pages 65–169, 1994.

Algorithm 1 Restricted Hartree-Fock

Input: nuclear coords $\boldsymbol{R}$, charges $\boldsymbol{Z}$, atomic basis functions $\tilde{\chi}_\mu$, num of electrons N
Compute nuclear-nuclear repulsion V_{NN}.
Compute integrals $S_{\mu\nu}$, $T_{\mu\nu}$, $V_{\mu\nu}^{\text{nucl}}$, $[\mu\nu|\lambda\sigma]$.
Construct orthonormal basis transformation matrix $\boldsymbol{X}$ from $\boldsymbol{S}$ using canonical method.
Initialize density matrix $\boldsymbol{P} = \mathbf{0}$.
Initialize Fock matrix $F_{\mu\nu} = T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}}$.
while $\boldsymbol{P}$ has not converged **do**
    Calculate transformed Fock matrix $\boldsymbol{F'} = \boldsymbol{X}^T \boldsymbol{F} \boldsymbol{X}$.
    Diagonalize $\boldsymbol{F'}$ to get eigenvectors $\boldsymbol{C'}$ and eigenvalues $\boldsymbol{\epsilon}$.
    Calculate $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{C'}$.
    Compute density matrix $P_{\mu\nu} = 2\sum_{i=1}^{N/2} C_{\mu i} C_{\nu i}$.
    Compute matrix $G_{\mu\nu} = \sum_{\lambda,\sigma} P_{\lambda\sigma} \left(2 \cdot [\mu\nu|\sigma\lambda] - [\mu\lambda|\sigma\nu]\right)$.
    Compute Fock matrix $F_{\mu\nu} = T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}} + G_{\mu\nu}$.
    Compute electronic energy $E_{ele} = \sum_{\mu,\nu} P_{\mu\nu} \cdot (T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}} + F_{\mu\nu})$.
end while
Compute total energy $E_{tot} = E_{ele} + V_{NN}$.
Output: $E_{tot}, \boldsymbol{\epsilon}, \mathbf{F}, \mathbf{P}, \mathbf{T}, \mathbf{V}^{\text{nucl}}$

[2] 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.

[3] 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.

[4] C David Sherrill. An introduction to hartree-fock molecular orbital theory. *Georgia inst. of technology*, 2000.

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

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