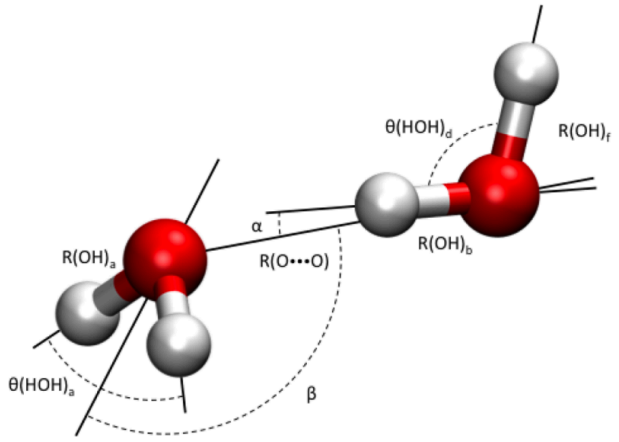
Chem 106: Computational Handout 6

Complete Basis Set Extrapolation: Water Dimer

Objectives: complete basis set extrapolation, additivity of correlation energy corrections

In This Exercise: we will consider the water dimer, the archetype of hydrogen bonding:



The best theoretical estimates (Lane *JCTC* **2013**, *9*, 316) place the electronic binding energy of this complex at –5.02 kcal/mol, which is in agreement with experiment when zero-point energy is taken into account. Here, we will see if we can reproduce this number with a small number of coupled-cluster (CC) calculations using complete basis set (CBS) extrapolation. Extrapolated energies are very useful for constructing benchmarks of small systems, which can be used to choose the best affordable method for a large system.

Which Program To Use: We will be using Gaussian 09, but other packages may well be more efficient, such as CFOUR, PSIFOUR, DALTON, ORCA, or MOLPRO. In G09, analytic gradients (but not frequencies) are available at the CCSD level. Neither analytic gradients nor frequencies are available at the CCSD(T) level. (Analytic derivatives are formulas that allow the derivative of energy to be calculated directly, as opposed to being estimated by finite differences.) Higher-order excitations are also not available in G09.

CCSD and CCSD(T) do not parallelize well beyond a few processors and have large memory requirements. However, because of symmetry and the small size of the molecules considered here, all the jobs can be performed in a few hours with moderate disk usage.

If you do have to run a large-scale CC job for a research project, the best strategy is to swap to a dedicated large volume, rather than the local node. On Odyssey, the dedicated large volume is a massive Lustre filesystem. For the duration of this class, you can access this system at `/n/regal/chem106`. You are allowed to use up to 1.2 PB (1 petabyte = 10^6 GB), which is (hopefully) more than enough for any CC job that can finish in finite time.

More information about `regal`: https://rc.fas.harvard.edu/resources/odyssey-storage/#regal

Alert! Files that are not modified for more than 90 days will be automatically deleted from `regal`.

Approach: We will compute water monomer and dimer using CCSD(T) with successively larger basis sets, and examine the convergence behavior. We will take a subset of this dataset and examine the performance of standard CBS extrapolation methods. We will also calculate the overall interaction

energy assuming the additivity of electron correlation energy corrections (HF and MP2 calculations are performed automatically during a CCSD(T) calculation).

Setup: standard. You will be generating your own output files, with the exception of that for the CSD(T)/cc-pV6Z calculation on water dimer (which requires at least a week). You will need the blank extrapolation Excel spreadsheet that is provided in the repository at `chem106_calculations/water/excel`.

Monomer Setup

1. Optimize the geometry of the monomer at CCSD/cc-pVTZ. This will get the geometry quite close to the true geometry. As we will see, the error incurred by this approximation is very small.

```
%nprocshared=4
%mem=3GB
#p cssd cc-pvtz opt freq=noraman

water monomer

0 1
O  -1.551007  -0.114520   0.000000
H  -1.934259   0.762503   0.000000
H  -0.599677   0.040712   0.000000
```

2. Take the final geometry of this optimization and create a template file called `water_monomer.gjf`:

```
%nprocshared=4
%mem=3GB
#p @method

water monomer CCSD/cc-pVTZ geometry

0 1
O  0.000000    0.118005    0.000000
H  0.753667   -0.472069    0.000000
H -0.753667   -0.471972    0.000000
```

3. You will use `sed` to generate input files from this template. `sed` stands for stream editor. You can think of `sed` as a search and replace program. For example:

```
~/chem106_calculations/water/prep $ sed s/@method/CCSD\(T\)\ cc-pVDZ/ water_monomer.gjf
%nprocshared=4
%mem=3GB
#p CCSD(T) cc-pVDZ

water monomer CCSD/cc-pVTZ geometry

0 1
O  0.000000    0.118005    0.000000
H  0.753667   -0.472069    0.000000
H -0.753667   -0.471972    0.000000
```

This command searches for (`s`) the string `@method` and replaces it with `CCSD(T) cc-pVDZ`. Note that both parentheses and the space must be escaped with a backslash `\`. `sed` reads in the file `water_monomer.gjf` and performs the search-and-replace, and sends the output to the screen (i.e., `water_monomer.gjf` itself is not modified). To save the result as a new file, *redirect* the output with `>`:

```
sed s/@method/CCSD\(T\)\ cc-pVDZ/ water_monomer.gjf > water_monomer-ccsd_t-dz.gjf
```

Alert! To modify the template file itself, use `sed -i s/foo/bar/ a.txt`. You should always check to make sure that you are actually performing the desired replacement, as this action cannot be undone. This command does not work on `macOS`. If you installed `GNU sed` in Exercise 0, then you can use the `-i` flag with the `gsed` command. The flag will work correctly on Odyssey.

Alert! The syntax used above, which is of the form `s/foo/bar/` will only replace the first instance of `foo` on every line. To replace *every* instance of `foo` with `bar`, add a `g` at the end, as in `s/foo/bar/g`. To make a replacement on a specific line, prepend the line number, as in `10s/foo/bar/`.

4. Create a script that will setup your jobs. Use vi to create a new file called `water_monomer_prep.sh`:

```
#!/bin/bash

sed s/@method/CCSD\(T\)\ cc-pVDZ/ water_monomer.gjf > water_monomer-ccsd_t-dz.gjf
sed s/@method/CCSD\(T\)\ cc-pVTZ/ water_monomer.gjf > water_monomer-ccsd_t-tz.gjf
…
```

This will let you copy and paste the repetitive search and replace action. (If you make a mistake in this file, you can use `sed` on it to repair the problem!)

We will need CCSD(T) calculations at cc-pVDZ (dz), cc-pVTZ (tz), cc-pVQZ (qz), cc-pV5Z (5z), and cc-pV6Z (6z). (The strings in parenthess are the abbreviations that should be used.)

5. Before a script can be run, it must have “execute” permissions:

```
chmod u+rwx water_monomer_prep.sh
```

6. Run the script: `./water_monomer_prep.sh`. Be sure your script does not overwrite the original template!

Dimer Setup:

1. We will use the frozen geometry at CCSD(T)/cc-pVQZ, which is part of the standard S22 benchmark for non-covalent complexes:

http://www.begdb.com/index.php?action=oneMolecule&state=show&id=82

For convenience, here is the template file:

```
%nprocshared=4
%mem=3GB
#p @method

water dimer CCSD(T)/cc-pVQZ geometry

0 1
O  -1.551007  -0.114520   0.000000
H  -1.934259   0.762503   0.000000
H  -0.599677   0.040712   0.000000
O   1.350625   0.111469   0.000000
H   1.680398  -0.373741  -0.758561
H   1.680398  -0.373741   0.758561
```

2. Use the same procedure as for the monomer to generate analogous files for the dimer.

3. Run all the jobs. Most of the jobs will be finished within several hours. Skip the calculation at cc-pV6Z (which would take about a week to run). The energies (hartree) for the this job are:

HF/cc-pV6Z: –152.140155589
MP2/cc-pV6Z: –152.73319889413
CCSD(T)/cc-pV6Z: –152.75202781

Extract Energies

We will need the Hartree–Fock (HF), MP2, and CCSD(T) energies. It is best to obtain these energies will full numeric precision, so please follow precisely the instructions below:

1. Extract the HF energies using `awk` to search for `SCF Done`. Since these are all single point jobs, there should only be one result per file. Use `sort` to get the energies ordered by basis set. I am hiding the answer below so you can try to figure out the right command first.

```
…
water_dimer-ccsd_t-dz.out -152.06253624 9
| sort -n -k2 -r
[ekwan@rcl ogin10 output]$ awk '/SCF Done/ {print FILENAME, $5}' water_dimer-ccsd_t*.out
```

2. Extract the MP2 energies by searching for the word `EUMP2`. Gaussian uses scientific notation of the form `1.23D+04`, so use the `awk` command `gsub` to change back to `E`. This will allow you to paste into Excel. Here is a hint to get you started:

```
awk '____ {gsub("D","E",$NF); print _____}' water_dimer-ccsd_t*.out
```

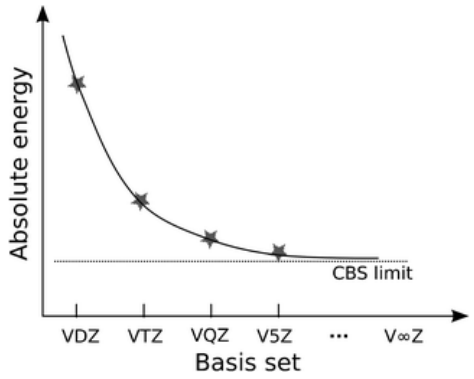
(Semicolons `;` separate commands so that multiple commands can be placed on the same line.)

3. Extract the CCSD(T) energies by searching for the word "`CCSD(T)=` ". You will need to perform the same replacement as in step 2. Remember to escape any necessary characters.

4. Paste the resulting energies for the monomer and dimer into the Excel spreadsheet `water_extrapolation.xlsx` in `chem106_calculations/water/excel`.

Extrapolating to the CBS Limit

As the basis set size is increased in a calculation, the energy decreases. In the limit of an infinite basis set, the best possible wavefunction will be found and the energy reaches a minimum called the **complete basis set (CBS)** limit. With the Dunning correlation consistent basis sets (and many other basis sets) it is found empirically that the CBS limit is approached smoothly:



Because each increase in the zeta level (i.e., the number of basis functions per atomic orbital) results in a large increase in computational cost (N^7 for CCSD(T)), it is generally impractical to perform calculations at the CBS limit. However, by plotting the energy as a function of zeta, we can project to the CBS limit. This is called "extrapolation to the CBS limit."

Many formulas have been proposed for the form of the relationship between energy and zeta (leading reference: Feller, *J. Chem. Phys.* **2013**, *138*, 074103). The choice of this functional form itself introduces some noise, but overall, one achieves an approximate increase in accuracy by 1–2 zeta levels. Here we will try these forms:

$$E(\zeta) = E_{CBS} + c_1 \exp(-c_2 t)$$
$$E(\zeta) = E_{CBS} + c_1 \zeta^{-3}$$

The exponential fit extracts three parameters, and therefore requires at least three data points. The cubic fit extracts two parameters. (ζ is 2 for cc-pVDZ, 3 for cc-pVTZ, etc.) In general, the double-zeta level can be unreliable and should be avoided in this analysis if possible. Here, we have the luxury of calculating up to cc-pV6Z. We will use lower zeta levels to estimate the energy at higher zeta levels to get a sense of the error introduced by different choices of functional form.

Note: my analysis is stored in the repository under `water extrapolation solutions.xlsx`. You should use `water extrapolation blank.xlsx` to perform your own analysis.

1. In the first tab (`monomer`), add the energies:
    - HF: cells C7-C11
    - MP2: cells B25-B29
    - CCSD(T): cells B43-B47

2. Calculate the MP2 correlation energy in cells C25-C29 as the decrease in energy over the corresponding HF energy at the same zeta level.

3. Calculate the additional correlation energy recovered by CCSD(T) in cells C43-C47.

4. Calculate the CCSD(T) correlation energy in cells M43-M47 as the decrease in energy over the corresponding HF energy at the same zeta level.

5. We will now perform CBS extrapolation for the HF energy using the two-point formula in the block of cells at D7. This fit will be based on the cc-pVTZ and cc-pVQZ energies (which is why D6 is marked `fit(3,4)`). We will minimize the relative squared error (E7-E11) between the true (C7-C11) and fitted (D7-D11) energies using Solver.

6. Enter in reasonable guesses in for E(CBS) and c_1 in E15 and E16.

7. Enter in the fitted energy E(ζ) in D7-D11. Use partial relative references to facilitate copying and pasting.

8. In E13, set the sum of squares to be the sum of E8 and E9 (these correspond to the desired zeta levels). Solver cannot handle small objective function values, so it would be a good idea to multiply this value by something like 1E10 to make it a big enough number (otherwise the minimization will not converge).

9. Use Solver to minimize the sum of squares by varying E(CBS) and c_1. Plot the true and fitted energies as a function of zeta to ensure that the fit converged.

Hint: On my computer, it is important to click on the text field for the objective cell or the cells to be optimized before pressing the "minimize" button that allows one to select the desired cells. Otherwise, Excel changes the wrong cell!

10. Calculate the error between the extrapolated energy and the true energy at ζ=5 and ζ=6 in E18 and E19.

11. Repeat this procedure for the other blocks. What is the effect of formula choice and fitting points on the accuracy?

Note: In the MP2 row, we are fitting the correlation energy, rather than the complete energy. This can have advantages in the multi-component extrapolation procedure that will be discussed next, particularly if different numbers of basis set points are available at each level of theory (the usual case). In the CCSD(T) row, we are similarly fitting the additional correlation energy recovered by CCSD(T). In the block at M40, we are fitting the correlation energy directly.

12. Repeat this procedure for the water dimer in the `dimer` tab.

Composite Extrapolation and Interaction Energy

To calculate the full extrapolated energy, we assume the additivity of energies as follows:

$$E_{extrapolated} = E_{CBS}\left(\mathrm{HF}\right) + E_{corr}\left(\mathrm{MP2}\right) + \Delta E_{corr}\left(\mathrm{CCSD(T)}\right)$$
$$E_{extrapolated} = E_{CBS}\left(\mathrm{HF}\right) + E_{corr}\left(\mathrm{CCSD(T)}\right)$$

The top formula is a three-step composite extrapolation, whose components can be based on two-, three-, or more point CBS extrapolations.

1. In the `extrap` tab, perform the two- and three-step extrapolation procedures. You should set the required energies as cell references so that the composite extrapolation will update if you change the underlying CBS extrapolation in the other tabs. To do this, click on the cell you want to fill with an energy, press `=`, click on the appropriate tab and cell that you need, and press *ENTER*.

2. What is the error incurred by the two- and three-step procedures?

3. Starting in row 24, add the raw energies that would be estimated from the indicated single-point calculations. Calculate their associated errors.

4. Is there a trend in the errors from the single point calculations?

5. In the two composite procedures we followed above, we obtained HF and MP2 energies up to cc-pV5Z and CCSD(T) energies up to cc-pVQZ, which is a reasonable model of the energies that would be available for a larger system. How does the accuracy compare to the "best" single point CCSD(T) calculation that would have been available at cc-pVQZ? How much accuracy did we gain in zeta levels?

6. In the first two tabs, we calculated the errors in the absolute energy/correlation energy associated with the CBS extrapolation procedure. What are the errors that are pertinent to the composite procedure we used above?

7. How can the errors in the absolute energies be relatively large, but the error in the interaction energy be relatively small?

8. Even the CCSD(T)/cc-pV6Z result does not quite reach the benchmark value. What are the remaining sources of error?

Discussion

In the limit of a complete one-electron basis (i.e., CBS) and a full multi-electron basis (i.e., all possible excitations), we solve the non-relativistic Schrodinger equation exactly. If other effects are small (relativistic, non-Born–Oppenheimer, etc.), then this should correspond to the experimental result. In reality, we cannot reach either of these limits. We approximate the truncation of the one-electron basis by CBS extrapolation and try to recover as much of the correlation energy from the multi-electron basis as possible.

This exercise has considered some of the issues in constructing accurate high-level energies for model systems. More thorough analyses along these lines often constitute full papers in computational chemistry journals. Such analyses would also include higher level excitations (up to CCSDTQ), the treatment of zero-point energy (including anharmonicity), better ground state geometries, etc. For an example of such an analysis, see the work of Allen and Schaefer (*J. Chem. Phys.* **2004**, *120*, 11586). In general, the assumption of the additivity of effects is extremely good.

Even without a full analysis here, the interaction energy is very close to the benchmark value, which is itself in good agreement with experimental data (Reisler *J. Chem. Phys.* **2011**, *134*, 211101). (A direct comparison with experiment is not possible with an accurate zero-point energy, which was not evaluated in this exercise.)

When deciding what method would be appropriate for a large system, one should consult the literature for any relevant benchmark numbers. The performance of DFT or other affordable methods can then be assessed against these benchmarks, as well as any available experimental data. If benchmark data are not available, they can be constructed using the methods described here.

Eugene Kwan
March 2017