Finding density functionals with ML

Kieron Burke and friends
UC Irvine Chemistry & Physics

http://dft.uci.edu
A. Review of DFT
The electronic structure problem

- Use atomic units
- Born-Oppenheimer approximation
- All non-relativistic (but added back in)
- Wavefunctions antisymmetric and normalized
- Only discuss ground-state electronic problem here, but many variations.

Hamiltonian for $N$ electrons in the presence of external potential $v(r)$:

$$\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V},$$

where the kinetic and elec-elec repulsion energies are

$$\hat{T} = -\frac{1}{2} \sum_{i=1}^{N} \nabla_i^2, \quad \hat{V}_{ee} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|},$$

and difference between systems is $N$ and the one-body potential

$$\hat{V} = \sum_{i=1}^{N} v(r_i)$$

Often $v(r)$ is electron-nucleus attraction

$$v(r) = -\sum_{\alpha} \frac{Z_{\alpha}}{|\mathbf{r} - \mathbf{R}_{\alpha}|}$$

where $\alpha$ runs over all nuclei, plus weak applied $E$ and $B$ fields.

$$\{ \hat{T} + \hat{V}_{ee} + \hat{V} \} \psi = E \psi, \quad E = \min_{\psi} \langle \psi | \hat{T} + \hat{V}_{ee} + \hat{V} | \psi \rangle$$
1964: HK theorem: There exists $F[n]$

Define *fictitious* non-interacting electrons satisfying:

$$\left\{-\frac{1}{2}\nabla^2 + v_S(r)\right\}\phi_j(r) = \epsilon_j \phi_j(r), \quad \sum_{j=1}^N |\phi_j(r)|^2 = n(r).$$

where $v_S(r)$ is *defined* to yield $n(r)$.

Define $T_s$ as the kinetic energy of the KS electrons, $U$ as their Hartree energy and

$$F = T + V_{ee} = T_s + U + E_{XC}$$

the remainder is the exchange-correlation energy.

Most important result of exact DFT:

$$v_S(r) = v(r) + \int d^3 r' \frac{n(r')}{|r-r'|} + v_{XC}[n](r), \quad v_{XC}(r) = \frac{\delta E_{XC}}{\delta n(r)}$$

Knowing $E_{XC}[n]$ gives closed set of self-consistent equations.

**Orbital-free DFT:** Approximate $T_s[n]$ directly, and go much, much faster.
FIG. 2. The number of DFT citations has exploded (as have ab initio methods). PBE is the number of citations of of Ref. [22], and B88 of Ref. [18]. Dark indicates papers using either of these approximations without citing the original papers, while other is all other DFT papers. All numbers are estimates.

Semiclassical work in progress

- Almost exact exchange at almost no cost
  - For 1d boxes, do LDA calculation, then evaluate semiclassical exchange.
- Turning points
  - Finally derived and proved formulas in presence of turning points
- Asymptotic expansion for correlation
  - Gives new version of PBE, more accurate for atoms

Feb 23, 2015  IPAM ML
B. Strong correlation

- Ongoing project with Steve White at UCI
- Apply DMRG to continuum problems
- Understand limitations and failures of standard DFT approximations
Chapter 1. Introduction

(a) 25 randomly chosen $64 \times 64$ pixel images from the Olivetti face database. (b) The mean and the first three principal component basis vectors (eigenfaces). Figure generated by pcaImageDemo.

When used as input to other statistical models, such low dimensional representations often result in better predictive accuracy, because they focus on the "essence" of the object, filtering out inessential features. Also, low dimensional representations are useful for enabling fast nearest neighbor searches and two dimensional projections are very useful for visualizing high dimensional data.

The most common approach to dimensionality reduction is called principal components analysis or PCA. This can be thought of as supervising a variant of (multi-output) linear regression, where we observe the high-dimensional response $y$, but the low-dimensional "cause" $z$. Thus the model has the form $z \rightarrow y$; we have to "invert" the arrow, and differentiate the latent low-dimensional $z$ from the observed high-dimensional $y$. See Section 12.1 for details.

Dimensionality reduction, and PCA in particular, has been applied in many different areas. Some examples include the following:

• In biology, it is common to use PCA to interpret gene microarray data, to account for the fact that each measurement is usually the result of many genes which are correlated in their behavior by the fact that they belong to different biological pathways.

• In natural language processing, it is common to use a variant of PCA called latent semantic analysis for document retrieval (see Section 27.2.2).

• In signal processing (e.g., of acoustic or neural signals), it is common to use ICA (which is a variant of PCA) to separate signals into their different sources (see Section 12.6).

• In computer graphics, it is common to project motion captured data to a low dimensional space, and use it to create animations. See Section 15.5 for one way to tackle such problems.

Figure 1.10  a) 25 randomly chosen $64 \times 64$ pixel images from the Olivetti face database. (b) The mean and the first three principal component basis vectors (eigenfaces). Figure generated by pcaImageDemo.
Machine learning

- Powerful branch of artificial intelligence
- Essentially fitting and interpolating
- Maps problem into much higher-dimension feature space, using a simple kernel
- Higher-dimension often means more linear
- Perform regression in feature space
- Project back to original problem
Kernel ridge regression

- Kernel ridge regression (KRR). Given \{x_j, f_j\}

\[
\hat{f}(x) = \sum_{j=1}^{M} \alpha_j k(x_j, x)
\]

\[
k(x, x') = \exp\left(-\|x - x'\|^2/(2\sigma^2)\right)
\]

- Minimize:

\[
C(\alpha) = \sum_{j=1}^{M} (\hat{f}(x_j) - f_j)^2 + \lambda^2 \|\alpha\|^2
\]

\[
\alpha = (K + \lambda^2 I)^{-1}f
\]
ML applications in electronic structure

• Most with Klaus Mueller of TU Berlin, computer science.

• ML now being applied directly to, e.g., molecular energies from geometries for drug design, many by Matthias Rupp (U. Basel)

• Our efforts are focused on finding $T_s[n]$ from examples, headed by John Snyder (Humboldt fellow at TU Berlin/MPI Halle)
Demo problem in DFT

1. \( N \) non-interacting same-spin fermions confined to 1d box

2. Define class of potential:
   \[
   v(x) = - \sum_{i=1}^{3} a_i \exp\left(-\frac{(x - b_i)^2}{2c_i^2}\right)
   \]

3. Represent the density on a grid with spacing \( \Delta x = 1/(G - 1) \)

4. ML-DFA for KE:
   \[
   \hat{T}(n) = \bar{T} \sum_{j=1}^{M} \alpha_j k(n_j, n)
   \]
Generate 2000 potentials. Solve for up to 4 electrons.
Performance for $T_s$

| $N$ | $M$ | $\lambda$ | $\sigma$ | $|\Delta T|$ | $|\Delta T|^{\text{std}}$ | $|\Delta T|^{\text{max}}$ |
|-----|-----|-----------|---------|-------------|-----------------|-----------------|
| 1   | 40  | $2.4 \times 10^{-5}$ | 238 | 3.3 | 3.0 | 23. |
|     | 60  | $1.0 \times 10^{-5}$ | 95 | 1.2 | 1.2 | 10. |
|     | 80  | $6.7 \times 10^{-6}$ | 48 | 0.43 | 0.54 | 7.1 |
|     | 100 | $3.4 \times 10^{-7}$ | 43 | 0.15 | 0.24 | 3.2 |
|     | 150 | $2.5 \times 10^{-7}$ | 33 | 0.060 | 0.10 | 1.3 |
|     | 200 | $1.7 \times 10^{-7}$ | 28 | 0.031 | 0.053 | 0.65 |
| 2   | 100 | $1.3 \times 10^{-7}$ | 52 | 0.13 | 0.20 | 1.8 |
| 3   | 100 | $2.0 \times 10^{-7}$ | 74 | 0.12 | 0.18 | 1.8 |
| 4   | 100 | $1.4 \times 10^{-7}$ | 73 | 0.078 | 0.14 | 2.3 |
| $1-4^\dagger$ | 400 | $1.8 \times 10^{-7}$ | 47 | 0.12 | 0.20 | 3.6 |

LDA $\sim 223$ kcal/mol, Gradient correction $\sim 159$ kcal/mol
**functional derivative?**

- Functionals are defined on infinite-dimensional spaces
- With finite interpolation, can always find bad directions
- Can we make a cruder definition that will work for our purposes?

\[
\frac{\delta T[n]}{\delta n(x)} = \mu - v(x)
\]

\[
\frac{1}{\Delta x} \nabla_n \hat{T}(n) = \sum_{j=1}^{M} \alpha_j' (n_j - n) k(n_j, n)
\]

\[
\alpha_j' = \frac{\alpha_j}{\sigma^2 \Delta x}
\]
Principal component analysis

\[ X' = (n_{j_1} - n, \ldots, n_{j_m} - n)^\top \]

\[ C = \frac{1}{m} X^\top X \]

\[ \lambda_j, x_j \]

\[ P_{m, \ell}(n) = V^\top V \]

\[ V = (x_1, \ldots, x_\ell)^\top \]
we apply ML methodology to a prototype density functional approximation. We use kernel ridge regression with a Gaussian kernel to approximate the non-interacting kinetic energy of 1$d$ multi-electron systems. We illustrate the accuracy of the ML–DFA in Fig. 1.

Projected functional derivative

\[ -P_{m,l}(n) \hat{\nabla}_n T(n)/\Delta x \]

with fewer than 100 training densities, we can achieve approximations that are more accurate for the molecular interactions from quantum mechanics and avoiding fitting to specific finite systems. But chemists typically use a few empirical functionals that are limited class of molecules. Non-empirical functionals can be considered controlled extrapolations that work well across a large range of systems, and be used to approximate exchange-correlation density functionals.

Present DFAs often fail for strongly correlated systems. Crucially, they rely on these density functional approximations as DFAs. The quality of the results depends on how well the methodology is adapted to apply ML to DFT problems. It is extremely accurate. We also define key technical concepts required to apply ML to DFT problems.

There is a never-ending search for improved XC approximations. ML is a powerful tool for finding patterns in high-dimensional spaces. It employs algorithms that are fuelled by this cultural divide. Passionate debates are needed to advance science. This empirical method has two parameters, set via cross-validation, and enables an orbital-free minimization of the total energy to find a self-consistent density. This empirical method has two parameters, set via cross-validation, and enables an orbital-free minimization of the total energy to find a self-consistent density. This empirical method has two parameters, set via cross-validation, and enables an orbital-free minimization of the total energy to find a self-consistent density.

We determine densities for which our new functional will fail or perform well. Finally, we use principle component analysis to extract accurate empirical derivatives from our functional, enabling an orbital-free minimization of the total energy to find a self-consistent density. This empirical method has two parameters, set via cross-validation, and enables an orbital-free minimization of the total energy to find a self-consistent density.

We illustrate the accuracy of the ML–DFA in Fig. 1. The successful construction of this functional is kept to a minimum but remains necessary to specify the basic mechanism and representation of data. Intuition is uniquely defined by the properties of any system, and be used to approximate exchange-correlation density functionals. It requires no human intuition. In principle, this general technique can be extended to multi-dimensional systems, and be used to approximate exchange-correlation density functionals. Using standard methods from machine learning, we introduce a novel technique for density functional approximation. We use kernel ridge regression with a Gaussian kernel to approximate the non-interacting kinetic energy of 1$d$ multi-electron systems. With fewer than 100 training densities, we can achieve approximations that are more accurate for the molecular interactions from quantum mechanics and avoiding fitting to specific finite systems.

We define key technical concepts required to apply ML to DFT problems. It is extremely accurate. We also define key technical concepts required to apply ML to DFT problems. It is extremely accurate.
Gradient descent search:

\[ n^{(j+1)} = n^{(j)} - \epsilon P_{m,\ell}(n^{(j)})(v + \nabla_n \hat{T}(n^{(j)})/\Delta x) \]

\[ N = 1 \]
\[ M = 100 \]

\[ \Delta T_j = 0.154 \text{ kcal/mol} \]
\[ \Delta T_j^{sc} = 6.53 \text{ kcal/mol} \]

Ratio = 43

\[ m = 15, \ell = 5 \]
SC density movie
Finding Density Functionals with Machine Learning

Is there a data-driven method that can be applied to approximate density functionals?

Density functional theory (DFT) is the most popular electronic structure calculation method in the world. In principle, the mapping of the electron density to energy is exact, but in practice, both the kinetic energy and the energy of the interaction between electrons must be approximated.

In the past few decades, development of both empirical and non-empirical functionals requires great intuition built on years of experience, as well as painstaking trial and error. Recently, some of us have approached this problem by a completely new data-driven method called machine learning approximation (MLA). Comparing to traditional approximation, MLA has (i) lower computational cost and (ii) significantly higher accuracy.

Machine Learning

Kernel Ridge Regression is a powerful kernel-based ML method to find non-linear pattern in high dimensional space.

- **training set** \((x, y)\), \(j = 1, \ldots, N\)
- **kernel trick**
  we define a mapping for input space to feature space \(\phi \circ f \in \mathbb{R}^F\)
- **hyperparameters**
  - \(\lambda\), \(\sigma\)
  - mean \(\mu\)
  - length scale \(\alpha\)

Model System

Consider a non-interacting many-fermion system subject to a smooth external potential in one dimension,

\[
\begin{align*}
H &= \sum \phi^{\dagger}(j) \phi(j), \quad \phi(j) = \phi(x, j),
\end{align*}
\]

with hard wall on both side. We are going to approximate the kinetic energy function \(f^{\dagger}\).

- **Local density approximation (LDA)**
  \[
  \psi_{\text{LDA}}(x) = \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} \rho_{\text{LDA}}(y) \text{d}y}
  \]

**Challenge of Finding Density**

This far, we have focused on the discussion of the performance of the MLA evaluated on exact densities. However, in order for a functional to be useful, it must also predict the ground-state density. An accurate functional derivative is necessary in order to solve Euler equation and yield an accurate density.
- Performed many 1d KS calculations of diatomics as function of bond length, using LDA with soft-Coulomb repulsion, including several with more than 2 electrons.
Constrained optimal density

- Convergence of constrained optimal density with # of training points.

FIG. 7. Difference between the constrained optimal density $\tilde{n}(x)$ and the KS density $n(x)$ for various numbers of training densities $N_T$. The error decreases uniformly for all $x$. The system is $H_2$ at equilibrium bond length. The inset shows the KS density.

*Kernels, Pre-Images and Optimization* John Snyder, Sebastian Mika, Kieron Burke, Klaus-Robert Müller, Chapter in Empirical Inference - Festschrift in Honor of Vladimir N. Vapnik (2013)
Types of errors in DFT

- $\Delta E_F = \tilde{E}_{xc}[n] - E_{xc}[n]$
- $\Delta E_D = \tilde{E}_{xc}[\tilde{n}] - \tilde{E}_{xc}[n]$
- $\Delta E = \Delta E_F + \Delta E_D$

Error analysis of energies in kcal/mol as a function of $R$ with different numbers of training data, on constrained optimal densities.

**FIG. 8.** The total error of the model and the functional- and density-driven errors $\Delta E_F$ and $\Delta E_D$ for $H_2$ with (a) 10 and (b) 20 training densities.
Functional derivatives and densities

• How can we get accurate densities from lousy derivatives?
• Once solution density is within interpolation manifold, simply constrain derivative to stay on that manifold
• Analogy:
  – Problem: find global minimum of 2D surface, given exact data along a 1D curve in that surface that passes through the minimum.
  – Solution: Make sure you stay on the path.
• PS: Inspired density-corrected DFT, which corrects many self-interaction errors!


Conceptual relationship

• ML works when
  a) There’s a rule
  b) Rule is too complicated for humans
  c) There’s data

• HK theorems say
  a) There is a functional
  b) It cannot be given explicitly, exactly
  c) Examples give exact values

• More important, practically:
  – In chemistry and materials, we only care about solutions to an absurdly small fraction of possible problems, i.e., Coulomb potentials at various positions, so underlying dimensionality of solutions is very small, just solving differential equation is hard.
Road map back to reality

Roadmap to 3d land

1d box
- model selection, projected functional derivatives, OF-DFT
- bond breaking, self-consistent densities

1d diatomics
- dimensionality, basis sets, representation, inversion symmetry

3d atoms, diatomics
- full symmetries, scaling

3d molecules
- scalability, data accumulation
- ab-initio MD, active learning

large systems, real applications
D. Simplified example

Kevin Vu

Li Li
Kevin’s paper: from functions to functionals

- Plot error as a function of hyperparameters
- Repeat for fitting $f(x) = \cos x$

Curves have roughly the same “valley” shape for all $N_T$
- Bottom of the valley is an order of magnitude deeper than the walls
- These valleys are nearly identical in shape for sufficiently large $N_T$, which indicates that this particular feature arises in a systematic manner as $N_T$ increases
D. Cross-fertilization

• All preliminary results
DMRG meets DFT meets ML

- Ran H$_4$ with fixed separations $b$ using DMRG
- Use 30 values of $b$ to train ML version of exact $F[n]$
- Yields accurate exact binding energy curve self-consistently.
• ML of functionals works in model cases to produce highly accurate approximate functionals
• Totally different approach from anything before
• ML can even
  – find accurate densities
  – say when it will work within tolerance (makes Klaus nervous)
  – break bonds
  – Do the full functional
• But
  – only demonstrated in 1d
  – Need to do arbitrary-sized system (representation question)
• Thanks to
  – Students: Li Li, John Snyder, Kevin Vu, Isabelle Pelaschier
  – Collaborators: Klaus Mueller, Matthias Rupp, Katia Hansen
  – Funders: NSF from chem, DMR, math