LOCAL GAUSSIAN PROCESS REGRESSION

SPENCER HILL, QUEEN'S UNIVERSITY

spencer.hill@queensu.ca

With support from Dr. Tucker Carrington Jr. and Dr. Sergei Manzhos

OBJECTIVE

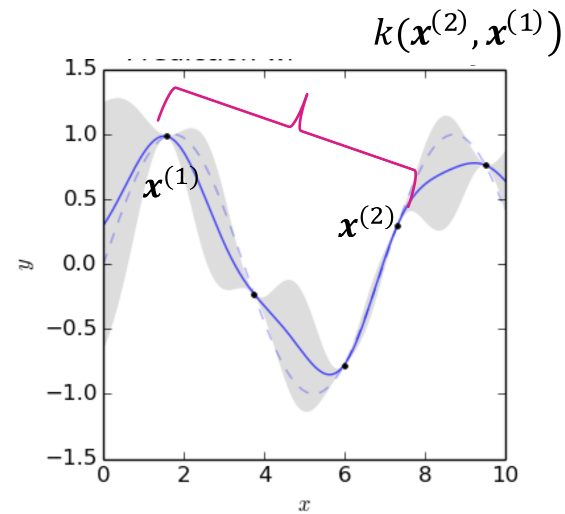
- Potential Energy Surfaces (PES) are required to calculate most chemical observables (i.e., reaction rates, etc.)
- Constructing a PES generally requires interpolating between known potential energy points in a multi-dimensional space
- The objective is to be able to compute a vibrational spectrum with errors approximately 1 cm^{-1}, a challenging task for analytic methods
- A popular machine learning method to accomplish this is Gaussian Process Regression (GPR)

GAUSSIAN PROCESS REGRESSION (GPR)

What is the expected value of f at $\boldsymbol{x}$ given the set $\{t^{(n)}, \boldsymbol{x}^{(n)}\}$?

Matrix K describes how correlated each pair of data points are

$$y(\boldsymbol{x}) = \boldsymbol{K}^{*}\boldsymbol{K}^{-1}\boldsymbol{t}$$



$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) + \delta & k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) & \cdots & k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(M)}) \\ k(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) & k(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}) + \delta & & k(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(M)}) \\ \vdots & & \ddots & \vdots \\ k(\boldsymbol{x}^{(M)}, \boldsymbol{x}^{(1)}) & k(\boldsymbol{x}^{(M)}, \boldsymbol{x}^{(2)}) & \cdots & k(\boldsymbol{x}^{(M)}, \boldsymbol{x}^{(M)}) + \delta \end{pmatrix}$$

$$\boldsymbol{K}^{*} = \left(k(\boldsymbol{x}, \boldsymbol{x}^{(1)}) \quad k(\boldsymbol{x}, \boldsymbol{x}^{(2)}) \quad \ldots \quad k(\boldsymbol{x}, \boldsymbol{x}^{(M)})\right) \qquad \boldsymbol{K}^{**} = k(\boldsymbol{x}, \boldsymbol{x})$$

$$(RBF)\colon\ k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^{2} exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{x}'|^{2}}{2l^{2}}\right) \rightarrow \prod_{i=1}^{D} exp\left(-\frac{|x_i - x_i'|^{2}}{2l_i^{2}}\right)$$

hyperparameters

Optimized l_i informs on relevance of the i-th variable

PROS AND CONS OF GPR[1]

Pros	Cons
• Demonstrated sufficiently low error with relatively few *ab initio* points • Simple to use and train, with few hyperparameters trained by maximizing the log marginal likelihood • Generality of method across multiple functions	• Computational complexity scales $O(n^3)$ with the number of training examples n • Space complexity scales $O(n^2)$ with n • Time and space complexity limit GPR to training problems with $n < 10^4$

1. J. Chem. Phys. **148**, 241702 (2018)

LOCAL GAUSSIAN PROCESS REGRESSION

We propose Local Gaussian Process Regression (LGPR), which leverages the correlation of the covariance function to reduce the computational and space complexity.

$\widehat{\boldsymbol{K}}$

$$\boldsymbol{K} = \begin{pmatrix} k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}\right) + \delta & k\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) & \cdots & 0 \\ k\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}\right) & k\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}\right) + \delta & & 0 \\ & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 + \delta \end{pmatrix}$$

For points with low correlation, the covariance will be close to zero

We can approximate $\boldsymbol{K}$ as the m-by-m matrix $\widehat{\boldsymbol{K}}$, leading to the modified equation

$$y(\boldsymbol{x}) = \boldsymbol{K}^{*} \widehat{\boldsymbol{K}^{-1}} \boldsymbol{t}$$

The m entries are determined by those with covariance above a threshold value

$$\boldsymbol{K}^{*} = \left(k\left(\boldsymbol{x}, \boldsymbol{x}^{(1)}\right) \quad k\left(\boldsymbol{x}, \boldsymbol{x}^{(2)}\right) \quad \ldots \quad k\left(\boldsymbol{x}, \boldsymbol{x}^{(\boldsymbol{m})}\right) \right), \; t^{(m)}$$

THEORETICAL BENEFITS

- LGPR has time complexity $O(n + m^3)$ and space complexity $O(n + m^2)$, which is caused by the need to compute covariance between a test point and each training point
- By constraining $m \ll n$, LGPR permits an arbitrarily large number of training points without dramatically increasing the computation time
- LGPR is embarrassingly parallel, decreasing the time required to make large numbers of predictions

LGPR IMPLEMENTATION AND METHOD

- LGPR was implemented using Python and the *sklearn* library.
- Euclidean distances between the test point and training points are computed and used to determine the m prior points
- It was found that optimizing the log-marginal likelihood of the hyperparameters for each prediction point $\boldsymbol{x}'$ was intractable for large numbers of predictions
    - Hyperparameters were optimized over a subset of the data and averaged across the entire dataset. This did not significantly increase the prediction errors
- A minimum bound on m was found to improve the prediction accuracy for regions with sparse training point distribution

H_2CO

- The potential for was computed for H_2CO by constructing a set of 120,000 points using a pseudo-random Sobol sequence and accepting the point $\boldsymbol{x}$ if

$$\frac{V_{max} - V(x) + \Delta}{V_{max} + \Delta} > b$$

$V(x)$ is the potential function, $\Delta = 500\ \mathrm{cm}^{-1}$, $V_{max} = 17\,000\ cm^{-1}$, and b is a random number in $[0,1]$

- 5000 training points were used for the full GPR and LGPR
- Vibrational spectra were computed with the Space-Fixed Gaussian Basis method of Manzhos and Carrington[2]

2. J. Chem. Phys. **145**, 224110 (2016).

H_2CO RESULTS

Average *m* value	Potential RMSE	Spectrum Mean Absolute Frequency	Spectrum RMSE
Full GPR	8.37 cm^{-1}	0.869	1.31
951	8.72 cm^{-1}	0.844	1.35
651	9.26 cm^{-1}	0.886	1.38
466	10.94 cm^{-1}	0.925	1.36

- LGPR performed comparably to the full GPR, and more importantly had Spectrum Mean Absolute Frequency and Root Mean Square Errors (RMSE) of approximately 1 cm^{-1}

6D AND 9D MORSE OPERATORS

- The potential was computed for 6 and 9-dimensional coupled morse operators, which for *k* dimensions predicts the potential of point Q according to,

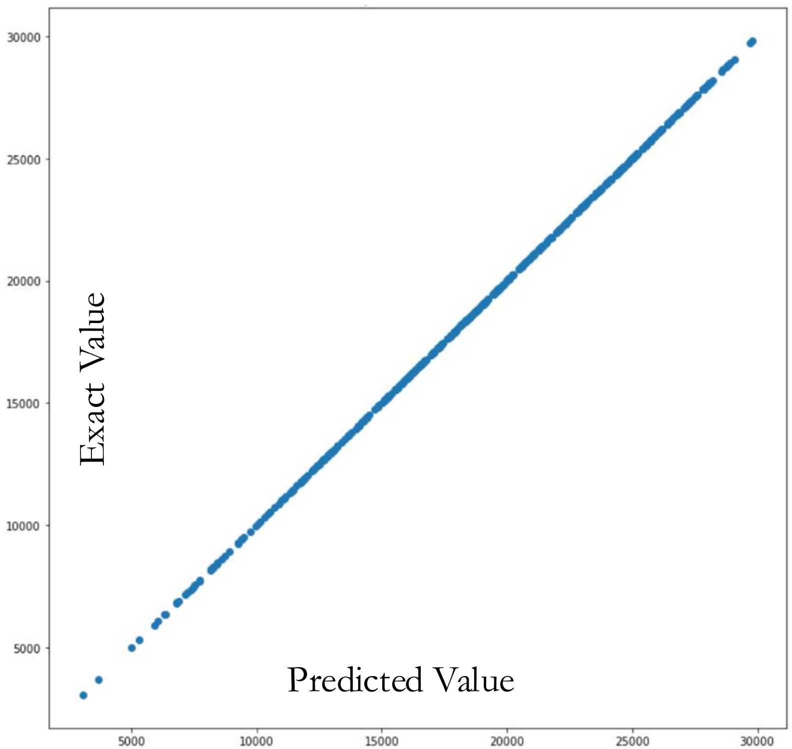
$$V(Q) = \sum_{i=1}^{k} D_e\left(1 - e^{-a(q_i - r_e)}\right) + \frac{D_e}{100}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left(1 - e^{-a(q_i - r_e)}\right)\left(1 - e^{-a(q_j - r_e)}\right)$$

where D_e has the value 37,255 cm^{-1}, *a* is 1.8677 inverse Angstrom, and r_e is 1.275 Angstrom

- A pseudo-random Sobol sequence was also used to construct the training point sets, which had 20 000 and 100 000 training points respectively for the 6 and 9-dimensional operators

6D AND 9D MORSE OPERATORS RESULTS

6D Morse Operator: Predicted vs. Exact

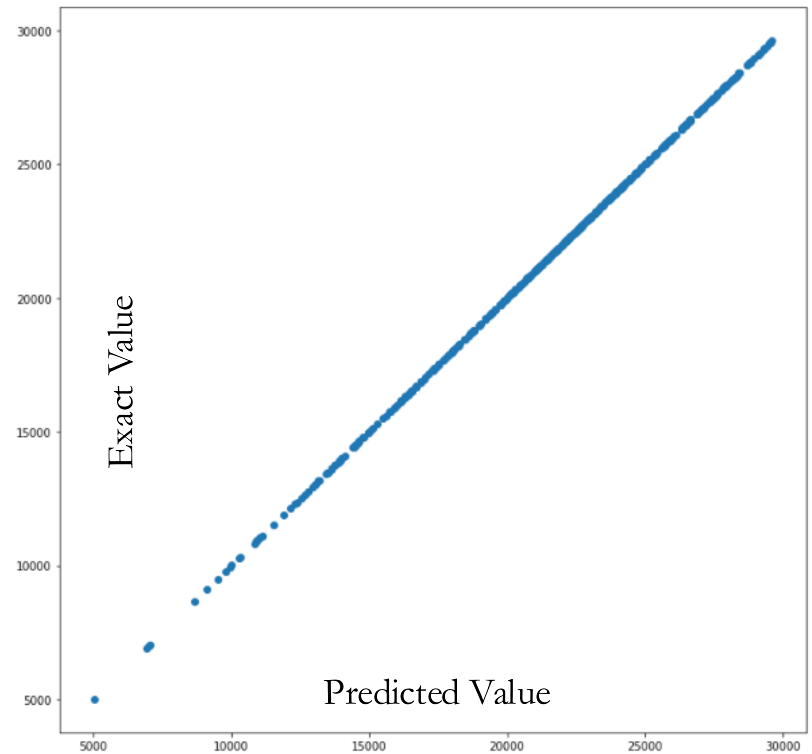


6D Morse Operator

Full GPR RMSE: 1.37 cm^{-1}

LGPR RMSE ($900 < m < 1000$): : 2.04 cm^{-1}

9D Morse Operator: Predicted vs. Exact



9D Morse Operator

Full GPR (20 000 training points) RMSE: 6.49 cm^{-1}

LGPR RMSE ($2000 < m < 2100$): 7.10 cm^{-1}

CONCLUSIONS

- We proposed LGPR, a local GPR method to reduce the computational and space complexity and permit larger numbers of training points
- LGPR accomplishes this by computing the covariance matrix for a subset of the data with high correlation to the test point
- LGPR was shown to be similarly accurate to GPR over H_2CO and 6 and 9-dimensional morse operators while reducing the required computation
- LGPR has the potential to be expanded to higher-dimensional computations that are currently intractable for conventional GPRs