Parameterizing Molecular Models via Gaussian Approximation Methods

Vasiliki Kyrieri

DEPARTMENT OF MATHEMATICS AND APPLIED MATHEMATICS, UNIVERSITY OF CRETE

Supervisors: Prof. Vagelis Harmandaris, Dr. Evangelia Kalligiannaki

Abstract

Molecular Dynamics (MD) simulations is a scientific field that focuses on the physical motion of atoms on a very small order of magnitude. Today, many chemical and biological processes, such as protein interactions, can be modeled through MD simulations to analyze their properties. The challenging part of studying MD simulations of complex systems refers to an as-accurate-as-possible prediction of the structure-property relationship at the microscopic level and the expensive calculations of the dynamic quantities due to the wide range of length and time scales. By decreasing the number of degrees of freedom, the new system can be used with fewer variables. This method, known as Coarse-Graining (CG), maps the atomistic particles into mesoscopic particles such as "superatoms". There exists a variety of methods to obtain the total force of the mesoscopic system, either parametric or non-parametric models. The Gaussian process regression (GPR) model is a flexible non-parametric family of models capable of approximating functions using relatively small data sets and is applied to a simple system, a methane system. Its results are compared with two more straightforward approximations. One with a parametric pair potential, the Lennard-Jones potential, and another with the Linear B-splines representation. We learn the approximate force fields with the Force Matching (FM) criterion of loss, which minimizes the average distance between the atomistic forces and the approximate CG forces.

1. Molecular-Dynamics

MD simulation is a technique for studying the movement of atoms in a classical many-body system. At the microscopic (atomistic) level, simulating molecular systems can be a time-consuming process due to the detailed calculations between all the atoms. A way to handle these challenges is to model the systems on the mesoscopic level. This means we study coarse-grained particles instead of atoms to decrease the length and time scales accessible by simulations by decreasing the degrees of freedom. Let us assume a system of N atoms with position vector $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_N), \mathbf{q} \in \mathbb{R}^{3N}$. The new CG coordinates are calculated through a linear mapping function $\mathbf{\Pi}$

$$\mathbf{\Pi} : \mathbb{R}^{3N} \to \mathbb{R}^{3M}, \quad \mathbf{q} \mapsto \mathbf{\Pi}(\mathbf{q}) = \mathcal{Q}$$

where $\mathcal{Q} = (\mathbf{Q}_1, \mathbf{Q}_2, .., \mathbf{Q}_M), \quad \mathcal{Q} \in \mathbb{R}^{3M} \quad (M < N)$ the new CG coordinates. In this study, the CG coordinates are given by the center of mass of the methane molecule (see figure 1). To define the new effective CG system, we must find a CG model that best represents the atomistic reference system, ideally in terms of its structural and dynamic properties.

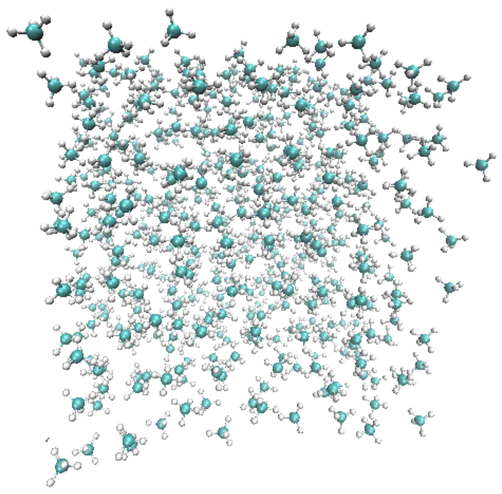

Figure 1: *Methane system*

2. Gaussian Process Regression Theory

GPR is a nonlinear, non-parametric regression helpful for interpolating between data points scattered in high-dimensional input space, based on Bayesian probability theory.

Assume a training data set comprising n_s observations $\mathbf{x}_n$, where $n = 1, 2, \cdots, n_s$ with the corresponding target values $\{\tilde{t}_n\}$. Our aim is to predict the value of $y : \mathbb{R}^d \to \mathbb{R}$, following the observation equation

$$\tilde{t} = y(\mathbf{x}) + \epsilon, \quad \epsilon \text{ is the data noise,}$$

where $y = y(\mathbf{x})$ is a one-dimensional continuous-time Gaussian process, $y(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, $y(\mathbf{x}) = \mathbf{x}^{tr}\mathbf{a}$, with $m(\mathbf{x}) = 0$. The covariance function $k(\mathbf{x}, \mathbf{x}')$ is the kernel of the Gaussian process and $\mathbf{a} = (a_1, \cdots, a_{n_s})^{tr}$ is the n_s-dimensional weight vector we want to approximate. Two approaches will be presented: the weight-space and the function-space views of the GPR.

Modelling functions: the weight space view

This approach is based on the posterior distribution of the weight vector and is calculated by the Bayesian rule applying the known prior weight distribution

$$p(\mathbf{a}|\tilde{\mathbf{t}}, \mathbf{X}) \propto p(\tilde{\mathbf{t}}|\mathbf{X}, \mathbf{a})p(\mathbf{a}) = N(\lambda^{-1}\mathbf{A}^{-1}\mathbf{X}\tilde{\mathbf{t}}, \mathbf{A}^{-1})$$

where $\mathbf{A} = \beta\mathbf{I} + \lambda^{-1}\mathbf{K}$, $\mathbf{K}$ is the covariance matrix with elements $\mathbf{K}_{mn} = k(\mathbf{x}_m, \mathbf{x}_n)$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_{n_s}), \mathbf{x}_i \in \mathbb{R}^d$.

Modelling functions: the function space view

In the weight space view of the previous section, we concentrate our interest on distributions over weights. In this approach, we focus directly on such distributions over functions. Our goal is to calculate the second-order statistics of $\tilde{t}_{n_s+1}|\tilde{\mathbf{t}}_{n_s} \sim N(m(\mathbf{x}_{n_s+1}), \sigma^2(\mathbf{x}_{n_s+1}))$. According to Bayesian rule

$$m(\mathbf{x}_{n_s+1}) = \mathbf{k}^{tr}\mathbf{C}_{n_s}^{-1}\tilde{\mathbf{t}}$$

where $\mathbf{x}_{n_s+1}$ is the new value with the corresponding target value t_{n_s+1}. $\mathbf{C}_{n_s}$ is the covariance matrix of the marginal distribution $p(\tilde{\mathbf{t}})$ and is given by

$$\mathbf{C}_{n_s+1} = \begin{pmatrix} \mathbf{C}_{n_s} & \mathbf{k} \\ \mathbf{k}^{tr} & c \end{pmatrix}, \quad c = k(\mathbf{x}_{n_s+1}, \mathbf{x}_{n_s+1}) + \lambda^{-1} \text{ (scalar)}$$

where λ is the precision of the noise, $\mathbf{k} = (k(\mathbf{x}_1, \mathbf{x}_{n_s+1}), \cdots, k(\mathbf{x}_{n_s}, \mathbf{x}_{n_s+1}))$ and the covariance matrix $\mathbf{C}_{n_s}$ has elements $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \lambda\delta_{nm}$. An expansion in radial basis function of the mean equation gives the following form:

$$m(\mathbf{x}_{n_s+1}) = \sum_{n=1}^{n_s} a_n k(\mathbf{x}_n, \mathbf{x}_{n_s+1})$$

where the weight elements a_n is the n^{th} of $\mathbf{C}_{n_s}^{-1}\tilde{\mathbf{t}}$.

Goal: to approximate $y(\mathbf{x})$ by $f(\mathbf{x})$ as a linear combination of n_s basis functions. $f(\mathbf{x}) = m(\mathbf{x}_n) = \sum_{i=1}^{n_s} a_i k(\mathbf{x}, \mathbf{x}_i)$ (see figure 2). Calculating the above expression is computationally expensive when n_s is large. Thus, we choose a sparser model with far fewer kernel basis functions than the input data points where the locations of these basis functions (which we call the representative set) need not coincide with the input data locations.

$$f(\mathbf{x}) = m(\mathbf{x}_n) = \sum_{i=1}^{n_p} a_i k(\mathbf{x}, \mathbf{x}_i).$$

where n_p is the number of basis points and x_i takes values of a grid. One very popular choice of a kernel is the radial basis function kernel

$$k(\mathbf{x}, \mathbf{x}_i) = \delta^2 \exp(-\frac{|\mathbf{x} - \mathbf{x}_i|^2}{2\theta^2})$$

where each $\mathbf{x}_i$ is a previously observed input value in $\mathbf{X}$ and the weights are collected in the vector

$$\mathbf{a} = [\mathbf{K} + \lambda\mathbf{I}]^{-1}\tilde{\mathbf{t}}_n.$$

where λ is the regularization parameter and the matrix $\mathbf{K} + \lambda\mathbf{I}$ corresponds to the matrix $\mathbf{C}_{n_s}$. *This equation shows that Gaussian process regression is equivalent to a linear regression model using basis functions k to project the inputs into a feature space.*

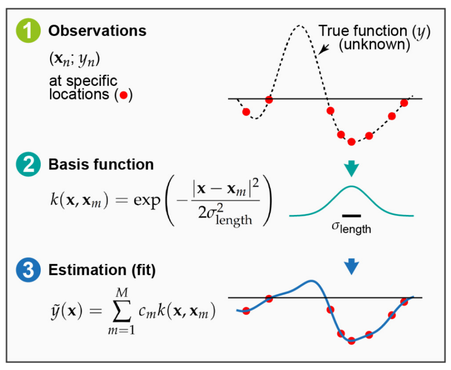

Figure 2: *The prediction of a GPR model*

Kernel Ridge Regression (KRR) model

The fitting of the Gaussian process regression model can also be written as a (kernel) ridge regression problem. That is, to find a by minimizing the loss (or cost) function

$$L = \sum_{n=1}^{n_s} \frac{[y_n - f(\mathbf{x}_n)]^2}{\sigma_i^2} + \lambda\mathbf{R},$$

where $\mathbf{R}$ is the regularization parameter, and $\sigma_i = 1$ is the relative importance parameter of individual data points. The regularization terms that we work with in this thesis are: $\mathbf{R} = \sum_{n,n'}^{n_p} a_n k(\mathbf{x}_n, \mathbf{x}_{n'}) a_{n'} = \mathbf{a}^{tr}\mathbf{K}_{n_p n_p}\mathbf{a}$ and $\mathbf{R} = \sum_{n,n'}^{n_p} a_n a_{n'} = \mathbf{a}^{tr}\mathbf{I}_{n_p n_p}\mathbf{a}$.

3. Coarse-grained molecular models with Gaussian approximation methods

For our methane system, we aim to approximate the total force as a sum of local terms. Thus, we define the Force Matching approach according to the $\{\mathbf{q}_i, \mathbf{F}_i\}_{i=1}^{n_s}$ observation/samples from all-atom simulations.

$$\min_{\mathbf{a}} \frac{1}{n_s} \sum_{l=1}^{n_s} \sum_{I=1}^{M} ||(\mathbf{F}o\Pi_I)(\mathbf{q}^{(l)}) - \mathbf{F}_{CG,I}(\Pi\mathbf{q}^{(l)}; \mathbf{a})||_3^2$$

where $(\mathbf{F}o\Pi)$ is known from samples and the second term $\mathbf{F}_{CG}$ contains the approximated values and is calculated by Newton's second law

$$\mathbf{F}_{CG,I}(\mathcal{Q}; \mathbf{a}) = -\nabla_{\mathbf{Q}_I} U_{CG}(\mathcal{Q}; \mathbf{a}), \quad \mathbf{F}_{CG,I}(\mathcal{Q}; \mathbf{a}) \in \mathbb{R}^3.$$

The minimization problem is in analogy to the affiliated with the loss function of the KKR chapter. The parametrized CG potential that describes the free energy surface U_{CG} is, in principle, a many-body potential of the 3M CG coordinates

$$\begin{aligned} U_{CG}\ &(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_M; \mathbf{a}) = \\ &\sum_{j<k} W_{dimer}(D_{dimer}(\mathbf{Q}_{j,\{\alpha\}}, \mathbf{Q}_{k,\{\beta\}}); \hat{\mathbf{a}}) \\ &+ \sum_{j,k,l} W_{trimer}(D_{trimer}(\mathbf{Q}_{j,\{\alpha\}}, \mathbf{Q}_{k,\{\beta\}}, \mathbf{Q}_{l,\{\gamma\}}); \mathbf{a}^*) \end{aligned}$$

where D_{dimer}, D_{trimer} are the descriptors that describe the interaction between two and three CG particles correspondingly. $D_{dimer}(\mathbf{Q}_j, \mathbf{Q}_k) = D(\mathbf{Q}_j, \mathbf{Q}_k) \in \mathbb{R}$ is the Euclidean distance between two CG particles $D(\mathbf{Q}_j, \mathbf{Q}_k) = r_{jk} = \sqrt{|Q_{j1} - Q_{k1}|^2 + |Q_{j2} - Q_{k2}|^2 + |Q_{j3} - Q_{k3}|^2}$ and the trimer descriptor is given by $D(\mathbf{Q}_j, \mathbf{Q}_k, \mathbf{Q}_l) = (r_{jk} + r_{jl}, (r_{jk} - r_{jl})^2, r_{kl}) \in \mathbb{R}^3, \quad j, k, l = 1, 2, \cdots, M$. The minimization problem has the form

$$\begin{aligned} \min_{\hat{\mathbf{a}}, \mathbf{a}^*} \frac{1}{n_s} \quad &\sum_{s=1}^{n_s} \sum_{I=1}^{M} ||(\mathbf{F}(\mathbf{Q}_I^s) - \sum_{j<k}^{M} \sum_{i=1}^{n_p} \hat{a}_i k(x_i, r_{jk}^s)(-\nabla_{\mathbf{Q}_I} r_{jk}^s) \\ &- \sum_{j<k<l}^{M} \sum_{\mathbf{i}}^{\mathbf{n_p}^*} \alpha_{\mathbf{i}}^* \sum_{z}^{3} \frac{\partial k^*(\mathbf{x_i}, (d_1(\cdot), d_2(\cdot), d_3(\cdot)))}{\partial d_z(\cdot)} (-\nabla_{\mathbf{Q}_I} d_z(\mathbf{Q}_j^s, \mathbf{Q}_k^s, \mathbf{Q}_l^s))||_3^2 \end{aligned} \tag{1}$$

where $\mathbf{Q}_I^s = (\Pi\mathbf{q}^s)_I$, $n_p, \#\mathbf{n_p}^*$ are the number of basis points $x_i, \mathbf{x_i}$ correspondingly. The second term of the problem is equal to the linear form $\hat{\mathbf{K}}\hat{\mathbf{a}}$, and the third term corresponds to the linear form $\mathbf{K}^*\mathbf{a}^*$. The two kernel functions are given by $\hat{k}(x_i, x) = \hat{\delta}^2 \exp(-\frac{|x - x_i|^2}{2\hat{\theta}^2})$ and $k^*(\mathbf{x}, \mathbf{x}') = \delta_*^2 \exp\left[\sum_{i=1}^{3}(-\frac{|x_i - x_i'|^2}{2(\theta_*^i)^2})\right]$, where $\hat{\delta}, \hat{\theta}, \delta_*, \theta_*^i$ are the hyperparameters of the model. To solve the minimization problem, a linear problem arises, $\mathbf{F} = \mathbf{Ka}$, where $\mathbf{F}$ denote the target forces and $\mathbf{K} = [\hat{\mathbf{K}}, \mathbf{K}^*], \mathbf{a} = [\hat{\mathbf{a}}, \mathbf{a}^*]^{tr}$.

We model the methane system in two ways: the first approach includes only the two-body interactions, and the second both the two- and three-body interactions.

4. Results of the Methane system with Dimer Descriptor only

Our system has 512 CH_4 methane molecules at the NVT ensemble at 100K temperature, and, in total, 7400 frames (observations) were collected.

When we estimate a regression model, the differences between the dependent variable's actual and "predicted" values can be measured by different measures. In this work, we report the chi-square error and the Wasserstein metric. We also present the corresponding measurements of the LJ and linear-splines models for comparison.

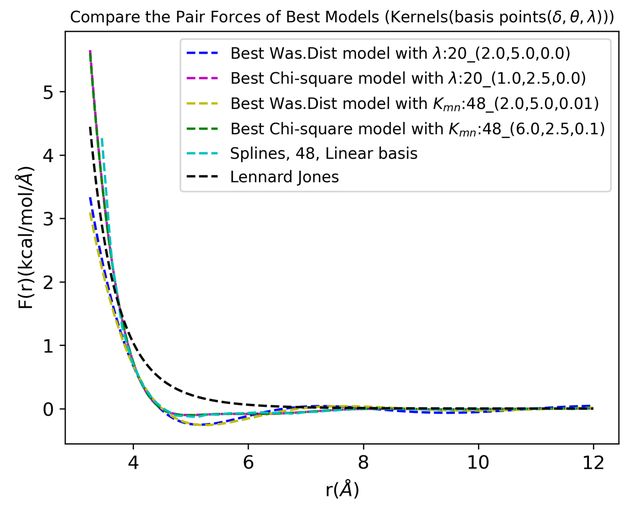

Figure 3: *Pair Forces plot for the methane system*

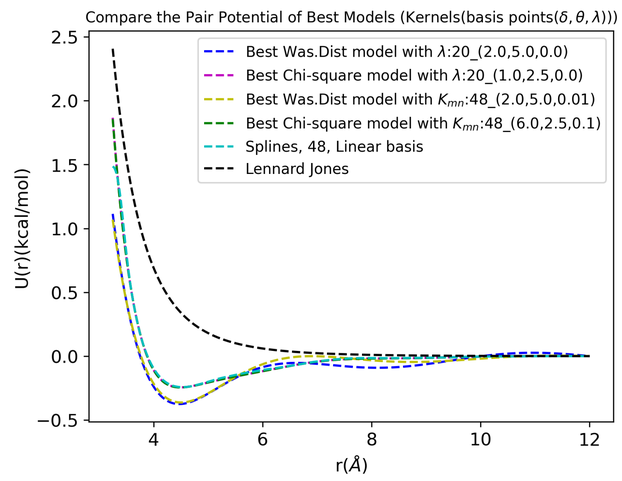

Figure 4: *Potential plot according to the Pair Forces values for the methane system*

(basis points, δ, θ, λ)	Wasserstein train	Wasserstein test	Chi-Square train	Chi-Square test
λI (20 2.0 5.0 0.0)	0.09	0.10	1.33	1.36
λI (20 1.0 2.5 0.0)	0.23	0.24	0.77	0.80
K_{mn} (48 2.0 5.0 0.01)	0.07	0.07	1.73	1.73
K_{mn} (48 6.0 2.5 0.1)	0.23	0.24	0.77	0.80
Splines 48	0.23	0.25	0.76	0.80
LJ	0.25	0.27	0.78	0.81

Table 1: *Models with minimum values of Chi-Square error and Wasserstein distance for 100 samples*

We tested the two-body contributions model for 100 samples. The training of the CG methane model was performed for: different choices for the number of basis points for each parameter; δ, θ, and for the λ parameter (in total, 108 different trained models). Moreover, we applied two regularization matrices: the first one consisted of the parameter λ and the identity matrix $\mathbf{I}$, and the second one of the parameter λ and the kernel matrix $\mathbf{R}$ of the basis points. The models presented in figures 3 & 4 and table 1 are the best given by the minimum value of the chi-square error and the Wasserstein distance for the different forms of the regularization matrix and the training and test dataset. We also tested these six models for 4000 samples showing similar results to the 100 sample models.

5. Conclusions and Future Work

[1] Gaussian kernel models produce results of similar quality to the splines method and they require an almost equal computational cost. **[2]** In the KRR method, the set of hyperparameters is a determining factor. **[3]** The models with a minimum value of Chi-square error seem to have better results.**[4]** The preparation of the trimer Kernel matrix costs computationally. **Thus,** we will have better results if we include the trimer term and learn better the hyperparameter.

6. Acknowledgements

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No [52], SISDECS, and from the project "Computational Modeling of Complex Molecular Systems" funded by Goodyear.

7. References

[1]"The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems", "E. Kalligiannaki and V. Harmandaris and M.A. Katsoulakis and P. Plech", 2015, **[2]** "Neural Network Potential Surfaces: A Comparison of two Approaches", Chazirakis, Anthony and Kirieri, Vassia and Sarris, Ilias-Marios and Kalligiannaki, Evangelia and Harmandaris, Vagelis, 2020, **[3]**"Many-body coarse-grained interactions using Gaussian approximation potentials", John, ST and Csányi, Gábor., 2017