Electronic supplementary information: Graphical Gaussian Process Regression Model for Aqueous Solvation Free Energy Prediction of Organic Molecules in Redox Flow Battery

Peiyuan Gao,[†] Xiu Yang,[*,‡] Yu-Hang Tang,[¶] Muqing Zheng,[‡] Amity Anderson,[†] Vijayakumar Murugesan,[*,†] Aaron Hollas,[†] and Wei Wang[*,†]

†*Pacific Northwest National Laboratory, Richland 99352, USA*

‡*Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015, USA*

¶*Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

E-mail: xiy518@lehigh.edu; vijay@pnnl.gov; wei.wang@pnnl.gov

GPR framework

We present a brief review of the GPR method adopted from Reference .[1,2] We denote the observation locations as $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ ($\boldsymbol{x}^{(i)} \in D, D \subseteq \mathbb{R}^d$) and the observed values of the QoI at these locations as $\boldsymbol{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(N)})^\top$ ($y^{(i)} \in \mathbb{R}$). For simplicity, we assume that $y^{(i)}$ are scalars. The GPR method aims to identify a GP $Y(\boldsymbol{x}, \omega) : D \times \Omega \to \mathbb{R}$ based on the input/output data set $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, where Ω is the sample space of a probability triple. Here, $\boldsymbol{x}$ can be considered as parameters for this GP, such that $Y(\boldsymbol{x}, \cdot) : \Omega \to \mathbb{R}$ is a

Gaussian random variable for any $\boldsymbol{x}$ in the set D. A GP $Y(\boldsymbol{x}, \omega)$ is usually denoted as

$$Y(\boldsymbol{x}) \sim \mathcal{GP}\left(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right), \tag{1}$$

where ω is not explicitly listed for brevity, $\mu(\cdot) : D \to \mathbb{R}$ and $k(\cdot, \cdot) : D \times D \to \mathbb{R}$ are the mean and covariance functions (also called *kernel* function), respectively:

$$\mu(\boldsymbol{x}) = \mathrm{E}\left\{Y(\boldsymbol{x})\right\}, \tag{2}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}\left\{Y(\boldsymbol{x}), Y(\boldsymbol{x}')\right\} = \mathrm{E}\left\{(Y(\boldsymbol{x}) - \mu(\boldsymbol{x}))(Y(\boldsymbol{x}') - \mu(\boldsymbol{x}'))\right\}. \tag{3}$$

The variance of $Y(\boldsymbol{x})$ is $k(\boldsymbol{x}, \boldsymbol{x})$, and its standard deviation is $\sigma(\boldsymbol{x}) = \sqrt{k(\boldsymbol{x}, \boldsymbol{x})}$. The covariance matrix, denoted as $\boldsymbol{C}$, is defined as $C_{ij} = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. For any $\boldsymbol{x}^* \in D$, the GPR prediction and variance are

$$\hat{y}(\boldsymbol{x}^*) = \mu(\boldsymbol{x}^*) + \boldsymbol{c}(\boldsymbol{x}^*)^\top \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \tag{4}$$

$$\hat{s}^2(\boldsymbol{x}^*) = \sigma^2(\boldsymbol{x}^*) - \boldsymbol{c}(\boldsymbol{x}^*)^\top \boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{x}^*), \tag{5}$$

where $\boldsymbol{c}(\boldsymbol{x}^*)$ is a vector of covariance: $(\boldsymbol{c}(\boldsymbol{x}^*))_i = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^*)$. Here $\hat{s}^2(\boldsymbol{x}^*)$ is also called the mean squared error (MSE) of the prediction because $\hat{s}^2(\boldsymbol{x}^*) = \mathrm{E}\left\{(\hat{y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*))^2\right\}$.[2] Consequently, $\hat{s}(\boldsymbol{x}^*)$ is called the root mean squared error (RMSE).

In practice, it is common to assume that $\mu(\boldsymbol{x})$ is a constant function, i.e., $\mu(\boldsymbol{x}) \equiv \mu$. Also, the most widely used kernels in scientific computing are the Matérn functions, especially its two special cases, i.e., exponential and squared-exponential (Gaussian) kernels. For example, the Gaussian kernel can be written as $k(\boldsymbol{\tau}) = \sigma^2 \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|_w^2\right)$, where the weighted norm is defined as $\|\boldsymbol{x} - \boldsymbol{x}'\|_w^2 = \sum_{i=1}^{d} \left(\frac{x_i - x_i'}{l_i}\right)^2$. Here, l_i $(i = 1, \ldots, d)$, the correlation lengths in the i direction, are constants.

In the GPR method with graph kernel, the mean and covariance functions $\mu(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}')$ are obtained by identifying their hyperparameters via maximizing the log marginal

likelihood L[3]

$$\ln L = -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{N}{2}\ln 2\pi. \tag{6}$$

Moreover, to account for the observation noise, one can assume that the noise is independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and variance δ^2, and replace $\boldsymbol{C}$ with $\boldsymbol{C} + \delta^2\boldsymbol{I}$. In this study, we assume that observations $\boldsymbol{y}$ are noiseless. If $\boldsymbol{C}$ is not invertible or its condition number is very large, one can add a small regularization term $\alpha\boldsymbol{I}$ (α is a small positive real number) to $\boldsymbol{C}$, which is equivalent to assuming there is an observation noise. In addition, $\hat{s}$ can be used in global optimization, or in the greedy algorithm to identify locations of additional observations.

Given a stationary covariance function, the covariance matrix $\boldsymbol{C}$ can be written as $\boldsymbol{C} = \sigma^2\boldsymbol{\Psi}$, where $\Psi_{ij} = \exp(-\frac{1}{2}\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_w^2)$. The estimators of μ and σ^2, denoted as $\hat{\mu}$ and $\hat{\sigma}^2$, are

$$\hat{\mu} = \frac{\mathbf{1}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{y}}{\mathbf{1}^\top\boldsymbol{\Psi}^{-1}\mathbf{1}}, \qquad \hat{\sigma}^2 = \frac{(\boldsymbol{y}-\mathbf{1}\hat{\mu})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\mathbf{1}\hat{\mu})}{m}, \tag{7}$$

where $\mathbf{1}$ is a constant vector consisting of 1s,[2] m is the total number of samples. It is also common to set $\mu = 0$.[3] The hyperparameters σ and l_i are identified by maximizing the log marginal likelihood in Eq. (6). The terms $\hat{y}(\boldsymbol{x}^*)$ and $\hat{s}^2(\boldsymbol{x}^*)$ in Eq. (4) take the following form:

$$\hat{y}(\boldsymbol{x}^*) = \hat{\mu} + \boldsymbol{\psi}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\mathbf{1}\hat{\mu}), \tag{8}$$

$$\hat{s}^2(\boldsymbol{x}^*) = \hat{\sigma}^2\left(1 - \boldsymbol{\psi}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}\right), \tag{9}$$

where $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{x}^*)$ is a (column) vector consisting of correlations between the observed data and the prediction, i.e., $\psi_i = \frac{1}{\sigma^2}k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^*)$.

References

(1) Abrahamsen, P. A review of Gaussian random fields and correlation functions. 1997.

(2) Forrester, A.; Keane, A.; Sòbester, A. *Engineering Design via Surrogate Modelling: A Practical Guide*; John Wiley & Sons, 2008.

(3) Williams, C. K.; Rasmussen, C. E. *Gaussian processes for machine learning*; MIT press Cambridge, MA, 2006; Vol. 2.