Supplemental material for: Error mitigation in variational quantum eigensolvers using tailored probabilistic machine learning

Tao Jiang,[1] John Rogers,[2] Marius S. Frank,[3] Ove Christiansen,[3] Yong-Xin Yao,[1,4,*] and Nicola Lanatà[5,6,†]

[1] *Ames Laboratory, Ames, Iowa 50011, USA*
[2] *Department of Physics and Astronomy, Texas A&M University, College Station, Texas 77845, USA*
[3] *Department of Chemistry, Aarhus University, 8000, Aarhus C, Denmark*
[4] *Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA*
[5] *School of Physics and Astronomy, Rochester Institute of Technology, 84 Lomb Memorial Drive, Rochester, New York 14623, USA*
[6] *Center for Computational Quantum Physics, Flatiron Institute, New York, New York 10010, USA*

(Dated: March 5, 2024)

In this supplemental material, we provide a derivation of Gaussian Process Regression (GPR) based on a functional integral perspective. Furthermore, we derive explicitly the customized prior based on a specific Variational Quantum Eigensolver (VQE) ansatz introduced in the main text.

I. PATH INTEGRAL FORMULATION OF PARAMETRIC GPR

Our goal is to learn a real valued function $\mathcal{E}(\boldsymbol{\theta})$ ($\boldsymbol{\theta} \in \mathbb{R}^d$) from a finite set of training data points:

$$D = \{(\boldsymbol{\theta}_\alpha, \mathcal{E}_\alpha, \sigma_\alpha) \,|\, \alpha = 1, .., n\}\,, \tag{1}$$

where each $\mathcal{E}_\alpha$ is the outcome of the evaluation of $\mathcal{E}$ for the input parameter $\boldsymbol{\theta}_\alpha$, which is assumed to be sampled from a gaussian distribution:

$$P(\mathcal{E}_\alpha|\boldsymbol{\theta}_\alpha) \propto \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\mathcal{E}(\boldsymbol{\theta}_\alpha) - \mathcal{E}_\alpha)^2\right\}, \tag{2}$$

i.e., the probability of evaluating $\mathcal{E}_1, .., \mathcal{E}_n$ (which are assumed to be independent) from a given underlying function $\mathcal{E}(\boldsymbol{\theta})$ is assumed to be:

$$\begin{aligned} P[D|\mathcal{E}] &\propto \prod_{\alpha=1}^{n} P(\boldsymbol{\theta}_\alpha|\mathcal{E}_\alpha) \\ &\propto \exp\left\{-\sum_{\alpha=1}^{n}\frac{1}{2\sigma_\alpha^2}(\mathcal{E}(\boldsymbol{\theta}_\alpha) - \mathcal{E}_\alpha)^2\right\}, \end{aligned} \tag{3}$$

Specifically, we aim to compute the so-called "posterior probability distribution" $P[\mathcal{E}|D]$, i.e., the probability that the function that we aim to learn is $\mathcal{E}(\boldsymbol{\theta})$, based on: (I) the data D at our disposal and (II) a gaussian "prior probability distribution" $P[\mathcal{E}]$, encoding our prior knowledge before having any training data points.

Our first goal is to define precisely the concept of a probability distribution over a space of functions. Following the path integral procedure, this can be accomplished by first considering a discrete finite mesh with uniform spacing ϵ, over a d-dimensional rectangle R:

$$M_\epsilon = \{\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_N\} \subset R \subset \mathbb{R}^d\,. \tag{4}$$

Over such discretized domain, probability measures can be rigorously represented as $p_\epsilon[\mathcal{E}]\mathcal{D}_\epsilon[\mathcal{E}]$, where:

$$p_\epsilon[\mathcal{E}] = p_\epsilon[\mathcal{E}(\boldsymbol{\theta}_1), .., \mathcal{E}(\boldsymbol{\theta}_N)] \tag{5}$$

is a standard N-dimensional probability function, and:

$$\mathcal{D}_\epsilon[\mathcal{E}] = \prod_{\boldsymbol{\theta} \in M_\epsilon} d\mathcal{E}(\boldsymbol{\theta}) \tag{6}$$

is the standard path integral measure.

A. The parametric prior

In our context of application, the prior probability distribution is designed to enforce the fact that $\mathcal{E}$ has to be of the following mathematical form:

$$\mathcal{E}(\boldsymbol{\theta}) = \sum_{s=1}^{S} \xi_s T_s(\boldsymbol{\theta})\,, \tag{7}$$

where $T_s : R \subset \mathbb{R}^d \to \mathbb{R}$ are known functions, while the coefficients ξ_s are unknown. This information can be encoded in the following probability distribution:

$$\begin{aligned} P_\epsilon^\eta[\mathcal{E}] \propto \int \prod_{r=1}^{S} d\xi_r \, e^{-\frac{\epsilon}{2\eta^2}\sum_{\boldsymbol{\theta}\in M_\epsilon}\left(\mathcal{E}(\boldsymbol{\theta}) - \sum_s \xi_s T_s(\boldsymbol{\theta})\right)^2} \\ \times\, e^{-\frac{t}{2}\epsilon\sum_{\boldsymbol{\theta}\in M_\epsilon}\mathcal{E}^2(\boldsymbol{\theta})}\,, \end{aligned} \tag{8}$$

where we have introduced the hyperparameter $t > 0$, whose role is to make the probability distribution normalizable by enforcing that the range of $\mathcal{E}$ is bounded as we are going to prove below. The parameter η will be considered in the limit as it approaches zero (i.e., we will take the limit $\eta \to 0$ later in our formalism).

Let us prove that $P_\epsilon^\eta[\mathcal{E}]$ is a normalizable gaussian probability distribution with zero mean for all finite values of η and t. By performing the gaussian integral in Eq. (8), we obtain that:

$$P_\epsilon^\eta[\mathcal{E}] \propto e^{-\frac{1}{2}\sum_{\boldsymbol{\theta},\boldsymbol{\theta}'\in M_\epsilon}\epsilon\left[t\mathbb{1} + \frac{1}{\eta^2}\Pi\right]_{\boldsymbol{\theta}\boldsymbol{\theta}'}\mathcal{E}(\boldsymbol{\theta})\mathcal{E}(\boldsymbol{\theta}')}\,, \tag{9}$$

* Corresponding author: ykent@iastate.edu
† Corresponding author: nxlsps@rit.edu

where $\mathbb{1}$ is the $N \times N$ identity matrix and:

$$\Pi = \mathbb{1} - T(T^\dagger T)^{-1} T^\dagger \tag{10}$$

$$T_{\boldsymbol{\theta} s} = T_s(\boldsymbol{\theta}) \,\forall\, s = 1, .., S,\; \boldsymbol{\theta} \in M_\epsilon\,. \tag{11}$$

Note that Π is an orthogonal projector and, therefore, it is positive semi-definite. It follows that Eq. (9) represents a normalizable zero-mean gaussian distribution $\forall\ t > 0$.

B. Posterior probability distribution

Let us assume to have a series of data D (see Eq. (1)), where $\boldsymbol{\theta}_\alpha \in M_\epsilon\ \forall\, \alpha = 1, .., n$. As explained in the main text (see before Eq.20), from Bayes' theorem it follows that the posterior conditional probability distribution for the function $\mathcal{E}$ is the following:

$$P^\eta_\epsilon[\mathcal{E}|D] \propto P^\eta_\epsilon[\mathcal{E}] e^{-\sum_{\alpha=1}^n \frac{1}{2\sigma_\alpha^2}(\mathcal{E}(\boldsymbol{\theta}_\alpha) - \mathcal{E}_\alpha)^2}\,, \tag{12}$$

which represents the probability distribution for the function $\mathcal{E}$, given the data set D and the prior $P^\eta_\epsilon[\mathcal{E}]$ (see Eqs. (3) and (8)).

C. Probabilistic predictions at a test point

We are interested in calculating quantities of the following form:

$$\begin{aligned}\langle \mathcal{E}^l(\boldsymbol{\theta})\rangle &= \int \mathcal{D}_\epsilon[\mathcal{E}] P^\eta_\epsilon[\mathcal{E}|D]\ (\mathcal{E}(\boldsymbol{\theta}))^l \\ &= \int \mathcal{D}_\epsilon[\mathcal{E}] P^\eta_\epsilon[\mathcal{E}]\, e^{-\sum_{\alpha=1}^n \frac{1}{2\sigma_\alpha^2}(\mathcal{E}(\boldsymbol{\theta}_\alpha) - \mathcal{E}_\alpha)^2} \mathcal{E}^l(\boldsymbol{\theta})\,,\end{aligned} \tag{13}$$

where $l \in \mathbb{N}$ and $\boldsymbol{\theta} \in M_\epsilon$ (which is assumed to be different from all of the $\boldsymbol{\theta}_\alpha$ in the training data set) is a so-called "test point," i.e., a point where we want to evaluate the probability distribution for $\mathcal{E}(\boldsymbol{\theta})$, based on our posterior probability distribution.

Eq. (13) can be conveniently rewritten by integrating out all variables except $\mathcal{E}(\boldsymbol{\theta}_1), .., \mathcal{E}(\boldsymbol{\theta}_\alpha)$ and $\mathcal{E}(\boldsymbol{\theta})$. From standard Gaussian identities, it follows that this gives the following expression:

$$\langle \mathcal{E}^l(\boldsymbol{\theta})\rangle = \frac{\int \left[\prod_{\alpha=1}^n d\mathcal{E}(\boldsymbol{\theta}_\alpha)\right] d\mathcal{E}(\boldsymbol{\theta})\, e^{-S^\eta_\epsilon - U}\ (\mathcal{E}(\boldsymbol{\theta}))^l}{\int \left[\prod_{\alpha=1}^n d\mathcal{E}(\boldsymbol{\theta}_\alpha)\right] d\mathcal{E}(\boldsymbol{\theta})\, e^{-S^\eta_\epsilon - U}}\,, \tag{14}$$

where:

$$U = \sum_{\alpha=1}^n \frac{1}{2\sigma_\alpha^2}(\mathcal{E}(\boldsymbol{\theta}_\alpha) - \mathcal{E}_\alpha)^2 \tag{15}$$

and:

$$\begin{aligned}S^\eta_\epsilon = &\frac{1}{2}\sum_{\alpha,\beta=1}^n [\bar{\mathcal{K}}^\eta_\epsilon]^{-1}_{\alpha,\beta}\, \mathcal{E}(\boldsymbol{\theta}_\alpha)\mathcal{E}(\boldsymbol{\theta}_\beta) \\ &+ \frac{1}{2}[\bar{\mathcal{K}}^\eta_\epsilon]^{-1}_{n+1,n+1}\, \mathcal{E}(\boldsymbol{\theta})\mathcal{E}(\boldsymbol{\theta}) \\ &+ \frac{1}{2}\sum_{\alpha=1}^n [\bar{\mathcal{K}}^\eta_\epsilon]^{-1}_{\alpha,n+1}\, \mathcal{E}(\boldsymbol{\theta}_\alpha)\mathcal{E}(\boldsymbol{\theta}) \\ &+ \frac{1}{2}\sum_{\beta=1}^n [\bar{\mathcal{K}}^\eta_\epsilon]^{-1}_{n+1,\beta}\, \mathcal{E}(\boldsymbol{\theta})\mathcal{E}(\boldsymbol{\theta}_\beta)\,,\end{aligned} \tag{16}$$

where $[\bar{\mathcal{K}}^\eta_\epsilon]$ is the $(n+1) \times (n+1)$ matrix with entries:

$$[\bar{\mathcal{K}}^\eta_\epsilon]_{\alpha,\beta} = K^\eta_\epsilon(\boldsymbol{\theta}_\alpha, \boldsymbol{\theta}_\beta) \quad \forall\, \alpha, \beta \in 1, .., n \tag{17}$$

$$[\bar{\mathcal{K}}^\eta_\epsilon]_{\alpha,n+1} = K^\eta_\epsilon(\boldsymbol{\theta}_\alpha, \boldsymbol{\theta}) \quad \forall\, \alpha \in 1, .., n \tag{18}$$

$$[\bar{\mathcal{K}}^\eta_\epsilon]_{n+1,\beta} = K^\eta_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_\beta) \quad \forall\, \beta \in 1, .., n \tag{19}$$

$$[\bar{\mathcal{K}}^\eta_\epsilon]_{n+1,n+1} = K^\eta_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}) \tag{20}$$

and

$$K^\eta_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int \mathcal{D}_\epsilon[\mathcal{E}] P^\eta_\epsilon[\mathcal{E}]\, \mathcal{E}(\boldsymbol{\theta})\, \mathcal{E}(\boldsymbol{\theta}')\ \ \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}' \in M_\epsilon \tag{21}$$

is the so-called "kernel function" of the prior distribution P^η_ϵ.

As discussed in the main text, we are specifically interested in calculating:

$$\bar{\mathcal{E}}^\eta_\epsilon(\boldsymbol{\theta}) = \int \mathcal{D}_\epsilon[\mathcal{E}] P^\eta_\epsilon[\mathcal{E}|D]\, \mathcal{E}(\boldsymbol{\theta}) \tag{22}$$

$$(\Sigma^\eta_\epsilon(\boldsymbol{\theta}))^2 = \int \mathcal{D}_\epsilon[\mathcal{E}] P^\eta_\epsilon[\mathcal{E}|D]\, \left(\mathcal{E}^2(\boldsymbol{\theta}) - \langle \mathcal{E}(\boldsymbol{\theta})\rangle^2\right), \tag{23}$$

where Eq. (22) represents our prediction for $\mathcal{E}(\boldsymbol{\theta})$ at any test point $\boldsymbol{\theta}$ and Eq. (23) represents the uncertainty of our prediction. These quantities can be conveniently evaluated by computing first the "partition function":

$$Z^\eta_\epsilon(\lambda) := \int \Big[\prod_{\alpha=1}^n d\mathcal{E}(\boldsymbol{\theta}_\alpha)\Big] d\mathcal{E}(\boldsymbol{\theta})\, e^{-S^\eta_\epsilon - U + \lambda \mathcal{E}(\boldsymbol{\theta})} \tag{24}$$

and subsequently using the following identities:

$$\bar{\mathcal{E}}^\eta_\epsilon(\boldsymbol{\theta}) = \partial_\lambda \ln(Z^\eta_\epsilon(\lambda)) \tag{25}$$

$$(\Sigma^\eta_\epsilon(\boldsymbol{\theta}))^2 = \partial^2_\lambda \ln(Z^\eta_\epsilon(\lambda))\,. \tag{26}$$

A direct calculation shows that:

$$\bar{\mathcal{E}}^\eta_\epsilon(\boldsymbol{\theta}) = \sum_{\alpha,\beta=1}^n K^\eta_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_\alpha)[\bar{K}^\eta_\epsilon]^{-1}_{\alpha\beta} \mathcal{E}_\beta \tag{27}$$

$$(\Sigma^\eta_\epsilon(\boldsymbol{\theta}))^2 = K^\eta_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}) - \sum_{\alpha,\beta=1}^n K^\eta_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_\alpha)[\bar{K}^\eta_\epsilon]^{-1}_{\alpha\beta} K^\eta_\epsilon(\boldsymbol{\theta}_\beta, \boldsymbol{\theta})\,, \tag{28}$$

where $\bar{K}^\eta_\epsilon$ is the $n \times n$ matrix with entries:

$$[\bar{K}^\eta_\epsilon]_{\alpha\beta} = K^\eta_\epsilon(\boldsymbol{\theta}_\alpha, \boldsymbol{\theta}_\beta) + \sigma^2_\alpha \delta_{\alpha\beta} \quad \forall\, \alpha, \beta \in 1, .., n\,. \tag{29}$$

D. Calculation of the Kernel function

As shown in the previous section, the GPR estimate of our prediction for $\mathcal{E}(\boldsymbol{\theta})$ and the corresponding uncertainty (see Eqs. (39) and (40), respectively) depend explicitly on η and ϵ through the Kernel function:

$$K^\eta_\epsilon(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \int \mathcal{D}_\epsilon[\mathcal{E}] P^\eta_\epsilon[\mathcal{E}]\, \mathcal{E}(\boldsymbol{\theta}_1)\mathcal{E}(\boldsymbol{\theta}_2)\,, \tag{30}$$

which is defined $\forall\, \boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in M_\epsilon$.

Since we aim to enforce Eq. (7) *exactly*, we need to evaluate Eq. (30) for $\eta \to 0$. In this limit we obtain:

$$\begin{aligned}
K_\epsilon(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= \lim_{\eta\to 0} K^\eta_\epsilon(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\\
&\propto \lim_{\eta\to 0} \int \mathcal{D}_\epsilon[\mathcal{E}] \int \prod_{r=1}^{S} d\xi_r\, e^{-\frac{\epsilon}{2\eta^2}\sum_{\boldsymbol{\theta}\in M_\epsilon}\left(\mathcal{E}(\boldsymbol{\theta})-\sum_s \xi_s T_s(\boldsymbol{\theta})\right)^2} e^{-\frac{t}{2}\epsilon\sum_{\boldsymbol{\theta}\in M_\epsilon}\mathcal{E}^2(\boldsymbol{\theta})}\, \mathcal{E}(\boldsymbol{\theta}_1)\,\mathcal{E}(\boldsymbol{\theta}_2)\\
&\propto \int \mathcal{D}_\epsilon[\mathcal{E}] \int \prod_{r=1}^{S} d\xi_r\, \delta\big(\mathcal{E}(\boldsymbol{\theta}) - \sum_s \xi_s T_s(\boldsymbol{\theta})\big)\, e^{-\frac{t}{2}\epsilon\sum_{\boldsymbol{\theta}\in M_\epsilon}\mathcal{E}^2(\boldsymbol{\theta})}\, \mathcal{E}(\boldsymbol{\theta}_1)\,\mathcal{E}(\boldsymbol{\theta}_2)\\
&= \int \prod_{r=1}^{S} d\xi_r\, e^{-\frac{t}{2}\epsilon\sum_{\boldsymbol{\theta}\in M_\epsilon}\left(\sum_{s=1}^S \xi_s T_s(\boldsymbol{\theta})\right)^2} \left(\sum_{s_1=1}^{S}\xi_{s_1}T_{s_1}(\boldsymbol{\theta}_1)\right)\left(\sum_{s_2=1}^{S}\xi_{s_2}T_{s_2}(\boldsymbol{\theta}_2)\right)\\
&= \sum_{s_1,s_2=1}^{S} T_{s_1}(\boldsymbol{\theta}_1)T_{s_2}(\boldsymbol{\theta}_2) \int \prod_{r=1}^{S} d\xi_r\, e^{-\frac{t}{2}\sum_{s,s'=1}^S \mathcal{A}^\epsilon_{ss'}\xi_s\xi_{s'}}\,\xi_{s_1}\xi_{s_2}\\
&= t^{-1}\sum_{s_1,s_2=1}^{S} \Delta^\epsilon_{s_1 s_2}\, T_{s_1}(\boldsymbol{\theta}_1)T_{s_2}(\boldsymbol{\theta}_2)\,,
\end{aligned} \tag{31}$$

where:

$$\Delta^\epsilon = [\mathcal{A}^\epsilon]^{-1} \tag{32}$$

$$\mathcal{A}^\epsilon_{ss'} = \epsilon \sum_{\boldsymbol{\theta}\in M_\epsilon} T_s(\boldsymbol{\theta})T_{s'}(\boldsymbol{\theta})\,. \tag{33}$$

The final step is to compute the Kernel function in the continuum limit $\epsilon \to 0$, which is given by the following equation:

$$\begin{aligned}
K(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= \lim_{\epsilon\to 0} K_\epsilon(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\\
&= t^{-1}\sum_{s_1,s_2=1}^{S} \Delta_{s_1 s_2}\, T_{s_1}(\boldsymbol{\theta})T_{s_2}(\boldsymbol{\theta})\,,
\end{aligned} \tag{34}$$

where:

$$\Delta = \mathcal{A}^{-1} \tag{35}$$

$$\begin{aligned}
\mathcal{A}_{ss'} &= \lim_{\epsilon\to 0}\mathcal{A}^\epsilon_{ss'} = \lim_{\epsilon\to 0}\epsilon\sum_{\boldsymbol{\theta}\in M_\epsilon} T_s(\boldsymbol{\theta})T_{s'}(\boldsymbol{\theta})\\
&= \int_R d\boldsymbol{\theta}\, T_s(\boldsymbol{\theta})T_{s'}(\boldsymbol{\theta})\,.
\end{aligned} \tag{36}$$

Note that the calculation of $K(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ in Eq. (31) becomes straightforward if the the functions $T_s(\boldsymbol{\theta})$ are replaced by an orthonormal basis $\tau_{\mathbf{k}}(\boldsymbol{\theta})$ of the same space with respect to the $L^2(R)$ metric, as in the main text. In fact, with such a choice we obtain that:

$$K(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = t^{-1}\sum_{\mathbf{k}=1}^{S} \tau_{\mathbf{k}}(\boldsymbol{\theta})\tau_{\mathbf{k}}(\boldsymbol{\theta})\,, \tag{37}$$

which is practically more convenient because: (i) there is no need to invert the matrix $\mathcal{A}$ (which may become prohibitive for high-dimensional spaces), and (ii) evaluating Eq. (37) involves a single summation rather than a double summation, which makes it less computationally demandint to evaluate.

E. Summary of final equations

In summary, by replacing the kernel function $K^\eta_\epsilon(\boldsymbol{\theta},\boldsymbol{\theta}')$ with:

$$K(\boldsymbol{\theta},\boldsymbol{\theta}') = \lim_{\epsilon\to 0}\lim_{\eta\to 0} K^\eta_\epsilon(\boldsymbol{\theta},\boldsymbol{\theta}')\,, \tag{38}$$

we obtain the equations shown in the main text, i.e.:

$$\bar{\mathcal{E}}(\boldsymbol{\theta}) = \sum_{\alpha,\beta=1}^{n} K(\boldsymbol{\theta},\boldsymbol{\theta}_\alpha)[\bar{K}]^{-1}_{\alpha\beta}\mathcal{E}_\beta \tag{39}$$

$$(\Sigma(\boldsymbol{\theta}))^2 = K(\boldsymbol{\theta}, \boldsymbol{\theta}) - \sum_{\alpha,\beta=1}^{n} K(\boldsymbol{\theta}, \boldsymbol{\theta}_\alpha) \bar{K}^{-1}_{\alpha\beta} K(\boldsymbol{\theta}_\beta, \boldsymbol{\theta})\,, \quad (40)$$

where $\bar{K}$ is the $n \times n$ matrix with entries:

$$[\bar{K}]_{\alpha\beta} = K(\boldsymbol{\theta}_\alpha, \boldsymbol{\theta}_\beta) + \sigma^2_\alpha \delta_{\alpha\beta} \quad \forall\, \alpha, \beta \in 1, .., n\,. \quad (41)$$