Appendix:

Optimize to Generalize in Gaussian Processes: An Alternative Objective Based on the Rényi Divergence

Introduction

This appendix contains all technical details in our main paper and some additional empirical results.

A The Variational Rényi Lower Bound

A.1 The Rényi Divergence

The Rényi's α-divergence between p and q is defined as [Rényi et al., 1961]

$$D_\alpha[p||q] = \frac{1}{\alpha - 1} \log \int p(\boldsymbol{w})^\alpha q(\boldsymbol{w})^{1-\alpha} d\boldsymbol{w}, \alpha \in [0, 1),$$

where $\boldsymbol{w}$ is the parameter for p, q. Let $q \coloneqq q(\boldsymbol{f}, \boldsymbol{U})$ and $p \coloneqq p(\boldsymbol{f}, \boldsymbol{U}, \boldsymbol{Y})$. In the context of $\mathcal{GP}$s, we have

$$\begin{aligned}
&D_\alpha[q(\boldsymbol{f}, \boldsymbol{U})||p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})] \\
&= \frac{1}{\alpha - 1} \log \int q(\boldsymbol{f}, \boldsymbol{U})^\alpha p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})^{1-\alpha} d\boldsymbol{U} d\boldsymbol{f} \\
&= \frac{1}{1-\alpha} \log P(\boldsymbol{Y})^{1-\alpha} - \frac{1}{1-\alpha} \log \int q(\boldsymbol{f}, \boldsymbol{U})^\alpha \left(p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y}) p(\boldsymbol{Y})\right)^{1-\alpha} d\boldsymbol{U} d\boldsymbol{f} \\
&= \log p(\boldsymbol{Y}) - \frac{1}{1-\alpha} \log \int q(\boldsymbol{f}, \boldsymbol{U}) \frac{\left(p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y}) p(\boldsymbol{Y})\right)^{1-\alpha}}{q(\boldsymbol{f}, \boldsymbol{U})^{1-\alpha}} d\boldsymbol{U} d\boldsymbol{f} \\
&= \log p(\boldsymbol{Y}) - \frac{1}{1-\alpha} \log \mathbb{E}_q \left[\left(\frac{p(\boldsymbol{f}, \boldsymbol{U}, \boldsymbol{Y})}{q(\boldsymbol{f}, \boldsymbol{U})} \right)^{1-\alpha} \right].
\end{aligned}$$

Therefore, the Rényi variational lower bound can be derived as

$$\mathcal{L}_\alpha(q;\boldsymbol{Y}) = \frac{1}{1-\alpha}\log \mathbb{E}_q\left[\left(\frac{p(\boldsymbol{f},\boldsymbol{U},\boldsymbol{Y})}{q(\boldsymbol{f},\boldsymbol{U})}\right)^{1-\alpha}\right]. \tag{1}$$

A.2 Mean-field Assumption

When we apply the Rényi divergence to $\mathcal{GP}$ and assume that $q(\boldsymbol{f},\boldsymbol{U}) = p(\boldsymbol{f}|\boldsymbol{U})q(\boldsymbol{U})$ (mean-field assumption), we can further obtain

$$\begin{aligned}
\mathcal{L}_\alpha(q;\boldsymbol{Y}) &\coloneqq \frac{1}{1-\alpha}\log \mathbb{E}_q\left[\left(\frac{p(\boldsymbol{f},\boldsymbol{U},\boldsymbol{Y})}{q(\boldsymbol{f},\boldsymbol{U})}\right)^{1-\alpha}\right] \\
&= \frac{1}{1-\alpha}\log \mathbb{E}_q\left[\left(\frac{p(\boldsymbol{Y}|\boldsymbol{f})\cancel{p(\boldsymbol{f}|\boldsymbol{U})}p(\boldsymbol{U})}{\cancel{p(\boldsymbol{f}|\boldsymbol{U})}q(\boldsymbol{U})}\right)^{1-\alpha}\right] \\
&= \frac{1}{1-\alpha}\log \int p(\boldsymbol{f}|\boldsymbol{U})q(\boldsymbol{U})\left(\frac{p(\boldsymbol{Y}|\boldsymbol{f})p(\boldsymbol{U})}{q(\boldsymbol{U})}\right)^{1-\alpha} d\boldsymbol{U}d\boldsymbol{f} \\
&= \frac{1}{1-\alpha}\log \int p(\boldsymbol{f}|\boldsymbol{U})q(\boldsymbol{U})^{\alpha}\Big(p(\boldsymbol{Y}|\boldsymbol{f})p(\boldsymbol{U})\Big)^{1-\alpha} d\boldsymbol{U}d\boldsymbol{f} \\
&= \frac{1}{1-\alpha}\log \int\int p(\boldsymbol{f}|\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}d\boldsymbol{f}q(\boldsymbol{U})^{\alpha}p(\boldsymbol{U})^{1-\alpha}d\boldsymbol{U}.
\end{aligned}$$

For simplicity, we drop the notation $\boldsymbol{Y}$ in the $\mathcal{L}_\alpha(q;\boldsymbol{Y})$. It can be easily shown that $p(\boldsymbol{f}|\boldsymbol{U}) = \phi(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{U}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$, where $\boldsymbol{Q} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}}$. Besides, we have $p(\boldsymbol{Y}|\boldsymbol{f}) = \phi(\boldsymbol{f}, \sigma_\epsilon^2 I)$. Therefore,

$$\begin{aligned}
&\int p(\boldsymbol{f}|\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}d\boldsymbol{f} \\
&= \int p(\boldsymbol{f}|\boldsymbol{U})(|2\pi\sigma_\epsilon^2 I|^{-0.5}e^{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{f})^T(\sigma_\epsilon^2 I)^{-1}(\boldsymbol{Y}-\boldsymbol{f})})^{1-\alpha}d\boldsymbol{f} \\
&= \frac{|2\pi\sigma_\epsilon^2 I|^{-0.5(1-\alpha)}}{|2\pi\sigma_\epsilon^2 I/(1-\alpha)|^{-0.5}}\int p(\boldsymbol{f}|\boldsymbol{U})\phi(\boldsymbol{f}, \frac{\sigma_\epsilon^2 I}{1-\alpha})d\boldsymbol{f} \\
&= \frac{|2\pi\sigma_\epsilon^2 I|^{-0.5(1-\alpha)}}{|2\pi\sigma_\epsilon^2 I/(1-\alpha)|^{-0.5}}\phi(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{U}, \frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}) \\
&= (2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2}}(\frac{1}{1-\alpha})^{\frac{N}{2}}\phi(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{U}, \frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}) \\
&= p_\alpha(\boldsymbol{Y}|\boldsymbol{U}).
\end{aligned}$$

A.3 Find the Optimal Member, q, of the Family of Approximate Densities $\mathcal{Q}$

Instead of treating $q(\boldsymbol{U})$ as a pool of free parameters, it is desirable to find the optimal $q^*(\boldsymbol{U})$ to maximize the lower bound. To proceed, we have,

$$\begin{aligned}
\mathcal{L}_\alpha(q) &= \frac{1}{1-\alpha} \log \int p_\alpha(\boldsymbol{Y}|\boldsymbol{U}) q(\boldsymbol{U})^\alpha p(\boldsymbol{U})^{1-\alpha} d\boldsymbol{U} \\
&= \frac{1}{1-\alpha} \log \int q(\boldsymbol{U}) (\frac{p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U})}{q(\boldsymbol{U})})^{1-\alpha} d\boldsymbol{U} \\
&= \frac{1}{1-\alpha} \log \mathbb{E}_q (\frac{p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U})}{q(\boldsymbol{U})})^{1-\alpha}
\end{aligned}$$

By taking derivative of $\mathcal{L}_\alpha(q)$ with respect to $q(\boldsymbol{U})$ and set it to 0, we can obtain the optimal expression of $q(\boldsymbol{U})$:

$$q^*(\boldsymbol{U}) \propto p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}).$$

Specifically,

$$q^*(\boldsymbol{U}) = \frac{p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U})}{\int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}) d\boldsymbol{U}}.$$

Therefore, we can obtain

$$\begin{aligned}
\mathcal{L}^*_\alpha(q; \boldsymbol{Y}) &= \frac{1}{1-\alpha} \log[\mathbb{E}_q (\frac{p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U})}{q(\boldsymbol{U})})]^{1-\alpha} \\
&= \log \mathbb{E}_q (\frac{p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U})}{q(\boldsymbol{U})}) \\
&= \log \int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}) d\boldsymbol{U}.
\end{aligned}$$

where $\mathcal{L}^*_\alpha(q; \boldsymbol{Y})$ is $\mathcal{L}_\alpha(q)$ with $q^*(\boldsymbol{U})$.

A.4 Finding the closed form

So far, we have shown that

$$\mathcal{L}^*_\alpha(q; \boldsymbol{Y}) = \log \int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}) d\boldsymbol{U}.$$

Our final goal is to simplify this integration and obtain our proposed lower bound.

It can be shown that

$$
\begin{aligned}
p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{\frac{1}{1-\alpha}} &= [(2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2}}(\frac{1}{1-\alpha})^{\frac{N}{2}}]^{\frac{1}{1-\alpha}}\phi(\boldsymbol{K_{f,U}}\boldsymbol{K_{U,U}^{-1}}\boldsymbol{U}, \frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K_{f,f}} - \boldsymbol{Q})^{\frac{1}{1-\alpha}} \\
&= [(2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2(1-\alpha)}}(\frac{1}{1-\alpha})^{\frac{N}{2(1-\alpha)}}]C\phi(\boldsymbol{K_{f,U}}\boldsymbol{K_{U,U}^{-1}}\boldsymbol{U}, \sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K_{f,f}} - \boldsymbol{Q}]),
\end{aligned}
$$

where $C = \frac{|2\pi(\frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{-0.5/(1-\alpha)}}{|2\pi(\sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K_{f,f}} - \boldsymbol{Q}])|^{-0.5}} = |2\pi(\frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}(1-\alpha)^{N/2}$. Since $p(\boldsymbol{U}) = \phi(\boldsymbol{0}, \boldsymbol{K_{U,U}})$, we have

$$
\begin{aligned}
\mathcal{L}_\alpha(q) &= \log \int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}) d\boldsymbol{U} \\
&= \log C_x \phi(\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K_{f,f}} - \boldsymbol{Q}] + \boldsymbol{K_{f,U}}\boldsymbol{K_{U,U}^{-1}}\boldsymbol{K_{U,f}}) \\
&= \log C_x \phi(\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K_{f,f}} - \boldsymbol{Q}] + \boldsymbol{Q}) \\
&= \log \phi(\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K_{f,f}}] + \alpha\boldsymbol{Q}) + \log C_x,
\end{aligned}
$$

where

$$
\begin{aligned}
C_x &= [(2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2(1-\alpha)}}(\frac{1}{1-\alpha})^{\frac{N}{2(1-\alpha)}}][|2\pi(\frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}(1-\alpha)^{N/2}] \\
&= (2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2(1-\alpha)}}(1-\alpha)^{\frac{-\alpha N}{2(1-\alpha)}}|2\pi(\frac{\sigma_\epsilon^2}{1-\alpha}I + \boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} \\
&= |\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} \\
&\approx \left\{1 + \frac{1-\alpha}{\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K_{f,f}} - \boldsymbol{Q}) + \mathcal{O}(\frac{(1-\alpha)^2}{\sigma_\epsilon^4})\right\}^{\frac{-\alpha}{2(1-\alpha)}}.
\end{aligned}
$$

The last equality comes from the variation of Jacobi's formula. The $\approx$ approximates well only when $\frac{1-\alpha}{\sigma_\epsilon^2}$ is "small". It can be seen that, when α is close to 1, our objective function contains the regularization term $\frac{1-\alpha}{\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})$, which is similar to the regularization term in $\mathcal{L}_{VI}$. The tuning parameter α controls how close $\boldsymbol{Q}$ is to $\boldsymbol{K_{f,f}}$ and hence it encourages densities q that place their mass on configurations of the latent variables that explain the observed data. This is also true for any $\alpha \in [0, 1)$ yet the regularization effect is conveyed through the determinant $|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}$.

B Other Bounds

In this section, we provide details on obtaining $\mathcal{L}_{jensen}$ and $\mathcal{L}_{VI}$.

$$\begin{aligned}
\mathcal{L}_\alpha(q) &:= \frac{1}{1-\alpha}\log\mathbb{E}_q\left[\left(\frac{p(\boldsymbol{f},\boldsymbol{U},\boldsymbol{Y}|\boldsymbol{\mathcal{Z}})}{q(\boldsymbol{f},\boldsymbol{U}|\boldsymbol{\mathcal{Z}})}\right)^{1-\alpha}\right] \\
&= \frac{1}{1-\alpha}\log\int\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})^{1-\alpha}q(\boldsymbol{U})d\boldsymbol{U} \\
&= \underbrace{\frac{1}{1-\alpha}\log\mathbb{E}_{q(\boldsymbol{U})}\left[\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})^{1-\alpha}\right]}_{\text{Rényi variational lower bound}} \qquad (2) \\
&\geq \frac{1}{1-\alpha}\left\{\mathbb{E}\log\left[\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})^{1-\alpha}\right]\right\} \\
&= \frac{1}{1-\alpha}\left\{\int q(\boldsymbol{U})\log\left[\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})^{1-\alpha}\right]d\boldsymbol{U}\right\} \\
&= \frac{1}{1-\alpha}\Bigg\{\int q(\boldsymbol{U})\log\left[q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})^{1-\alpha}\right] \\
&\qquad + q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)d\boldsymbol{U}\Bigg\} \\
&= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})] + \frac{1}{1-\alpha}\left\{\int q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)d\boldsymbol{U}\right\} \qquad (3) \\
&\geq -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})] + \frac{1}{1-\alpha}\left\{\int q(\boldsymbol{U})\int p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})\log\left(p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}\right)d\boldsymbol{f}d\boldsymbol{U}\right\} \\
&= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})] + \frac{1}{1-\alpha}\left\{\int q(\boldsymbol{U})\int p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})\log\left(p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}\right)d\boldsymbol{f}d\boldsymbol{U}\right\} \\
&= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})] + \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{U})}\left[\log p(\boldsymbol{Y}|\boldsymbol{f})\right] = \mathcal{L}_{VI}. \qquad (4)
\end{aligned}$$

Here,

$$\begin{aligned}
\mathcal{L}_\alpha(q) &= \frac{1}{1-\alpha}\log\mathbb{E}_{q(\boldsymbol{U})}\left[\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})^{1-\alpha}\right], \\
\mathcal{L}_{Jensen} &= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})] + \frac{1}{1-\alpha}\left\{\int q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)d\boldsymbol{U}\right\}, \\
\mathcal{L}_{VI} &= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})] + \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{U})}\left[\log p(\boldsymbol{Y}|\boldsymbol{f})\right].
\end{aligned}$$

It can be seen that $\mathcal{L}_\alpha(q) \geq \mathcal{L}_{Jensen} \geq \mathcal{L}_{VI}$. Therefore, $\mathcal{L}_{Jensen}$ is decreasing as $\alpha \to 1$. This implies $\frac{1}{1-\alpha}\left\{\int q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{\mathcal{Z}})d\boldsymbol{f}\right)d\boldsymbol{U}\right\}$ is decreasing as $\alpha \to 1$. Alternatively, one can take a derivative with respect to α and conclude that the aforementioned function is decreasing.

C Computation

We elaborate the modified BBMM approach here. By scrutinizing our objective function, defined below,

$$\begin{aligned}\mathcal{L}_\alpha(q^*) &= \log\int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)}p(\boldsymbol{U})d\boldsymbol{U} = \\ &\log\left\{\mathcal{N}\Big(\boldsymbol{0},\sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q}\Big)\right\} + \log\left|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\right|^{\frac{-\alpha}{2(1-\alpha)}}.\end{aligned} \tag{5}$$

we can see that the computational complexity is dominated by the first term, which has the same complexity as the exact $\mathcal{GP}$, and the determinant term. The detailed computing procedure is provided as follows. We rewrite the function above as

$$\mathcal{L}_\alpha(q^*) = \log|2\pi\boldsymbol{\Xi}|^{-\frac{1}{2}} - \frac{1}{2}\boldsymbol{Y}^T\boldsymbol{\Xi}^{-1}\boldsymbol{Y} + \log C_x \tag{6}$$

where matrices $\boldsymbol{\Xi} \coloneqq \sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q}$ and $C_x = \left|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\right|^{\frac{-\alpha}{2(1-\alpha)}}$.

In Eq. (6), two expensive terms $\log|\boldsymbol{\Xi}|$ and $\boldsymbol{\Xi}^{-1}\boldsymbol{Y}$ can be efficiently estimated by the Batched Conjugate Gradients Algorithm (mBCG) [Gardner et al., 2018] with some modifications. The remaining work is to estimate the determinant term. First, we can write it as

$$\log C_x = \log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} = \log|\frac{\boldsymbol{\Xi}}{\sigma_\epsilon^2} + \frac{1-2\alpha}{\sigma_\epsilon^2}\boldsymbol{Q}|^{\frac{-\alpha}{2(1-\alpha)}}.$$

By the matrix determinant lemma, we have

$$
\begin{aligned}
\log|\frac{\boldsymbol{\Xi}}{\sigma_\epsilon^2}+\frac{1-2\alpha}{\sigma_\epsilon^2}\boldsymbol{Q}| &= \log|\frac{1}{\sigma_\epsilon^2}||\boldsymbol{\Xi}+(1-2\alpha)\boldsymbol{Q}| = \log|\frac{1}{\sigma_\epsilon^2}||\boldsymbol{\Xi}+(1-2\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}}| \\
&= \log|\frac{1}{\sigma_\epsilon^2}||\frac{1}{(1-2\alpha)}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}+\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}}\boldsymbol{\Xi}^{-1}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}|+\log|(1-2\alpha)\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}|+\log|\boldsymbol{\Xi}|.
\end{aligned}
$$

In this equation, $\log|\boldsymbol{\Xi}|$ is already available as aforementioned. Therefore, only $\boldsymbol{\Xi}^{-1}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}$ is expensive to compute. Similarly, we resort to the CG algorithm to overcome this difficulty. Overall, the resulting matrix is of dimension $M\times M$ (note that $M \ll N$) and is cheap to compute.

On Computing Inverse

$\boldsymbol{\Xi}^{-1}\boldsymbol{Y}$ can be calculated by the conjugate gradient (CG) algorithm. Specifically, we solve the following quadratic optimization problem

$$
\boldsymbol{\Xi}^{-1}\boldsymbol{Y} = \arg\min_{\boldsymbol{u}}\left(\frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Xi}\boldsymbol{u}-\boldsymbol{u}^T\boldsymbol{Y}\right).
$$

Furthermore, CG can be extended to return a matrix output. Let $\boldsymbol{\Theta}=[\boldsymbol{Y}\quad \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}]$, then we can compute both $\boldsymbol{\Xi}^{-1}\boldsymbol{Y}$ and $\boldsymbol{\Xi}^{-1}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}$ by solving

$$
\boldsymbol{\Xi}^{-1}\boldsymbol{\Theta} = \arg\min_{\boldsymbol{U}}\left(\frac{1}{2}\boldsymbol{U}^T\boldsymbol{\Theta}\boldsymbol{U}-\boldsymbol{U}^T\boldsymbol{\Theta}\right).
$$

On Computing Determinant

$\log|\boldsymbol{\Xi}|$ can be computed in two ways. First, we can use pivoted Cholesky decomposition. Second, we can use Lanczos algorithm. When running Lanczos algorithm, we only need to return the Tridiagonal matrix T and we have $\log|\boldsymbol{\Xi}| = \mathrm{Tr}(\log T)$.

On Computing Gradient

Let $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t]$ be a set of vectors where $\boldsymbol{z}_i$ is drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Then we can use mBCG to compute $\boldsymbol{\Xi}^{-1}\boldsymbol{Z}$ and calculate gradient as

$$\mathrm{Tr}\left(\boldsymbol{\Xi}^{-1}\frac{d\boldsymbol{\Xi}}{d\boldsymbol{w}}\right) \approx \frac{1}{t}\sum_{i=1}^{t}(\boldsymbol{z}_i^T\boldsymbol{\Xi}^{-1})\left(\frac{d\boldsymbol{\Xi}}{d\boldsymbol{w}}\boldsymbol{z}_i\right).$$

where $\boldsymbol{w} = (\sigma_\epsilon, \boldsymbol{\theta})$ is our model parameters. Please refer to Gardner et al. [2018] for the detailed implementation.

D Convergence Results

D.1 An Upper Bound

Lemma 1. *Suppose we have two positive semi-definite (PSD) matrices A and B such that $A - B$ is also a PSD matrix, then $|A| \geq |B|$. Furthermore, if A and B are positive definite (PD), then $B^{-1} \geq A^{-1}$.*

The proof of this Lemma can be found in any matrix theory textbook. Based on this lemma, we can compute a data-dependent upper bound on the log-marginal likelihood [Titsias, 2014].

Claim 1. $\log p(\boldsymbol{Y}) \leq \log \frac{1}{|2\pi((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\alpha Tr(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})\boldsymbol{I}+\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}} := \mathcal{L}_{upper}$.

Proof. Since

$$\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I} = (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I} \succeq (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \sigma_\epsilon^2\boldsymbol{I} \succeq 0,$$

where $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} \geq \mathbf{0}, \forall \boldsymbol{x}$. Then, we can obtain $|\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I}| \geq$

$|(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \sigma_\epsilon^2\boldsymbol{I}|$ since they are both PSD matrix. Therefore,

$$\frac{1}{|2\pi(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}} \leq \frac{1}{|2\pi((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}}.$$

Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the eigen-decomposition of $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}$. This decomposition exists since the matrix is PD. Then

$$\begin{aligned}\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{Y} = \boldsymbol{z}^T\boldsymbol{\Lambda}\boldsymbol{z} = \sum_{i=1}^N \lambda_i z_i^2 \leq \lambda_{max}\sum_{i=1}^N z_i^2 = \lambda_{max}\|\boldsymbol{z}\|^2 \\ = \lambda_{max}\|\boldsymbol{Y}\|^2 \leq \sum_{i=1}^N \lambda_i \|\boldsymbol{Y}\|^2 \leq \mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\|\boldsymbol{Y}\|^2,\end{aligned}$$

where $\boldsymbol{z} = \boldsymbol{U}^T\boldsymbol{Y}$, $\{\lambda_i\}_{i=1}^N$ are eigenvalues of $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}$ and $\lambda_{max} = \max(\lambda_1, \ldots, \lambda_N)$. Therefore, we have $\boldsymbol{Y}^T(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{Y} \leq \mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\|\boldsymbol{Y}\|^2 = \mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{Y}^T\boldsymbol{Y}$. Apparently, $\alpha\boldsymbol{Y}^T(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{Y} \leq \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{Y}^T\boldsymbol{Y}$. Therefore, we can obtain

$$\begin{aligned}&\boldsymbol{Y}^T(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I})\boldsymbol{Y} \leq \boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \sigma_\epsilon^2\boldsymbol{I})\boldsymbol{Y} + \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{Y}^T\boldsymbol{Y} \\ &= \boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{I} + \sigma_\epsilon^2\boldsymbol{I})\boldsymbol{Y}.\end{aligned}$$

Based on this inequality, it is easy to show that

$$e^{-\frac{1}{2}\boldsymbol{Y}^T(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}} \leq e^{-\frac{1}{2}\boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{I} + \sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}}.$$

Finally, we obtain

$$\begin{aligned}&\frac{1}{|2\pi(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{Y}^T(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}} \\ &\leq \frac{1}{|2\pi((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q} + \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{I} + \sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}}.\end{aligned}$$

$\square$

We will use this upper bound to prove our main theorem.

D.2 Rate of Convergence and Related Lemmas

Claim 2. $-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} \leq \frac{\alpha}{2(1-\alpha)}\log\left(\frac{Tr(\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}))}{N}\right)^N.$

Proof. Based on the inequality of arithmetic and geometric means, we have

$$\frac{\mathrm{Tr}(M)}{N} \geq |M|^{1/N},$$

given an positive semi-definite matrix M with dimension N. Therefore, we can obtain

$$|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{1/N} \leq \frac{\mathrm{Tr}(\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}))}{N}.$$

By some simple algebra manipulation, we will obtain

$$\frac{\alpha}{2(1-\alpha)}\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})| \leq \frac{\alpha}{2(1-\alpha)}\log\left(\frac{\mathrm{Tr}(\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}))}{N}\right)^N.$$

□

We first provide a lower bound and an upper bound on the Rényi divergence.

Lemma 2. *For any set of $\{\boldsymbol{x}_i\}_{i=1}^N$, if the output $\{y_i\}_{i=1}^N$ are generated according to some generative model, then*

$$\begin{aligned}&-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} \leq \mathbb{E}_y\Bigg[D_\alpha[p||q]\Bigg]\\ &\leq -\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{\alpha\, Tr(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}{2\sigma_\epsilon^2}.\end{aligned} \tag{7}$$

Proof. We have

$$\begin{aligned}&\mathbb{E}_y\Bigg[D_\alpha[p||q]\Bigg]\\ &= \mathbb{E}_y\Bigg[\log p(\boldsymbol{Y}) - \log\phi(\boldsymbol{0},\sigma_\epsilon^2 I+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}) - \log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}\Bigg]\\ &= -\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \mathbb{E}_y\Bigg[\log\frac{\phi(\boldsymbol{0},\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I})}{\phi(\boldsymbol{0},\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})}\Bigg].\end{aligned}$$

It is apparent that the lower bound to (7) is

$$-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}},$$

since the KL divergence is non-negative. We then provide an upper bound to (7). We have

$$\begin{aligned}
&-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}+\mathbb{E}_y\left[\log\frac{\phi(\boldsymbol{0},\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I})}{\phi(\boldsymbol{0},\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})}\right]\\
&=-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}\\
&\quad-\frac{N}{2}+\frac{1}{2}\log\left(\frac{|\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}|}{|\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I}|}\right)+\frac{1}{2}\mathrm{Tr}\Big((\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})^{-1}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I})\Big)\\
&\leq-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}-\frac{N}{2}+\frac{1}{2}\mathrm{Tr}\Big((\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})^{-1}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I})\Big).
\end{aligned}$$

This inequality follows from the fact that $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I}\succeq\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}$. Since

$$\begin{aligned}
&\frac{1}{2}\mathrm{Tr}\Big((\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})^{-1}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I})\Big)\\
&=\frac{1}{2}\mathrm{Tr}(\boldsymbol{I})+\frac{1}{2}\mathrm{Tr}\left((\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})^{-1}(\tilde{\boldsymbol{K}})\right)\\
&\leq\frac{N}{2}+\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})\lambda_1((\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q})^{-1})/2\\
&\leq\frac{N}{2}+\frac{\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}{2\sigma_\epsilon^2},
\end{aligned}$$

where $\tilde{\boldsymbol{K}}=\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\sigma_\epsilon^2\boldsymbol{I}-\Big(\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}\Big)$ and $\lambda_1(\boldsymbol{M})$ is the largest eigenvalue of an arbitrary matrix M. We apply the Hölder's inequality for schatten norms to the second last inequality. Therefore, we obtain the upper bound as follow.

$$-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}+\frac{\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}{2\sigma_\epsilon^2}.$$

$\square$

As $\alpha\to 1$, we recover the bounds for the KL divergence. Specifically, we get the lower bound $\frac{\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}{2\sigma_\epsilon^2}$ and upper bound $\frac{\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}{\sigma_\epsilon^2}$ [**?**].

Lemma 3. *Given a symmetric positive semidefinite matrix $\boldsymbol{K_{f,f}}$, if M columns are selected to form a Nyström approximation such that the probability of selecting a subset of columns Z is proportional to the determinant of the principal submatrix formed by these columns and the matching rows, then*

$$\mathbb{E}_Z\bigg[Tr(\boldsymbol{K_{f,f}} - \boldsymbol{Q})\bigg] \leq (M+1)\sum_{m=M+1}^{N} \lambda_m(\boldsymbol{K_{f,f}}).$$

This lemma is proved in [?]. Following this lemma and by Lemma 2, we can show that

$$\begin{aligned}
&\mathbb{E}_Z\bigg[-\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}\bigg] \\
&= \mathbb{E}_Z\bigg[\frac{\alpha}{2(1-\alpha)}\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})|\bigg] \\
&\leq \mathbb{E}_Z\bigg[\frac{\alpha}{2(1-\alpha)}\log\bigg(\frac{\mathrm{Tr}(\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q}))}{N}\bigg)^N\bigg] \\
&\leq \frac{\alpha N}{2(1-\alpha)}\log\mathbb{E}_Z\bigg[\bigg(\frac{\mathrm{Tr}(\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q}))}{N}\bigg)\bigg] \\
&\leq \frac{\alpha N}{2(1-\alpha)}\log\bigg\{1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{(M+1)\sum_{m=M+1}^{N}\lambda_m(\boldsymbol{K_{f,f}})}{N}\bigg\}.
\end{aligned}$$

As $\alpha \to 1$, this bound becomes $\frac{1}{2\sigma_\epsilon^2}(M+1)\sum_{m=M+1}^{N}\lambda_m(\boldsymbol{K_{f,f}})$. Following the inequality and lemma above, we can obtain the following corollary.

Corollary 1.

$$\mathbb{E}_{Z\sim v}[Tr(\boldsymbol{K_{f,f}} - \boldsymbol{Q})] \leq (M+1)\sum_{m=M+1}^{N}\lambda_m(\boldsymbol{K_{f,f}}) + 2Nv\epsilon.$$

This inequality is from [?]. Using this fact, we can show that

$$
\begin{aligned}
&\mathbb{E}_{Z\sim v}\left[-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}\right]\\
&\leq \frac{\alpha}{2(1-\alpha)}\mathbb{E}_{Z\sim v}\left[\log\left(\frac{\mathrm{Tr}(\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}))}{N}\right)^N\right]\\
&\leq \frac{\alpha N}{2(1-\alpha)}\log\left[1+\frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)\sum_{m=M+1}^N\lambda_m(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}})+2Nv\epsilon]}{N}\right].
\end{aligned}
$$

The next theorem is based on a lemma. We will prove this lemma first.

Lemma 4.

$$
D_\alpha[p||q]\leq -\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}+\|\boldsymbol{Y}\|^2\,\frac{\alpha\,Tr(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}{\sigma_\epsilon^4+\alpha\sigma_\epsilon^2\,Tr(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}
$$

where $\tilde{\lambda}_{max}$ is the largest eigenvalue of $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}$.

Proof. Based on Claim 1, we have

$$
\begin{aligned}
\mathcal{L}_{upper}&=\log\frac{1}{|2\pi((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}}e^{-\frac{1}{2}\boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})\boldsymbol{I}+\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}}\\
&\leq -\frac{1}{2}\log|(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\sigma_\epsilon^2\boldsymbol{I}|-\frac{N}{2}\log(2\pi)-\frac{1}{2}\boldsymbol{Y}^T((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\alpha\tilde{\lambda}_{max}\boldsymbol{I}+\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{Y}\\
&:=\mathcal{L}'_{upper},
\end{aligned}
$$

using the fact that $\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})\geq\tilde{\lambda}_{max}$. Then, we have

$$
\begin{aligned}
&\mathcal{L}'_{upper}-\mathcal{L}_\alpha(q)\\
&=-\log|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}\\
&\quad+\frac{1}{2}\boldsymbol{Y}^T\Big(((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\sigma_\epsilon^2\boldsymbol{I})^{-1}-((1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\alpha\tilde{\lambda}_{max}\boldsymbol{I}+\sigma_\epsilon^2\boldsymbol{I})^{-1}\Big)\boldsymbol{Y}.
\end{aligned}
$$

Let $(1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}+\alpha\boldsymbol{Q}+\sigma_\epsilon^2\boldsymbol{I}=\boldsymbol{V}\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}\boldsymbol{V}^T$ be the eigenvalue decomposition and denote by

$\gamma_1 \geq \ldots \geq \gamma_N$ all eigenvalues. Then we can obtain

$$
\begin{aligned}
&\frac{1}{2}(\boldsymbol{V}^T\boldsymbol{Y})^T\bigg(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-1} - (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}} + \alpha\tilde{\lambda}_{max}\boldsymbol{I})^{-1}\bigg)(\boldsymbol{V}^T\boldsymbol{Y}) \\
&= \frac{1}{2}\boldsymbol{z'}^T\bigg(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-1} - (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}} + \alpha\tilde{\lambda}_{max}\boldsymbol{I})^{-1}\bigg)\boldsymbol{z'} \\
&= \frac{1}{2}\sum_i z_i'^2 \frac{\alpha\tilde{\lambda}_{max}}{\gamma_i^2 + \alpha\gamma_i\tilde{\lambda}_{max}} \\
&\leq \frac{1}{2}\left\|\boldsymbol{Y}\right\|^2 \frac{\alpha\tilde{\lambda}_{max}}{\gamma_N^2 + \alpha\gamma_N\tilde{\lambda}_{max}} \\
&\leq \frac{1}{2}\left\|\boldsymbol{Y}\right\|^2 \frac{\alpha\tilde{\lambda}_{max}}{\sigma_\epsilon^4 + \alpha\sigma_\epsilon^2\tilde{\lambda}_{max}},
\end{aligned}
$$

where $\boldsymbol{z'} = \boldsymbol{V}^T\boldsymbol{Y}$. Therefore, we have

$$
\begin{aligned}
D_\alpha[p||q] &\leq -\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{1}{2}\left\|\boldsymbol{Y}\right\|^2 \frac{\alpha\tilde{\lambda}_{max}}{\sigma_\epsilon^4 + \alpha\sigma_\epsilon^2\tilde{\lambda}_{max}} \\
&\leq -\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{1}{2}\left\|\boldsymbol{Y}\right\|^2 \frac{\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})}{\sigma_\epsilon^4 + \alpha\sigma_\epsilon^2\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})}.
\end{aligned}
$$

□

For simplicity, we split our main theorem into two theorems and prove them separately.

Theorem 1. *Suppose N data points are drawn i.i.d from input distribution $p(\boldsymbol{x})$ and $k(\boldsymbol{x},\boldsymbol{x}) \leq v, \forall \boldsymbol{x} \in \mathcal{X}$. Sample M inducing points from the training data with the probability assigned to any set of size M equal to the probability assigned to the corresponding subset by an ϵ k-Determinantal Point Process (k-DPP) [**?**] with $k = M$. If $\boldsymbol{Y}$ is distributed according to a sample from the prior generative model, with probability at least $1-\delta$,*

$$
\begin{aligned}
D_\alpha[p||q] \leq \alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{2\delta\sigma_\epsilon^2} + \\
\frac{1}{\delta}\frac{\alpha}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon]}{N}\right]^N.
\end{aligned}
$$

where λ_m are the eigenvalues of the integral operator $\mathcal{K}$ associated to kernel, k and $p(\boldsymbol{x})$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{Z|\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{Y}}\left[D_\alpha[p||q]\right]\right]\right] \\
&\le \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{Z|\boldsymbol{X}}\left[-\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{\alpha \mathrm{Tr}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})}{2\sigma_\epsilon^2}\right]\right] \\
&\le \mathbb{E}_{\boldsymbol{X}}\left[\frac{\alpha N}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)\sum_{m=M+1}^{N}\lambda_m(\boldsymbol{K_{f,f}}) + 2Nv\epsilon]}{N}\right] + \right. \\
&\quad \left. \alpha\frac{(M+1)\sum_{m=M+1}^{N}\lambda_m(\boldsymbol{K_{f,f}}) + 2Nv\epsilon}{2\sigma_\epsilon^2}\right] \\
&\le \frac{\alpha N}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon]}{N}\right] + \\
&\quad \alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{2\sigma_\epsilon^2}.
\end{aligned}$$

By the Markov's inequality, we have the following bound with probability at least $1-\delta$ for any $\delta \in (0,1)$.

$$\begin{aligned}
D_\alpha[p||q] \le{}& \alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{2\delta\sigma_\epsilon^2} + \\
&\frac{1}{\delta}\frac{\alpha}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon]}{N}\right]^N.
\end{aligned}$$

□

As $\alpha \to 1$, we obtain the bound for the KL divergence.

Theorem 2. *Suppose N data points are drawn i.i.d from input distribution $p(\boldsymbol{x})$ and $k(\boldsymbol{x},\boldsymbol{x}) \le v, \forall \boldsymbol{x} \in \mathcal{X}$. Sample M inducing points from the training data with the probability assigned to any set of size M equal to the probability assigned to the corresponding subset by an ϵ k-Determinantal Point Process (k-DPP) [**?**] with $k = M$. With probability at least $1-\delta$,*

$$\begin{aligned}
D_\alpha[q||p] \le{}& \frac{1}{\delta}\frac{\alpha}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon]}{N}\right]^N + \\
&\alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{2\delta\sigma_\epsilon^2}\frac{\|\boldsymbol{Y}\|^2}{\sigma_\epsilon^2}
\end{aligned}$$

where $C = N\sum_{m=M+1}^{\infty}\lambda_m$ and λ_m are the eigenvalues of the integral operator $\mathcal{K}$ associated to kernel, k and $p(\boldsymbol{x})$.

Proof. Using lemma in appendix, we have

$$\begin{aligned}
D_\alpha[p||q] &\leq -\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{1}{2}\|\boldsymbol{Y}\|^2 \frac{\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})}{\sigma_\epsilon^4 + \alpha\sigma_\epsilon^2\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})} \\
&\leq -\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{1}{2}\frac{\|\boldsymbol{Y}\|^2}{\sigma_\epsilon^2} \frac{\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})}{\sigma_\epsilon^2 + \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})} \\
&\leq -\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \frac{1}{2}\frac{\|\boldsymbol{Y}\|^2}{\sigma_\epsilon^2} \frac{\alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})}{\sigma_\epsilon^2}.
\end{aligned}$$

Following the same argument in the proof of Theorem 1, we have

$$\frac{\alpha}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon]}{N}\right]^N + \\ \alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{2\sigma_\epsilon^2}\frac{\|\boldsymbol{Y}\|^2}{\sigma_\epsilon^2}.$$

□

As $\alpha \to 1$, we reach the bound for the KL divergence.

D.3 Smooth Kernel

Proof. We know $\frac{C(M+1)}{2\delta\sigma_\epsilon^2} < \frac{1}{N^{\gamma+1}}$. By Theorem 2, we can obtain the following bound

$$D_\alpha[p||q] \leq 2\alpha\frac{R}{\sigma_\epsilon^2}\frac{1}{N^\gamma} + \frac{1}{\delta}\frac{\alpha}{2(1-\alpha)}\log\left[1 + (1-\alpha)\Big(\frac{4\delta}{N^{\gamma+2}}\Big)\right]^N.$$

□

D.4 Non-smooth Kernel

For the Matérn $r + \frac{1}{2}$, $\lambda_m \asymp \frac{1}{m^{2r+2}}$ kernel, where $\asymp$ means "asymptotically equivalent to", we can obtain $\sum_{m=M+1}^{\infty}\lambda_m = \mathcal{O}(\frac{1}{M^{2r+1}})$. Let $\sum_{m=M+1}^{\infty}\lambda_m \leq A\frac{1}{M^{2r+1}}$. Then by Theorem 1, we have

$$\alpha \frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv_0\epsilon}{2\delta\sigma_\epsilon^2}\frac{\|\boldsymbol{Y}\|^2}{\sigma_\epsilon^2} \leq \alpha \frac{(M+1)NA\frac{1}{M^{2k+1}} + 2Nv_0\epsilon}{2\delta\sigma_\epsilon^2}\frac{RN}{\sigma_\epsilon^2}$$
$$= \frac{\alpha R}{2\delta\sigma_\epsilon^4}\Big(\frac{(M+1)N^2A}{M^{2r+1}} + 2N^2v_0\epsilon\Big).$$

In order to let $\lim_{N\to\infty}\frac{(M+1)N^2}{M^{2r+1}} \to 0$, we require $M = N^t$ (t will be clarified shortly). Therefore,

$$\frac{(M+1)N^2A}{M^{2r+1}} = \frac{(N^t+1)N^2A}{N^{(2r+1)t}} \leq \frac{A}{N^{2rt-2}}.$$

Let $2rt - 2 \geq \gamma$, then $t \geq \frac{\gamma+2}{2r}$. Therefore, we have

$$\frac{\alpha R}{2\sigma_\epsilon^4}\Big(\frac{(M+1)N^2A}{M^{2r+1}} + 2N^2v_0\epsilon\Big) \leq \frac{\alpha R}{N^\gamma\sigma_\epsilon^2} + \frac{\alpha RA}{2\delta\sigma_\epsilon^4N^\gamma}.$$

Another term in the bound can also be simplified as

$$\frac{\alpha}{2(1-\alpha)}\log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)C + 2Nv_0\epsilon]}{N}\right]^N \leq \frac{\alpha N}{2(1-\alpha)}\log\left[1 + (1-\alpha)\Big(\frac{A+2\delta}{\sigma_\epsilon^2N^{\gamma+2}}\Big)\right].$$

It can be seen that we require more inducing points ($\mathcal{O}(N^t)$) when we are using non-smooth kernels and t decreases as we increase the smoothness (i.e., r) of the Matérn kernel.

E More Results

We provide more detailed results in this section. In table 1, we report the negative loss (NL) of each method.

Table 1: NL of all models on many datasets. The NL is calculated over 30 experiments with different initial points.

Dataset	EGP	SGP	PEP	Rényi
Bike	0.41 ± 0.02	0.15 ± 0.03	0.10 ± 0.01	0.45 ± 0.01
C-MAPSS	-1.00 ± 0.01	-1.45 ± 0.01	-1.55 ± 0.02	-0.01 ± 0.01
Protein	2.15 ± 0.01	1.42 ± 0.01	1.97 ± 0.05	2.99 ± 0.04
Traffic	-0.42 ± 0.01	-0.47 ± 0.02	-1.07 ± 0.01	-0.20 ± 0.04
Battery	2.23 ± 0.06	2.10 ± 0.02	2.17 ± 0.02	2.55 ± 0.01

References

Alshraideh, H. and E. Khatatbeh (2014). A gaussian process control chart for monitoring autocorrelated process data. *Journal of Quality Technology 46*(4), 317–322.

Asuncion, A. and D. Newman (2007). Uci machine learning repository.

Bhattacharya, A., D. Pati, Y. Yang, et al. (2019). Bayesian fractional posteriors. *Annals of Statistics 47*(1), 39–66.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Bui, T., D. Hernández-Lobato, J. Hernandez-Lobato, Y. Li, and R. Turner (2016). Deep gaussian processes for regression using approximate expectation propagation. In *International conference on machine learning*, pp. 1472–1481.

Bui, T. D., J. Yan, and R. E. Turner (2017). A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. *The Journal of Machine Learning Research 18*(1), 3649–3720.

Chen, H., L. Zheng, R. Al Kontar, and G. Raskutti (2020). Stochastic gradient descent in correlated settings: A study on gaussian processes. *Neural Information Processing Systems*.

Currin, C., T. Mitchell, M. Morris, and D. Ylvisaker (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association 86*(416), 953–963.

Daley, R. (1993). *Atmospheric data analysis*. Number 2. Cambridge university press.

Damianou, A. and N. Lawrence (2013). Deep gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207–215.

Deisenroth, M. and S. Mohamed (2012). Expectation propagation in gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 2609–2617.

Dowsland, K. A. and J. Thompson (2012). Simulated annealing. *Handbook of natural computing*, 1623–1655.

Frigola, R., F. Lindsten, T. B. Schön, and C. E. Rasmussen (2013). Bayesian inference and learning in gaussian process state-space models with particle mcmc. In *Advances in Neural Information Processing Systems*, pp. 3156–3164.

Furrer, R., M. G. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics 15*(3), 502–523.

Gardner, J., G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, pp. 7576–7586.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. CRC Press.

Gramacy, R. B. and D. W. Apley (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics 24*(2), 561–578.

Gramacy, R. B. and B. Haaland (2016). Speeding up neighborhood search in local gaussian process prediction. *Technometrics 58*(3), 294–303.

Gramacy, R. B. and H. Lian (2012). Gaussian process single-index models as emulators for computer experiments. *Technometrics 54*(1), 30–41.

Grünwald, P. (2012). The safe bayesian. In *International Conference on Algorithmic Learning Theory*, pp. 169–183. Springer.

Guinness, J. (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics 60*(4), 415–429.

Havasi, M., J. M. Hernández-Lobato, and J. J. Murillo-Fuentes (2018). Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. In *Advances in Neural Information Processing Systems*, pp. 7506–7516.

Hensman, J., N. Fusi, and N. D. Lawrence (2013). Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*.

Hensman, J., A. G. Matthews, M. Filippone, and Z. Ghahramani (2015). Mcmc for variationally sparse gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 1648–1656.

Hoang, T. N., Q. M. Hoang, and B. K. H. Low (2015). A unifying framework of anytime sparse gaussian process regression models with stochastic variational inference for big data. In *ICML*, pp. 569–578.

Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley (2013). Stochastic variational inference. *The Journal of Machine Learning Research 14*(1), 1303–1347.

Jacot, A., F. Gabriel, and C. Hongler (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*.

Jones, B. and R. T. Johnson (2009). Design and analysis for the gaussian process model. *Quality and Reliability Engineering International 25*(5), 515–524.

Joseph, V. R., L. Gu, S. Ba, and W. R. Myers (2019). Space-filling designs for robustness experiments. *Technometrics 61*(1), 24–37.

Journel, A. G. and C. J. Huijbregts (1978). *Mining geostatistics*, Volume 600. Academic press London.

Kaufman, C. G., M. J. Schervish, and D. W. Nychka (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association 103*(484), 1545–1555.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464.

Krishna, A., V. R. Joseph, S. Ba, W. A. Brenneman, and W. R. Myers (2020). Robust experimental designs for model calibration. *arXiv preprint arXiv:2008.00547*.

Lalchand, V. and C. E. Rasmussen (2019). Approximate inference for fully bayesian gaussian process regression. *arXiv preprint arXiv:1912.13440*.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Liu, H., Y.-S. Ong, X. Shen, and J. Cai (2018). When gaussian process meets big data: A review of scalable gps. *arXiv preprint arXiv:1807.01065*.

Martinez-Cantin, R. (2014). Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits. *J. Mach. Learn. Res. 15*(1), 3735–3739.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in applied probability 5*(3), 439–468.

Matthews, A. G. d. G., M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani (2018). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Miller, J. W. and D. B. Dunson (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*.

Plumlee, M. (2019). Computer model calibration with confidence and consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(3), 519–545.

Plumlee, M., C. Erickson, B. Ankenman, and E. Lawrence (2020). Composite grid designs for adaptive computer experiments with fast inference. *Biometrika*.

Rana, S., C. Li, S. Gupta, V. Nguyen, and S. Venkatesh (2017). High dimensional bayesian optimization with elastic gaussian process. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2883–2891. JMLR. org.

Rényi, A. et al. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Ripley, B. D. (1981). *Spatial statistics*, Volume 575. John Wiley & Sons.

Rose, K., E. Gurewitz, and G. Fox (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters 11*(9), 589–594.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical science*, 409–423.

Schölkopf, B., A. J. Smola, F. Bach, et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Snelson, E. and Z. Ghahramani (2006). Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pp. 1257–1264.

Snoek, J., H. Larochelle, and R. P. Adams (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp. 2951–2959.

Srinivas, N., A. Krause, S. M. Kakade, and M. Seeger (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics 8*, 1–19.

Sung, C.-L., Y. Hung, W. Rittase, C. Zhu, and C. Jeff Wu (2020). A generalized gaussian process model for computer experiments with binary time series. *Journal of the American Statistical Association 115*(530), 945–956.

Takapoui, R. and H. Javadi (2016). Preconditioning via diagonal scaling. *arXiv preprint arXiv:1610.03871*.

Thompson, P. D. (1956). Optimum smoothing of two-dimensional fields 1. *Tellus 8*(3), 384–393.

Titsias, M. and N. D. Lawrence (2010). Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 844–851.

Titsias, M. K. (2014). Variational inference for gaussian and determinantal point processes.

Tran, D., R. Ranganath, and D. M. Blei (2015). The variational gaussian process. *arXiv preprint arXiv:1511.06499*.

Tuo, R. and W. Wang (2020). Kriging prediction with isotropic matern correlations: robustness and experimental designs. *Journal of Machine Learning Research 21*(187), 1–38.

Wang, K. A., G. Pleiss, J. R. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson (2019). Exact gaussian processes on a million data points. *arXiv preprint arXiv:1903.08114*.

Wang, W., R. Tuo, and C. Jeff Wu (2019). On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 1–27.

Wei, P., F. Liu, and C. Tang (2018). Reliability and reliability-based importance analysis of structural systems using multiple response gaussian process model. *Reliability Engineering & System Safety 175*, 183–195.

Yang, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*.

Yue, X. and R. A. Kontar (2020a). Joint models for event prediction from time series and survival data. *Technometrics*, 1–10.

Yue, X. and R. A. Kontar (2020b). Why non-myopic bayesian optimization is promising and how far should we look-ahead? a study via rollout. In *International Conference on Artificial Intelligence and Statistics*, pp. 2808–2818. PMLR.

Zhang, Q., P. Chien, Q. Liu, L. Xu, and Y. Hong (2021). Mixed-input gaussian process emulators for computer experiments with a large number of categorical levels. *Journal of Quality Technology 53*(4), 410–420.