Appendix

Xubo Yue, Raed Al Kontar

1 Proof of Claims in Sec. 2.1

Claim 1. $\lim_{\alpha\to 1} D_\alpha[p||q] = KL[p||q]$.

Proof. Applying the L'Hopital rule, we have

$$\begin{aligned}
\lim_{\alpha\to 1} D_\alpha[p||q] &= \lim_{\alpha\to 1} \frac{1}{\alpha-1}\log\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta} \\
&= \lim_{\alpha\to 1} \frac{1}{\frac{d}{d\alpha}(\alpha-1)} \frac{d}{d\alpha}\log\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta} \\
&= \lim_{\alpha\to 1} \frac{d}{d\alpha}\log\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta}
\end{aligned}$$

By the Leibniz's rule, we have

$$\begin{aligned}
\lim_{\alpha\to 1} D_\alpha[p||q] &= \lim_{\alpha\to 1} \frac{d}{d\alpha}\log\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta} \\
&= \lim_{\alpha\to 1} \frac{\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha}[\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta}} \\
&= \frac{\int p(\boldsymbol{\theta})[\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\
&= \int p(\boldsymbol{\theta}) \log\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= KL[p||q].
\end{aligned}$$

$\square$

Claim 2. $\mathcal{L}_0(q;\boldsymbol{y}) = \log P(\boldsymbol{y})$.

Proof. This is trivial, just let $\alpha = 0$. $\square$

Claim 3. $\mathcal{L}_{VI} = \lim_{\alpha\to 1}\mathcal{L}_\alpha(q;\boldsymbol{y}) \le \mathcal{L}_{\alpha_+}(q;\boldsymbol{y}) \le \log P(\boldsymbol{y}) \le \mathcal{L}_{\alpha-}(q;\boldsymbol{y}), \forall \alpha_+ \in (0,1), \alpha_- < 0$.

Proof. The first equality follows from the Claim 1. The left inequality can be obtained by the Jensen inequality. $\square$

2 The Variational Rényi Lower Bound

When we apply the VR bound to $\mathcal{GP}$ and assume that $q(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{\mathcal{Z}}) = p(\boldsymbol{f}|\boldsymbol{U}, \boldsymbol{\mathcal{Z}})q(\boldsymbol{U})$, we can further obtain

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

It has been shown that $p(\boldsymbol{f}|\boldsymbol{U}, \boldsymbol{\mathcal{Z}}) = \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{U}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$, 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 $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}, \sigma_\epsilon^2 I)$. Therefore,

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

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. This can be achieved by the special case of the Hölder inequality (i.e., Lyapunov inequality).

Then we have,

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

The optimal $q(\boldsymbol{U})$ is

$$
q^*(\boldsymbol{U}) \propto p(\boldsymbol{y}|\boldsymbol{U}, \boldsymbol{\mathcal{Z}})^{1/(1-\alpha)} p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}}).
$$

Specifically,

$$
q^*(\boldsymbol{U}) = \frac{p(\boldsymbol{y}|\boldsymbol{U}, \boldsymbol{\mathcal{Z}})^{1/(1-\alpha)} p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}})}{\int p(\boldsymbol{y}|\boldsymbol{U}, \boldsymbol{\mathcal{Z}})^{1/(1-\alpha)} p(\boldsymbol{U}|\boldsymbol{\mathcal{Z}}) d\boldsymbol{U}}.
$$

It can be shown that

$$
\begin{aligned}
&p(\boldsymbol{y}|\boldsymbol{U}, \boldsymbol{\mathcal{Z}})^{\frac{1}{1-\alpha}} \\
&= [(2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2}} (\frac{1}{1-\alpha})^{\frac{N}{2}}]^{\frac{1}{1-\alpha}} \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}} \boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1} \boldsymbol{U}, \frac{\sigma_\epsilon^2}{1-\alpha} I + \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})^{\frac{1}{1-\alpha}} \\
&= [(2\pi\sigma_\epsilon^2)^{\frac{\alpha N}{2(1-\alpha)}} (\frac{1}{1-\alpha})^{\frac{N}{2(1-\alpha)}}] C \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}} \boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1} \boldsymbol{U}, \sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}]),
\end{aligned}
$$

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

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

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". Therefore, the lower bound can be expressed as

$$
\begin{aligned}
&\mathcal{L}_\alpha(q;\boldsymbol{y}) \\
&\approx \log\mathcal{N}(\boldsymbol{0},\sigma_\epsilon^2 I+(1-\alpha)[\boldsymbol{K_{f,f}}]+\alpha\boldsymbol{Q})+\log\left\{1+\frac{1-\alpha}{\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K_{f,f}}-\boldsymbol{Q})+\mathcal{O}(\frac{(1-\alpha)^2}{\sigma_\epsilon^4})\right\}^{\frac{-\alpha}{2(1-\alpha)}},
\end{aligned}
$$

given that $\alpha \in (0,1)$. While this form is attractive, it is not practically useful since when $1-\alpha$ is "large", the approximation does not work well. In the analysis section, we will instead use $|\boldsymbol{I}+\frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}}-\boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}$ to prove the convergence result.

3 The Data-dependent Upper Bound

Lemma 4. *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 $B^{-1} \geq A^{-1}$.*

This lemma has been proved in (Horn and Johnson, 2012). Based on this lemma, we can compute a data-dependent upper bound on the log-marginal likelihood (Titsias, 2014).

Claim 5. $\log p(\boldsymbol{y}) \leq \log \frac{1}{|2\pi((1-\alpha)\boldsymbol{K_{f,f}}+\alpha\boldsymbol{Q}+\sigma_\epsilon^2\boldsymbol{I})|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{y}^T((1-\alpha)\boldsymbol{K_{f,f}}+\alpha\boldsymbol{Q}+\alpha\, Tr(\boldsymbol{K_{f,f}}-\boldsymbol{Q})\boldsymbol{I}+\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{y}} := \mathcal{L}_{upper}.$

Proof. Since

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

where $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} \geq \boldsymbol{0}, \forall\boldsymbol{x}$. Then, we can obtain $|\boldsymbol{K_{f,f}}+\sigma_\epsilon^2\boldsymbol{I}| \geq |(1-\alpha)\boldsymbol{K_{f,f}}+$

$\alpha\boldsymbol{Q} + \sigma_\epsilon^2 \boldsymbol{I}|$ 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

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

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 $\boldsymbol{y}^T(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{y} \leq \mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\left\|\boldsymbol{y}\right\|^2 = \mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{y}^T\boldsymbol{y}$. 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

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

□

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

4 Detailed Proof of Convergence Result

Let $q \coloneqq q(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{\mathcal{Z}})$ and $p \coloneqq p(\boldsymbol{f}, \boldsymbol{U}, \boldsymbol{y}|\boldsymbol{\mathcal{Z}})$.

Claim 6. $-\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} \leq \frac{\alpha}{2(1-\alpha)} \log\left(\frac{Tr(\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}))}{N}\right)^N.$

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 7. *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*

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

Proof. We have

$$\begin{aligned} &\mathbb{E}_y\Big[\mathrm{VR}[q||p]\Big] \\ &= \mathbb{E}_y\Big[\log p(\boldsymbol{y}) - \log\mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q}) - \log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}}\Big] \\ &= -\log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} + \mathbb{E}_y\Big[\log\frac{\mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I})}{\mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{I} + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q})}\Big]. \end{aligned}$$

It is apparent that the lower bound to (1) is

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

since the KL divergence is non-negative. We then provide an upper bound to (1). We have

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

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

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

where $\tilde{\boldsymbol{K}}=\boldsymbol{K_{f,f}}+\sigma_\epsilon^2\boldsymbol{I}-(\sigma_\epsilon^2\boldsymbol{I}+(1-\alpha)\boldsymbol{K_{f,f}}+\alpha\boldsymbol{Q})$ 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.

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

□

As $\alpha\to 1$, we recover the bounds for the KL divergence. Specifically, we get the lower bound $\frac{\mathrm{Tr}(\boldsymbol{K_{f,f}}-\boldsymbol{Q})}{2\sigma_\epsilon^2}$ and upper bound $\frac{\mathrm{Tr}(\boldsymbol{K_{f,f}}-\boldsymbol{Q})}{\sigma_\epsilon^2}$ (Burt et al., 2019).

Lemma 8. *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*

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

This lemma is proved in (Belabbas and Wolfe, 2009). Following this lemma and by Lemma 6, we can show that

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

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 9.

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

This inequality is from (Burt et al., 2019). Using this fact, we can show that

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

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

Lemma 10. *Then,*

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

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

Proof. Based on Claim 5, we have

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

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

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

Let $(1-\alpha)\boldsymbol{K_{f,f}} + \alpha\boldsymbol{Q} + \sigma_\epsilon^2\boldsymbol{I} = \boldsymbol{V}\boldsymbol{\Lambda_\alpha}\boldsymbol{V}^T$ be the eigenvalue decomposition and denote by $\gamma_1 \geq \ldots \geq \gamma_N$ all eigenvalues. Then we can obtain

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

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

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

□

Theorem 11. *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) (Belabbas and Wolfe, 2009) with $k = M$. If $\boldsymbol{y}$ is distributed according to a sample from the prior generative model, with probability at least $1-\delta$,*

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

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

Proof. We have

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

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

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

□

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

Theorem 12. *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) (Belabbas and Wolfe, 2009) with $k = M$. With probability at least $1 - \delta$,*

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

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

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

Following the same argument in the proof of Theorem 11, 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.

5 Other Properties

5.1 Generalization

The VR bound encompasses wide ranges of bounds. For example, the lower bound encapsulates the popular ELBO.

Claim 13. $\lim_{\alpha\to 1}\mathcal{L}_\alpha(q;\boldsymbol{y}) = \log\mathcal{N}(\mathbf{0},\sigma_\epsilon^2\boldsymbol{I}+\boldsymbol{Q}) - \frac{1}{2\sigma_\epsilon^2}Tr(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}) = \mathcal{L}_{VI}$.

Proof. It is easy to see that

$$\lim_{\alpha\to 1}\log\mathcal{N}(\mathbf{0},\sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}] + \alpha\boldsymbol{Q}) = \log\mathcal{N}(\mathbf{0},\sigma_\epsilon^2\boldsymbol{I}+\boldsymbol{Q}),$$
$$\lim_{\alpha\to 1}\log\left\{1+\frac{1-\alpha}{\sigma_\epsilon^2}Tr(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}) + \mathcal{O}(\frac{(1-\alpha)^2}{\sigma_\epsilon^4})\right\}^{\frac{-\alpha}{2(1-\alpha)}} = \log e^{-\frac{1}{2\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q})}.$$

Therefore, we have

$$\lim_{\alpha\to 1}\mathcal{L}_\alpha(q;\boldsymbol{y}) = \log\mathcal{N}(\mathbf{0},\sigma_\epsilon^2\boldsymbol{I}+\boldsymbol{Q}) - \frac{1}{2\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}-\boldsymbol{Q}) = \mathcal{L}_{VI}.$$

□

References

Belabbas, M.-A. and Wolfe, P. J. (2009). Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374.

Burt, D. R., Rasmussen, C. E., and Van Der Wilk, M. (2019). Rates of convergence for sparse variational gaussian process regression. *arXiv preprint arXiv:1903.03571*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

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