An Alternative Gaussian Process Objective Based on the Rényi Divergence

Xubo Yue

Industrial and Operations Engineering, University of Michigan, Ann Arbor

and

Raed Al Kontar

Industrial and Operations Engineering, University of Michigan, Ann Arbor

July 27, 2021

Abstract

We introduce an alternative closed-form objective function for parameter estimation in the Gaussian process ($\mathcal{GP}$) based on the Rényi α-divergence. This objective offers a structured and tunable balance between model-fit and prior regularization and therefore is capable of controlling the enforced regularization on the objective. We then show that the Rényi divergence from the true $\mathcal{GP}$ posterior can be made arbitrarily small under the proposed objective and derive convergence rates for a class of smooth and non-smooth kernels. Experiments on a wide range of real-life engineering applications show that the proposed objective is promising and can deliver significant improvement over several state-of-the-art $\mathcal{GP}$ approaches.

Keywords: Gaussian process, Rényi α-divergence, variational inference, real-life applications

1 Introduction

The Gaussian process ($\mathcal{GP}$, also known as kriging) is a collection of random variables, any finite number of which has a joint Gaussian distribution [Sacks et al., 1989, Currin et al., 1991]. It is widely used to reconstruct functions based on their scattered observations. In literature, $\mathcal{GP}$s were originally used to tackle regression problems in meteorology [Thompson, 1956, Daley, 1993], geostatistics [Matheron, 1973, Journel and Huijbregts, 1978] and spatial statistics [Ripley, 1981].

Over the past two decades, $\mathcal{GP}$ theory and its application has seen great success in various statistics areas. These include experimental design [Krishna et al., 2020, Gramacy and Apley, 2015, Joseph et al., 2019], Bayesian optimization [Snoek et al., 2012, Rana et al., 2017, Wang et al., 2019b], computer experiments and calibration [Kennedy and O'Hagan, 2001, Plumlee et al., 2020, Plumlee, 2019, Sung et al., 2020, Gramacy, 2020], reliability [Wei et al., 2018], reinforcement learning and bandits [Srinivas et al., 2009] and recently deep learning [Damianou and Lawrence, 2013, Bui et al., 2016, Matthews et al., 2018]. Indeed, this success is due to the many desirable properties $\mathcal{GP}$s possess, such as their uncertainty quantification capability and highly flexible model priors where prior knowledge can often be readily accommodated in the mean and covariance function. This progress was also observed on a theoretical level. Matthews et al. [2018] prove that a fully connected, feedforward network will converge to a $\mathcal{GP}$ as the network width goes to infinity. This exciting work has brought upon many insightful connections between $\mathcal{GP}$s and deep neural networks [Jacot et al., 2018, Yang, 2019]. Chen et al. [2020] show that mini-batch stochastic gradient descent can be applied in correlated settings, specifically within a $\mathcal{GP}$; a result that allowed scaling $\mathcal{GP}$s far beyond what is currently possible. Wang et al. [2019b] derived uniform error bounds for $\mathcal{GP}$s trained using a Matérn kernel. These bounds were then used to find generalization bounds for Bayesian optimization and sequential experimental design [Martinez-Cantin, 2014, Yue and Kontar, 2020b, Tuo and Wang, 2020].

In this paper, we focus on parameter estimation within $\mathcal{GP}$s through defining an objective for obtaining parameter estimates. Hereby, we start by presenting our main result. Detailed notation will be further highlighted in the coming sections. More specifically, we introduce an alternative objective for a $\mathcal{GP}$ that aims at minimizing the Rényi α-divergence between the true and approximated posterior. Let $\phi(\boldsymbol{Y}|\cdot,\cdot)$ denote a multivariate Gaussian density for the set of observations $\boldsymbol{Y}$ and $|\cdot|$ a determinant operator. The alternative objective is given as

$$\mathcal{L}_\alpha(q^*) = \log\{\phi(\boldsymbol{Y}|\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q})\} + \log|\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}},$$

where $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$ is the full covariance matrix, σ_ϵ^2 is the noise parameter, $\boldsymbol{Q} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}}$ is a Nyström low-rank approximation of the exact covariance matrix $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$ and $\boldsymbol{U}$ is a collection of latent variables. Here $\alpha \in [0,1)$ is a tuning parameter.

Our proposed objective is a lower bound on the commonly used marginal likelihood function, given as $\log p(\boldsymbol{Y}|\boldsymbol{X})$ and contains a rich family of $\mathcal{GP}$ inference models including variational inference (VI). This bound offers a structured and tunable balance between model-fit and prior regularization and therefore is capable of controlling the enforced regularization on the objective function.

From a theoretical aspect, we show that the Rényi divergence from the true $\mathcal{GP}$ posterior can be made arbitrarily small and derive convergence rates of our bound under a smooth and non-smooth kernel (refer to Sec. 7). More importantly, we illustrate the superior performance of our proposed objective over state-of-the-art $\mathcal{GP}$ approaches on a wide variety of engineering applications and datasets.

1.1 Organization

We organize the paper as follows. In Sec. 2, we conduct a detailed literature review. In Sec. 3, we briefly review related background knowledge. We then provide the Rényi variational objective for $\mathcal{GP}$s in Sec. 4, and its underlying motivation in Sec. 5. The optimization algorithm and predictive distribution are presented in Sec. 6. Theoretical properties of our objective are investigated in Sec. 7. In Sec. 8 we provide numerical experiments over a range of engineering applications to demonstrate the advantages of our method. Our experiments include a traffic, battery, house electric, bike and turbofan engine dataset. We conclude our paper in Sec. 9. We note that we defer most derivations to the appendix and only highlight the main results.

2 Related Work

Though it is by no means an exhaustive list, recent advances in model estimation for Gaussian processes can be roughly split into five main trends (for further details see the recent survey in Liu et al. [2018]). **First**, sampling methods such as Markov chain Monte Carlo (MCMC) [Gramacy and Lian, 2012, Frigola et al., 2013, Hensman et al., 2015] and Hamiltonian Monte Carlo [Havasi et al., 2018] have been extensively studied. However, a sampling approximation is usually computationally intensive. Notably, a recent comparison study [Lalchand and Rasmussen, 2019] shows that variational inference (VI) can achieve remarkable performance compared to sampling approaches while the former has better theoretical properties and can be fitted into many existing efficient optimization frameworks. **Second**, expectation propagation (EP) [Deisenroth and Mohamed, 2012] is an iterative local message passing method designed for approximate Bayesian inference. Based on this approach, Bui et al. [2017] propose the power EP (PEP) framework to learn $\mathcal{GP}$s and demonstrate that PEP encapsulates a rich family of approximated $\mathcal{GP}$s such as FITC and DTC [Bui et al., 2017]. Though accurate and promising, the EP family, in general, is not guaranteed to converge [Bishop, 2006]. **Third**, variational inference is an approach to estimate probability densities through efficient optimization algorithms [Hoffman et al., 2013, Hoang et al., 2015, Blei et al., 2017]. It approximates intractable posterior distributions using a tractable distributional family $\mathcal{Q}$. This approximation in turn yields a lower bound that is optimized to learn model parameters. VI has caught the most attention compared to the other approximate inference algorithms due to ease of use and elegant theoretical properties. **Fourth**, there has been a recent push on utilizing GPU acceleration and distributed computing to optimize the log-marginal likelihood in $\mathcal{GP}$s. Such approaches leverage Blackbox Matrix-Matrix multiplication, distributed Cholesky factorization and kernel partitioning [Gardner et al., 2018a,b, Wang et al., 2019a]. **Lastly**, some literature consider low rank or sparse approximation techniques [Gramacy and Haaland, 2016] and covariance tapering [Furrer et al., 2006, Kaufman et al., 2008]. Besides those trends, some notable work also approximate Gaussian process using Vecchia's approximation method [Guinness, 2018] and stochastic partial differential equation approximation methods

[Lindgren et al., 2011].

3 Notation and Brief Review

We start by introducing some notations and briefly review the Gaussian process. Assume we have collected N training data points $\boldsymbol{Y} = [y_i]_{i=1}^N$ with corresponding D-dimensional inputs $\boldsymbol{X} = [\boldsymbol{x}_i]_{i=1}^N$, where $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^D$. We decompose the output as $y_i = f(\boldsymbol{x}_i) + \epsilon_i$, where $f(\cdot)$ is a $\mathcal{GP}$ and $\epsilon_i(\cdot)$ denotes additive noise with zero mean and σ_ϵ^2 variance. The $\mathcal{GP}$ places a prior over functions such that $p(\boldsymbol{f}|\boldsymbol{X}) = \phi(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}})$ where $\boldsymbol{f} = [f_1, ..., f_N]$ is a vector of latent function values $f_i = f(\boldsymbol{x}_i)$ and $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} \coloneqq \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ is a covariance matrix whose entries are determined by a covariance function $k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$ parameterized through $\boldsymbol{\theta}$. Here we note that for notational simplicity we assume zero mean $\mathcal{GP}$ and hereon we neglect conditioning on the input.

Often the end goal of a $\mathcal{GP}$ is to predict output $\boldsymbol{f}^*$ given new inputs $\boldsymbol{X}^*$. To do so, the predictive distribution is attained via $p(\boldsymbol{f}^*|\boldsymbol{Y}) = \int p(\boldsymbol{f}^*|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{Y})d\boldsymbol{f}$ and is given as

$$\boldsymbol{f}^*|\boldsymbol{Y} \sim \mathcal{N}\Big(\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}}[\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I}]^{-1}\boldsymbol{Y}, \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*} - \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}}[\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \sigma_\epsilon^2\boldsymbol{I}]^{-1}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}^*}\Big).$$

Here $p(\boldsymbol{f}^*|\boldsymbol{f})$ is the conditional prior derived from the $\mathcal{GP}$ prior $\boldsymbol{f}, \boldsymbol{f}^* \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} & \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}^*} \\ \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}} & \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*} \end{pmatrix}\right)$ and $p(\boldsymbol{f}|\boldsymbol{Y})$ is the posterior of $\boldsymbol{f}$.

Given the predictive distribution, it is clear that good parameter estimation of $(\sigma_\epsilon, \boldsymbol{\theta})$ is imperative to $\mathcal{GP}$s. Perhaps the most popular approach for parameter estimation is by directly maximizing the well-known marginal log-likelihood function $p(\boldsymbol{Y}) = \int p(\boldsymbol{Y}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f}$. Given that $\boldsymbol{Y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma_\epsilon^2\boldsymbol{I})$ the log-marginal likelihood can be written as

$$\mathcal{L}_{marginal} \coloneqq \log p(\boldsymbol{Y}) = \log \phi(\boldsymbol{Y}|\boldsymbol{0}, \sigma_\epsilon^2 I + \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}). \tag{1}$$

However, under most well developed covariance functions, $\mathcal{L}_{marginal}$ is highly non-linear and non-convex. As a result, $\mathcal{GP}$s are vulnerable to obtaining parameter estimates with bad generalization power. For instance, it is not uncommon to have critical points in the objective that interpret data as pure noise. The goal of this work is to provide an alternative objective that encourages parameter estimates with improved generalization power to new data.

4 An Alternative $\mathcal{GP}$ Objective

Instead of maximizing the marginal likelihood we offer an alternative objective that features a trade-off between model fit and prior regularization and allows fine-tuning this trade-off. We start by introducing this alternative objective and then discuss its advantages and motivation.

We follow the general philosophy of variational inference which turns inference into an optimization problem, where an optimal density (q^*), relative to some distance measure, is chosen from a distributional family ($\mathcal{Q}$) to approximate a target distribution - here the posterior of a $\mathcal{GP}$. To do so, we augment our probability space by M continuous latent variables $\boldsymbol{U} = [u(\boldsymbol{z}_i)]_{i=1}^M$ observed at inputs $\boldsymbol{\mathcal{Z}} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_M]^T$. We assume that $\boldsymbol{U}$ are drawn from the same $\mathcal{GP}$ prior $p(\boldsymbol{f})$. $\boldsymbol{\mathcal{Z}}$ may be a subset of the input ($\boldsymbol{X}$) or some free parameters, often referred to as pseudo-inputs or inducing points [Snelson and Ghahramani, 2006], to be optimized over.

Notice that from the augmented joint model $p(\boldsymbol{Y}, \boldsymbol{f}, \boldsymbol{U})$ we still reach the same marginal likelihood in (1) through marginalization $p(\boldsymbol{Y}) = \int p(\boldsymbol{Y}, \boldsymbol{f}, \boldsymbol{U}) d\boldsymbol{f} d\boldsymbol{U} = \int p(\boldsymbol{Y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{U}) p(\boldsymbol{U}) d\boldsymbol{f} d\boldsymbol{U}$ where $p(\boldsymbol{f}|\boldsymbol{U}) = \phi(\boldsymbol{f}|\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}} \boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1} \boldsymbol{U}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$ and $\boldsymbol{Q} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{U}} \boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1} \boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}})$. This hints to the fact that one may let $p(\boldsymbol{U})$ be a distribution that adds some level of flexibility to the model.

4.1 The α-ELBO

Exploiting this added flexibility we now take a variational route to approximate the joint posterior over the latent variables $p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})$. We use the Rényi's α-divergence as a distance measure. This in turn will lead to an our proposed objective.

The Rényi's α-divergence, first proposed in [Rényi et al., 1961], is a distance measure between two probability density functions (p and q) of a continuous random variable.

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

This divergence contains a rich family of distance measures such as KL-divergence, Bhattacharyya coefficient and χ^2-divergence. Also, $D_\alpha[q||p]$ is continuous and non-decreasing on $\alpha \in [0,1)$.

In the context of $\mathcal{GP}$s, our goal is to find an optimal posterior density $q^\star$ over the latent variables $\boldsymbol{f}, \boldsymbol{U}$ belonging to some distributional family $\mathcal{Q}$, by minimizing the Rényi α-divergence between the variational density $q(\boldsymbol{f}, \boldsymbol{U})$ and the target posterior $p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})$

$$q^*(\boldsymbol{f}, \boldsymbol{U}) \coloneqq \operatorname*{arg\,min}_{q(\boldsymbol{f},\boldsymbol{U})\in\mathcal{Q}} D_\alpha[q(\boldsymbol{f}, \boldsymbol{U})||p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})].$$

Through some algebraic manipulations (Appendix A.1), one can find that

$$D_\alpha[q(\boldsymbol{f}, \boldsymbol{U})||p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})] = \log p(\boldsymbol{Y}) - \mathcal{L}_\alpha(q)\,. \tag{2}$$

Such that

$$\mathcal{L}_\alpha(q) \coloneqq \frac{1}{1-\alpha}\log \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{U})}\left[\left(\frac{p(\boldsymbol{f}, \boldsymbol{U}, \boldsymbol{Y})}{q(\boldsymbol{f}, \boldsymbol{U})}\right)^{1-\alpha}\right]. \tag{3}$$

Following (3) and since $D_\alpha[q||p] \geq 0$, we have that $\mathcal{L}_\alpha(q) \leq \mathcal{L}_{marginal}$ is a lower bound on the log-marginal likelihood and maximizing $\mathcal{L}_\alpha(q)$ will equivalently minimize $D_\alpha[q||p]$.

In order to maximize, $\mathcal{L}_\alpha(q)$, one first needs to find the optimal density. To this end, we exploit a mean-field assumption $q(\boldsymbol{f}, \boldsymbol{U}) = p(\boldsymbol{f}|\boldsymbol{U})q(\boldsymbol{U})$. This in turn poses $q(\boldsymbol{U})$ as the variational density to be optimized.

Under this mean-field assumption (See Appendix A.2),

$$\begin{aligned}\mathcal{L}_\alpha(q) &= \frac{1}{1-\alpha}\log \int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})q(\boldsymbol{U})^\alpha p(\boldsymbol{U})^{1-\alpha} d\boldsymbol{U} \\ &= \frac{1}{1-\alpha}\log \mathbb{E}_{q(\boldsymbol{U})} p_\alpha(\boldsymbol{Y}|\boldsymbol{U})q(\boldsymbol{U})^{\alpha-1} p(\boldsymbol{U})^{1-\alpha}\,,\end{aligned} \tag{4}$$

where $p_\alpha(\boldsymbol{Y}|\boldsymbol{U}) = \int p(\boldsymbol{f}|\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha} d\boldsymbol{f}$. We slightly abuse notation as $p_\alpha(\boldsymbol{Y}|\boldsymbol{U})$ is not a probability density anymore. Our goal next goal is to find

$$q^*(\boldsymbol{U}) \coloneqq \operatorname*{arg\,min}_{q(\boldsymbol{U})} D_\alpha[p(\boldsymbol{f}|\boldsymbol{U})q(\boldsymbol{U})||p(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{Y})] = \operatorname*{arg\,max}_{q(\boldsymbol{U})} \mathcal{L}_\alpha(q)$$

Fortunately this can be optimally solved in closed form (See Appendix A.3)

$$q^*(\boldsymbol{U}) \propto p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}) \tag{5}$$

Now, plugging in $q^*(\boldsymbol{U})$ to (4), we reach our final result, $\mathcal{L}_\alpha(q^*)$ that denotes the lower bound under the optimal q and is given as (see Appendix A.4)

$$\begin{aligned}\mathcal{L}_\alpha(q^*) &= \log \int p_\alpha(\boldsymbol{Y}|\boldsymbol{U})^{1/(1-\alpha)} p(\boldsymbol{U}) d\boldsymbol{U} = \\ &\log \left\{ \mathcal{N}\Big(\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K_{f,f}} + \alpha \boldsymbol{Q}\Big) \right\} + \log \Big| \boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q}) \Big|^{\frac{-\alpha}{2(1-\alpha)}}.\end{aligned} \tag{6}$$

Through scrutinizing (6) we directly observe that the new lower bound unifies components from the exact covariance ($\boldsymbol{K_{f,f}}$) and an approximate covariance ($\boldsymbol{Q} = \boldsymbol{K_{f,U}} \boldsymbol{K_{U,U}^{-1}} \boldsymbol{K_{U,f}}$), also known as the Nyström approximation. Interestingly, when $\alpha = 0$, the marginal log-likelihood is recovered, $\mathcal{L}_0 = \mathcal{L}_{marginal} = \log p(\boldsymbol{Y})$. While, for $\alpha \to 1$, we recover the traditional VI bound obtained from maximizing a KL divergence distance measure. This indeed is a direct consequence of the fact that $\lim_{\alpha \to 1} D_\alpha[p||q] = KL[p||q]$ [Titsias and Lawrence, 2010, Tran et al., 2015, Liu et al., 2018, Yue and Kontar, 2020a].

By maximizing (6) with respect to $(\sigma_\epsilon, \boldsymbol{\theta})$ and possibly $\boldsymbol{\mathcal{Z}}$ one can obtain a set of parameters that minimize the Rényi's α-divergence with the true posterior. Hereon we refer to our bound as the Rényi $\mathcal{GP}$ or the α-ELBO as it is an evidence lower bound (ELBO) on the marginal log-likelihood.

5 Why use $\mathcal{L}_\alpha$

In this section we shed light on the advantages of our alternative bound from two related perspectives: (1) trade-off between model fit and fidelity to the prior class; (2) fractional posteriors which raise the likelihood to some power.

To understand why $\mathcal{L}_\alpha(q^*)$ is a promising objective, we first define the lower bounds below. Here $\mathcal{L}_{VI} = \lim_{\alpha \to 1} \mathcal{L}_\alpha(q)$ while $\mathcal{L}_{Jensen}$ is obtained from a direct Jensen's inequality on $\mathcal{L}_\alpha(q)$ (see appendix B).

$$\mathcal{L}_\alpha(q) = \frac{1}{1-\alpha}\log \mathbb{E}_{q(\boldsymbol{U})}\Big[p_\alpha(\boldsymbol{Y}|\boldsymbol{U})q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U})^{1-\alpha}\Big]$$

$$\mathcal{L}_{Jensen} = \underbrace{\frac{1}{1-\alpha}\int q(\boldsymbol{U})\log p_\alpha(\boldsymbol{Y}|\boldsymbol{U})d\boldsymbol{U}}_{\text{Model fit}}\underbrace{-KL[q(\boldsymbol{U})||p(\boldsymbol{U})]}_{\text{Prior regularization}}$$

$$\mathcal{L}_{VI} = \lim_{\alpha\to 1}\mathcal{L}_\alpha(q) = \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{U})}\left[\log p(\boldsymbol{Y}|\boldsymbol{f})\right] - KL[q(\boldsymbol{U})||p(\boldsymbol{U})]$$

where $\mathcal{L}_{marginal} \geq \mathcal{L}_\alpha(q) \geq \mathcal{L}_{Jensen} \geq \mathcal{L}_{VI}$ holds true for any $\alpha \in [0,1)$. We focus on $\mathcal{L}_\alpha(q)$ as $\mathcal{L}_\alpha(q^*)$ is a by-product from optimizing $\mathcal{L}_\alpha(q)$. Note that $KL[q(\boldsymbol{U})||p(\boldsymbol{U})] = KL[q(\boldsymbol{U},\boldsymbol{f})||p(\boldsymbol{U},\boldsymbol{f})]$ since $q(\boldsymbol{f},\boldsymbol{U}) = p(\boldsymbol{f}|\boldsymbol{U})q(\boldsymbol{U})$. Also, $p(\boldsymbol{Y}|\boldsymbol{f}) = p(\boldsymbol{Y}|\boldsymbol{f},\boldsymbol{U})$.

One can directly observe that the bounds mirror the trade-off between the likelihood and prior. For instance, in $\mathcal{L}_{Jensen}$, the first term denotes the model fit and encourages the density of the latent variables to place probability mass on configurations that best explain the observed data; this often induces a rugged objective with many local critical points each with a specific interpretation of the data. Whereas the KL term is a regularizer that encourages latent variables close to the prior class. Intrinsically this regularization restricts the complexity of the estimated posterior density and hence offers a trade-off between the fit and complexity of the latent variable estimators. A good trade-off is critical for generalization to new data [Schölkopf et al., 2002]. Here α plays the role of tuning this enforced regularization and is data dependent. This in turn allows data to speak for themselves. Note that in a $\mathcal{GP}$, the prior is imposed via the kernel. Ultimately, prior regularization is encouraging kernel hyper-parameters that satisfy both the prior class while at the meantime suiting the observed data.

Indeed, $\mathcal{GP}$ literature has shown that inference via $\mathcal{L}_{marginal}$ can be advantageous on some datasets and VI on others [Lalchand and Rasmussen, 2019, Wang et al., 2019a, Chen et al., 2020, Wang et al., 2019a]. For instance, Rainforth et al. [2018] question the implicit assumptions that a tighter bound on the marginal likelihood are better objectives and show that tighter bounds can be detrimental to the process of learning. This literature also sheds light on the benefits of $\mathcal{L}_\alpha(q)$ and the need to tune prior regularization much like tuning the regularization parameter in ridge regression or LASSO.

Another insight on the advantages of $\mathcal{L}_\alpha(q)$ is through fractional posteriors (often referred to as tempered or inexact posteriors) where a likelihood is raised to a some fractional power. Indeed, fractional posteriors have gained renewed interest in recent years within Bayesian statistics due to their empirical success in improving generalization and their robustness to model mis-specifications [Bhattacharya et al., 2019, Miller and Dunson, 2018, Grünwald, 2012]. For instance, Miller and Dunson [2018] show that raising likelihood to a well-chosen power induces robustness to a mismatch between the model used and the true generating data process. Specifically, under specific regularity conditions, they show that fractional likelihoods are asymptotically equivalent to conditioning on having the empirical distribution of the observed data close to the empirical distribution of data sampled from the model, with respect to a relative entropy distance.

In our context, using (5), our posterior over the latent variables is given as

$$q^*(\boldsymbol{f}, \boldsymbol{U}) = p(\boldsymbol{f}|\boldsymbol{U})q^*(\boldsymbol{U}) \propto p(\boldsymbol{f}, \boldsymbol{U})\Big[\int p(\boldsymbol{f}|\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}d\boldsymbol{f}\Big]^{\frac{1}{1-\alpha}}.$$

Notice that the data likelihood $p(\boldsymbol{Y}|\boldsymbol{f})$ is raised to the power of $1-\alpha$. As $\alpha \to 1$, the likelihood $p(\boldsymbol{Y}|\boldsymbol{f})$ will be flattened and its impact reduced. Therefore, as $\alpha \to 1$, the induced regularization will prefer q to be more spread out across configurations of the hidden variables and not only concentrated around ones that best explain the observed data. This decreased dependence on the likelihood also leads to a smoother loss surface where sharp critical points (often anomalies in the data) are smoothened out. In the meantime, excessive smoothing can obscure meaningful solutions. Hence the role of α in achieving this balance.

As a side note, in our simulations and case studies we find two interesting observations. (1) $\alpha \approx 0.5$ often leads to best results (2) the optimal α yields significantly flatter solutions. Flatness has been linked to improves generalization to unobserved data [Chaudhari et al., 2019].

6 Computation & Prediction

6.1 Computation

The recent work of Chen et al. [2020] theoretically shows that mini-batch stochastic gradient descent (SGD) can be used for optimizing $\mathcal{GP}$s. This result in turn allows scaling $\mathcal{GP}$s to very large data size regimes. For instance, in their work, a $\mathcal{GP}$ with one million data points can be trained within half an hour on a standard laptop. In the experimental section, we use SGD to estimate all parameters. One drawback of Chen et al. [2020] is that in the prediction phase, using SGD implies only using a batch of the data to perform predictions. This is sub-optimal, specifically since prediction is a one-shot problem unlike the iterative procedure of learning parameters. To overcome this difficulty, we modify and employ the recently proposed algorithm - Blackbox Matrix-Matrix multiplication [Gardner et al., 2018a,b, Wang et al., 2019a]. The BBMM is an efficient approach to optimize $\mathcal{L}_{marginal}$. This algorithm offers a fast way to calculate the predictive distribution using conjugate gradients (CG), pivoted Cholesky decomposition and parallel computing. Though BBMM is less efficient than SGD, we only require predictions once after parameter estimates are obtained. Therefore, BBMM is a viable method at the prediction stage. Indeed, in our experiments, BBMM acquires predictions within seconds. We defer the detailed BBMM algorithm into Appendix C. An R code is attached in the supplementary file.

6.2 Prediction

Assume that $p(y|\boldsymbol{U}) = \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},U}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{U}, \sigma_\epsilon^2\boldsymbol{I} + (1-\alpha)(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}))$

After estimating all parameters through maximizing $\mathcal{L}_\alpha(q^*)$, we can predict predict the output at a new input point $\boldsymbol{x}^*$. The predictive distribution is given by

$$\begin{aligned} p(y^\star|\boldsymbol{X}, \boldsymbol{x}^*, \boldsymbol{Y}) &= \int p(y^*|\boldsymbol{f}^*)p(\boldsymbol{f}^*|\boldsymbol{X}, \boldsymbol{x}^*, \boldsymbol{Y})d\boldsymbol{f}^* \\ &= \int p(y^*|\boldsymbol{f}^*)p(\boldsymbol{f}^*|\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{x}^*, \boldsymbol{Y})p(\boldsymbol{U}|\boldsymbol{X}, \boldsymbol{x}^*, \boldsymbol{Y})d\boldsymbol{U}d\boldsymbol{f}^* \\ &= \int p(y^*|\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{x}^*, \boldsymbol{Y})p(\boldsymbol{U}|\boldsymbol{X}, \boldsymbol{x}^*, \boldsymbol{Y})d\boldsymbol{U}. \end{aligned}$$

Therefore, we obtain (see Appendix E)

$$p(y^*|\boldsymbol{X}, \boldsymbol{x}^*, \boldsymbol{Y}) = \mathcal{N}(\boldsymbol{A}\Xi^{-1}\boldsymbol{Y}, \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*} + \sigma_\epsilon^2\boldsymbol{I} - \boldsymbol{A}\Xi^{-1}\boldsymbol{A}^T), \tag{7}$$

where $\Xi := \sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha\boldsymbol{Q}$, $\boldsymbol{A} = \boldsymbol{K}_{\boldsymbol{f}^*,U}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}^*}$ and $\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*}$ denotes the covariance matrix evaluated at $\boldsymbol{x}^*$. Consequently, the predicted trajectories have mean $\boldsymbol{A}\Xi^{-1}\boldsymbol{Y}$ and variance $\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*} + \sigma_\epsilon^2 I - \boldsymbol{A}\Xi^{-1}\boldsymbol{A}^T$.

7 Theoretical Properties

In this section, we study the rate of convergence of our algorithm.

7.1 A Data-dependent Upper Bound

In order to derive convergence rates, we first need to obtain a data-dependent upper bound on the marginal likelihood. Titsias [2014] provides a bound based on the KL divergence. We can generalize this bound into (details in Appendix D.1)

$$\mathcal{L}_{upper} \geq \mathcal{L}_{marginal} = \log \frac{1}{|2\pi\boldsymbol{\Xi}|^{\frac{1}{2}}} - \frac{1}{2}\boldsymbol{Y}^T\Big(\boldsymbol{\Xi} + \alpha\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})\boldsymbol{I}\Big)^{-1}\boldsymbol{Y}. \tag{8}$$

7.2 Rate of Convergence

Given the upper bound, we provide the rate of convergence of the proposed α-ELBO (Appendix D.2).

Theorem 1. *Suppose N data points are drawn independently from input distribution $p(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}) \leq v_0, \forall \boldsymbol{x} \in \mathcal{X}$. 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 $1 - \delta$,*

$$D_\alpha[q||p] \leq \frac{\alpha}{2\delta(1-\alpha)} \log \left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)C + 2Nv_0\epsilon]}{N}\right]^N + \alpha\frac{(M+1)C + 2Nv_0\epsilon}{2\delta\sigma_\epsilon^2}\frac{\|\boldsymbol{Y}\|^2}{\sigma_\epsilon^2}.$$

Furthermore, if $\boldsymbol{Y}$ is distributed according to a sample from the prior generative model, then with probability at least $1 - \delta$,

$$D_\alpha[p||q] \leq \alpha\frac{(M+1)C + 2Nv_0\epsilon}{2\delta\sigma_\epsilon^2} + \frac{1}{\delta}\frac{\alpha}{2(1-\alpha)} \log \left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)C + 2Nv_0\epsilon]}{N}\right]^N.$$

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 and $p(\boldsymbol{x})$.

Theorem 1 implies that $D_\alpha[p||q]$ can be made arbitrarily small with high probability. The rate of convergence can be controlled by sample size, number of inducing variables, decay rate of eigenvalues and tuning parameter α. See Sec. 7.3 for more details.

7.3 Consequences

Based on Theorem 1, we can derive the convergence rate for both smooth (e.g., the square exponential kernel) and non-smooth (e.g., Matérn) kernels.

7.3.1 Smooth Kernel

We will provide a convergence result with the square exponential (SE) kernel. The m-th eigenvalue of kernel operator is $\lambda_m = v\sqrt{2a/A}B^{m-1}$, where $a = 1/(4\sigma_\epsilon^2)$, $b = 1/(2\ell^2)$, $c = \sqrt{a^2+2ab}$, $A = a+b+c$ and $B = b/A$. ℓ is the length parameter, v is signal variance and σ_ϵ is the noise parameter. We can obtain $\sum_{m=M+1}^{\infty} \lambda_m = \frac{v\sqrt{2a}}{(1-B)\sqrt{A}}B^M$.

Corollary 2. *Suppose* $\|\boldsymbol{Y}\|^2 \leq RN$*, where* R *is a constant. Fix* $\gamma > 0$ *and take* $\epsilon = \frac{\delta\sigma_\epsilon^2}{vN^{\gamma+2}}$*. Assume the input data is normally distributed and regression in performed with a SE kernel. With probability* $1-\delta$*,*

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

when inference is performed with $M = \frac{(3+\gamma)\log N + \log\eta}{\log(B^{-1})}$*, where* $\eta = \frac{v\sqrt{2a}}{a\sqrt{A}\sigma_\epsilon^2\delta(1-B)}$.

This corollary is proved in Appendix D.3. It has two implications. First, it implies that the number of inducing points should be of order $\mathcal{O}(\log N)$ (i.e., sparse). In a high dimensional input space, following a similar proof, we can show that this order becomes $\mathcal{O}(\log^D N)$. Second, the tuning parameter α plays an important role in controlling convergence rates.

7.3.2 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 (Appendix D.4)

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

Let $M = N^t$ (t will be clarified shortly) and $2rt - 2 \geq \gamma$, then $t \geq \frac{\gamma+2}{2r}$. Therefore, we have (Appendix D.4)

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

Another term in the bound can also be simplified as

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

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.

8 Experiments

We benchmark our model with recent state-of-the-art methods: (1) the exact inference procedure for $\mathcal{GP}$s (EGP) [Wang et al., 2019a, Chen et al., 2020]. This method directly optimizes the exact likelihood function $\mathcal{L}_{\text{marginal}}$. We use SGD to estimate parameters and use BBMM to obtain predictions; (2) the stochastic variational $\mathcal{GP}$ (SGP) [Hoffman et al., 2013, Hensman et al., 2013]. This method performs stochastic VI to the exact $\mathcal{GP}$ and optimizes the derived variational lower bound; (3) the power expectation propagation (PEP) [Bui et al., 2017] with optimal α values.

8.1 A Toy Example

We first investigate the performance of our method on well known simulated functions with 1,000 data points in various dimensions. Data is from the Virtual Library of Simulation Experiments (`http://www.sfu.ca/~ssurjano/index.html`). The testing functions are Gramacy & Lee function ($D = 1$), Branin-Hoo function ($D = 2$) and Griewank-D function ($D \geq 2$). For each dataset, we randomly split 60% data as training sets and 40% as testing sets. We set the number of inducing points to be 50. Throughout the experiment, we use the Matérn kernel. For each function, we run our model 30 times with different $\alpha \in [0.2, 0.8]$ and initial parameters. The performance of each model is measured by Root Mean Square Error (RMSE).

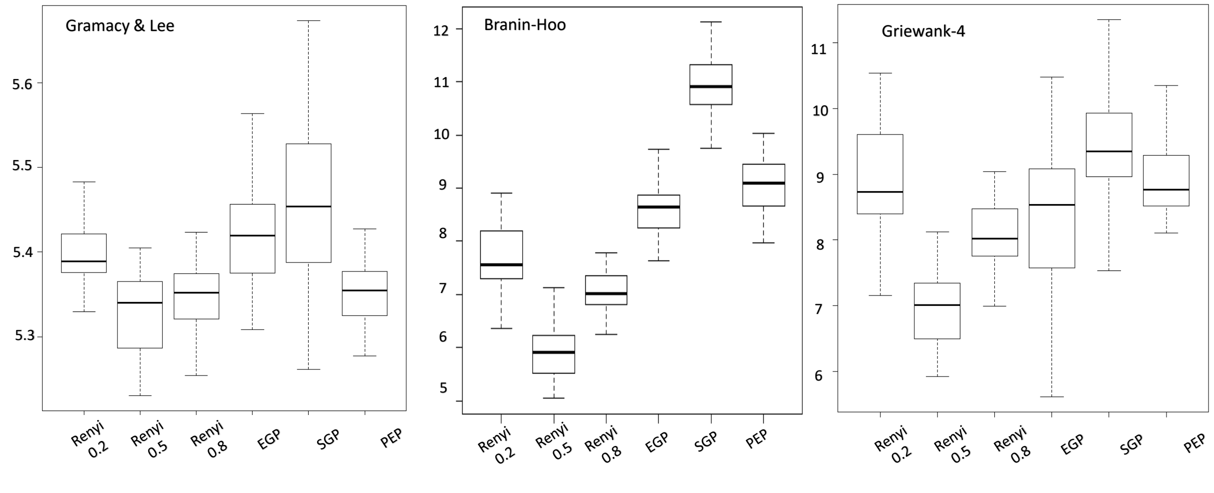

Figure 1: Boxplots of RMSE on toy functions.

We report results from Gramacy & Lee function, Branin-Hoo function and Griewank-4 function in Figure 1. The results clearly indicate that our model, in general, has the smallest RMSE among all benchmark models. When $\alpha \approx 0.5$, we achieve the smallest RMSE. Here, we note that EGP outperforms the SGP in some cases while the opposite happens sometimes. This observation aligns with the conclusion made by Wang et al. [2019a]: SGP is not necessarily better than EGP and vice versa. Overall, it can be seen that regularization parameter tuning is necessary in the context of $\mathcal{GP}$s. Interestingly, when α is around 0.2, the RMSE is compromised. This evidences the danger of ambitiously tightening the the lower bound (as $\alpha = 0$ recovers $\mathcal{L}_{marginal}$) and the need to tune this tightness.

8.2 Real Data

We benchmark the Rényi $\mathcal{GP}$ on a range of datasets that include the (1) *Bike*, (2) *Traffic* (3) *PM2.5* and (4) *House electric* datasets form the UCI data repository [Asuncion and Newman, 2007] (`https://archive.ics.uci.edu/ml/datasets.php`). (5) Battery data from the General Motors Onstar System and (6) *C-MAPSS* aircraft turbofan engines dataset provided by the National Aeronautics and Space Administration (NASA) (`https://ti.arc.nasa.gov/tech/dash/groups/pcoe/`).

Table 1: RMSE of all models on different datasets. The RMSE is calculated over 30 replications with different initial points. For the Rényi $\mathcal{GP}$, we also report the optimal α value. The NLL values are reported in Appendix F.

Dataset	N	EGP	SGP	PEP (optimal α)	Rényi	Optimal α
Bike	17,389	13.41 ± 1.23	16.93 ± 3.33	13.76 ± 2.35	$\mathbf{10.99 \pm 1.33}$	0.50
C-MAPSS	33,727	16.11 ± 1.15	17.45 ± 1.66	15.09 ± 2.01	$\mathbf{12.87 \pm 0.41}$	0.45
PM2.5	43,824	11.74 ± 0.81	15.85 ± 1.03	10.83 ± 0.94	$\mathbf{8.02 \pm 0.55}$	0.55
Traffic	48,204	15.42 ± 1.42	17.47 ± 1.42	15.17 ± 1.05	$\mathbf{12.85 \pm 1.40}$	0.50
Battery	104,046	20.16 ± 1.06	29.96 ± 1.09	21.33 ± 2.04	$\mathbf{9.90 \pm 1.10}$	0.50
House Electric	1,311,539	26.35 ± 1.22	24.27 ± 1.25	18.35 ± 1.15	$\mathbf{17.01 \pm 0.91}$	0.50

Table 2: RMSE of Rényi $\mathcal{GP}$ with different M and α values. The batch size is 1152.

House Electric	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	PEP (optimal α)
$M = 128$	21.87 ± 1.69	18.88 ± 0.91	$\mathbf{17.69 \pm 0.99}$	18.92 ± 0.93	20.12 ± 1.00	19.45 ± 1.78
$M = 256$	20.00 ± 1.23	18.33 ± 0.77	$\mathbf{17.01 \pm 0.94}$	18.41 ± 0.95	19.97 ± 1.03	19.27 ± 1.62
$M = 512$	20.12 ± 1.07	18.26 ± 0.82	18.00 ± 0.99	$\mathbf{16.56 \pm 0.69}$	19.99 ± 1.05	17.29 ± 0.86
$M = 1152$	21.05 ± 0.91	$\mathbf{16.33 \pm 0.85}$	17.14 ± 1.01	17.03 ± 1.02	19.77 ± 0.97	16.99 ± 0.65

Table 3: Sharpness of final solutions and RMSE

$M = 256$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	EGP	SGP	PEP (optimal α)
Sharpness	2.25 ± 0.12	0.53 ± 0.07	$\mathbf{0.11 \pm 0.04}$	0.42 ± 0.09	2.01 ± 0.10	2.57 ± 0.11	2.13 ± 0.08	0.58 ± 0.08
RMSE	20.77 ± 1.01	18.26 ± 1.00	$\mathbf{17.35 \pm 0.94}$	18.31 ± 0.88	20.05 ± 1.21	21.16 ± 1.01	20.31 ± 0.81	17.98 ± 0.41

Table 4: Running Time Comparison (minutes)

Dataset	N	EGP	SGP	PEP	Rényi
House Electric (with learning inducing points)	1,311,539	66.2 ± 0.7	1269.3 ± 9.3	123.1 ± 8.2	129.6 ± 8.5
House Electric (without learning inducing points)	1,311,539	66.2 ± 0.7	64.2 ± 1.3	67.3 ± 1.1	69.6 ± 0.8

The size of these datasets ranges from 17,389 to 1,311,539 data points. For each dataset, we randomly split 60% data as training sets and 40% as testing sets and replicate the experiment 30 times. All data are standardized to have mean 0 and variance 1. We study the effect of α on the prediction performance. For each dataset, we optimize our α-ELBO with different $\alpha \in \{0.1, 0.35, \dots, 0.65, 0.9\}$ and select the optimal α with the smallest RMSE.

Inference: For α-ELBO, SGP and PEP, we use SGD with batch size 1024 and $M = 1024$ to optimize all hyperparameters (this is the recommended setting for M in Wang et al. [2019a], Bui et al. [2017]). For EGP, we use batch size 64 with learning rate 0.01. The number of epoch is set to be 100. **Prediction:** For mBCG algorithm, we use a diagonal-scaling-preconditioning-matrix to stabilize the algorithm and boost convergence speed [Takapoui and Javadi, 2016]. In mBCG, the maximum number of iterations is set to be $10N$.

8.3 Results and Discussion

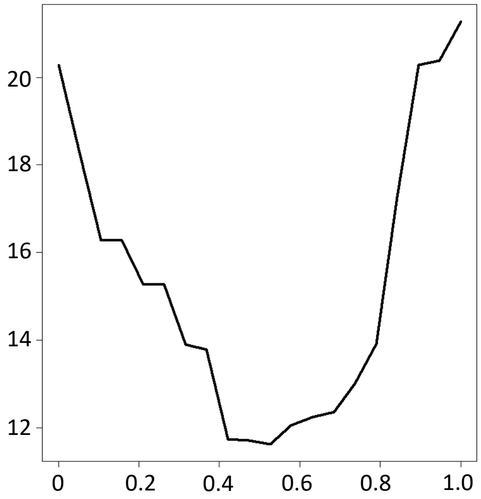

Figure 2: RMSE vs. α using Bike Dataset

Experimental results are reported in Table 1. The performance of each model is measured by RMSE. The RMSE is calculated over 20 experiments with different initial points. We also report the negative log-likelihood (NLL) in Appendix F. Based on Figure 1 and Table 1, we can obtain some important insights. **First**, the results indicate that Rényi $\mathcal{GP}$ achieves the smallest RMSE among all benchmarks on all datasets ranging from small data regimes ($N \approx 17,000$) to large data regimes ($N \approx 1,300,000$). **Second**, by empirical observations, we do find that experiments with α near 0.5 have superior performance. Intuitively, a smaller α is on par with purely maximizing the likelihood and hence might overfit to the data at hand. On the other hand, a larger α might enforce excessive prior regularization and obscures meaningful critical points in the marginal likelihood function. A moderate α (close to 0.5) balances this dilemma and provides better results. This argument can be further supported by Figure 1. We uniformly sample 20 values of α ranging from 0.05 to

0.95, and plot the mean RMSE with respect to the corresponding α for the Bike dataset. The plot demonstrates that intermediate α values are often best. **Third**, EGP seems to outperform SGP on some data sets while the opposite happens on others; both while being inferior to PEP and α-ELBO. This confirms that neither marginal or sparse/variational inference is always superior over the other as the extent of regularization needed is data dependent. We also observe that PEP sometimes does not converge as seen in Tables 1 and 2 where standard deviations for RMSE of PEP is sometimes large. This is not surprising as PEP is a heuristic that currently lacks theoretical backing. **Furthermore**, the advantages of our model become increasingly significant when the sample size increases. This reveals that controlling smoothness and the tightness of the ELBO is necessary and promising when we have big and high dimensional data. Lastly, we report the running time of all models in Table 4. As shown in the Table, the training time for α-ELBO is comparable to PEP on the House Electric dataset with 1.3 million data points (note that all methods are trained with 100 epochs). They are slower than EGP since EGP does not need to learn the distribution of inducing points. When we uniformly distribute inducing points and do not optimize them, the running times are similar to EGP. Notably, since PEP and α-ELBO are significantly faster than SGP, even when we tune α 10 times, the running times are still comparable.

8.4 Different Choice of α and M

We conduct a sensitivity analysis with different α, M on the large scale House Electric dataset from the UCI repository. By employing SGD, the α-ELBO can be trained efficiently using one RTX-2080 GPU. Table 2 reports the RMSE of all models. We use Matérn 3/2 kernel (we also observed that the square exponential kernel delivered similar performance). The results confirm the benefits of the α-ELBO regardless of M. We also observe the previous results where a moderate range of $\alpha \in [0.4, 0.6]$ can deliver promising results.

8.5 Flatness

We perform a simple experiment on the House Electric dataset. Using the flatness measure in Keskar et al. [2016], we obtain the **flatness** of solutions obtained under different α's. Results are reported in Table 3 (more results can be found in Appendix F). *We find a very interesting phenomena: the optimal α yields significantly flatter solutions. Indeed, flatness has been recently linked with improved generalization* [Chaudhari et al., 2019]. This increase in flatness of solutions allows for robust solutions that safeguard against overfitting and improve generalization.

9 Conclusion

We introduce an alternative objective for obtaining parameter estimates in $\mathcal{GP}$s, based on the Rényi α-divergence. This bound offers a structured and tunable balance between model-fit and prior regularization and therefore is capable of controlling the enforced regularization on the objective function. Through many numerical studies, we demonstrate that our proposed objective improves the prediction performance of a $\mathcal{GP}$ over several state of the art inference techniques.

We hope our work spurs interest in the merits of the Rényi based inference in $\mathcal{GP}$s and the notion of controlling regularization via tuning the tightness of the bound on the marginal likelihood. Further, deriving generalization bounds based on the α-ELBO can be an interesting future direction and may help explain the phenomena of obtaining flatter solutions. Along this direction, devising optimization algorithms that specifically target critical points in flat neighborhoods in $\mathcal{GP}$s may also pose strong potential in further pushing the generalization capabilities of $\mathcal{GP}$s.

References

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

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

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

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

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

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

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

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

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

C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. 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, 1991.

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

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

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

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

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

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

J. R. Gardner, G. Pleiss, R. Wu, K. Q. Weinberger, and A. G. Wilson. Product kernel interpolation for scalable gaussian processes. *arXiv preprint arXiv:1802.08903*, 2018b.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

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

F. Lindgren, H. Rue, and J. Lindström. 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, 2011.

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

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

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

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

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

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

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

T. Rainforth, A. R. Kosiorek, T. A. Le, C. J. Maddison, M. Igl, F. Wood, and Y. W. Teh. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.

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

A. Rényi et al. 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, 1961.

B. D. Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 1981.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

X. Yue and R. A. Kontar. 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*, pages 2808–2818. PMLR, 2020b.

Appendix:

An Alternative Gaussian Process 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

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

Therefore, the Rényi variational lower bound can be derived as

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

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

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

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

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

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 $p(\boldsymbol{U}) = \phi(\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}})$, we have

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

where

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

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 $\frac{1-\alpha}{\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$, which is similar to the regularization term in $\mathcal{L}_{VI}$. The tuning parameter α controls how close $\boldsymbol{Q}$ is to $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$ 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) &\coloneqq \frac{1}{1-\alpha}\log\mathbb{E}_q\Bigg[\bigg(\frac{p(\boldsymbol{f},\boldsymbol{U},\boldsymbol{Y}|\boldsymbol{Z})}{q(\boldsymbol{f},\boldsymbol{U}|\boldsymbol{Z})}\bigg)^{1-\alpha}\Bigg] \\
&= \frac{1}{1-\alpha}\log\int\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{Z})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{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{Z})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{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{Z})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{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{Z})d\boldsymbol{f}\right)q(\boldsymbol{U})^{\alpha-1}p(\boldsymbol{U}|\boldsymbol{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{Z})^{1-\alpha}\right] \\
&\qquad + q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{Z})d\boldsymbol{f}\right)d\boldsymbol{U}\bigg\} \\
&= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{Z})] + \frac{1}{1-\alpha}\bigg\{\int q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{Z})d\boldsymbol{f}\right)d\boldsymbol{U}\bigg\} \qquad (3) \\
&\geq -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{Z})] + \frac{1}{1-\alpha}\bigg\{\int q(\boldsymbol{U})\int p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{Z})\log\Big(p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha})\Big)d\boldsymbol{f}d\boldsymbol{U}\bigg\} \\
&= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{Z})] + \frac{1}{1-\alpha}\bigg\{\int q(\boldsymbol{U})\int p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{Z})\log\Big(p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha})\Big)d\boldsymbol{f}d\boldsymbol{U}\bigg\} \\
&= -KL[q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{Z})] + \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{U})}\left[\log p(\boldsymbol{Y}|\boldsymbol{f})\right] = \mathcal{L}_{VI}. \qquad (4)
\end{aligned}
$$

Here,

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

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 $\frac{1}{1-\alpha}\bigg\{\int q(\boldsymbol{U})\log\left(\int p(\boldsymbol{Y}|\boldsymbol{f})^{1-\alpha}p(\boldsymbol{f}|\boldsymbol{U},\boldsymbol{Z})d\boldsymbol{f}\right)d\boldsymbol{U}\bigg\}$ 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,

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

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 $\boldsymbol{\Xi} := \sigma_\epsilon^2 I + (1-\alpha)\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} + \alpha \boldsymbol{Q}$ 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 $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then we can use mBCG to compute $\boldsymbol{\Xi}^{-1}\boldsymbol{Z}$ and calculate gradient as

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

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

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,

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

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

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

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

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

Based on this inequality, it is easy to show that

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

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.

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*

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

Proof. We have

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

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

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

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

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

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

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

□

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 $\frac{\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})}{\sigma_\epsilon^2}$ [Burt et al., 2019].

Lemma 3. *Given a symmetric positive semidefinite matrix $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$, 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*

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

This lemma is proved in [Belabbas and Wolfe, 2009]. Following this lemma and by Lemma 2, we can show that

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

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

Corollary 1.

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

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

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

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

Lemma 4. *Then,*

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

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

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

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

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

□

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) [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 $1-\delta$,*

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

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

Proof. We have

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

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

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

□

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 $k(\boldsymbol{x},\boldsymbol{x})\leq v,\forall\boldsymbol{x}\in\mathcal{X}$. 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 $1-\delta$,*

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

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

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

□

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

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

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,

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

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

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

Another term in the bound can also be simplified as

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

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 Prediction

We have

$$p(\boldsymbol{U}|\boldsymbol{X},\boldsymbol{x}^*,\boldsymbol{Y})=\mathcal{N}(\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}^*}\boldsymbol{\Xi}^{-1}\boldsymbol{Y},\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}-\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}^*}\boldsymbol{\Xi}^{-1}\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{U}}).$$

Also,

$$p(y^*|\boldsymbol{X},\boldsymbol{U},\boldsymbol{x}^*,\boldsymbol{Y})=\mathcal{N}(\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{U},\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*}+\sigma_\epsilon^2\boldsymbol{I}-\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}^*}).$$

by Gaussian identity. Combining above equations,

$$p(y^*|\boldsymbol{X},\boldsymbol{x}^*,\boldsymbol{Y})=\mathcal{N}(\boldsymbol{A}\boldsymbol{\Xi}^{-1}\boldsymbol{Y},\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*}+\sigma_\epsilon^2\boldsymbol{I}-\boldsymbol{A}\boldsymbol{\Xi}^{-1}\boldsymbol{A}^T),\tag{8}$$

where $\boldsymbol{A}=\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{U}}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{U}}^{-1}\boldsymbol{K}_{\boldsymbol{U},\boldsymbol{f}^*}$ and $\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*}$ denotes the covariance matrix evaluated at $\boldsymbol{x}^*$. Consequently, the predicted trajectories have mean $\boldsymbol{A}\boldsymbol{\Xi}^{-1}\boldsymbol{Y}$ and variance $\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*}+\sigma_\epsilon^2I-\boldsymbol{A}\boldsymbol{\Xi}^{-1}\boldsymbol{A}^T$.

F More Results

We provide more detailed results in this section. In table 1, we report the negative loss of each method. In Table 2, we report the sharpness measures of solutions found by each method.

Table 1: NLL of all models on many datasets. The NLL is calculated over 20 experiments with different initial points. For the Rényi $\mathcal{GP}$, we also report the optimal α value.

Dataset	EGP	SGP	PEP	Rényi	Optimal α
Bike	0.41 ± 0.02	0.15 ± 0.03	0.10 ± 0.01	-0.28 ± 0.01	0.50
C-MAPSS	-1.00 ± 0.01	-1.45 ± 0.01	-1.55 ± 0.02	-2.03 ± 0.01	0.45
PM2.5	2.04 ± 0.03	1.55 ± 0.03	1.83 ± 0.08	1.02 ± 0.04	0.55
Traffic	-0.42 ± 0.01	-0.47 ± 0.02	-1.07 ± 0.01	-0.85 ± 0.04	0.50
Battery	2.23 ± 0.06	2.10 ± 0.02	2.17 ± 0.02	1.60 ± 0.01	0.50

Table 2: Sharpness of final solutions and RMSE (Battery Data)

$M = 512$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	EGP	SGP	PEP (optimal α)
Sharpness	13.70 ± 1.05	10.11 ± 0.97	$\mathbf{4.38 \pm 1.31}$	9.98 ± 0.79	15.22 ± 1.06	19.99 ± 0.82	27.12 ± 1.44	18.43 ± 0.61
RMSE	18.29 ± 0.83	15.17 ± 0.55	$\mathbf{9.90 \pm 1.10}$	14.25 ± 1.02	19.88 ± 1.34	20.16 ± 1.06	29.96 ± 1.09	21.33 ± 2.04

References

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

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

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

A. Rényi et al. 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, 1961.

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