BAYESIAN LEARNING WITH DEEP Q-EXPONENTIAL PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by deep neural networks, the deep Gaussian process (DGP) generalizes the standard GP by stacking multiple layers of GPs. Despite the enhanced expressiveness, GP, as an L_2 regularization prior, tends to be over-smooth and sub-optimal for inhomogeneous subjects, such as images with edges. Recently, Q-exponential process (Q-EP) has been proposed as an L_q relaxation to GP and demonstrated with more desirable regularization properties through a parameter $q > 0$ with $q = 2$ corresponding to GP. Sharing the similar tractability of posterior and predictive distributions with GP, Q-EP can also be stacked to improve its modeling flexibility. In this paper, we generalize Q-EP to deep Q-EP to enjoy both proper regularization and improved expressiveness. The generalization is realized by introducing shallow Q-EP as a latent variable model and then building a hierarchy of the shallow Q-EP layers. Sparse approximation by inducing points and scalable variational strategy are applied to facilitate the inference. We demonstrate the numerical advantages of the proposed deep Q-EP model by comparing with multiple state-of-the-art deep probabilistic models.

Keywords: Deep Models, Inhomogeneous Subjects, Regularization, Latent Representation, Model Expressiveness

1 INTRODUCTION

Gaussian process (GP Rasmussen & Williams, 2005; J. M. Bernardo & Smith, 1998) has gained enormous successes and been widely used in statistics and machine learning community. With its flexibility in learning functional relationships (Rasmussen & Williams, 2005) and latent representations (Titsias & Lawrence, 2010), and capability in tractable uncertainty quantification, GP has become one of the most popular non-parametric modeling tools. Facilitated by the sparse approximation (Titsias, 2009) and scalable variational inferences (SVGP Hensman et al., 2015; Salimbeni & Deisenroth, 2017), GP has been popularized for a variety of high-dimensional learning tasks. Neal (1996) in his seminal work discovered that Bayesian neural networks with infinite width converged to GP with certain kernel function. Inspired by the advancement of deep learning (Goodfellow et al., 2016), Damianou & Lawrence (2013) pioneered in generalizing GP with deep structures, hence named deep GP. Ever since then, there has been a large volume of follow-up works including deep convolutional GP (Blomqvist et al., 2020), deep sigma point process (DSPP Jankowiak et al., 2020b), deep image prior (Ulyanov et al., 2020), deep kernel process (Aitchison et al., 2021), deep variational implicit process (Ortega et al., 2023), deep horseshoe GP (Castillo & Randrianarisoa, 2024), and various applications (Dutordoir et al., 2020; Li et al., 2021; Jones et al., 2023).

Despite its flexibility, GP, as an L_2 regularization method, tends to produce random candidate functions that are over-smooth and thus sub-optimal for modeling inhomogeneous objects with abrupt changes or sharp contrast. To address this issue, an L_q based stochastic process, Q-exponential process (Q-EP Li et al., 2023), has recently been proposed to impose flexible regularization through a parameter $q > 0$, which includes GP as a special case when $q = 2$. Similarly as Lasso inducing sparsity for regression, $q = 1$ is often adopted for Q-EP to impose stronger regularization than GP to properly capture dramatic changes in certain portions of inhomogeneous data, e.g., edges in an image. Different from other L_1 based priors such as Laplace random field (Podgórski & Wegener, 2011; Kozubowski et al., 2013) and Besov process (Lassas et al., 2009; Dashti et al., 2012),

Q-EP shares with GP the unique tractability of posterior and predictive distributions (Theorem 3.5 of Li et al., 2023), which essentially permits a deep generalization by stacking multiple stochastic mappings (Damianou & Lawrence, 2013).

Motivated by the improved expressiveness of deep GP and the flexible regularization of Q-EP, in this work we generalize Q-EP to *deep Q-EP* to enhance the capability of Q-EP in modeling inhomogeneous data. On one hand, by stacking multiple layers of Q-EP mappings, deep Q-EP becomes more capable of characterizing complex latent representations than the standard Q-EP. On the other hand, inherited from Q-EP, deep Q-EP maintains the control of regularization through the parameter $q > 0$, whose smaller values impose stronger regularization, more amenable than (deep) GP to preserve inhomogenous traits such as edges in an image. First, we introduce the building block, shallow Q-EP model, which can be regarded as a kernelized latent variable model (LVM) (Lawrence, 2003; Titsias & Lawrence, 2010). Such shallow model is also viewed as a stochastic mapping F from input (or latent) variables X to output variables Y defined by a kernel. Then as in Lawrence & Moore (2007); Damianou & Lawrence (2013), we extend such mapping by stacking multiple shallow Q-EP layers to form a hierarchy for the deep Q-EP. Sparse approximation by inducing points (Titsias, 2009) is adopted for the variational inference of deep Q-EP. A theoretic barricade for developing the evidence lower bound (ELBO) in the setting of Q-EP is that the power in the exponent of its density makes many involved expectations intractable. We solve this challenge by taking advantage of Jensen's inequality. The inference procedure, as in deep GP, can be efficiently implemented in `GPyTorch` (Gardner et al., 2018).

Connection to existing works Our proposed deep Q-EP is closely related to deep GP (Damianou & Lawrence, 2013) and two other works, deep kernel learning (DKL-GP Wilson et al., 2016) and DSPP (Jankowiak et al., 2020b). Deep Q-EP generalizes deep GP with a parameter $q > 0$ to control the regularization (See Figure 1 for its effect on learning representations) and includes deep GP as a special case for $q = 2$. DKL-GP combines the deep learning architectures (neural networks) with the non-parametric flexibility of kernel methods (GP). The GP part can also be replaced by Q-EP to generate new methods like DKL-QEP (See Section 5.4.) DSPP is motivated by parametric GP models (PPGPR Jankowiak et al., 2020a) and applies sigma point approximation or quadrature-like integration to the predictive distribution. The majority of popular deep probabilistic models rely on GP. This is one of the few developed out of a non-Gaussian stochastic process. Our proposed work on deep Q-EP has multi-fold contributions to deep probabilistic models:

1. We propose a novel deep probabilistic model based on Q-EP that generalizes deep GP with flexibility of regularization for handling data inhomogeneity.
2. We develop the variational inference for deep Q-EP and efficiently implement it.
3. We demonstrate numerical advantages of deep Q-EP in modeling inhomogeneous data by comparing with state-of-the-art deep probabilistic models.

The rest of the paper is organized as follows. Section 2 introduces the background of Q-EP. We then develop shallow Q-EP in Section 3 as the building block for deep Q-EP in Section 4. In these two sections, we highlight the importance of posterior tractability in the development and some obstacles in deriving the variational lower bounds. In Section 5 we demonstrate the numerical advantages by comparing with multiple deep probabilistic models in various learning tasks. Finally, we conclude with some discussion on the limitation and potential improvement in Section 6.

2 Background: Q-exponential Processes

2.1 Multivariate Q-exponential Distribution

Based on L_q regularization, the univariate *q-exponential distribution* (Dashti et al., 2012) with an inexact density (not normalized to 1), $\pi_q(u) \propto \exp\left(-\frac{1}{2}|u|^q\right)$, is one of the following exponential power (EP) distributions $\mathrm{EP}(\mu, \sigma, q)$ with $\mu = 0, \sigma = 1$:

$$p(u|\mu, \sigma, q) = \frac{q}{2^{1+1/q}\sigma\Gamma(1/q)} \exp\left\{-\frac{1}{2}\left|\frac{u-\mu}{\sigma}\right|^q\right\}.$$

This family includes normal distribution $\mathcal{N}(\mu, \sigma^2)$ for $q = 2$ and Laplace distribution $L(\mu, b)$ with $\sigma = 2^{-1/q}b$ for $q = 1$ as special cases.

Li et al. (2023) generalize the univariate q-exponential random variable to a multivariate random vector on which a stochastic process can be defined with two requirements by the Kolmogorov' extension theorem (Øksendal, 2003): i) **exchangeability** of the joint distribution, i.e. $p(\mathbf{u}_{1:N}) = p(\mathbf{u}_{\tau(1:N)})$ for any finite permutation τ; and ii) **consistency** of marginalization, i.e. $p(\mathbf{u}_1) = \int p(\mathbf{u}_1, \mathbf{u}_2) d\mathbf{u}_2$.

Suppose a function $u(x)$ is observed at N locations, $x_1, \cdots, x_N \in \mathcal{D} \subset \mathbb{R}^d$. Li et al. (2023) find a consistent generalization, named *multivariate q-exponential distribution*, for $\mathbf{u} = (u(x_1), \cdots, u(x_N))$ from the family of elliptic contour distributions (Johnson, 1987; Fang & Zhang, 1990).

Definition 1. *A multivariate q-exponential distribution for a random vector $\mathbf{u} \in \mathbb{R}^N$, denoted as $\mathrm{q{-}ED}_N(\boldsymbol{\mu}, \mathbf{C})$, has the following density*

$$p(\mathbf{u}|\boldsymbol{\mu}, \mathbf{C}, q) = \frac{q}{2}(2\pi)^{-\frac{N}{2}} |\mathbf{C}|^{-\frac{1}{2}} r(\mathbf{u})^{(\frac{q}{2}-1)\frac{N}{2}} \exp\left\{-\frac{r^{\frac{q}{2}}}{2}\right\}, \quad r = (\mathbf{u}-\boldsymbol{\mu})^{\mathsf{T}} \mathbf{C}^{-1} (\mathbf{u}-\boldsymbol{\mu}). \quad (1)$$

Remark 1. *If taken negative logarithm, the density of $\mathrm{q-ED}$ in (1) yields a quantity dominated by some weighted L_q norm of $\mathbf{u}-\boldsymbol{\mu}$, i.e. $\frac{1}{2} r^{\frac{q}{2}} = \frac{1}{2} \|\mathbf{u}-\boldsymbol{\mu}\|_{\mathbf{C}}^q$. From the optimization perspective, $\mathrm{q-ED}$, when used as a prior, imposes L_q regularization in obtaining the maximum a posterior (MAP).*

The following proposition describes the role of matrix $\mathbf{C}$ in characterizing the covariance between the components (Li et al., 2023).

Proposition 2.1. *If $\mathbf{u} \sim \mathrm{q{-}ED}_N(\boldsymbol{\mu}, \mathbf{C})$, then we have*

$$\mathrm{E}[\mathbf{u}] = \boldsymbol{\mu}, \qquad \mathrm{Cov}(\mathbf{u}) = \frac{2^{\frac{2}{q}}\Gamma(\frac{N}{2}+\frac{2}{q})}{N\Gamma(\frac{N}{2})}\mathbf{C} \doteq N^{\frac{2}{q}-1}\mathbf{C}, \quad as \quad N \to \infty.$$

2.2 Q-EXPONENTIAL PROCESS AND MULTI-OUTPUT Q-EP

Li et al. (2023) prove that the multivariate q-exponential random vector $\mathbf{u} \sim \mathrm{q{-}ED}_N(0, \mathbf{C})$ satisfies the conditions of Kolmogorov's extension theorem hence it can be generalized to a stochastic process. For this purpose, we scale it by a factor $N^{\frac{1}{2}-\frac{1}{q}}$ so that its covariance is asymptotically finite (refer to Proposition 2.1). If $\mathbf{u} \sim \mathrm{q{-}ED}_N(0, \mathbf{C})$, then we denote $\mathbf{u}^* := N^{\frac{1}{2}-\frac{1}{q}}\mathbf{u} \sim \mathrm{q{-}ED}_N^*(0, \mathbf{C})$ as a *scaled* q-exponential random variable. With a covariance (symmetric and positive-definite) kernel $\mathcal{C} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, we define the following *q-exponential process (Q-EP)* based on the scaled q-exponential distribution $\mathrm{q{-}ED}_N^*(0, \mathbf{C})$.

Definition 2. *A (centered) q-exponential process $u(x)$ with kernel $\mathcal{C}$, $\mathrm{q{-}}\mathcal{EP}(0, \mathcal{C})$, is a collection of random variables such that any finite set, $\mathbf{u} := (u(x_1), \cdots u(x_N))$, follows a scaled multivariate q-exponential distribution $\mathrm{q{-}ED}^*(0, \mathbf{C})$, where $\mathbf{C} = [\mathcal{C}(x_i, x_j)]_{N\times N}$. If $\mathcal{C} = \mathcal{I}$, then u is said to be* marginally identical but uncorrelated (m.i.u.).

Remark 2. *When $q = 2$, $\mathrm{q{-}ED}_N(\boldsymbol{\mu}, \mathbf{C})$ reduces to $\mathcal{N}_N(\boldsymbol{\mu}, \mathbf{C})$ and $\mathrm{q{-}}\mathcal{EP}(0, \mathcal{C})$ becomes $\mathcal{GP}(0, \mathcal{C})$. When $q \in [1, 2)$, $\mathrm{q{-}}\mathcal{EP}(0, \mathcal{C})$ lends flexibility to modeling functional data with more regularization than GP. In practice, $q = 1$ is often adopted for faster posterior convergence (Agapiou et al., 2021; Lan et al., 2023) and the capability of preserving inhomogeneous features (rough functional data, edges in image, etc). Refer to Figure 1 for the regularization effect of q.*

One caveat of Q-EP is that uncorrelation (identity covariance) does not imply independence except for the special Gaussian case ($q = 2$). For multiple Q-EPs, $(u_1(x), \cdots, u_D(x))$, we usually do not assume them independent because their joint distribution is difficult to work with (due to the lack of additivity in the exponential part of density function (1)). Rather, uncorrelation is a preferable assumption. In general, we define multi-output (multivariate) Q-EPs through matrix vectorization.

Definition 3. *A multi-output (multivariate) q-*exponential process*, $u(\cdot) = (u_1(\cdot), \cdots, u_D(\cdot))$, each $u_j(\cdot) \sim \mathrm{q-}\mathcal{EP}(\mu_j, \mathcal{C}_x)$, is said to have association $\mathbf{C}_t$ if at any finite locations, $\mathbf{x} = \{x_n\}_{n=1}^N$, $\mathrm{vec}([u_1(\mathbf{x}), \cdots, u_D(\mathbf{x})]_{N\times D}) \sim \mathrm{q-ED}_{ND}(\mathrm{vec}(\boldsymbol{\mu}), \mathbf{C}_t \otimes \mathbf{C}_x)$, where we have $u_j(\mathbf{x}) = [u_j(x_1), \cdots, u_j(x_N)]^{\mathsf{T}}$, for $j = 1, \ldots, D$, $\boldsymbol{\mu} = [\mu_1(\mathbf{x}), \cdots, \mu_D(\mathbf{x})]_{N\times D}$ and $\mathbf{C}_x = [\mathcal{C}_x(x_n, x_m)]_{N\times N}$. We denote $u \sim \mathrm{q-}\mathcal{EP}(\mu, \mathcal{C}_x, \mathbf{C}_t)$. In particular, $\{u_j(\cdot)\}$ are m.i.u. if $\mathbf{C}_t = \mathbf{I}_D$.*

To improve the modeling expressiveness of Q-EP, we stack m.i.u. multi-output Q-EPs to build a deep Q-EP, similarly as constructing deep GP with multiple GP layers. For this purpose, we first introduce Bayesian (multivariate) regression with Q-EP priors.

2.3 Bayesian Regression with Q-EP Priors

Given data $\mathbf{x} = \{x_n\}_{n=1}^N$ and $\mathbf{y} = \{y_n\}_{n=1}^N$, we consider the generic Bayesian regression model:

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{q{-}ED}_N(0, \Sigma), \\ f &\sim \mathrm{q{-}}\mathcal{EP}(0, \mathcal{C}). \end{aligned} \tag{2}$$

It is proved in Theorem 3.5 of Li et al. (2023) that the posterior (predictive) distribution is analytically tractable when both the prior and the likelihood are Q-EP, which is one of the keys for the deep generalization of Q-EP.

Theorem 2.1. *For the regression model* (2)*, the posterior distribution of* $f(x_*)$ *at* x_* *is*

$$f(x_*)|\mathbf{y}, \mathbf{x}, x_* \sim \mathrm{q{-}ED}(\boldsymbol{\mu}^*, \mathbf{C}^*), \quad \boldsymbol{\mu}^* = \mathbf{C}_*^\mathsf{T}(\mathbf{C}+\Sigma)^{-1}\mathbf{y}, \quad \mathbf{C}^* = \mathbf{C}_{**} - \mathbf{C}_*^\mathsf{T}(\mathbf{C}+\Sigma)^{-1}\mathbf{C}_*,$$

where $\mathbf{C} = \mathcal{C}(\mathbf{x}, \mathbf{x})$*,* $\mathbf{C}_* = \mathcal{C}(\mathbf{x}, x_*)$*, and* $\mathbf{C}_{**} = \mathcal{C}(x_*, x_*)$*.*

Denote $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_Q]_{N\times Q}$, $\mathbf{F} = [f_1(\mathbf{X}), \cdots, f_D(\mathbf{X})]_{N\times D}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_D]_{N\times D}$. With m.i.u. Q-EP priors as in Definition (3) imposed on $f := (f_1, \cdots, f_D)$, we now consider the following multivariate regression problem:

$$\begin{aligned} \text{likelihood}: \quad & \mathrm{vec}(\mathbf{Y})|\mathbf{F} \sim \mathrm{q{-}ED}_{ND}(\mathrm{vec}(\mathbf{F}), \mathbf{I}_D \otimes \Sigma), \\ \text{prior on latent function}: \quad & f \sim \mathrm{q{-}}\mathcal{EP}(0, \mathcal{C}, \mathbf{I}_D). \end{aligned} \tag{3}$$

Based on the additivity of $\mathrm{q{-}ED}$ (as a special elliptic contour) random variables (Fang & Zhang, 1990), we can find the marginal of $\mathbf{Y}$ by noticing that $\mathbf{Y} = \mathbf{F} + \boldsymbol{\varepsilon}$ with $\mathrm{vec}(\boldsymbol{\varepsilon}) \sim \mathrm{q{-}ED}(\mathbf{0}, \mathbf{I}_D \otimes \Sigma)$:

$$\text{marginal likelihood}: \quad \mathrm{vec}(\mathbf{Y})|\mathbf{X} \sim \mathrm{q{-}ED}_{ND}(\mathbf{0}, \mathbf{I}_D \otimes (\mathbf{C}+\Sigma)). \tag{4}$$

3 Shallow Q-EP Model

In this section we introduce the shallow (1-layer) Q-EP model which serves as a building block for the deep Q-EP model to be developed in Section 4. We start with the the marginal model (4) that can be identified as a latent variable model (LVM) (Lawrence, 2003) with specified kernel. This defines a shallow Q-EP model. Then we develop variational infererence with sparse approximation for such model (Titsias & Lawrence, 2010) and stack multiple layers to build the deep Q-EP.

Note the marginal model (4) of $\mathbf{Y}|\mathbf{X}$ can be viewed as a stochastic mapping (Theorem 2.1 of Li et al., 2023):

$$\tilde{f}: \mathbf{X} \to \mathbf{Y} = R\mathbf{L}_\mathbf{X}\mathbf{S},$$

where $R^q \sim \chi^2(N)$, $\mathbf{L}_\mathbf{X}$ is the Cholesky factor of $\mathbf{C}_\mathbf{X} + \Sigma$ whose value depends on $\mathbf{X}$, and $\mathbf{S} := [S_1, \cdots, S_D] \sim \mathrm{Unif}(\prod_{d=1}^D \mathcal{S}^{N+1})$, i.e. each S_d is uniformly distributed on an N-dimensional unit-sphere $\mathcal{S}^{N+1}$.

Note $\mathbf{X}$ is an input variable in the supervised learning, and could also be a latent variable in the unsupervised learning. In the latter case, the shallow Q-EP model (4) of $\mathbf{Y}|\mathbf{X}$ can be regarded an LVM obtained by integrating out the latent function $\mathbf{F}$ in model (3), which is a linear mapping in probabilistic PCA (Tipping & Bishop, 1999) and a multi-output GP in GP-LVM (Lawrence, 2003; 2005). GP can be replaced by Q-EP to impose flexible regularization on the input (latent) space, and hence we propose the shallow Q-EP model as also a Q-EP LVM.

For the convenience of exposition, we set $\Sigma = \beta^{-1}\mathbf{I}_N$ and denote $\mathbf{K} := \mathbf{C}_\mathbf{X} + \Sigma$. We adopt the following automatic relevance determination (ARD) kernel as in Titsias & Lawrence (2010), e.g. squared exponential (SE), to determine the dominant dimensions in the input (latent) space:

$$\mathbf{K} = [k(\mathbf{x}_n, \mathbf{x}_m)]_{N\times N}, \quad k(\mathbf{x}_n, \mathbf{x}_m) = \alpha^{-1}\exp\left\{-\frac{1}{2}(\mathbf{x}_n - \mathbf{x}_m)^\mathsf{T}\,\mathrm{diag}(\boldsymbol{\gamma})(\mathbf{x}_n - \mathbf{x}_m)\right\}. \tag{5}$$

3.1 BAYESIAN SHALLOW Q-EP

Like Titsias & Lawrence (2010), we adopt a prior for the input (latent) variable $\mathbf{X}$ and introduce the following Bayesian shallow Q-EP model:

$$\begin{aligned} \text{marginal likelihood}: &\quad \mathrm{vec}(\mathbf{Y})|\mathbf{X} \sim \mathrm{q{-}ED}(\mathbf{0}, \mathbf{I}_D \otimes \mathbf{K}), \\ \text{prior on input/latent variable}: &\quad \mathrm{vec}(\mathbf{X}) \sim \mathrm{q{-}ED}(\mathbf{0}, \mathbf{I}_{QD}). \end{aligned} \tag{6}$$

Compared with the optimization method (Lawrence, 2003), the Bayesian training procedure is robust to overfitting and can automatically determine the intrinsic dimensionality of the nonlinear input (latent) space (Titsias & Lawrence, 2010) by thresholding the correlation length-scale γ.

For more practical applications, we use variational Bayes, instead of Markov Chain Monte Carlo (MCMC), to train the shallow Q-EP model (6). The variational inference for this model is much more complicated than GP-LVM because the log-likelihood (3) is no longer represented as a quadratic form of data. It should be noted that many expectations in the evidence lower bound (ELBO) are no longer analytically tractable with a general power q in the exponent of the density (1), which makes it much more challenging to derive a computable ELBO. We solve this issue with the help of Jensen's inequality.

For variational Bayes, we approximate the posterior distribution $p(\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})$ with the uncorrelated $\mathrm{q{-}ED}$:

$$q(\mathbf{X}) \sim \mathrm{q{-}ED}(\boldsymbol{\mu}, \mathrm{diag}(\{\mathbf{S}_n\})),$$

where each covariance $\mathbf{S}_n$ is of size $D \times D$ and can be chosen as a diagonal matrix for convenience.

To speed up the computation, sparse variational approximation (Titsias, 2009; Lawrence & Moore, 2007) is adopted by introducing the inducing points $\tilde{\mathbf{X}} \in \mathbb{R}^{M \times Q}$ with their function values $\mathbf{U} = [f_1(\tilde{\mathbf{X}}), \cdots, f_D(\tilde{\mathbf{X}})] \in \mathbb{R}^{M \times D}$. Hence the marginal likelihood $p(\mathbf{Y}|\mathbf{X})$ in (6) can be augmented to a joint distribution of several $\mathrm{q{-}ED}$ random variables:

$$p(\mathbf{Y}|\mathbf{X}) \propto p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \tilde{\mathbf{X}})p(\mathbf{U}|\tilde{\mathbf{X}}),$$

where $p(\mathrm{vec}(\mathbf{F})|\mathbf{U}, \mathbf{X}, \tilde{\mathbf{X}}) \sim \mathrm{q{-}ED}(\mathrm{vec}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{U}), \mathbf{I}_D \otimes (\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}))$ and $p(\mathrm{vec}(\mathbf{U})|\tilde{\mathbf{X}}) \sim \mathrm{q{-}ED}(\mathbf{0}, \mathbf{I}_D \otimes \mathbf{K}_{MM})$.

Denote by $\varphi(r; \Sigma, D) := -\frac{D}{2}\log|\Sigma| + \frac{ND}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$. With the variational distribution $q(\mathbf{F}, \mathbf{U}, \mathbf{X}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X})q(\mathbf{U})q(\mathbf{X})$ for $q(\mathbf{U}) \sim \mathrm{q{-}ED}(\mathbf{M}, \mathrm{diag}(\{\boldsymbol{\Sigma}_d\}))$, the following final ELBO is obtained by the two-stage approach in (SVGP Hensman et al., 2015) (Refer to Section A.1 for details):

$$\begin{aligned} \log p(\mathbf{Y}) \geq & \mathcal{L}(q) = \int q(\mathbf{X})q(\mathbf{U})p(\mathbf{F}|\mathbf{U}, \mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{U})p(\mathbf{X})}{q(\mathbf{U})q(\mathbf{X})} d\mathbf{F}d\mathbf{U}d\mathbf{X} \\ \geq & h^*(\mathbf{Y}, \mathbf{X}) - \mathrm{KL}_{\mathbf{U}}^* - \mathrm{KL}_{\mathbf{X}}^*, \\ h^*(\mathbf{Y}, \mathbf{X}) = & \varphi(r_{\mathbf{Y}}; \beta^{-1}\mathbf{I}_N, D), \\ r_{\mathbf{Y}} = & r(\mathbf{Y}, \Psi_1\mathbf{K}_{MM}^{-1}\mathbf{M}) + \beta\mathrm{tr}(\mathbf{M}^\mathsf{T}\mathbf{K}_{MM}^{-1}(\Psi_2 - \Psi_1^\mathsf{T}\Psi_1)\mathbf{K}_{MM}^{-1}\mathbf{M}) \\ & + \beta D[\psi_0 - \mathrm{tr}(\mathbf{K}_{MM}^{-1}\Psi_2)] + \beta\sum_{d=1}^{D}\mathrm{tr}(\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_d\mathbf{K}_{MM}^{-1}\Psi_2), \\ -\mathrm{KL}_{\mathbf{U}}^* = & \frac{1}{2}\sum_{d=1}^{D}\log|\boldsymbol{\Sigma}_d| + \varphi\left(\mathrm{tr}(\mathbf{M}^\mathsf{T}\mathbf{K}_{MM}^{-1}\mathbf{M}) + \sum_{d=1}^{D}\mathrm{tr}(\boldsymbol{\Sigma}_d\mathbf{K}_{MM}^{-1}); \mathbf{K}_{MM}, D\right), \\ -\mathrm{KL}_{\mathbf{X}}^* = & \frac{1}{2}\sum_{n=1}^{N}\log|\mathbf{S}_n| + \varphi\left(\mathrm{tr}(\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\mu}) + \sum_{n=1}^{N}\mathrm{tr}(\mathbf{S}_n); \mathbf{I}_N, Q\right), \end{aligned} \tag{7}$$

where $\psi_0 = \mathrm{tr}(\langle\mathbf{K}_{NN}\rangle_{q(\mathbf{X})})$, $\Psi_1 = \langle\mathbf{K}_{NM}\rangle_{q(\mathbf{X})}$, and $\Psi_2 = \langle\mathbf{K}_{MN}\mathbf{K}_{NM}\rangle_{q(\mathbf{X})}$.

Remark 3. *When $q = 2$, $\varphi(r; \Sigma, D) = -\frac{D}{2}\log|\Sigma| - \frac{1}{2}r$ with $r = r(\mathbf{Y}, \Psi_1\mathbf{K}_{MM}^{-1}\mathbf{M})$ becomes the log-density of matrix normal $\mathcal{MN}_{N\times D}(\Psi_1\mathbf{K}_{MM}^{-1}\mathbf{M}, \beta^{-1}\mathbf{I}_N, \mathbf{I}_D)$. Then the ELBO (7) reduces to the ELBO as in Equation (7) of (SVGP Hensman et al., 2015) with an extra term $\beta\mathrm{tr}(\mathbf{M}^\mathsf{T}\mathbf{K}_{MM}^{-1}(\Psi_2 - \Psi_1^\mathsf{T}\Psi_1)\mathbf{K}_{MM}^{-1}\mathbf{M})$. The computational complexity, $\mathcal{O}(NM^2)$, remains the same as GP-LVM (Titsias & Lawrence, 2010).*

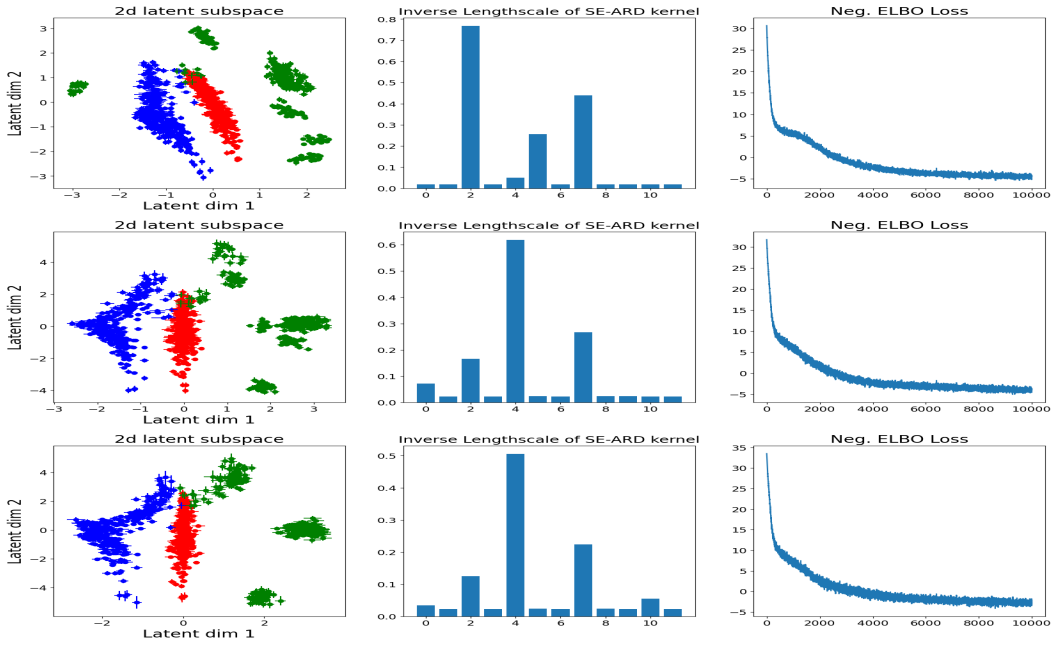


Figure 1: 2d latent space of multi-phase oil-flow dataset: contrasting GP-LVM ($q = 2$) (top row) with two shallow Q-EPs for $q = 1.25$ (middle row) and $q = 1$ (bottom row). Smaller q tends to contract the latent space and hence regularizes the learned latent representation, an effect similarly existing among ridge regression, elastic-net, and Lasso.

We demonstrate the behavior of shallow Q-EP as an LVM in unsupervised learning and contrast it with GP-LVM using the canonical multi-phase oil-flow dataset (Titsias & Lawrence, 2010) that consists of 1000 observations (12-dimensional) corresponding to three different phases of oil-flow. Figure 1 visualizes the 2d latent subspaces identified with two most dominant latent dimensions found by GP-LVM (top) and two shallow Q-EP models with $q = 1.25$ (middle) and $q = 1$ (bottom) respectively. The vertical and horizontal bars indicate axis aligned uncertainty around each latent point. As GP-LVM corresponds to a shallow Q-EP with $q = 2$, the parameter $q > 0$ controls a regularization effect of shallow Q-EP: the smaller q leads to more regularization on the learned latent representations and hence yields clusters more aggregated, as illustrated by the green class in the first column of Figure 1. The two types of models also differ in the dominant relevant dimensions: $(2, 5, 7)$ for GP-LVM versus $(2, 4, 7)$ for QEP-LVM. Note, the ELBO loss of shallow Q-EP converges slightly faster than that of GP-LVM in this example, yet their final values are not comparable because two models have different densities.

4 Deep Q-EP Model

In this section, we construct the deep Q-EP model by stacking multiple shallow Q-EP layers introduced in Section 3, similarly as building deep GP with GP-LVMs (Damianou & Lawrence, 2013). More specifically, we consider a hierarchy of L shallow Q-EP layers (6) as follows:

$$\begin{aligned}
y_{nd} &= f_d^0(\mathbf{x}_n^1) + \varepsilon_{nd}^0, \quad d = 1, \cdots, D_0, \quad \mathbf{x}_n^1 \in \mathbb{R}^{D_1}, \\
x_{nd}^1 &= f_d^1(\mathbf{x}_n^2) + \varepsilon_{nd}^1, \quad d = 1, \cdots, D_1, \quad \mathbf{x}_n^2 \in \mathbb{R}^{D_2}, \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
x_{nd}^{L-1} &= f_d^{L-1}(\mathbf{z}_n) + \varepsilon_{nd}^{L-1}, \quad d = 1, \cdots, D_{L-1}, \quad \mathbf{z}_n \in \mathbb{R}^{D_L},
\end{aligned}$$

where $\boldsymbol{\varepsilon}^\ell \sim \mathrm{q}{-}\mathrm{ED}(\mathbf{0}, \Gamma^\ell)$, $f^\ell \sim \mathrm{q}{-}\mathcal{EP}(0, k^\ell, I_{D_\ell})$ for $\ell = 0, \cdots, L-1$ and we identify $\mathbf{Y} = \mathbf{X}^0$ and $\mathbf{Z} = \mathbf{X}^L$.

Consider the prior $\mathbf{Z} \sim \mathrm{q{-}ED}(\mathbf{0}, \mathbf{I}_{ND_L})$. The joint probability, augmented with the inducing points $\tilde{\mathbf{X}}^\ell$ and the associated function values $\mathbf{U}^\ell = [f_d^\ell(\tilde{\mathbf{X}}^\ell)]_{d=1}^{D_\ell}$, is decomposed as

$$p(\{\mathbf{X}^\ell, \mathbf{F}^\ell, \mathbf{U}^\ell\}_{\ell=0}^{L-1}, \mathbf{Z}) = \prod_{\ell=0}^{L-1} p(\mathbf{X}^\ell|\mathbf{F}^\ell) p(\mathbf{F}^\ell|\mathbf{U}^\ell, \mathbf{X}^{\ell+1}) p(\mathbf{U}^\ell) \cdot p(\mathbf{Z}).$$

And we use the following variational distribution

$$\mathcal{Q} = \prod_{\ell=0}^{L-1} p(\mathbf{F}^\ell|\mathbf{U}^\ell, \mathbf{X}^{\ell+1}) q(\mathbf{U}^\ell) q(\mathbf{X}^{\ell+1}), \quad q(\mathbf{X}^{\ell+1}) = \mathrm{q{-}ED}(\boldsymbol{\mu}^{\ell+1}, \mathrm{diag}(\{\mathbf{S}_n^{\ell+1}\})).$$

Then the ELBO becomes

$$\begin{aligned}\mathcal{L}(\mathcal{Q}) &= \int_{\{\mathbf{F}^\ell, \mathbf{U}^\ell, \mathbf{X}^{\ell+1}\}_{\ell=0}^{L-1}} \mathcal{Q} \log \frac{p(\{\mathbf{X}^\ell, \mathbf{F}^\ell, \mathbf{U}^\ell\}_{\ell=0}^{L-1}, \mathbf{Z})}{\prod_{\ell=0}^{L-1} q(\mathbf{U}^\ell) q(\mathbf{X}^{\ell+1})} \\ &= h_0 - \mathrm{KL}_{\mathbf{U}^0} + \sum_{\ell=1}^{L-1} [h_\ell - \mathrm{KL}_{\mathbf{U}^\ell} + \mathcal{H}_q(\mathbf{X}_\ell)] - \mathrm{KL}_{\mathbf{Z}},\end{aligned}$$

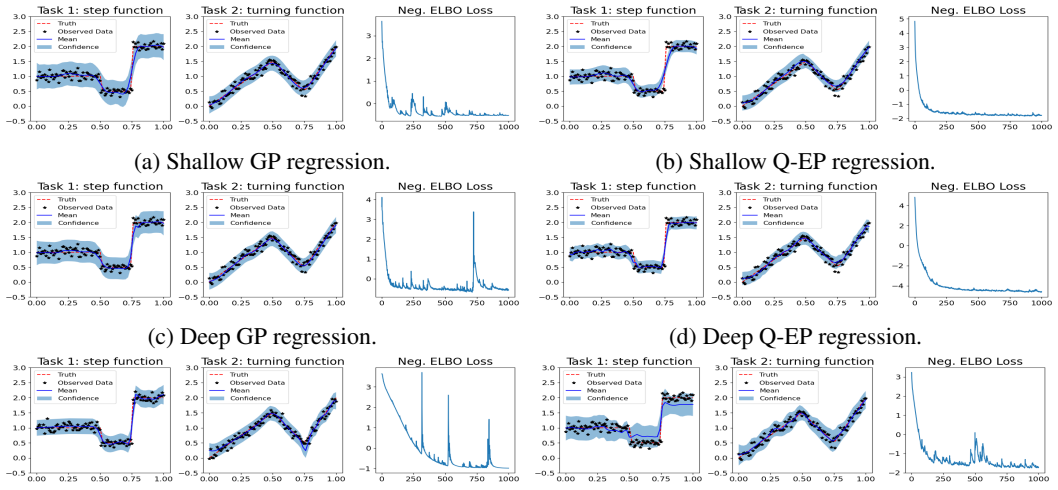
where $h_\ell = \left\langle \log p(\mathbf{X}^\ell|\mathbf{F}^\ell) \right\rangle_{q(\mathbf{F}^\ell) q(\mathbf{X}^{\ell+1}) q(\mathbf{X}^\ell)}$ with $q(\mathbf{X}^0) = q(\mathbf{Y}) \equiv 1$. Based on the previous bound (7), we have for $\ell = 1, \cdots, L-1$ (Refer to Section A.2 for details):

$$\begin{aligned}h_0 \geq & h^*(\mathbf{Y}, \mathbf{X}^1), \\ h_\ell \geq & h^*(\mathbf{X}^\ell, \mathbf{X}^{\ell+1}) = \varphi(r_{\boldsymbol{\mu}^\ell}; \Gamma^\ell, D_\ell), \\ r_{\boldsymbol{\mu}^\ell} = & r(\boldsymbol{\mu}^\ell, \Psi_1^\ell (\mathbf{K}_{MM}^\ell)^{-1} \mathbf{M}^\ell) + \mathrm{tr}((\mathbf{M}^\ell)^\mathsf{T} (\mathbf{K}_{MM}^\ell)^{-1} (\Psi_2^\ell - (\Psi_1^\ell)^\mathsf{T} (\Gamma^\ell)^{-1} \Psi_1^\ell) (\mathbf{K}_{MM}^\ell)^{-1} \mathbf{M}^\ell) \\ & + D_\ell [\psi_0^\ell - \mathrm{tr}((\mathbf{K}_{MM}^\ell)^{-1} \Psi_2^\ell)] + \sum_{d=1}^{D_\ell} \mathrm{tr}((\mathbf{K}_{MM}^\ell)^{-1} \boldsymbol{\Sigma}_d^\ell (\mathbf{K}_{MM}^\ell)^{-1} \Psi_2^\ell) \\ & + \mathrm{tr}((\mathbf{I}_{D_\ell} \otimes (\Gamma^\ell)^{-1}) \, \mathrm{diag}(\{\mathbf{S}_n^\ell\})), \\ -\mathrm{KL}_{\mathbf{U}^\ell}^* = & \frac{1}{2} \sum_{d=1}^{D_\ell} \log |\boldsymbol{\Sigma}_d^\ell| + \varphi \left(\mathrm{tr}((\mathbf{M}^\ell)^\mathsf{T} (\mathbf{K}_{MM}^\ell)^{-1} \mathbf{M}^\ell) + \sum_{d=1}^{D_\ell} \mathrm{tr}(\boldsymbol{\Sigma}_d^\ell (\mathbf{K}_{MM}^\ell)^{-1}); \mathbf{K}_{MM}^\ell, D_\ell \right), \\ \mathcal{H}_q(\mathbf{X}_\ell) \geq & \frac{1}{2} \sum_{n=1}^{N} \log |\mathbf{S}_n^\ell|, \\ -\mathrm{KL}_{\mathbf{Z}}^* = \geq & \frac{1}{2} \sum_{n=1}^{N} \log |\mathbf{S}_n^L| + \varphi \left(\mathrm{tr}((\boldsymbol{\mu}^L)^\mathsf{T} \boldsymbol{\mu}^L) + \sum_{n=1}^{N} \mathrm{tr}(\mathbf{S}_n^L); \mathbf{I}_N, D_L \right),\end{aligned}$$

where $\psi_0^\ell = \mathrm{tr}((\Gamma^\ell)^{-1} \langle \mathbf{K}_{NN}^\ell \rangle_{q(\mathbf{X}^{\ell+1})})$, $\Psi_1^\ell = \langle \mathbf{K}_{NM}^\ell \rangle_{q(\mathbf{X}^{\ell+1})}$, and $\Psi_2^\ell = \langle \mathbf{K}_{MN}^\ell \mathbf{K}_{NM}^\ell \rangle_{q(\mathbf{X}^{\ell+1})}$.

5 Numerical Experiments

In this section, we compare our proposed deep Q-EP with deep GP (DGP Damianou & Lawrence, 2013), deep kernel learning with GP (DKL-GP Wilson et al., 2016), and deep sigma point process (DSPP Jankowiak et al., 2020b) using simulated and benchmark datasets. In simulations, deep Q-EP model manifests unique features in properly modeling inhomogeneous data with abrupt changes or sharp contrast. For benchmark regression and classification problems, deep Q-EP demonstrates superior or comparable numerical performance. In most cases, 2 layer structure is sufficient for deep Q-EP to have superior or comparable performance compared with deep GP, and DSPP. A large feature extracting neural network (DNN with structure $D_L - 1000 - 500 - 50 - D_0$) is employed before one GP layer for DKL-GP unless stated otherwise. The Matérn kernel ($\nu = 1.5$) is adopted for all the models with trainable hyperparameters (magnitude and correlation strength) and $q = 1$ is chosen in Q-EP and deep Q-EP models for handling data inhomogeneity. All the models are implemented in `GPyTorch` (Gardner et al., 2018) and the codes will be released.



(a) Shallow GP regression.

(b) Shallow Q-EP regression.

(c) Deep GP regression.

(d) Deep Q-EP regression.

(e) DKL-GP regression.

(f) DSPP regression.

Figure 2: Comparing deep Q-EP (2d) with cutting-edge deep models including deep GP (2c), DKL-GP (2e) and DSPP (2f) on modeling a 2d-output time series.

5.1 Time Series Regression

We first consider a simulated 2-dimensional time series from Li et al. (2023), one with step jumps and the other with sharp turnings, whose true trajectories are as follows:

$$\begin{aligned} u_{\mathrm{J}}(t) &= 1, \quad t\in[0,1]; \quad 0.5, \quad t\in(1,1.5]; \quad 2, \quad t\in(1.5,2]; \quad 0, \quad otherwise; \\ u_{\mathrm{T}}(t) &= 1.5t, \quad t\in[0,1]; \quad 3.5-2t, \quad t\in(1,1.5]; \quad 3t-4, \quad t\in(1.5,2]; \quad 0, \quad otherwise. \end{aligned}$$

We generate time series $\{\mathbf{y}_i\}_{i=1}^N$ by adding Gaussian noises to the true trajectories evaluated at $N = 100$ evenly spaced points $t_i \in [0,2]$, i.e., $\mathbf{y}_i^* = [u_J(t_i), u_T(t_i)]^{\mathsf{T}} + \boldsymbol{\varepsilon}_i$, $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(\mathbf{0}, \sigma^2\mathbf{I}_2)$, $with\ \sigma = 0.1$, $i = 1, \cdots, N$. Then we make prediction over 50 points evenly spread over $[0,2]$.

Abrupt changes exist in these time series have for either values or directions, hence pose challenges for standard GP as an L_2 penalty based regression method. As shown in Figure 2, results by both deep GP and deep Q-EP are comparatively better than their shallow (one-layer) versions. Among these models, deep Q-EP yields the most accurate prediction and the tightest uncertainty bound (refer to Table B.1) due to its L_1 regularization feature that is more suitable to capture these abrupt changes. The loss of (deep) Q-EP model may not be comparable to those for other models because they are based on different probability distributions, and yet it converges faster and and more stably than GP (and the other two benchmark deep probabilistic models), supporting its advantage in convergence (Remark 2). Both DKL-GP and DSPP suffer from slow convergence and unstable training. As seen in Table B.1 comparing mean of absolute error (MAE), standard deviation (STD) of variational distribution and coefficient of determination (R^2), their results possess larger standard errors from repeated experiments, even though few individual runs may yield better results than Deep Q-EP.

5.2 UCI Regression Dataset

Next, we test deep Q-EP on a series of benchmark regression datasets (Wilson et al., 2016; Jankowiak et al., 2020b) from UCI machine learning repository. They are selected to represent data at different scales. As in Table 1, for most cases, deep Q-EP demonstrates superior or comparable performance measured by testing data in terms of MAE (accuracy), STD (uncertainty) and NLL because the Q-EP prior provides crucial regularization for datasets where sparse regression is more appropriate. Note, the marginal likelihood (NLL) values are not comparable among different models (with distinct probability distributions) and only listed for reference. As the data volume increases, DNN feature extractor starts to catch up so that DKL-GP surpasses the vanilla deep Q-EP in the song dataset. Note, the GP component of DKL can be replaced with Q-EP to regularize the model. In our

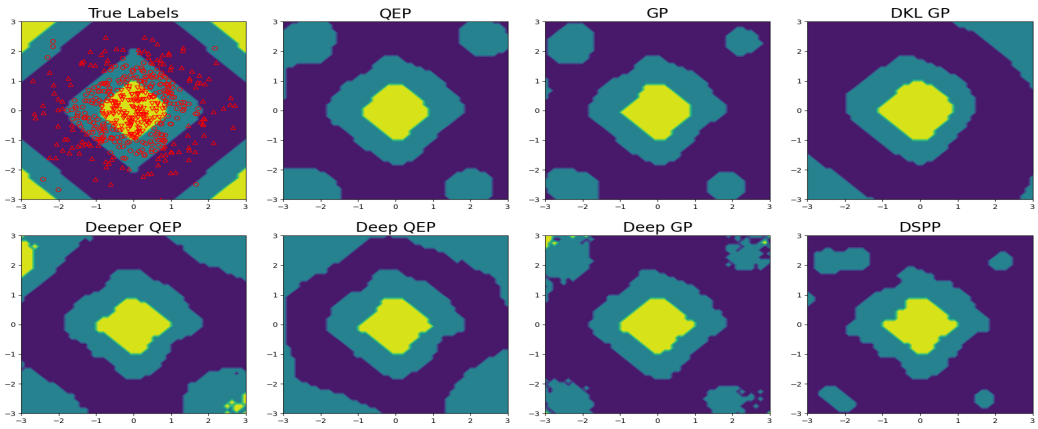


Figure 3: Comparing shallow (1-layer), deep (2-layer) and deeper (3-layer) Q-EPs with GP, deep GP, DKL-GP and DSPP on a classification problem defined on annular rhombus. Circles, upper and lower triangles label three classes in the training data.

experiment, the resulting DKL-QEP beats DKL-GP with (MAE, STD, NLL)$= (0.327, 0.009, 0.59)$ on the *protein* dataset. We will explore DKL-QEP further in Section 5.4.

Table 1: Regression on UCI datasets: mean of absolute error (MAE), standard deviation (STD) of variational distribution and negative logarithm of marginal likelihood (NLL) values by various deep models. Each result of the upper part is averaged over 10 experiments with different random seeds; values in the lower part are standard errors of these repeated experiments.

		Deep GP			Deep Q-EP			DKL-GP			DSPP		
Dataset	N, d	MAE	STD	NLL	MAE	STD	NLL	MAE	STD	NLL	MAE	STD	NLL
gas	2565, 128	0.19	0.06	0.4	**0.14**	**0.03**	-0.6	0.93	0.07	2.23	0.33	0.35	18.54
parkinsons	5875, 20	**8.17**	0.61	168.12	8.49	**0.38**	13	10.01	0.57	11.82	9.63	0.84	549.92
elevators	16599, 18	0.0639	0.014	-1.04	**0.0636**	**0.011**	-0.87	0.099	0.02	-0.29	0.09	0.09	0.52
protein	45730, 9	0.39	0.05	0.76	**0.35**	**0.014**	0.7	0.37	0.02	0.77	0.48	0.21	100.66
song	515345, 90	0.38	0.011	0.69	0.4	0.011	0.92	**0.35**	**0.008**	0.63	0.43	0.2	261.3
gas	2565, 128	0.07	0.02	0.16	0.03	0.01	0.24	0.36	0.02	1.04	0.24	0.13	22.06
parkinsons	5875, 20	1.38	0.16	97.06	1.74	0.11	3.42	1.55	0.25	4.89	1.51	0.29	349.22
elevators	16599, 18	3e-4	3e-4	7e-3	4e-4	3e-5	6e-3	0.06	0.05	1.32	0.02	0.02	0.64
protein	45730, 9	5e-3	4e-3	7e-3	5e-3	5e-4	0.01	0.09	6e-3	0.19	0.04	0.02	52.21
song	515345, 90	2e-3	1e-9	4e-3	0.04	3e-4	0.09	4e-3	1e-3	0.01	0.03	0.05	266.2

5.3 CLASSIFICATION

Consider a simulated classification problem with labels created on annular regions of a rhombus:

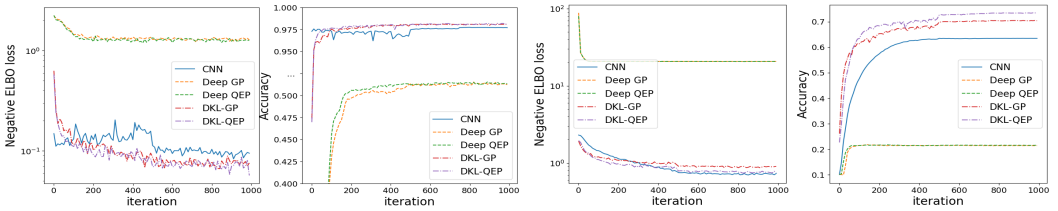
$$y_i = [\cos(0.4 * u * \pi \|\mathbf{x}_i\|_1)] + 1, \quad u \sim \text{Unif}[0,1], \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2), \quad i = 1, \cdots, N,$$

where $[x]$ rounds x to the nearest integer. We generate $N = 500$ random data points according to the formula which results in 3 classes' labels as illustrated in the leftmost panel of Figure 3. Note, the class regions have clear shapes with edges and are not simply connected. Q-EP and deep Q-EP are superior than their GP rivals in modeling such inhomogeneous data. Indeed, Figure 3 shows that even with small amount of data, Q-EP has better decision boundaries than GP and a 3-layer deeper Q-EP yields the best result closest to the truth among all the models. On the contrary, (deep) GP tends to yield round and over-smooth decision boundaries because of its L_2 nature. This is further illustrated in Figure B.1 with more fine details revealed by the logits. Note, it is understandable that none of these models characterizes the correct boundary around the corners due to the absence of data. Table B.2 compares their performance on testing data in terms of classification accuracy (ACC), area under ROC curve (AUC) and deep Q-EP achieves the highest accuracy.

We also compare deep Q-EP with other deep probabilistic models on several benchmark classification datasets with different sizes from UCI machine learning repository. Table 2 summarizes the comparison results in terms of ACC, AUC and NLL. Deep Q-EP still excels in most cases or has comparable performance, further supporting its advantage in the classification task.

Table 2: Classification on UCI datasets: accuracy (ACC), area under ROC curve (AUC) and negative logarithm of marginal likelihood (NLL) values by various deep models. Each result of the upper part is averaged over 10 experiments with different random seeds; values in the lower part are standard errors of these repeated experiments.

		Deep GP			Deep Q-EP			DKL-GP			DSPP		
Dataset	N, d, k	ACC	AUC	NLL	ACC	AUC	NLL	ACC	AUC	NLL	ACC	AUC	NLL
haberman	306, 3, 2	0.727	0.46	7.16	**0.732**	**0.505**	6.44	0.702	0.43	6.93	0.716	0.496	31.58
tic-tac-toe	957, 27, 2	0.971	0.52	67.57	**0.972**	0.53	48.69	0.922	**0.67**	15.8	0.736	0.5	430.25
car	1728, 21, 4	**0.99**	**0.9999**	501.9	0.983	0.999	1237.08	0.929	0.98	46.71	0.758	0.85	4.6e4
seismic	2583, 24, 2	0.931	0.28	11.75	**0.934**	0.44	10.69	0.931	0.44	9.43	0.849	**0.52**	3.7e4
nursery	12959, 27, 5	**0.9996**	**0.97**	2.1e5	**0.9996**	0.95	1.1e4	0.486	0.7	2.7e3	0.717	0.84	1.5e5
haberman	306, 3, 2	0.01	0.08	0.68	0.02	0.07	0.61	0.04	0.09	1	0.03	0.05	50.73
tic-tac-toe	957, 27, 2	0.02	0.08	20.68	0.04	0.37	13.25	0.19	0.15	4.22	0.23	0.44	73.5
car	1728, 21, 4	9e-3	2e-4	65.67	7e-3	1e-3	572.46	0.09	0.03	15.41	0.22	0.18	2.6e4
seismic	2583, 24, 2	0.002	0.02	1.25	0.0	0.1	0.9	0.006	0.08	1.48	0.27	0.13	1.7e4
nursery	12959, 27, 5	6e-8	0.04	4.5e4	6e-8	0.03	2.5e3	0.36	0.31	6e3	0.18	0.08	1e5



(a) MNIST digit classification.

(b) CIFAR-10 classification.

Figure 4: Comparing DKL-QEP and DKL-GP with CNN on two benchmark classification problems.

5.4 Image Classification

Finally, we test the proposed models on some benchmark image classification datasets, MNIST (60,000 training and 10,000 testing 28×28 handwritten digits) and CIFAR-10 (50,000 training and 10,000 testing 32×32 color images with 10 classes). As shown in Figure 4, while deep GP and deep Q-EP have mediocre classification accuracy, deep kernel learning (DKL Wilson et al., 2016) with CNN (common structure for these benchmarks) prefixed as a feature extractor works much better in both tasks. On MNIST dataset, DKL-GP has a 98.14% and DKL-QEP achieves a 98.19% test accuracy, improving vanilla CNN with 97.69% accuracy. On CIFAR-10, DKL-GP has accuracy 70% and DKL-QEP improves it to 73.4%, both having a good margin of advantage compared with vanilla CNN with 63.46%. Note, here we choose a relatively small CNN to demonstrate the improvement by adopting DKL with Q-EP even better than DKL-GP.

6 Conclusion

In this paper, we generalize Q-EP to deep Q-EP, which includes deep GP as a special case. Moreover, deep Q-EP inherits the flexible regularization controlled a parameter $q > 0$, which is advantageous in learning latent representations and modeling data inhomogeneity. We first generalize Bayesian GP-LVM to Bayesian QEP-LVM (as shallow Q-EP layer) and develop the variational inference for it. Then we stack multiple shallow Q-EP layer to build the deep Q-EP model. The novel deep model demonstrates numerical benefits in various learning tasks and can be combined with neural network for better characterizing complex latent representations in different data applications.

As common in GP and NN models, we do observe multi-modality of the posterior distributions, especially in the hyper-parameter spaces. Sub-optimal solutions can appear in the stochastic training process. These issues can be alleviated by dispersed or diversified initialization, or with adaptive training schedulers. One potential application of deep Q-EP is the inverse learning, similarly as done by deep GP (Jin et al., 2017; Abraham & Deo, 2023). Theory of the contraction properties (Finocchio & Schmidt-Hieber, 2023) is also an interesting research direction.

REFERENCES

Kweku Abraham and Neil Deo. Deep gaussian process priors for bayesian inference in nonlinear inverse problems. 12 2023. URL https://arxiv.org/pdf/2312.14294.pdf.

Sergios Agapiou, Masoumeh Dashti, and Tapio Helin. Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli*, 27(3):1616 – 1642, 2021. doi: 10.3150/20-BEJ1285. URL https://doi.org/10.3150/20-BEJ1285.

Laurence Aitchison, Adam Yang, and Sebastian W Ober. Deep kernel processes. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 130–140. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/aitchison21a.html.

Kenneth Blomqvist, Samuel Kaski, and Markus Heinonen. Deep convolutional gaussian processes. In *Machine Learning and Knowledge Discovery in Databases*, pp. 582â–597. Springer International Publishing, 2020. ISBN 9783030461478. doi: 10.1007/978-3-030-46147-8_35. URL http://dx.doi.org/10.1007/978-3-030-46147-8_35.

Ismaël Castillo and Thibault Randrianarisoa. Deep horseshoe gaussian processes. 03 2024. URL https://arxiv.org/pdf/2403.01737.pdf.

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar (eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL https://proceedings.mlr.press/v31/damianou13a.html.

Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, may 2012. doi: 10.3934/ipi.2012.6.183. URL https://doi.org/10.3934%2Fipi.2012.6.183.

Vincent Dutordoir, Mark van der Wilk, Artem Artemev, and James Hensman. Bayesian image classification with deep convolutional gaussian processes. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1529–1539. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/dutordoir20a.html.

K. Fang and Y.T. Zhang. *Generalized Multivariate Analysis*. Science Press, 1990. ISBN 9780387176512. URL https://books.google.com/books?id=WibvAAAAMAAJ.

Gianluca Finocchio and Johannes Schmidt-Hieber. Posterior contraction for deep gaussian process priors. *Journal of Machine Learning Research*, 24(66):1–49, 2023. URL http://jmlr.org/papers/v24/21-0556.html.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/27e8e17134dd7083b050476733207ea1-Paper.pdf.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 351–360, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/hensman15.html.

A. P. Dawid J. M. Bernardo, J. O. Berger and A. F. M. Smith. Regression and classification using gaussian process priors. *Bayesian Statistics*, 6:475–501, 1998. doi: 130.203.136.95/viewdoc/summary?doi=10.1.1.156.1910. URL http://130.203.136.95/viewdoc/summary?doi=10.1.1.156.1910.

Martin Jankowiak, Geoff Pleiss, and Jacob Gardner. Parametric Gaussian process regressors. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4702–4712. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/jankowiak20a.html.

Martin Jankowiak, Geoff Pleiss, and Jacob Gardner. Deep sigma point processes. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 789–798. PMLR, 03–06 Aug 2020b. URL https://proceedings.mlr.press/v124/jankowiak20a.html.

Ming Jin, Andreas C. Damianou, P. Abbeel, and Costas J. Spanos. Inverse reinforcement learning via deep gaussian process. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume abs/1512.08065, 2017. URL https://api.semanticscholar.org/CorpusID:4670729.

Mark E. Johnson. *Multivariate Statistical Simulation*, chapter 6 Elliptically Contoured Distributions, pp. 106–124. Probability and Statistics. John Wiley & Sons, Ltd, 1987. ISBN 9781118150740. doi: https://doi.org/10.1002/9781118150740.ch6. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118150740.ch6.

Andrew Jones, F. William Townes, Didong Li, and Barbara E. Engelhardt. Alignment of spatial genomics data using deep gaussian processes. *Nature Methods*, 20(9):1379–1387, August 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01972-2. URL http://dx.doi.org/10.1038/s41592-023-01972-2.

Tomasz J. Kozubowski, Krzysztof Podgórski, and Igor Rychlik. Multivariate generalized laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113:59–72, 2013. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2012.02.010. URL https://www.sciencedirect.com/science/article/pii/S0047259X12000516. Special Issue on Multivariate Distribution Theory in Memory of Samuel Kotz.

Shiwei Lan, Mirjeta Pasha, Shuyi Li, and Weining Shen. Spatiotemporal besov priors for bayesian inverse problems. 06 2023. URL https://arxiv.org/pdf/2306.16378.pdf.

Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant bayesian inversion and besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009.

Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/9657c1fffd38824e5ab0472e022e577e-Paper.pdf.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(60):1783–1816, 2005. URL http://jmlr.org/papers/v6/lawrence05a.html.

Neil D. Lawrence and Andrew J. Moore. Hierarchical gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, ICML 2007. ACM, June 2007. doi: 10.1145/1273496.1273557. URL http://dx.doi.org/10.1145/1273496.1273557.

Shuyi Li, Michael O'Connor, and Shiwei Lan. Bayesian learning via q-exponential process. In *Proceedings of the 37th Conference on Neural Information Processing Systems*. NeurIPS, 12 2023. URL https://arxiv.org/pdf/2210.07987.pdf. arxiv:2210.07987.

Yikuan Li, Shishir Rao, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Gholamreza Salimi-Khorshidi, Mohammad Mamouei, Thomas Lukasiewicz, and Kazem Rahimi. Deep bayesian gaussian processes for uncertainty estimation in electronic health records. *Scientific Reports*, 11(1), October 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-00144-6. URL http://dx.doi.org/10.1038/s41598-021-00144-6.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer New York, 1996. ISBN 9781461207450. doi: 10.1007/978-1-4612-0745-0. URL http://dx.doi.org/10.1007/978-1-4612-0745-0.

Bernt Øksendal. *Stochastic Differential Equations*. Springer Berlin Heidelberg, 2003. doi: 10.1007/978-3-642-14394-6. URL https://doi.org/10.1007%2F978-3-642-14394-6.

Luis A. Ortega, Simon Rodriguez Santana, and Daniel Hern
'andez-Lobato. Deep variational implicit processes. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8aeSJNbmbQq.

Krzysztof Podgórski and Jörg Wegener. Estimation for stochastic models driven by laplace motion. *Communications in Statistics - Theory and Methods*, 40(18):3281–3302, sep 2011. doi: 10.1080/03610926.2010.499051. URL https://doi.org/10.1080%2F03610926.2010.499051.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. doi: 10.7551/mitpress/3206.001.0001. URL https://doi.org/10.7551%2Fmitpress%2F3206.001.0001.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8208974663db80265e9bfe7b222dcb18-Paper.pdf.

Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 09 1999. ISSN 1369-7412. doi: 10.1111/1467-9868.00196. URL https://doi.org/10.1111/1467-9868.00196.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL https://proceedings.mlr.press/v5/titsias09a.html.

Michalis Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 844–851, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/titsias10a.html.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128(7):1867–1888, Jul 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01303-4. URL https://link.springer.com/content/pdf/10.1007/s11263-020-01303-4.pdf.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 370–378, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/wilson16.html.

Supplement Document for "Deep Q-Exponential Processes"

A Computation of Variational Lower Bounds

A.1 Shallow Q-EP

The variational lower bound for the log-evidence is

$$\log p(\mathbf{Y}) \geq \mathcal{L}(q) := \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X} = \tilde{\mathcal{L}}(q) - \mathrm{KL}(q(\mathbf{X})\|p(\mathbf{X})),$$

where the first term $\tilde{\mathcal{L}}(q) = \int q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{X}) d\mathbf{X}$ is intractable and hence difficult to bound.

A.1.1 Lower bound for the marginal likelihood

To address such intractability issue and speed up the computation, sparse variational approximation (Titsias, 2009; Lawrence & Moore, 2007) is adopted by introducing a set of inducing points $\tilde{\mathbf{X}} \in \mathbb{R}^{M\times Q}$ with their function values $\mathbf{U} = [f_1(\tilde{\mathbf{X}}), \cdots, f_D(\tilde{\mathbf{X}})] \in \mathbb{R}^{M\times D}$. Hence the marginal likelihood $p(\mathbf{Y}|\mathbf{X})$ defined in (6) can be augmented to the following joint distribution each being a $\mathrm{q}-\mathrm{ED}$:

$$p(\mathbf{Y}|\mathbf{X}) \propto p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \tilde{\mathbf{X}})p(\mathbf{U}|\tilde{\mathbf{X}}),$$

where we have $\mathrm{vec}(\mathbf{U})|\tilde{\mathbf{X}} \sim \mathrm{q}-\mathrm{ED}(\mathbf{0}, \mathbf{I}_D \otimes \mathbf{K}_{MM})$ and the conditional distribution

$$\mathrm{vec}(\mathbf{F})|\mathbf{U}, \mathbf{X}, \tilde{\mathbf{X}} \sim \mathrm{q}-\mathrm{ED}(\mathrm{vec}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{U}), \mathbf{I}_D \otimes (\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN})). \quad (8)$$

The inducing points $\tilde{\mathbf{X}}$ are regarded as variational parameters and hence they are dropped from the following probability expressions. We then approximate $p(\mathbf{F}, \mathbf{U}|\mathbf{X}) \propto p(\mathbf{F}|\mathbf{U}, \mathbf{X})p(\mathbf{U})$ with $q(\mathbf{F}, \mathbf{U}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X})q(\mathbf{U})$ in another variational Bayes as follows

$$\begin{aligned} \log p(\mathbf{Y}|\mathbf{X}) &\geq \int q(\mathbf{F}, \mathbf{U}) \log \frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{U}, \mathbf{X})p(\mathbf{U})}{q(\mathbf{F}, \mathbf{U})} d\mathbf{F} d\mathbf{U} \\ &= \int p(\mathbf{F}|\mathbf{U})q(\mathbf{U}) d\mathbf{U} \log p(\mathbf{Y}|\mathbf{F}) d\mathbf{F} + \int q(\mathbf{U}) \log \frac{p(\mathbf{U})}{q(\mathbf{U})} d\mathbf{U}. \end{aligned} \quad (9)$$

Different from Titsias (2009); Titsias & Lawrence (2010) using the variational calculus, (SVGP Hensman et al., 2015) computes the marginal likelihood ELBO (9) in two stages. Instead of the variational free form, we follow Hensman et al. (2015) to use the variational distribution for $\mathbf{U}$ of the following format conjugate to $p(\mathbf{F}|\mathbf{U})$:

$$q(\mathbf{U}) \sim \mathrm{q}-\mathrm{ED}(\mathbf{M}, \mathrm{diag}(\{\boldsymbol{\Sigma}_d\})). \quad (10)$$

Noticing that $\mathbf{F}|\mathbf{U}$ follows a conditional q-exponential (8), we can obtain the variational distribution of $\mathbf{F}$, $q(\mathbf{F})$, by marginalizing $\mathbf{U}$ out as follows

$$\begin{aligned} q(\mathbf{F}) &= \int q(\mathbf{F}, \mathbf{U}) d\mathbf{U} = \int p(\mathbf{F}|\mathbf{U})q(\mathbf{U}) d\mathbf{U} \\ \sim &\mathrm{q}-\mathrm{ED}(\mathrm{vec}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{M}), \\ &\mathbf{I}_D \otimes (\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}) + \mathrm{diag}(\{\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_d\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\})). \end{aligned}$$

Therefore, the variational lower bound of the marginal likelihood (9) becomes

$$\log p(\mathbf{Y}|\mathbf{X}) \geq \langle \log p(\mathbf{Y}|\mathbf{F}) \rangle_{q(\mathbf{F})} - \mathrm{KL}(q(\mathbf{U})\|p(\mathbf{U})).$$

Note, $\log p(\mathbf{Y}|\mathbf{F})$ is not a random quadratic form in general and hence the expectation in the first term has no explicit formula. Denote by $\log p(\mathbf{Y}|\mathbf{F}) = \varphi(r(\mathbf{Y}, \mathbf{F}))$, where $\varphi(r) := \frac{DN}{2}\log\beta + \frac{ND}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$ is convex for $q \in (0, 2]$, and $r(\mathbf{Y}, \mathbf{F}) = \mathrm{vec}(\mathbf{Y} - \mathbf{F})^{\mathsf{T}}(\beta^{-1}\mathbf{I}_{ND})^{-1}\mathrm{vec}(\mathbf{Y} - \mathbf{F}) = \beta\mathrm{tr}((\mathbf{Y} - \mathbf{F})(\mathbf{Y} - \mathbf{F})^{\mathsf{T}})$ is a quadratic form of random variable $\mathbf{Y}$. Therefore, by Jensen's inequality, we can bound from below as

$$\langle \log p(\mathbf{Y}|\mathbf{F}) \rangle_{q(\mathbf{F})} = \langle \varphi(r(\mathbf{Y}, \mathbf{F})) \rangle_{q(\mathbf{F})} \geq \varphi(\langle r(\mathbf{Y}, \mathbf{F}) \rangle_{q(\mathbf{F})}).$$

where we can calculate the expectation of the quadratic form $r(\mathbf{Y}, \mathbf{F})$ as

$$\begin{aligned}\langle r(\mathbf{Y}, \mathbf{F})\rangle_{q(\mathbf{F})} =& r(\mathbf{Y}, \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{M}) + \beta D \mathrm{tr}(\mathbf{K}_{NN} - \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}) \\ &+ \beta \sum_{d=1}^{D} \mathrm{tr}(\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_d\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}).\end{aligned}$$

Denote by $h(\mathbf{Y}, \mathbf{X}) = \langle\langle \log p(\mathbf{Y}|\mathbf{F})\rangle_{q(\mathbf{F})}\rangle_{q(\mathbf{X})}$. Then we solve the intractable expectation by another Jensen's inequality

$$h(\mathbf{Y}, \mathbf{X}) \geq \varphi(\langle\langle r(\mathbf{Y}, \mathbf{F})\rangle_{q(\mathbf{F})}\rangle_{q(\mathbf{X})}) =: h^*(\mathbf{Y}, \mathbf{X}).$$

Define $\psi_0 = \mathrm{tr}(\langle \mathbf{K}_{NN}\rangle_{q(\mathbf{X})})$, $\Psi_1 = \langle \mathbf{K}_{NM}\rangle_{q(\mathbf{X})}$, and $\Psi_2 = \langle \mathbf{K}_{MN}\mathbf{K}_{NM}\rangle_{q(\mathbf{X})}$. Further we calculate the expectations of quadratic terms similarly

$$\begin{aligned}\langle\langle r(\mathbf{Y}, \mathbf{F})\rangle_{q(\mathbf{F})}\rangle_{q(\mathbf{X})} =& \langle r(\mathbf{Y}, \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{M})\rangle_{q(\mathbf{X})} + \beta D[\psi_0 - \mathrm{tr}(\mathbf{K}_{MM}^{-1}\Psi_2)] \\ &+ \beta \sum_{d=1}^{D} \mathrm{tr}(\mathbf{K}_{MM}^{-1}\boldsymbol{\Sigma}_d\mathbf{K}_{MM}^{-1}\Psi_2), \\ \langle r(\mathbf{Y}, \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{M})\rangle_{q(\mathbf{X})} =& r(\mathbf{Y}, \Psi_1\mathbf{K}_{MM}^{-1}\mathbf{M}) + \beta \mathrm{tr}(\mathbf{M}^\mathsf{T}\mathbf{K}_{MM}^{-1}(\Psi_2 - \Psi_1^\mathsf{T}\Psi_1)\mathbf{K}_{MM}^{-1}\mathbf{M}).\end{aligned} \tag{11}$$

We also need to compute the K-L divergence $\mathrm{KL}_{\mathbf{U}} := \mathrm{KL}(q(\mathbf{U})\|p(\mathbf{U}))$

$$\mathrm{KL}_{\mathbf{U}} = \int q(\mathbf{U}) \log q(\mathbf{U}) d\mathbf{U} - \int q(\mathbf{U}) \log p(\mathbf{U}) d\mathbf{U} = -\mathcal{H}_q(\mathbf{U}) - \langle \log p(\mathbf{U})\rangle_{q(\mathbf{U})}.$$

Denote by $r = \mathrm{vec}^\mathsf{T}(\mathbf{U} - \mathbf{M})^\mathsf{T} \mathrm{diag}(\{\boldsymbol{\Sigma}_d\})^{-1}\mathrm{vec}^\mathsf{T}(\mathbf{U} - \mathbf{M})$. Then $\log q(\mathbf{U}) = -\frac{1}{2}\sum_{d=1}^{D} \log|\boldsymbol{\Sigma}_d| + \frac{MD}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$. From (Proposition A.1. of Li et al., 2023) we know that $r^{\frac{q}{2}} \sim \chi^2(MD)$. Therefore

$$\begin{aligned}\mathcal{H}_q(\mathbf{U}) &= \frac{1}{2}\sum_{d=1}^{D} \log|\boldsymbol{\Sigma}_d| + \frac{MD}{2}\left(\frac{q}{2} - 1\right)\frac{2}{q}\mathcal{H}(\chi^2(MD)) + \frac{MD}{2} \\ &= \frac{1}{2}\sum_{d=1}^{D} \log|\boldsymbol{\Sigma}_d| + \frac{MD}{2}\left(1 - \frac{2}{q}\right)\left[\frac{MD}{2} + \log\left(2\Gamma\left(\frac{MD}{2}\right)\right) + \left(1 - \frac{MD}{2}\right)\psi\left(\frac{MD}{2}\right)\right] + \frac{MD}{2}.\end{aligned}$$

Denote by $\varphi_0(r) := -\frac{D}{2}\log|\mathbf{K}_{MM}| + \frac{MD}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$. Then by Jensen's inequality again

$$\begin{aligned}\langle \log p(\mathbf{U})\rangle_{q(\mathbf{U})} &= \langle \varphi_0(\mathrm{tr}(\mathbf{U}^\mathsf{T}\mathbf{K}_{MM}^{-1}\mathbf{U}))\rangle_{q(\mathbf{U})} \geq \varphi_0(\langle \mathrm{tr}(\mathbf{U}^\mathsf{T}\mathbf{K}_{MM}^{-1}\mathbf{U})\rangle_{q(\mathbf{U})}), \\ \langle \mathrm{tr}(\mathbf{U}^\mathsf{T}\mathbf{K}_{MM}^{-1}\mathbf{U})\rangle_{q(\mathbf{U})} &= \mathrm{tr}(\mathbf{M}^\mathsf{T}\mathbf{K}_{MM}^{-1}\mathbf{M}) + \sum_{d=1}^{D} \mathrm{tr}(\boldsymbol{\Sigma}_d\mathbf{K}_{MM}^{-1}).\end{aligned}$$

The elements of ψ_0, Ψ_1 and Ψ_2 can be computed as

$$\begin{aligned}\psi_0^n &= \int k(\mathbf{x}_n, \mathbf{x}_n)\mathrm{q{-}ED}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{S}_n)d\mathbf{x}_n, \\ (\Psi_1)_{nm} &= \int k(\mathbf{x}_n, \mathbf{z}_m)\mathrm{q{-}ED}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{S}_n)d\mathbf{x}_n, \\ (\Psi_2^n)_{mm'} &= \int k(\mathbf{x}_n, \mathbf{z}_m)k(\mathbf{z}_{m'}, \mathbf{x}_n)\mathrm{q{-}ED}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{S}_n)d\mathbf{x}_n.\end{aligned}$$

With ARD SE kernel (5), we have $\psi_0 = N\alpha^{-1}$. While the integration in Ψ_1 and Ψ_2 is intractable in general, we can compute them using Monte Carlo approximation. Alternatively, we approximate

$$\begin{aligned}(\Psi_1)_{nm} &\approx \alpha^{-1}\exp\left\{-\frac{1}{2}\langle(\mathbf{x}_n - \mathbf{z}_m)^\mathsf{T}\mathrm{diag}(\boldsymbol{\gamma})(\mathbf{x}_n - \mathbf{z}_m)\rangle_{q(\mathbf{x}_n)}\right\} \\ &= \alpha^{-1}\exp\left\{-\frac{1}{2}[(\boldsymbol{\mu}_n - \mathbf{z}_m)^\mathsf{T}\mathrm{diag}(\boldsymbol{\gamma})(\boldsymbol{\mu}_n - \mathbf{z}_m) + \mathrm{tr}(\mathrm{diag}(\boldsymbol{\gamma})\mathbf{S}_n)]\right\}, \\ (\Psi_2^n)_{mm'} &\approx \alpha^{-2}\exp\left\{-\frac{1}{2}\sum_{\tilde{m}=m,m'}(\boldsymbol{\mu}_n - \mathbf{z}_{\tilde{m}})^\mathsf{T}\mathrm{diag}(\boldsymbol{\gamma})(\boldsymbol{\mu}_n - \mathbf{z}_{\tilde{m}})) + \mathrm{tr}(\mathrm{diag}(\boldsymbol{\gamma})\mathbf{S}_n)\right\}.\end{aligned}$$

If we use the ARD linear form, $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T} \operatorname{diag}(\boldsymbol{\gamma})\mathbf{x}'$, then we have

$$\psi_0^n = \operatorname{tr}(\operatorname{diag}(\boldsymbol{\gamma})(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\mathsf{T} + \mathbf{S}_n)), \quad (\Psi_1)_{nm} = \boldsymbol{\mu}_n^\mathsf{T} \operatorname{diag}(\boldsymbol{\gamma})\mathbf{z}_m,$$
$$(\Psi_2^n)_{mm'} = \mathbf{z}_m^\mathsf{T} \operatorname{diag}(\boldsymbol{\gamma})(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\mathsf{T} + \mathbf{S}_n) \operatorname{diag}(\boldsymbol{\gamma})\mathbf{z}_{m'}.$$

A.1.2 Lower bound for the K-L divergence added terms

Lastly, we need to compute the K-L divergence

$$\mathrm{KL}(q(\mathbf{X})\|p(\mathbf{X})) = \int q(\mathbf{X}) \log q(\mathbf{X}) d\mathbf{X} - \int q(\mathbf{X}) \log p(\mathbf{X}) d\mathbf{X} = -\mathcal{H}_q(\mathbf{X}) - \langle \log p(\mathbf{X}) \rangle_{q(\mathbf{X})}.$$

Denote by $r = \operatorname{vec}(\mathbf{X} - \boldsymbol{\mu})^\mathsf{T} \operatorname{diag}(\{\mathbf{S}_n\})^{-1} \operatorname{vec}(\mathbf{X} - \boldsymbol{\mu})$. Then $\log q(\mathbf{X}) = -\frac{1}{2}\sum_{n=1}^N \log|\mathbf{S}_n| + \frac{NQ}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$. From (Proposition A.1. of Li et al., 2023) we know that $r^{\frac{q}{2}} \sim \chi^2(NQ)$. Therefore

$$\mathcal{H}_q(\mathbf{X}) = \frac{1}{2}\sum_{n=1}^N \log|\mathbf{S}_n| + \frac{NQ}{2}\left(\frac{q}{2} - 1\right)\frac{2}{q}\mathcal{H}(\chi^2(NQ)) + \frac{NQ}{2}$$
$$= \frac{1}{2}\sum_{n=1}^N \log|\mathbf{S}_n| + \frac{NQ}{2}\left(1 - \frac{2}{q}\right)\left[\frac{NQ}{2} + \log\left(2\Gamma\left(\frac{NQ}{2}\right)\right) + \left(1 - \frac{NQ}{2}\right)\psi\left(\frac{NQ}{2}\right)\right] + \frac{NQ}{2}.$$

Denote by $\varphi_0(r) := \frac{NQ}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$. Then similarly by Jensen's inequality

$$\langle \log p(\mathbf{X}) \rangle_{q(\mathbf{X})} = \langle \varphi_0(\operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{X})) \rangle_{q(\mathbf{X})} \geq \varphi_0(\langle \operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) \rangle_{q(\mathbf{X})}),$$
$$\langle \operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) \rangle_{q(\mathbf{X})} = \operatorname{tr}(\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\mu}) + \sum_{n=1}^N \operatorname{tr}(\mathbf{S}_n).$$

A.2 Deep Q-EP

We only consider the hierarchy of two QEP-LVMs because the general L-layers follows by induction:

$$\begin{aligned} y_{nd} &= f_d^Y(\mathbf{x}_n) + \varepsilon_{nd}^Y, \quad d = 1, \cdots, D, \quad \mathbf{x}_n \in \mathbb{R}^Q, \\ x_{nq} &= f_q^X(\mathbf{z}_n) + \varepsilon_{nq}^X, \quad q = 1, \cdots, Q, \quad \mathbf{z}_n \in \mathbb{R}^{Q_Z}, \end{aligned} \tag{12}$$

where $\boldsymbol{\varepsilon}^Y \sim \mathrm{q{-}ED}(\mathbf{0}, \Gamma^Y)$, $\boldsymbol{\varepsilon}^X \sim \mathrm{q{-}ED}(\mathbf{0}, \Gamma^X)$, $f^Y \sim \mathrm{q{-}}\mathcal{EP}(0, k^Y)$ and $f^X \sim \mathrm{q{-}}\mathcal{EP}(0, k^X)$. Consider the prior $\mathbf{Z} \sim \mathrm{q{-}ED}(\mathbf{0}, \mathbf{I}_{NQ_Z})$. The variational inference for $p(\mathbf{Z}|\mathbf{Y})$ requires maximizing the following ELBO

$$\log p(\mathbf{Y}) \geq \mathcal{L}(\mathcal{Q}) := \int_{\mathbf{Z},\mathbf{F}^X,\mathbf{X},\mathbf{F}^Y} \mathcal{Q} \log \frac{p(\mathbf{Y}, \mathbf{F}^Y, \mathbf{X}, \mathbf{F}^X, \mathbf{Z})}{\mathcal{Q}}, \tag{13}$$

where the joint probability can be decomposed

$$p(\mathbf{Y}, \mathbf{F}^Y, \mathbf{X}, \mathbf{F}^X, \mathbf{Z}) = p(\mathbf{Y}|\mathbf{F}^Y)p(\mathbf{F}^Y|\mathbf{X}) \cdot p(\mathbf{X}|\mathbf{F}^X)p(\mathbf{F}^X|\mathbf{Z})p(\mathbf{Z})$$

Similarly as in Section 3.1, sparse variational approximation (Titsias & Lawrence, 2010) is adopted to introduce inducing points $\tilde{\mathbf{X}} \in \mathbb{R}^{M\times Q}, \tilde{\mathbf{Z}} \in \mathbb{R}^{M\times Q_Z}$ with associated function values $\mathbf{U}^Y \in \mathbb{R}^{M\times D}, \mathbf{U}^X \in \mathbb{R}^{M\times Q}$ respectively. Hence the augmented probability replaces the joint probability:

$$\begin{aligned} p(\mathbf{Y}, \mathbf{F}^Y, \mathbf{X}, \mathbf{F}^X, \mathbf{Z}, \mathbf{U}^Y, \mathbf{U}^X) =& p(\mathbf{Y}|\mathbf{F}^Y)p(\mathbf{F}^Y|\mathbf{U}^Y, \mathbf{X})p(\mathbf{U}^Y|\tilde{\mathbf{X}}) \cdot \\ & p(\mathbf{X}|\mathbf{F}^X)p(\mathbf{F}^X|\mathbf{U}^X, \mathbf{Z})p(\mathbf{U}^X|\tilde{\mathbf{Z}})p(\mathbf{Z}), \end{aligned}$$

where $\mathbf{F}^Y$ and $\mathbf{U}^Y$ are drawn from the same Q-EP; and similarly are $\mathbf{F}^X$ and $\mathbf{U}^X$. Now we specify the approximation distribution as

$$\mathcal{Q} = p(\mathbf{F}^Y|\mathbf{U}^Y, \mathbf{X})q(\mathbf{U}^Y)q(\mathbf{X}) \cdot p(\mathbf{F}^X|\mathbf{U}^X, \mathbf{Z})q(\mathbf{U}^X)q(\mathbf{Z}).$$

and choose $q(\mathbf{U}^Y)$ and $q(\mathbf{U}^X)$, and $q(\mathbf{X})$ and $q(\mathbf{Z})$ to be uncorrelated $\mathrm{q{-}ED}$'s:

$$q(\mathbf{U}^Y) \sim \mathrm{q{-}ED}(\mathbf{M}^Y, \operatorname{diag}(\{\boldsymbol{\Sigma}_d^Y\})), \quad q(\mathbf{U}^X) \sim \mathrm{q{-}ED}(\mathbf{M}^X, \operatorname{diag}(\{\boldsymbol{\Sigma}_d^X\})),$$
$$q(\mathbf{X}) \sim \mathrm{q{-}ED}(\boldsymbol{\mu}^X, \operatorname{diag}(\{\mathbf{S}_n^X\})), \quad q(\mathbf{Z}) \sim \mathrm{q{-}ED}(\boldsymbol{\mu}^Z, \operatorname{diag}(\{\mathbf{S}_n^Z\})).$$

Then the ELBO (13) becomes

$$\mathcal{L}(\mathcal{Q}) := \int_{\mathbf{Z},\mathbf{U}^X,\mathbf{F}^X,\mathbf{X},\mathbf{U}^Y,\mathbf{F}^Y} \mathcal{Q} \log \frac{p(\mathbf{Y}|\mathbf{F}^Y)p(\mathbf{U}^Y)p(\mathbf{X}|\mathbf{F}^X)p(\mathbf{U}^X)p(\mathbf{Z})}{q(\mathbf{U}^Y)q(\mathbf{X})q(\mathbf{U}^X)q(\mathbf{Z})}$$
$$= h(\mathbf{Y},\mathbf{X}) - \mathrm{KL}_{\mathbf{U}^Y} + h(\mathbf{X},\mathbf{Y}) - \mathrm{KL}_{\mathbf{U}^X} + \mathcal{H}_q(\mathbf{X}) - \mathrm{KL}_{\mathbf{Z}},$$

where we have

$$h(\mathbf{Y},\mathbf{X}) = \left\langle \log p(\mathbf{Y}|\mathbf{F}^Y) \right\rangle_{q(\mathbf{F}^Y)q(\mathbf{X})}, \quad h(\mathbf{X},\mathbf{Z}) = \left\langle \log p(\mathbf{X}|\mathbf{F}^X) \right\rangle_{q(\mathbf{F}^X)q(\mathbf{X})q(\mathbf{Z})}.$$

Note, $h(\mathbf{Y},\mathbf{X}) \geq h^*(\mathbf{Y},\mathbf{X})$ is the same as in the bound (7) for Bayesian LVM. However, $h(\mathbf{X},\mathbf{Z})$ has an extra integration with respect to $q(\mathbf{X})$. Replacing $\mathbf{X}$ with $\mathbf{Z}$ and $\mathbf{Y}$ with $\mathbf{X}$ in (11), we compute

$$\langle r(\mathbf{X}, \Psi_1(\mathbf{K}_{MM}^X)^{-1}\mathbf{U}^X)\rangle_{q(\mathbf{X})} = r(\boldsymbol{\mu}^X, \Psi_1(\mathbf{K}_{MM}^X)^{-1}\mathbf{U}^X) + \mathrm{tr}((\mathbf{I}_D \otimes (\Gamma^X)^{-1})\,\mathrm{diag}(\{\mathbf{S}_n^X\})).$$

Therefore we have a updated bound for $h(\mathbf{X},\mathbf{Z}) \geq h^*(\mathbf{X},\mathbf{Z}) = \varphi(r_{\boldsymbol{\mu}^X}; \Gamma^X, Q)$, where

$$\begin{aligned} r_{\boldsymbol{\mu}^X} =& r(\boldsymbol{\mu}^X, \Psi_1(\mathbf{K}_{MM}^X)^{-1}\mathbf{M}^X) + \mathrm{tr}((\mathbf{M}^X)^\mathsf{T}(\mathbf{K}_{MM}^X)^{-1}(\Psi_2^X - \Psi_1^\mathsf{T}(\Gamma^X)^{-1}\Psi_1)(\mathbf{K}_{MM}^X)^{-1}\mathbf{M}^X) \\ &+ Q[\psi_0 - \mathrm{tr}((\mathbf{K}_{MM}^X)^{-1}\Psi_2^X)] + \sum_{d=1}^{Q} \mathrm{tr}((\mathbf{K}_{MM}^X)^{-1}\boldsymbol{\Sigma}_d^X(\mathbf{K}_{MM}^X)^{-1}\Psi_2^X) \\ &+ \mathrm{tr}((\mathbf{I}_Q \otimes (\Gamma^X)^{-1})\,\mathrm{diag}(\{\mathbf{S}_n^X\})). \end{aligned}$$

Finally, we have

$$\mathcal{H}_q(\mathbf{X}) \geq \frac{1}{2}\sum_{n=1}^{N} \log|\mathbf{S}_n^X|, \; -\mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z})) \geq \frac{1}{2}\sum_{n=1}^{N} \log|\mathbf{S}_n^Z| + \varphi_0(\mathrm{tr}((\boldsymbol{\mu}^Z)^\mathsf{T}\boldsymbol{\mu}^Z) + \sum_{n=1}^{N}\mathrm{tr}(\mathbf{S}_n^Z)),$$

where $\varphi_0(r) := \frac{NQ_Z}{2}\left(\frac{q}{2} - 1\right)\log r - \frac{1}{2}r^{\frac{q}{2}}$.

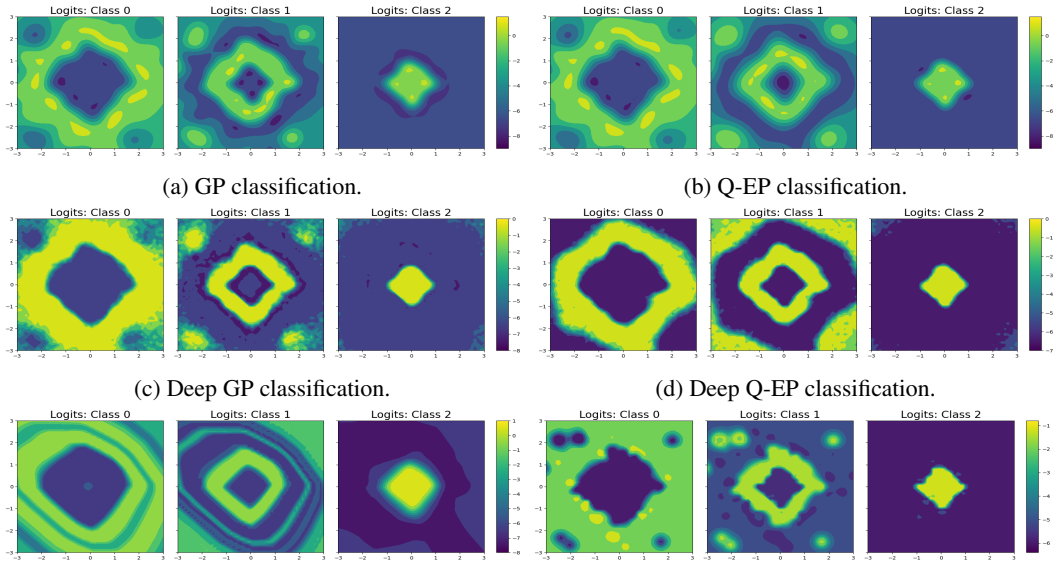
B More Numerical Results

B.1 Time Series

Table B.1: Regression on simulated time series: mean of absolute error (MAE), standard deviation (STD) of variational distribution, coefficient of determination (R^2), negative logarithm of marginal likelihood (NLL) and running time by various deep models. Each result of the upper part is averaged over 10 experiments with different random seeds; values after $\pm$ are standard errors of these repeated experiments.

Model	MAE	STD	R^2	NLL	time
Deep GP	0.058 $\pm$ 0.040	0.180 $\pm$ 0.051	0.951 $\pm$ 0.061	-1.437 $\pm$ 0.615	45.310 $\pm$ 0.915
Deep QEP	**0.055** $\pm$ 0.009	**0.111** $\pm$ 0.005	**0.965** $\pm$ 0.012	-1.790 $\pm$ 0.183	45.647 $\pm$ 1.449
DKL-GP	0.329 $\pm$ 0.344	0.170 $\pm$ 0.046	-0.284 $\pm$ 1.696	9.536 $\pm$ 15.014	13.992 $\pm$ 0.736
DSPP	0.216 $\pm$ 0.052	0.223 $\pm$ 0.057	0.728 $\pm$ 0.101	12.523 $\pm$ 10.026	40.953 $\pm$ 1.109

B.2 Classification



(a) GP classification. (b) Q-EP classification.

(c) Deep GP classification. (d) Deep Q-EP classification.

(e) DKL-GP classification. (f) DSPP classification.

Figure B.1: Comparing Q-EP (B.1b) and deep Q-EP (B.1d) with GP (B.1a), deep GP (B.1c), DKL-GP (B.1e) and DSPP (B.1f) on a classification problem defined on annular rhombus.

Table B.2: Classification on simulated annual rhombus: accuracy (ACC), area under ROC curve (AUC), negative logarithm of marginal likelihood (NLL) and running time by various deep models. Each result of the upper part is averaged over 10 experiments with different random seeds; values after $\pm$ are standard errors of these repeated experiments.

Model	ACC	AUC	NLL	time
GP	0.810 ± 0	**0.940** ± 0	17.673 ± 0	20.622 ± 0.346
Deep GP	0.825 ± 0.026	0.905 ± 0.012	534.782 ± 69.768	124.486 ± 2.978
QEP	0.834 ± 0	0.935 ± 0	4.670 ± 0	20.442 ± 0.559
Deep QEP	**0.856** ± 0.015	0.878 ± 0.019	96.736 ± 7.865	124.752 ± 0.575
DKL-GP	0.664 ± 0.196	0.732 ± 0.200	17.094 ± 5.533	23.874 ± 0.316
DSPP	0.744 ± 0.023	0.829 ± 0.056	588.543 ± 302.576	108.076 ± 1.725

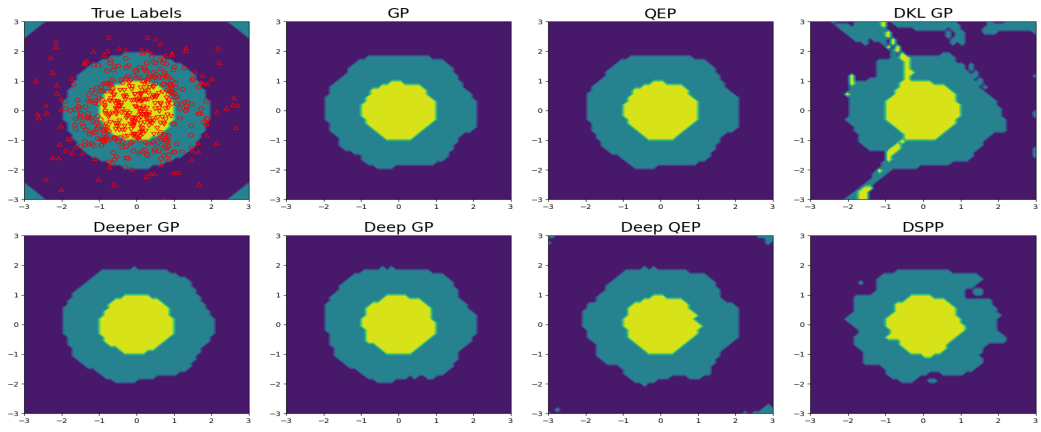
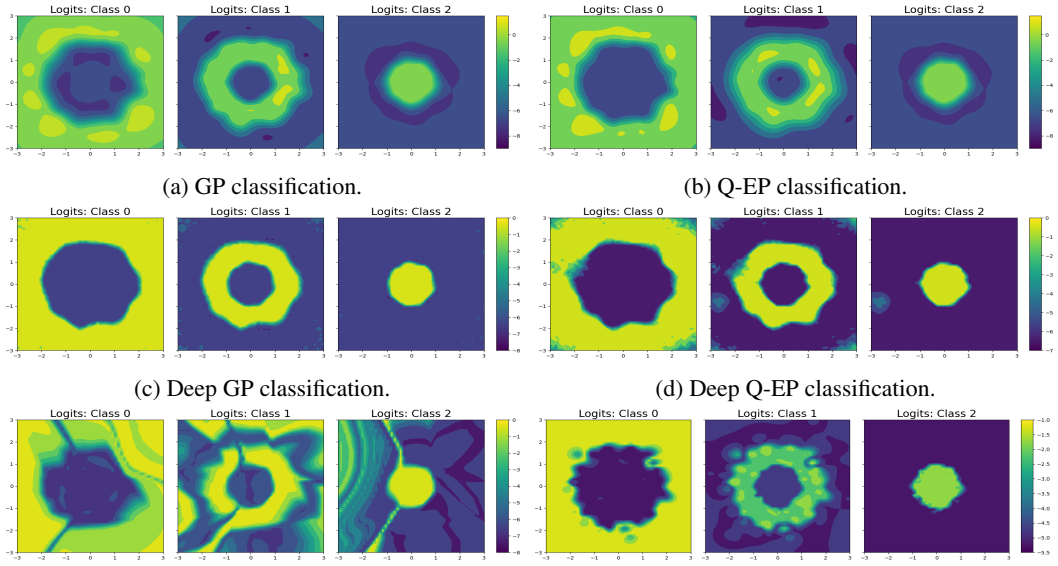


Figure B.2: Comparing shallow (1-layer) and deep (2-layer) Q-EPs with GP, deep GP, deeper GP (3-layer), DKL-GP and DSPP on a classification problem defined on annulus. Circles, upper and lower triangles label three classes in the training data.



(a) GP classification. (b) Q-EP classification.

(c) Deep GP classification. (d) Deep Q-EP classification.

(e) DKL-GP classification. (f) DSPP classification.

Figure B.3: Comparing Q-EP (B.3b) and deep Q-EP (B.3d) with GP (B.3a), deep GP (B.3c), DKL-GP (B.3e) and DSPP (B.3f) on a classification problem defined on annulus.

Table B.3: Classification on simulated annulus: accuracy (ACC), area under ROC curve (AUC), negative logarithm of marginal likelihood (NLL) and running time by various deep models. Each result of the upper part is averaged over 10 experiments with different random seeds; values after ± are standard errors of these repeated experiments.

Model	ACC	AUC	NLL	time
GP	0.951 ± 0	0.989 ± 0	18.821 ± 0	49.425 ± 1.728
Deep GP	**0.953** ± 0.03	0.991 ± 0.001	467.216 ± 45.845	199.600 ± 10.871
QEP	0.952 ± 0	0.985 ± 0	4.598 ± 0	49.301 ± 1.283
Deep QEP	0.950 ± 0.008	**0.992** ± 0.003	123.726 ± 12.965	197.677 ± 12.354
DKL-GP	0.854 ± 0.080	0.941 ± 0.099	19.039 ± 4.223	34.329 ± 0.918
DSPP	0.922 ± 0.026	0.970 ± 0.008	621.152 ± 297.205	166.974 ± 2.839