3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–14

Decoupled Sparse Gaussian Processes Components : Separating Decision Making from Data Manifold Fitting

Sebastian G. Popescu S.POPESCU16@IMPERIAL.AC.UK
David J. Sharp DAVID.SHARP@IMPERIAL.AC.UK
Ben Glocker B.GLOCKER@IMPERIAL.AC.UK
Imperial College London
James H. Cole
University College London JAMES.COLE@UCL.AC.UK

1. Introduction

Gaussian processes (GPs) are data-efficient Bayesian non-parametric models that offer calibrated uncertainty quantification and are robust to overfitting. Their drawback resides in the computational complexity for inverting the covariance matrix, which is cubic in computation and quadratic in memory. To solve this issue, Hensman et al. (2013) proposed an "inducing point" framework scalable to large data, obtaining posterior formulas conditioned on these artificial points. However, this also scales supralinearly with regards to inducing point numbers. Van der Wilk (2019) have shown that the parametric mean can only have M degrees of freedom, where M is the number of inducing points. On this note, recent work (Cheng and Boots, 2017; Salimbeni et al., 2018) have introduced a dual representation of Sparse GP (SGPs) in Reproducing Kernel Hilbert Space (RKHS), choosing a different set of basis for the mean, respectively the variance component. The degrees of freedom of the mean posterior equation can be increased at a linear cost with regards to number of inducing points. Similar in scope, Shi et al. (2020) decompose a SGP into two orthogonal components in prior function space and impose variational posteriors with different sets of inducing points for each component.

In this paper we make the following contributions:

- Propose a new RKHS parametrization of SGPs that separates the basis for the parametric, respectively non-parametric components on a SGPs as introduced in Hensman et al. (2017).
- Empirically show that this formulation results in a set of inducing points that has the task of devising the decision boundaries and another set which exclusively focuses on fitting the training data manifold.
- Show the equivalence between Bayesian Kernel Ridge Regression (BKRR) and SGPs in posterior function space. Based on insights from BKRR we propose a parametrization of the inducing points' variance as a function of their location.

2. Background

We defer the notation conventions and introduction to SGPs to the supplementary material, while introducing the necessary theoretical background for our method.

2.1. Disentangling Uncertainties in Sparse Gaussian Processes

As noted by Hensman et al. (2017), the variational GP posterior over function values can be divided into two components:

$$f(\cdot) = h(\cdot) + g(\cdot) \tag{1}$$

$$h(\cdot) = \mathcal{N}(h|0, K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}) \tag{2}$$

$$g(\cdot) = \mathcal{N}(g|K_{nm}K_{mm}^{-1}m, K_{nm}K_{mm}^{-1}SK_{mm}^{-1}K_{mn}) \tag{3}$$

The variance pertaining to $g(\cdot)$ will be denoted as within-data uncertainty as it is given by the basis functions that control the shape of the mean, thereby it could be interpreted as the variance stemming from the parametric side of the SGP. The variance of $h(\cdot)$ captures the shift from within to outside the data manifold as it is given by the current number of inducing points and will be denoted as distributional uncertainty.

2.2. Decoupled GPs

Cheng and Boots (2017) introduced the dual formulation of Sparse Gaussian Processes in RKHS. The authors motivated the Decoupled GPs by the $\mathcal{O}(M^3)$ complexity. They argue that the posterior mean can be decoupled from the posterior variance, when the goal is rather enhancing the predictive mean rather than obtaining better approximations to the real posterior variance of the full GP.

We define an RKHS $\mathbb{H}$ to be a Hilbert space with the reproducing property: $\forall x \in \mathbb{X}$, $\exists \phi_x \in \mathbb{H}$ such that $\forall f \in \mathbb{H}, f(x) = \phi_x^T f$. A $\mathbb{GP}(m, k)$ is equivalent to a Gaussian measure ν on a Banach space $\mathbb{B}$ which has a RKHS $\mathbb{H}$. There is a mean functional $\nu \in \mathbb{H}$ and a bounded positive semi-definite linear operator $\Sigma : \mathbb{H} \rightarrow \mathbb{H}$ such that for any $x, x^* \in \mathbb{X}, \exists \phi_x, \phi_{x^*} \in \mathbb{H}$, we can then write $m(x) = \phi_x^T \nu$ and $k(x, x^*) = \phi_x^T \Sigma \phi_{x^*}$

With the definition of the dual of GPs in mind, we can now proceed to characterize a subspace parametrization as follows: $\mu = \phi_Z a$ and $\Sigma = I + \phi_Z A \phi_Z^T$, where $a \in \mathbb{R}^M$ and $A \in \mathbb{R}^{M*M}$.

3. Decoupled GP Components

Our motivation is to get a better approximation of the true distributional variance of a full GP with minimum computational and memory overhead. With this in mind, our functional basis parametrization for $h(\cdot)$ is:

$$\mu_h = \phi_{Z\alpha}\mathbb{K}_\alpha^{-1}m \tag{4}$$

$$\Sigma_h = \phi_{Z\alpha}\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}\phi_{Z\alpha}^T \tag{5}$$

whereas for the $g(\cdot)$ component we impose the following parametrization:

$$\mu_g = 0 \tag{6}$$

$$\Sigma_g = I - \phi_Z\mathbb{K}_Z^{-1}\phi_Z^T \tag{7}$$

where $Z = \{\alpha, \beta\}$. The additional β inducing point locations are introduced to obtain a better fit in terms of variance to the data manifold. This comes without the introduction of additional variational parameters m_β, S_β that need to be optimized.

Consequently, we have the following maximization objective:

$$max\mathbb{L}_\theta(q(f)) = max_{q(f),\theta} \int q(f) log \frac{p_\theta(y|f)p(f)}{q(f)} df \tag{8}$$

$$max\mathbb{E}_{q(h),q(g)}[logp_\theta(y|h+g)] - \mathbb{KL}[q(f)||p(f)] \tag{9}$$

where $q(f) = \mathbb{N}(h+g|\mu_h, \Sigma_h + \Sigma_g)$, respectively the Kullback-Liebler divergence has the the following analytic form:

$$\mathbb{KL}[q(f)||p(f)] = 0.5 * \left[\text{Tr}(SK_\alpha^{-1}) + m^T K_\alpha^{-1} m - log \frac{|K_\alpha - K_{\alpha,Z} K_Z^{-1} K_{Z,\alpha} + S|}{K_\alpha} \right] \tag{10}$$

Through this RKHS parametrization, we have decoupled the inducing points of $h(\cdot)$ and $g(\cdot)$. In all experiments we use the same kernel hyperparameters for both components.

4. Equivalence between Bayesian Kernel Ridge Regression and Sparse Gaussian Processes

Kanagawa et al. (2018) provide an in-depth analysis on connections between kernel methods and GPs. In this section, we seek to establish connections between BKRR and SGPs.

At testing time, the parameterization for Decoupled SGP Components (DSGPC) has the following form:

$$\tilde{U}_h = K_{n,\alpha} K_\alpha^{-1} m \tag{11}$$

$$\tilde{\Sigma}_h = K_{n,\alpha} K_\alpha^{-1} S K_\alpha^{-1} K_{\alpha,n} \tag{12}$$

$$\tilde{\Sigma}_g = K_{n,n} - K_{n,Z} K_{Z,Z}^{-1} K_{Z,n} \tag{13}$$

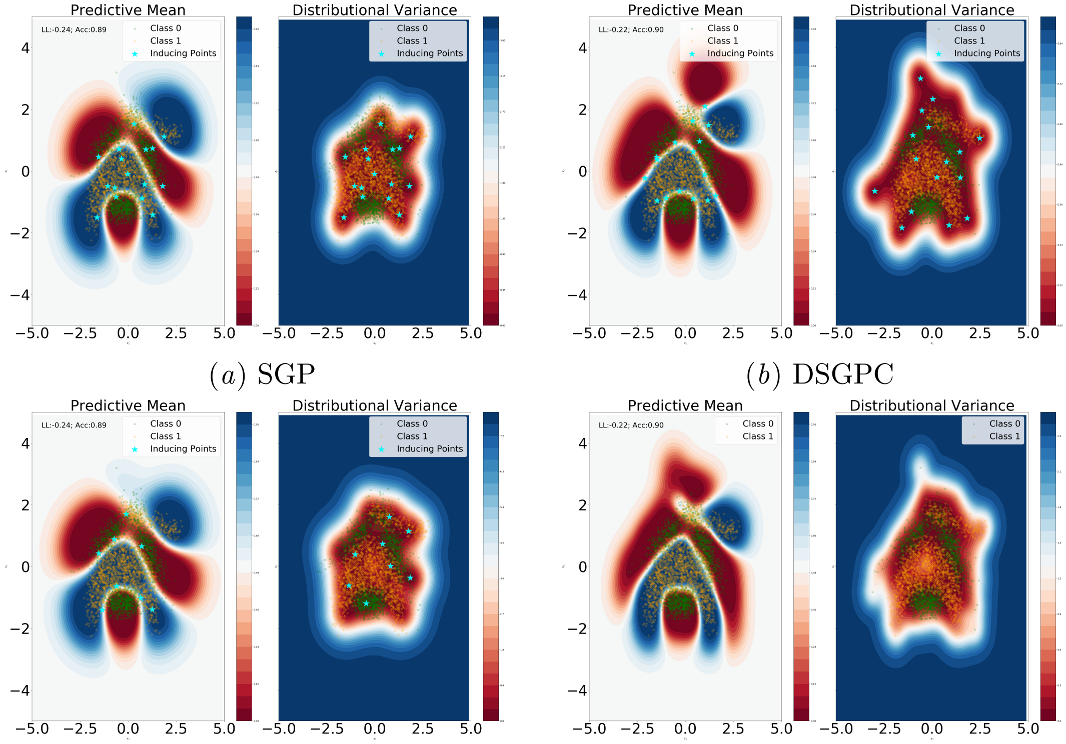
BKRR assumes the following model $p(Y|\beta, X) = X\beta + \epsilon$, where $\beta \sim \mathbb{N}(0, \frac{\sigma^2}{\lambda})$ and we define $\epsilon \sim \mathbb{N}(0, \sigma^2)$. One can obtain an analytic expression for the posterior $p(\beta|D) \sim \mathbb{N}(\tilde{\beta}, \tilde{\Sigma})$ with $\tilde{\beta} = (X^T X + \lambda\mathbb{I})^{-1} X^T Y$ and $\tilde{\Sigma} = \sigma^2 (X^T X + \lambda\mathbb{I})^{-1}$.To obtain the predictive mean and variance for a new point x^* we need to integrate out $p(y^*|x^*, D) = \int p(y^*|x^*, \beta) p(\beta|D) d\beta = \mathbb{N}(\phi(x^*)^T \tilde{\beta}, \sigma^2 + \phi(x^*)^T \Sigma \phi(x^*))$. After introducing the analytic equations for $p(\beta|D)$ we obtain the final form:

$$m(y^*) = K(x^*)(K + \lambda\mathbb{I})^{-1} Y \tag{14}$$

$$v(y^*) = \frac{\sigma^2}{\lambda} k(x^*, x^*) - \frac{\sigma^2}{\lambda} K(x^*)(K + \lambda\mathbb{I})^{-1} K(x^*) \tag{15}$$

One can easily notice that the above equations correspond to equations 11 and 13. A variant of equation 12 can be obtained as follows:

$$Var_Y(m(y^*)) = K(x^*)(K+\lambda\mathbb{I})^{-1}\left[\sigma^2 X(X^TX+\lambda\mathbb{I})^{-1}X^T+\sigma^2\mathbb{I}\right](K+\lambda\mathbb{I})^{-1}K(x^*) \quad (16)$$



(a) SGP

(b) DSGPC

(c) Solve-GP

(d) Full-GP

Figure 1: Predictive mean and distributional variance for m trained on ”banana” dataset.

4.1. BKRR inspired parameterization of inducing points’ variance

From equation 16 we can notice that $\sigma^2 X(X^TX+\lambda\mathbb{I})^{-1}X^T+\sigma^2\mathbb{I}$ basically stands for S in the conditional posterior variance (see eqn.23 in supplement). Rather than introducing additional parameters that need to be optimized, we can instead use this formulation of within-data manifold variance stemming from BKRR. This translates into the the following posterior for inducing point values u:

$$q(u) = \mathbb{N}(m, \sigma^2 Z_\alpha(Z_\alpha^T Z_\alpha + \lambda\mathbb{I}_{m_\alpha})^{-1}Z_\alpha^T + \sigma^2\mathbb{I}_{M_\alpha}) \quad (17)$$

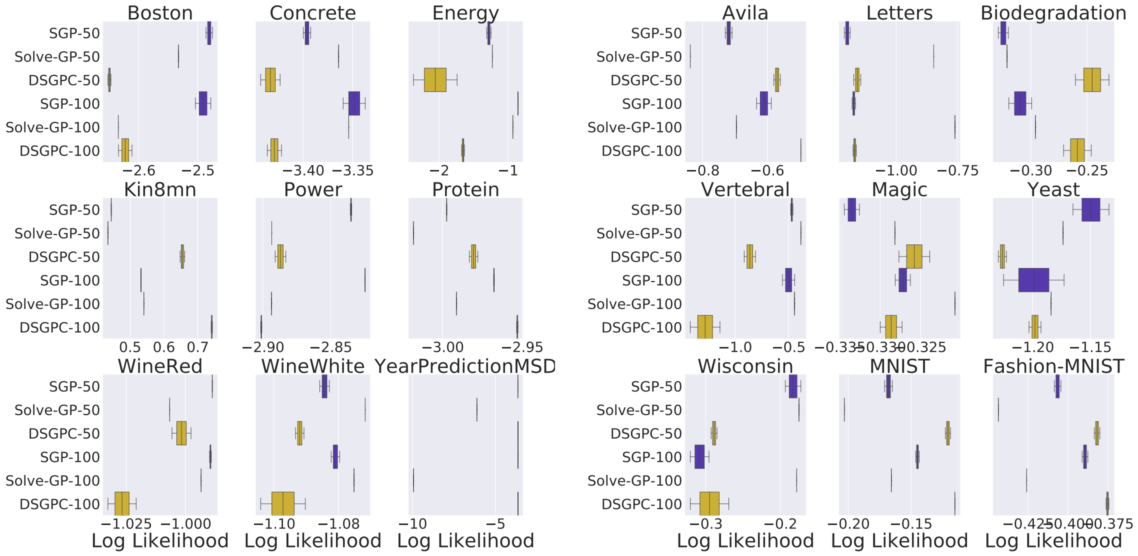
where λ can be interpreted as the additional jitter added for a stable Cholesky decomposition.

5. DSGPC separate decision making from fitting the data manifold

Since the α variational parameters are essentially tasked to construct the decision boundaries, the α inducing point locations are optimized to be in locations closer to the decision boundaries. Conversely, the β inducing points locations are only tasked to ensure low distributional variance within the data manifold, hence their locations will be spread out evenly across the training set, thereby resulting in a behaviour more similar to that of a full GP (Figure 1). The α and β inducing points' locations of Solve-GP are concentrated around the decision boundaries.

6. Results

We evaluate our models on a range of regression and classification benchmark tasks from the UCI machine learning dataset repository, alongside MNIST and Fashion-MNIST. All of the experiments were run with the same initializations and with 50 and 100 inducing points with the goal of comparing our method to SGP (Hensman et al., 2013) and Solve-GP (Shi et al., 2020).

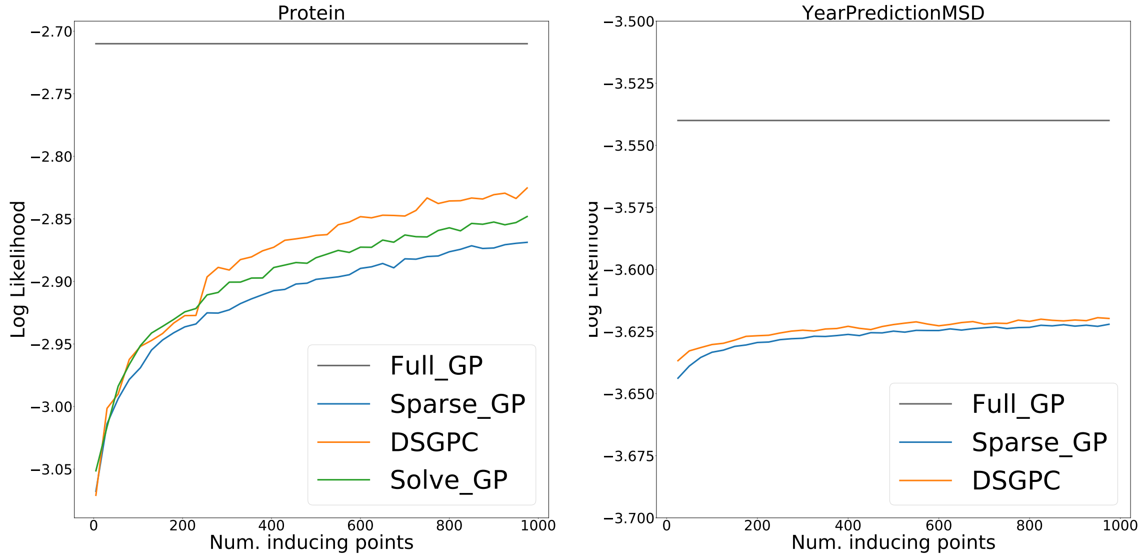


(a) Log-lik. Regression

(b) Log-lik. Classification

Figure 2: **All subplots**: 10 different runs of each model with different initialization seeds are taken into the composition of each boxplot. Higher values indicate better model fit.

For large scale classification datasets, such as Avila (n=20,867; d=10), MNIST(n=10,000; d=784) and Fashion-MNIST(n=10,000; d=784), we can observe that the DSGPC is surpassing its counterparts in all scenarios, whereas for Letters (n=20,000; d=16), Solve-GP provides improved performance. For large scale regression we use the "YearPrediction-MSD" dataset (n=515,345; d=90). DSGPC and SGP obtain relatively similar results with Solve-GP lagging behind. Furthermore, we investigative the convergence of models to the full GP log likelihood on the testing set based on number of inducing points (Figure 3). For "Protein", DSGPC and Solve-GP achieve faster convergence rates compared to SGP. Lastly, we were not able to successfully train Solve-GP on "YearPredictionMSD". DSGPC manages to be closer to the Full GP solution on this dataset as well.



(a) (b)

Figure 3: Optimisation behaviour of SGP, SolveGP and DSGPC for varying number of inducing points compared to the full GP.

7. Discussion

We have presented a dual formulation of SGPs in RKHS that is capable of separating the task of devising decision boundaries from the task of fitting the data manifold from the point of view of inducing point locations. This results in α inducing point locations which are situated close to the decision boundaries, whereas the β inducing point locations are centred in high density areas. This comes in stark contrast to the case of coupled inducing points, where the inducing point locations are a mix of the aforementioned scenarios.

The introduction of an additional amount of inducing points does not add up to a doubling of variational parameters, as the β inducing points do not warrant values in function space. Furthermore, explicitly linking the inducing points' variance to their location results in a model where we only have to learn the mean variational parameters. Future work should explore subspace basis formulations for learning the β inducing point locations to further reduce parameter numbers. Another prospective research avenue resides in applying DSGPC in the context of hierarchical GPs with Wasserstein-2 kernels (Popescu et al., 2020). The separation of inducing points' tasks should in theory improve out-of-distribution detection due to the enhanced data manifold fit in the first layer. In the supplementary material, we have also included an initial extension of our work to DGPs and intuitively show why the hidden layers of DGPs do not require a high number of β inducing points.

References

Ching-An Cheng and Byron Boots. Variational inference for gaussian process models with linear complexity. In *Advances in Neural Information Processing Systems*, pages 5184–

5194, 2017.

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

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

James Hensman, Nicolas Durrande, Arno Solin, et al. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–151, 2017.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Sebastian Popescu, David Sharp, James Cole, and Ben Glocker. Hierarchical gaussian processes with wasserstein-2 kernels. *arXiv preprint arXiv:2010.14877*, 2020.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Hugh Salimbeni, Ching-An Cheng, Byron Boots, and Marc Deisenroth. Orthogonally decoupled variational gaussian processes. In *Advances in neural information processing systems*, pages 8711–8720, 2018.

Jiaxin Shi, Michalis Titsias, and Andriy Mnih. Sparse orthogonal variational inference for gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 1932–1942. PMLR, 2020.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Mark Van der Wilk. *Sparse Gaussian process approximations and applications*. PhD thesis, University of Cambridge, 2019.

Supplementary Materials

Sparse Gaussian Processes

We denote the output vector Y, where each entry Y_i is a noisy observation of the function $F(x_i)$ for all input points $X = (x_i)_{i=1}^n$. We place a $GP(m,k)$ prior on the stochastic function F. We introduce inducing points $Z = (z_i)_{i=1}^m$ with inducing point function values $U = (u_i)_{i=1}^m$. Under standard Gaussian identities we have

$$p(Y|F) \sim \mathcal{N}(Y|F,\beta) \tag{18}$$

$$p(F|U;X,Z) \sim \mathcal{N}(F|K_{nm}K_{mm}^{-1}U, K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}; X, Z) \tag{19}$$

$$p(U;Z) \sim \mathcal{N}(U|0, K_{mm}^E) \tag{20}$$

By the definition of a sparse GP, the joint density is $p(Y,F,U) = p(F|U)p(U)\prod_{i=1}^n p(Y_i|F_i)$. We follow the variational inference framework introduced in Hensman et al. (2013) and maximize the lower bound on the marginal likelihood

$$L = E_{q(F,U)}[log\frac{p(Y,F,U)}{q(F,U)}] \tag{21}$$

where the variational posterior is choosen as $q(F,U) = p(F|U;X,Z)q(U)$, where $q(U) = \mathcal{N}(U|m,S)$. Here, m and S are free variational parameters. Due to the Gaussian nature of both terms we can marginalize U to arrive at $q(F) = \int p(F|U)q(U) = \mathcal{N}(F|\tilde{U},\tilde{\Sigma})$ where

$$\tilde{U} = K_{n,m}K_{mm}^{-1}m \tag{22}$$

$$\tilde{\Sigma} = K_{nn} - K_{nm}K_{mm}^{-1}[K_{mm} - S]K_{mm}^{-1}K_{mn} \tag{23}$$

Then the evidence lower bound (ELBO) can be rewritten as:

$$L = E_{q(F)}\left(logp(Y|F)\right) - E_{q(U)}\left(log\frac{q(U)}{p(U)}\right) \tag{24}$$

Derivation of Kullback-Liebler Divergence

We remind that the Kullback-Liebler divergence between two Gaussian measures $q = \mathbf{N}(m_q,\Sigma_q)$ and $p = \mathbf{N}(m_p,\Sigma_p)$ is given by:

$$KL(q||p) = 0.5 * \left[\mathrm{Tr}[\Sigma_p^{-1}\Sigma_q] + (m_p - m_q)^T\Sigma_p^{-1}(m_p - m_q) + \log\frac{|\Sigma_p|}{|\Sigma_q|}\right] \tag{25}$$

We now adapt this formula to suit our posterior $q = \mathbf{N}(\phi_{Z\alpha}\mathbb{K}_\alpha^{-1}m, I - \phi_Z\mathbb{K}_Z^{-1}\phi_Z^T + \phi_{Z\alpha}\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}\phi_{Z\alpha}^T)$, respectively the prior $p = \mathbf{N}(0,\mathbb{I})$.

Trace term:

$$\mathrm{Tr}[\Sigma_p^{-1}\Sigma_q] = \mathrm{Tr}\left[\mathbb{I}^{-1}\left[\mathbb{I} - \phi_Z\mathbb{K}_Z^{-1}\phi_Z^T + \phi_{Z\alpha}\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}\phi_{Z\alpha}^T\right]\right] \tag{26}$$

$$= \mathrm{Tr}\left[\mathbb{I}\right] - \mathrm{Tr}\left[\phi_Z\mathbb{K}_Z^{-1}\phi_Z^T\right] + \mathrm{Tr}\left[\phi_{Z\alpha}\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}\phi_{Z\alpha}^T\right] \tag{27}$$

$$= \mathrm{Tr}\left[\mathbb{I}\right] - \mathrm{Tr}\left[\mathbb{I}\right] + \mathrm{Tr}\left[\mathbb{K}_\alpha^{-1}S\right] \tag{28}$$

Quadratic term:

$$(\phi_{Z\alpha}\mathbb{K}_\alpha^{-1}m)^T\mathbb{I}^{-1}(\phi_{Z\alpha}\mathbb{K}_\alpha^{-1}m) = m^TK_\alpha^{-1}m \tag{29}$$

Log Determinants term:

$$\log|\Sigma_q| = \log|\mathbb{I} - \phi_Z\mathbb{K}_Z^{-1}\phi_Z^T + \phi_{Z\alpha}\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}\phi_{Z\alpha}^T| \tag{30}$$

$$= \log|\mathbb{I} - \phi_Z\mathbb{K}_Z^{-1}\phi_Z^T + \phi_{Z\alpha}\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}\phi_{Z\alpha}^T|\frac{|K_\alpha|}{|K_\alpha|} \tag{31}$$

$$= \log\frac{|K_\alpha - K_{\alpha,Z}K_Z^{-1}K_{Z,\alpha} + K_\alpha\mathbb{K}_\alpha^{-1}S\mathbb{K}_\alpha^{-1}K_\alpha|}{|K_\alpha|} \tag{32}$$

$$= \log\frac{|K_\alpha - K_{\alpha,Z}K_Z^{-1}K_{Z,\alpha} + S|}{|K_\alpha|} \tag{33}$$

Derivation of posterior for BKRR

BKRR assumes the following model $p(Y|\beta, X) = X\beta + \epsilon$, where w$\beta \sim \mathbb{N}(0, \frac{\sigma^2}{\lambda})$ and we define $\epsilon \sim \mathbb{N}(0, \sigma^2)$.

From a parameter space view, we are trying to solve the following problem:

$$p(\beta|D) = \frac{p(D|\beta)p(\beta)}{p(D)} \tag{34}$$

$$p(\beta|x_{1,\ldots,n}, y_{1,\ldots,n}) = \frac{\prod_{i=1}^n p(y_{i|\beta,x_i})}{Z} \tag{35}$$

$$p(\beta|x_{1,\ldots,n}, y_{1,\ldots,n}) = \frac{1}{Z}\prod_{i=1}^n e^{-\frac{1}{2\sigma^2}(y_i-x_i\beta)^2}e^{-\frac{\lambda}{2\sigma^2}|\beta|^2} \tag{36}$$

$$p(\beta|x_{1,\ldots,n}, y_{1,\ldots,n}) = \frac{1}{Z}e^{-\frac{1}{2\sigma^2}\left[(y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta\right]} \tag{37}$$

$$p(\beta|x_{1,\ldots,n}, y_{1,\ldots,n}) = \frac{1}{Z}e^{\frac{-1}{2}\left[\frac{1}{\sigma^2}y^Ty+\frac{1}{\sigma^2}\beta^T(x^Tx+\lambda\mathbb{I})\beta-\frac{2}{\sigma^2}\beta^TX^Ty\right]} \tag{38}$$

Finally, one can recognize the last equation as being a Gaussian with mean $\tilde{\beta} = (X^TX + \lambda\mathbb{I})^{-1}X^TY$, respectively variance $\Sigma = \sigma^2(X^TX + \lambda\mathbb{I})^{-1}$

Derivation of variance of BKRR predictive mean

$$Var_Y(m(y^*)) = Var(K(x^*)(K + \lambda\mathbb{I})^{-1}Y) \tag{39}$$

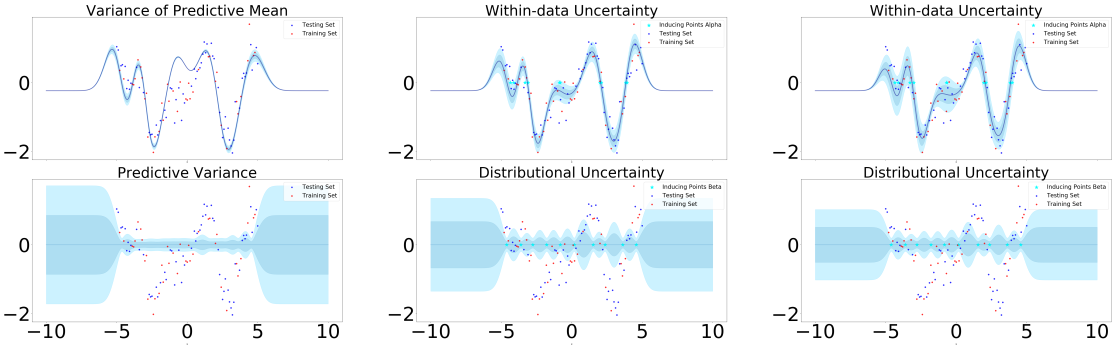
$$= K(x^*)(K + \lambda\mathbb{I})^{-1}Var(X\beta + \epsilon)(K + \lambda\mathbb{I})^{-1}K(x^*) \tag{40}$$

$$= K(x^*)(K + \lambda\mathbb{I})^{-1}\left[XVar(\beta)X^T + \sigma^2\mathbb{I}\right](K + \lambda\mathbb{I})^{-1}K(x^*) \tag{41}$$

$$= K(x^*)(K + \lambda\mathbb{I})^{-1}[\sigma^2X(X^TX + \lambda\mathbb{I})^{-1}X^T + \sigma^2\mathbb{I}]K(x^*)(K + \lambda\mathbb{I})^{-1} \tag{42}$$

Comparison between BKRR and DSGPC

We can observe a similar behaviour both in within-data and distributional variance between BKRR (Figure 4A) and DSGPC(Figure 4C). The kernel hyperparameters of the BKRR were



(a) Bayesian KRR (b) DSGPC-Free S (c) DSGPC-Parameterised S

Figure 4: "DSGPC-Free S" represents a DSGPC with the variance of the α inducing points kept free floating, whereas "DSGPC-Parameterised S" represents the model introduced in the main paper. Likelihood variance is not added.

not optimised. Additionally, we have included a variant of DSGPC (Figure 4B) where we keep the S term free-floating in a similar manner to SGP (Hensman et al., 2013). There are no visible differences in the two methods, motivating the usage of the parameterised S based on inducing points locations to further reduce variational parameters that need to be optimized.

Collapsed Lower bound for Decoupled Sparse Gaussian Process Components

We now derive the collapsed DSGPC lower bound by obtaining an analytic expression for the optimal $q(u)^*$.

$$\mathbb{E}_{q(h)}\mathbb{E}_{q(h)}log(y|h+g,\sigma^2) - KL[q(h)|p(h)] - KL[q(g)|p(g)] \tag{43}$$

By expanding the first term we obtain:

$$\begin{aligned}\mathbb{E}_{q(h)}log(y|h+g,\sigma^2) &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y-h-g)^T(y-h-g) &(44)\\ &= c - \frac{1}{2\sigma^2}\left[y^Ty + h^Th + g^Tg - 2y^Th\right. &(45)\\ &= \log\mathbf{N}(y|h,\sigma^2) - \frac{1}{2\sigma^2}Tr(\Sigma_g) &(46)\end{aligned}$$

We now plug equation 46 into equation 43 to obtain the collapsed bound

$$\begin{aligned}&\log\mathbf{N}(y|h,\sigma^2) - \frac{1}{2\sigma^2}Tr(\Sigma_g) - KL[q(h)|p(h)] - KL[q(g)|p(g)] &(47)\\ &\leq \log\int\mathbf{N}(y|h,\sigma^2)p(h)dh - \frac{1}{2\sigma^2}Tr(\Sigma_g) - KL[q(g)|p(g)] &(48)\\ &= \log\mathbf{N}(y|0,Q_{ff}+\sigma^2) - \frac{1}{2\sigma^2}Tr(K_{nn}-Q_{ff}) - \frac{1}{2\sigma^2}Tr(\Sigma_g) &(49)\end{aligned}$$

The optimal $q(u)^{*}$ can be represented as $\mathbf{N}(y|h,\sigma^2)p(h)$. By change of variable, respectively $h = K_{fz}K_{zz}^{-1}u$ we get the equivalent:

$$q(u)^{*} = \mathbf{N}(y|K_{fz}K_{zz}^{-1}u,\sigma^2)p(u) \tag{50}$$

One can notice that they have the same form as the ones in Titsias (2009).

$$\sigma^{-2}K_{zz}[K_{zz} + \sigma^{-2}K_{zn}K_{nz}]^{-1}K_{zn}y \tag{51}$$

$$K_{zz}[K_{zz} + \sigma^{-2}K_{zn}K_{nz}]^{-1}K_{zz} \tag{52}$$

We can notice that the optimal varaitional mean and variance are with respect to $Z = \{\alpha, \beta\}$.

Decoupling extended to hierarchical architectures

Motivation: In the case of DGPs, separating the inducing point locations for $g(\cdot)$ and $h(\dot)$ is motivated by the fact that most sampled realisations of intermediate layers are centred around 0. Hence, for inliers it suffices a small amount of inducing points to capture the data manifold at these intermediate layers (Figure 5). In contrast, there is a need of significantly more inducing point locations for the parametric part ($h(\cdot)$) so as to devise complex patterns.

We commence by laying the foundations of DGPs, subsequently introduce the new RKHS parametrization for intermediate layers followed by a derivation of the ELBO for this model which we entitle Decoupled Deep Gaussian Processes Components (DDGPC).

A Deep Gaussian Process (Damianou and Lawrence, 2013) is defined as a stack of shallow GPs acting as the prior:

$$p(y) = \underbrace{p(Y|F_L)}_{\text{likelihood}} \underbrace{\prod_{l=1}^{L} p(F_l|F_{l-1}, U_l; Z_{l-1})p(U_l)}_{\text{prior}} \tag{53}$$

where for brevity of notation we denote $Z_0 = X$. The sparse GPs between hidden layers are treated as being noiseless. As the prior is analytically intractable to integrate, Salimbeni and Deisenroth (2017) have suggested to sample from each hidden layer of the DGP in order to obtain unbiased stochastic gradients.

We introduce a factorized posterior between layers and dimensions of the following form:

$$q(F_L, \{U_l\}_{l=1}^{L}) = p(F_L|U_L; Z_{L-1}) \prod_{l=1}^{L} q(U_l) \tag{54}$$

where $q(U_l)$ is taken to be a multivariate Gaussian with mean m_l and variance $S_l = \sigma^2 Z_l(Z_l^T Z_l + \lambda\mathbb{I}_m)^{-1}Z_l^T + \sigma^2\mathbb{I}_M$, where λ can be interpreted as the jitter noise added for stable Cholesky decomposition.

Considering that most sampled points in the hidden layers of a DGP are closely centred around 0, we propose an alternative basis parameterization for the hidden layers of a DGP. The parametrization for the l-th layer $h_l(\cdot)$ is:

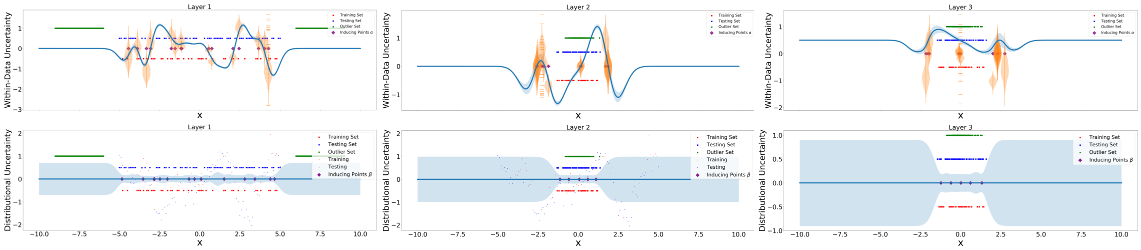


Figure 5: Decoupled DGP Components with 25 α inducing points, respectively 5 β inducing points for hidden layers; Decomposition of uncertainty into within-data (parametric) and distributional (non-parametric). Violin plots represent the variational distributions of the inducing points, with the x-axis denoting the location, and the y-axis the mean value of the inducing point. The x axis is taken to be the space of the previous layer. Red and blue scatter dots are sampled data points from the training, respectively testing set at each layer. **Remark** Red and blue dots are identified as within the data manifold in the hidden layers (distributional variance almost equal to 0) even with just 5 β inducing points, when we use 25 α inducing points.

$$\mu_{h_l} = \phi_{\alpha_l} \mathbb{K}_{\alpha_l}^{-1} m_l \tag{55}$$

$$\Sigma_{h_l} = \phi_{\alpha_l} \mathbb{K}_{\alpha_l}^{-1} S_l \mathbb{K}_{\alpha_l}^{-1} \phi_{\alpha_l}^{T} \tag{56}$$

whereas for the $g(\cdot)$ GP component we impose the following parametrization:

$$\mu_{g_l} = 0 \tag{57}$$

$$\Sigma_{g_l} = I - \phi_{\beta_l} \mathbb{K}_{\beta_l}^{-1} \phi_{\beta_l}^{T} \tag{58}$$

We can immediately notice that at testing time the computation cost of Σ_{g_l} is greatly reduce from $\mathcal{O}(M_{\alpha_l} + M_{\beta_l})^3$ to $\mathcal{O}((M_{\beta_l})^3$. In our experiments we only consider $M_{\alpha_l} >> M_{\beta_l}$

The KL divergence for the l-th layer is given by:

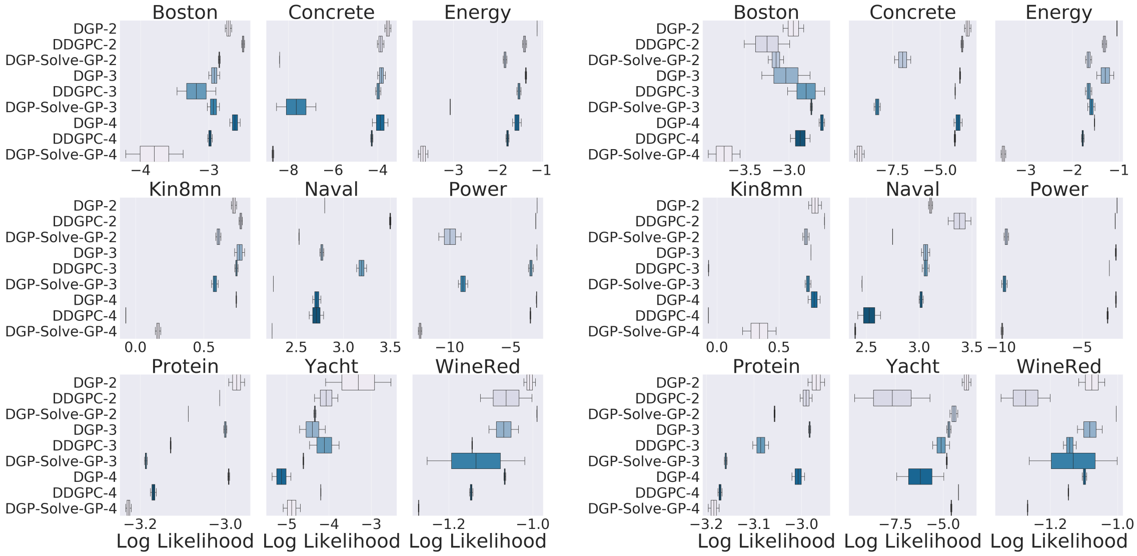
$$\mathbb{KL}[q(f_l)||p(f_l)] = 0.5 * \left[\mathrm{Tr}(S_l K_{\alpha_l}^{-1}) + m_l^T K_{\alpha_l}^{-1} m_l - log \frac{|K_{\alpha_l} - K_{\alpha_l,\beta_l} K_{\beta_l}^{-1} K_{\beta_l,\alpha_l} + S_l|}{K_{\alpha_l}} \right] \tag{59}$$

For the first GP layer in the hierarchy we keep the parametrization introduced in section 3 so as to capture the full extent of the data manifold.

Lastly, the ELBO is the following:

$$max \mathbb{E}_{[q(h_l),q(g_l)]_{l=1,..,L}} [log p_\theta (y|h_L + g_L)] - \sum_{l=1}^{L} \mathbb{KL}[q(f_l)||p(f_l)] \tag{60}$$

We provide some exploratory results on UCI regression tasks (Figure 6). We can notice a relative under performance of DDGPC and DGP-SolveGP in relation to DGPs. Experiments were performed under a similar training routine, limited to 50,000 iterations to gain an intuition of their convergence under a strict training regime. Future work should expand on these initial results to accommodate for adaptive training regimes until convergence.



(*a*) Num. ind. pts. 50

(*b*) Num. ind. pts. 100

Figure 6: **All subplots**: 10 different runs of each model with different initialization seeds are taken into the composition of each boxplot. Higher values indicate better model fit.

Additional details for UCI experiments

For all datasets we randomly selected 20% as the testing set, with the remainder being used for training. All implementations use the RBF kernel with automatic relevance determination. For the shallow GP experiments in the main paper we use 50 and 100 inducing points for each model. For the DGP experiments, in terms of model architecture, we use 2 hidden units for each hidden level, with 50, respectively 100 inducing points. For DDGPC we use 10 β inducing points for the hidden layers. All models are optimized for 50,000 iterations with a mini-batch of size 32 and the learning rate is set to 0.001. Results are provided for 2, 3 and 4 layers.

Kernel All Euclidean space kernels used the standard ARD RBF, using a lengthscale parameter per input dimension, initialized to 1.351. We initialize the variance of the kernel with the same value.

Likelihood For regression tasks the likelihood variance was initialized to 1.0

Model Name	Variational Parameters	Inducing Points	Function Space Orthogonality	Task Orthogonality
SGP (Hensman et al., 2013)	$M + \frac{M(M+1)}{2}$	M	✗	✗
SOLVE-GP (Shi et al., 2020)	$M_\alpha + M_\beta + \frac{M_\alpha(M_\alpha+1)}{2} + \frac{M_\beta(M_\beta+1)}{2}$	$M_\alpha + M_\beta$	✓	✗
DSGPC (current work)	M_α	$M_\alpha + M_\beta$	✓Prior; ✗Posterior	✓

Table 1: Task Orthogonality denotes the separation of fitting the data manifold from devising the decision boundaries. We assume a Cholesky decomposition parameterization of the variance.

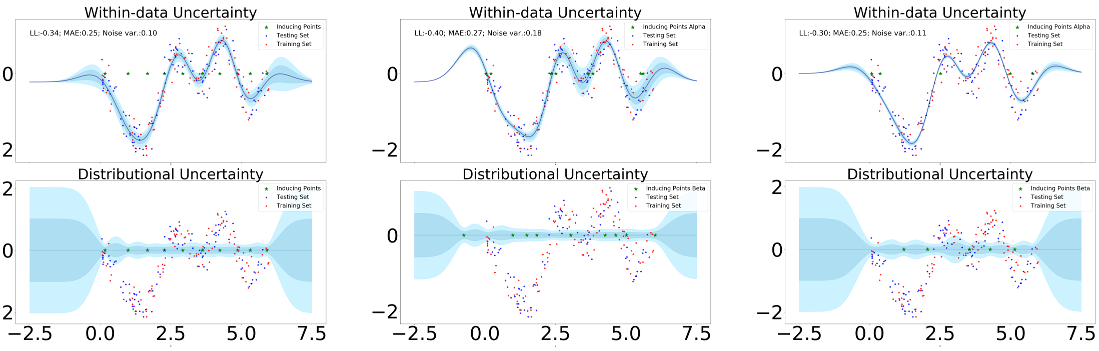
Inducing points We initialize the inducing point locations to the k-means of the training data for the first layer, whereas for hidden layers the locations are uniformly sampled in the interval [-1.0,1.0].

Variational parameters We initialize the mean to a zero column vector, whereas the variance is given by the lower triangular Cholesky decomposition which is initialized by the identity matrix for the last layer. For intermediate layers, the Cholesky decomposition is initialized by the identity matrix multiplied by 0.0001.

Optimization on UCI datasets All parameters were optimized using the Adam optimizer with a learning rate of 0.001. We used a batch size of 32 and trained for 50,000 iterations.

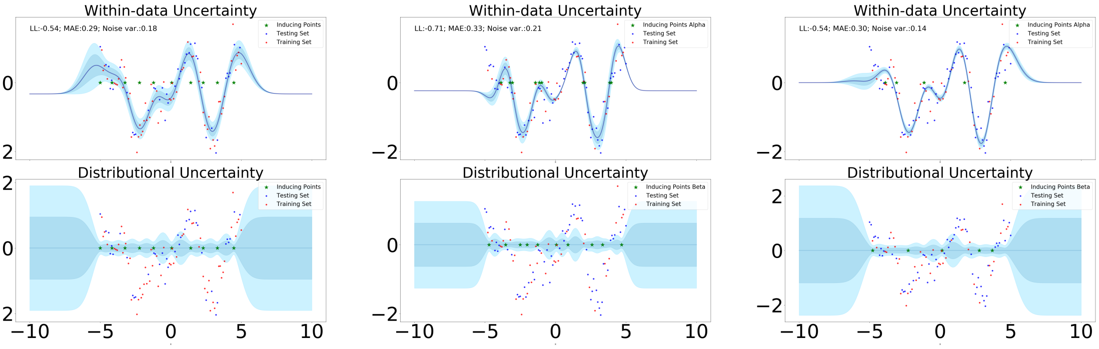
Additional tables and figures

In this subsection we provide a table showcasing the differences between the models considered in this paper, respectively additional figures on toy regression and classification task to highlight differences in function space and inducing points' locations.



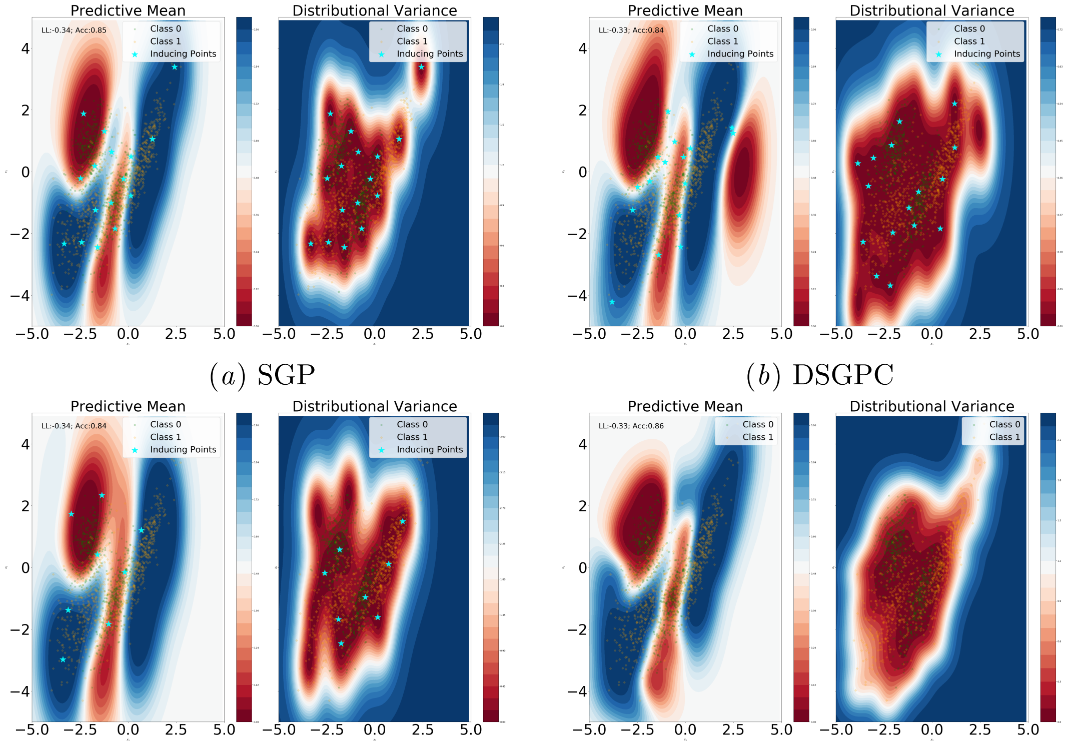
(*a*) SGP (*b*) DSGPC (*c*) SOLVE-GP

Figure 7: Predictive mean and distributional variance for Coupled and Decoupled SGPs trained on ”snelson” dataset. Likelihood variance is not added.



(*a*) SGP (*b*) DSGPC (*c*) SOLVE-GP

Figure 8: Predictive mean and distributional variance for models trained on toy regression task. Likelihood variance is not added.



(*a*) SGP

(*b*) DSGPC

(*c*) SOLVE-GP

(*d*) Full-GP

Figure 9: Predictive mean and distributional variance for models trained on toy classification task.