NATIONAL UNIVERSITY OF SINGAPORE

Deep Gaussian Process Regression (DGPR)

Haibin Yu A0123903N

February 12, 2017

1 Why Deep Gaussian Process

A Gaussian Process Regression model is equivalent to an infinitely wide neural network with single hidden layer and similarly a DGP is a multi-layer neural network with multiple infinitely wide hidden layers [Neal, 1995]. DGPs employ a hierarchical structural of GP mappings and therefore are arguably more flexible, have a greater capacity to generalize, and are able to provide better predictive performance [Damianou, 2015].

Then it comes into my mind that why we would like to proceed to be deep and what are the benefits about being deep. It has been argued that the addition of non-linear hidden layers can also potentially overcome practical limitations of shallow GPs [Bui et al., 2016]. So what are the limitations exactly?

Actually a GPR model is fully specified by a mean function $\mathbb{E}[\cdot]$ and the covariance function $\mathrm{cov}[\cdot,\cdot]$. Conventionally we manually set the mean function to be $\mathbf{0}$. Then we can say that a GPR model is fully specified by its covariance function which also can be denoted as the kernel.

Let us briefly examine the priors on functions encoded by some commonly used kernels

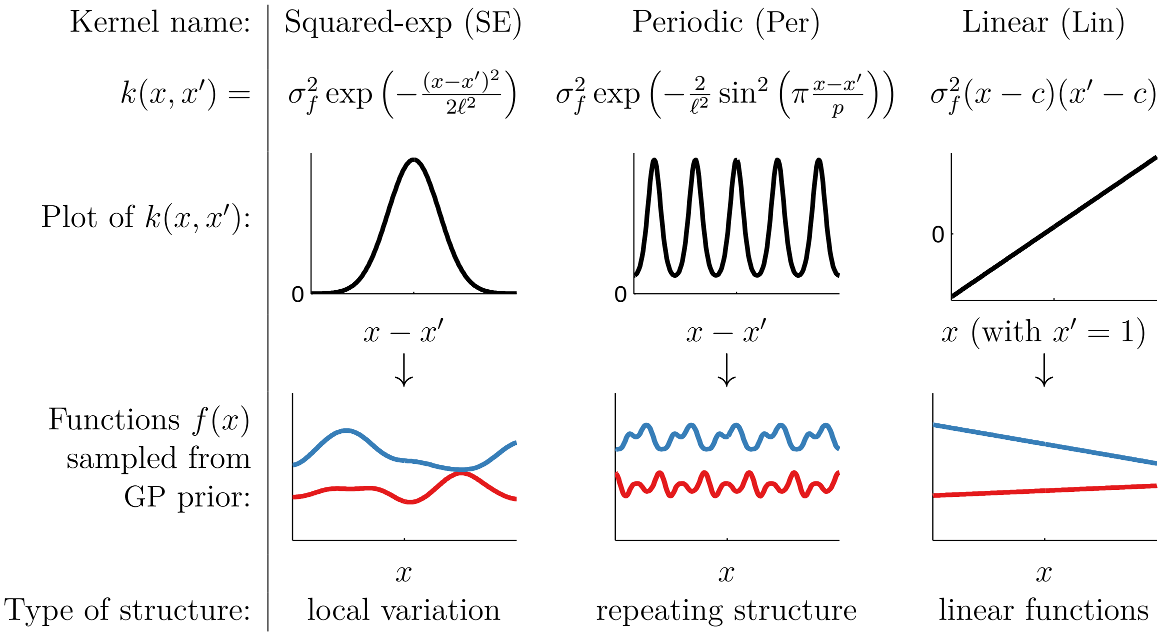


Figure 1: Sampled functions with three simple kernels

Undeniably, kernel plays an extremely important role in defining the capability of GPR, it there anything that DGPR model to do with the kernel, such as DGPR is related with deep kernel which generates more flexible, more capability of GPR model?

The definition of a DGPR model is a distribution on functions constructed by composing functions drawn from GP priors. An example of DGPR is a composition of vector-valued functions, with each function drawn independently from GP priors:

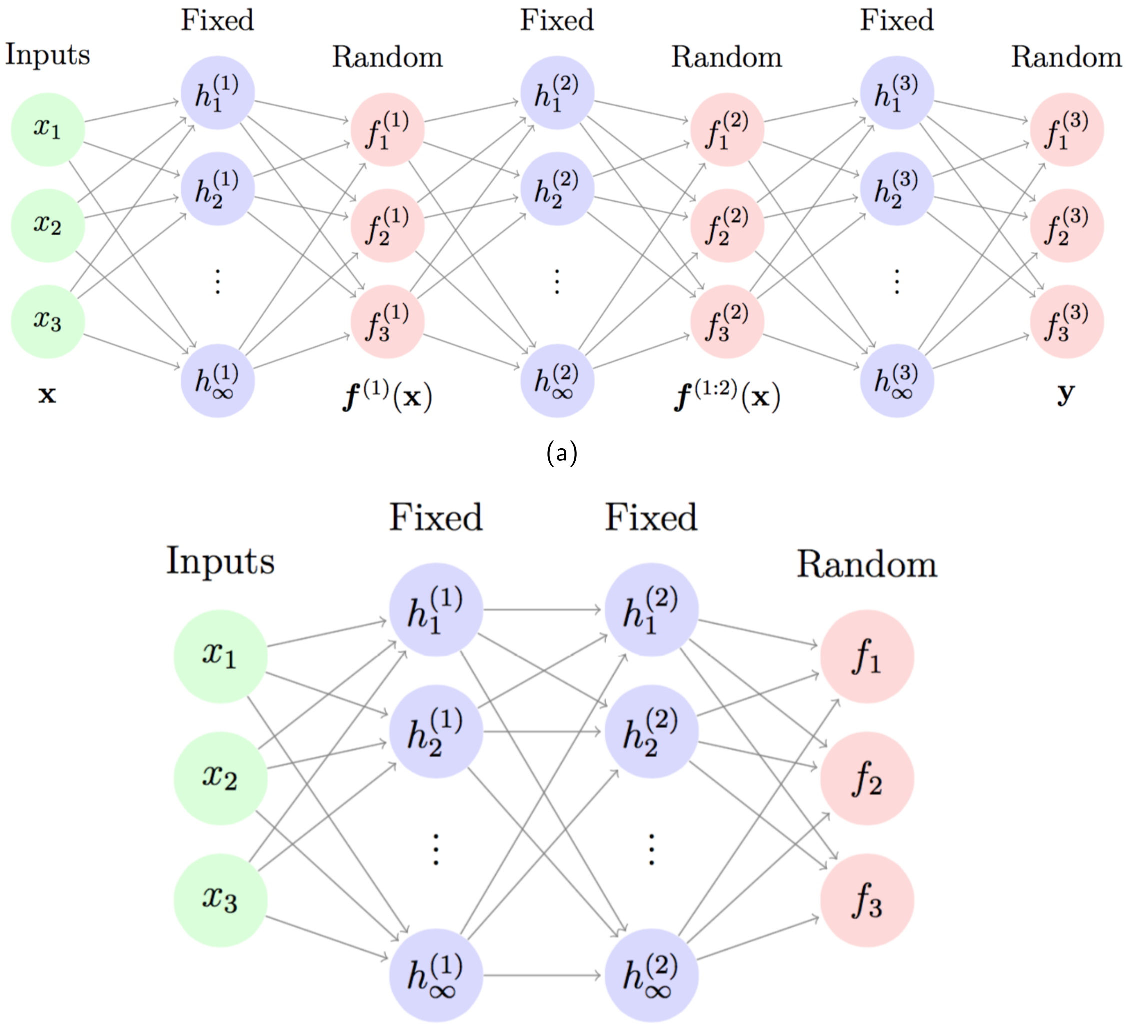
$$\mathbf{y} = f_l(f_{l-1}(\dots f_1(\mathbf{x}))) + \varepsilon \tag{1}$$

where we assume each function in the composition $f_i(\cdot)$ is itself a draw from a Gaussian process. In other words we develop our full probabilistic model through process composition. There is one appealing property for process composition called **Kolmogorov extension theorem** which guarantees that a suitably "consistent" collection of finite-dimensional distributions will define a stochastic process. The conditions required by the theorem are trivially satisfied by any stochastic process. Therefore, process composition resembles defining a new stochastic process. In other words, the process composition can provide richer class of process priors (somehow).

2 DGPR and Deep Kernel GPR

2.1 DGPR

Though we intuitively think that we can view DGPR as a GPR with a deep kernel, however, there is some discrepancies with what [Duvenaud, 2014] has proposed. According to what he has argued, DGPR is not the same as Deep Kernel GPR: Deep kernels imply a fixed representation as opposed to a prior over functions.



(a)

(b)

Figure 2: (a) Graphical framework for a 3-layer DGPR model. (b) Graphical framework for a deep kernel GPR

To be more specifically, if we follow the weight space view of GPR we can actually re-write DGPR as:

$$y = w_3\phi_3(w_2\phi_2(w_1\phi_1(x))) \tag{2}$$

where w_3, w_2 and w_1 are unknown and they are treated as random variables. In the context of GPR model, these variables are marginalized out. Then for each layer l_i we actually have a representation of that particular layer which is a GPR.

$$\begin{aligned} f_1(x) = w_1\phi_1(x) \\ f_2(x) = w_2\phi_2(x) \\ f_3(x) = w_3\phi_3(x) \end{aligned} \tag{3}$$

The example above of DGPR is a composition of vector-valued functions, with each function drawn independently from GP priors:

$$F(x) = f_3(f_2(f_1(x))) \text{ where each } f_i \text{ is a GP} \tag{4}$$

2.2 Deep Kernel GPR

As for the deep kernel GPR model, according to what [Duvenaud, 2014] has argued, it corresponds to a shallow GPR (1-layer GPR) with a more complicated kernel. Still using the weight space view of GPR, we could rewrite it as:

$$y = w\phi\left(\sigma_3(\theta_2\sigma_2(\theta_1\sigma_1(x)))\right) \tag{5}$$

Then after marginalizing out the latent variable w, we get

$$y = \mathcal{N}\left(0, k(g(x), g(x')\right) \ where\ g(x) = \sigma_3(\theta_2\sigma_2(\theta_1\sigma_1(x))) \tag{6}$$

2.3 Comparison

In the context of deep kernel GPR model, the latent input $g(x)$ follows a parametric form which [Duvenaud, 2014] mention it is a fixed representation; while in the DGPR case, $g(x)$ is actually drawn from a GP prior function. [Duvenaud, 2014] also mentioned that unless we can richly parametrize these deep kernels, their capacity to learn an appropriate representation will be limited in comparison to more flexible models such as deep neural network or DGPRs.

3 Existing DGP Model

The first DGP model was proposed by [Damianou and Lawrence, 2013] based on the variational inference method for GPR proposed by [Titsias, 2009]. In this scheme, a variational approximation over both latent functions and hidden variables is chosen such that a free energy is both computationally and analytically tractable. [Damianou and Lawrence, 2013] also mentioned that this is a hierarchical implementation of variational GPLVM where extending the work on variational GPLVM proposed by [Titsias and Lawrence, 2010]. The general framework is shown as the following figure:

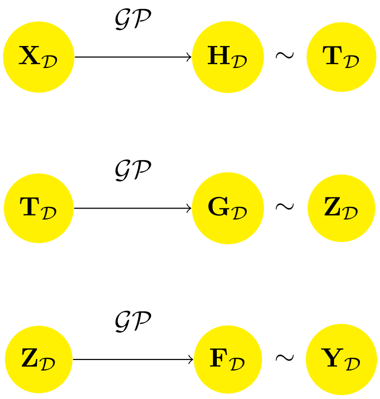


Figure 3: DGP model based on variational GPLVM

Critically, as a variational distribution over the hidden variables is used in this approach, in addition to one over the inducing outputs, the number of variational parameters increases linearly with the number of training data points which hinders the use of this method for large scale datasets. Furthermore, initialization for this scheme is a known issue, even for a modest number of data points. Another issue is in terms of prediction for regression task, it utilized an ad-hoc way modeling the predictive distribution with resort to the lower bound to approximate its predictive distribution.

4 Modified DGP

Since introducing a variational distribution to hidden variables every layer will incur the size of optimization problem grows with the number of training points. Instead we would like to propose a process composition which only requires a variational distribution over the inducing outputs, whilst removing the parameter scaling problem of [Damianou and Lawrence, 2013] as well as enabling stochastic or parallel learning on various SGPR frameworks (FITC, PITC, PIC and LMA) in DGP.

Compared with Figure 3, the modified model could be denoted as

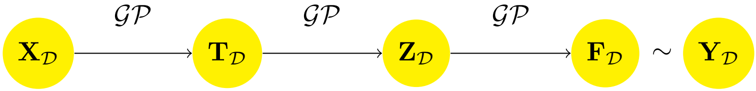

$\mathbf{X}_{\mathcal{D}}$
$\mathcal{GP}$
$\mathbf{T}_{\mathcal{D}}$
$\mathcal{GP}$
$\mathbf{Z}_{\mathcal{D}}$
$\mathcal{GP}$
$\mathbf{F}_{\mathcal{D}}$
$\sim$
$\mathbf{Y}_{\mathcal{D}}$

Figure 4: Modified DGP (we omit inducing variables for simplicity)

5 Interpretation of the Lower Bound

The log marginal likelihood of the modified DGP model becomes

$$\log p(\mathbf{Y}_{\mathcal{D}}) = \log \int p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{G}_{\mathcal{U}},\mathbf{T}_{\mathcal{D}})p(\mathbf{G}_{\mathcal{U}})p(\mathbf{T}_{\mathcal{D}}|\mathbf{H}_{\mathcal{U}})p(\mathbf{H}_{\mathcal{U}}) \tag{7}$$

To make it even simpler, we take a two-layer DGP as example then the log marginal distribution could be written as:

$$\log p(\mathbf{Y}_{\mathcal{D}}) = \log \int p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{U}}) \tag{8}$$

Still by applying our variational inference (VI) framework, the lower bound to the log marginal likelihood would become

$$\mathcal{L} = \int \mathcal{Q} \log \frac{p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{U}})}{\mathcal{Q}} \tag{9}$$

where $\mathcal{Q} = p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})q(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})q(\mathbf{Z}_{\mathcal{U}})$ is our basic assumption, hence the lower bound could be written as:

$$\begin{aligned}\mathcal{L} &= \int p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})q(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})q(\mathbf{Z}_{\mathcal{U}}) \log \frac{p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})p(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{U}}}{p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})q(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})q(\mathbf{Z}_{\mathcal{U}})} \\ &= \int p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})q(\mathbf{F}_{\mathcal{U}})p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})q(\mathbf{Z}_{\mathcal{U}}) \log \frac{p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}})\cancel{p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})}p(\mathbf{F}_{\mathcal{U}})\cancel{p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})}p(\mathbf{Z}_{\mathcal{U}})}{\cancel{p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}})}q(\mathbf{F}_{\mathcal{U}})\cancel{p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}})}q(\mathbf{Z}_{\mathcal{U}})} \\ &= \langle \log p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}}) \rangle_{\mathcal{Q}} - \mathrm{KL}(q(\mathbf{F}_{\mathcal{U}})||p(\mathbf{F}_{\mathcal{U}})) - \mathrm{KL}(q(\mathbf{Z}_{\mathcal{U}})||p(\mathbf{Z}_{\mathcal{U}}))\end{aligned} \tag{10}$$

Then for the likelihood part $\langle \log p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}}) \rangle_{\mathcal{Q}}$ could be calculated forwardly:

$$\begin{aligned} p(\mathbf{Z}_{\mathcal{D}}|\mathbf{Z}_{\mathcal{U}}) &= \mathcal{N}(\mathbf{Z}_{\mathcal{D}}|\mathbf{K}_{\mathbf{X}_{\mathcal{D}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{Z}_{\mathcal{U}}, \mathbf{K}_{\mathbf{X}_{\mathcal{D}}\mathbf{X}_{\mathcal{D}}} - \mathbf{K}_{\mathbf{X}_{\mathcal{D}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{K}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\mathbf{X}_{\mathcal{D}}}) \\ q(\mathbf{Z}_{\mathcal{U}}) &= \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}_{\mathcal{U}}}, \mathbf{\Sigma}_{\mathbf{Z}_{\mathcal{U}}}) \\ q(\mathbf{Z}_{\mathcal{D}}) &= \mathcal{N}(\mathbf{Z}_{\mathcal{D}}|\boldsymbol{\mu}_{\mathbf{Z}_{\mathcal{D}}}, \mathbf{\Sigma}_{\mathbf{Z}_{\mathcal{D}}})\end{aligned} \tag{11}$$

where $\boldsymbol{\mu}_{\mathbf{Z}_{\mathcal{D}}} = \mathbf{K}_{\mathbf{X}_{\mathcal{D}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\boldsymbol{\mu}_{\mathbf{Z}_{\mathcal{U}}}$ and $\mathbf{\Sigma}_{\mathbf{Z}_{\mathcal{D}}} = \mathbf{K}_{\mathbf{X}_{\mathcal{D}}\mathbf{X}_{\mathcal{D}}} - \mathbf{K}_{\mathbf{X}_{\mathcal{D}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{K}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\mathbf{X}_{\mathcal{D}}} + \mathbf{K}_{\mathbf{X}_{\mathcal{D}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{\Sigma}_{\mathbf{Z}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\widetilde{\mathbf{X}}_{\mathcal{U}}}\mathbf{K}_{\widetilde{\mathbf{X}}_{\mathcal{U}}\mathbf{X}_{\mathcal{D}}}$.

$$\begin{aligned} p(\mathbf{F}_{\mathcal{D}}|\mathbf{F}_{\mathcal{U}},\mathbf{Z}_{\mathcal{D}}) &= \mathcal{N}(\mathbf{F}_{\mathcal{D}}|\mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{F}_{\mathcal{U}}, \mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\mathbf{Z}_{\mathcal{D}}} - \mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{K}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\mathbf{Z}_{\mathcal{D}}}) \\ q(\mathbf{F}_{\mathcal{U}}) &= \mathcal{N}(\boldsymbol{\mu}_{\mathbf{F}_{\mathcal{U}}}, \mathbf{\Sigma}_{\mathbf{F}_{\mathcal{U}}}) \\ q(\mathbf{F}_{\mathcal{D}}) &= \mathcal{N}(\mathbf{F}_{\mathcal{D}}|\boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}, \mathbf{\Sigma}_{\mathbf{F}_{\mathcal{D}}})\end{aligned} \tag{12}$$

where $\boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}} = \mathbb{E}_{\mathbf{Z}_{\mathcal{D}}}\left[\mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\right]\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\boldsymbol{\mu}_{\mathbf{F}_{\mathcal{U}}}$

and $\mathbf{\Sigma}_{\mathbf{F}_{\mathcal{D}}} = \mathbb{E}_{\mathbf{Z}_{\mathcal{D}}}\left[\mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\mathbf{Z}_{\mathcal{D}}} - \mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{K}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\mathbf{Z}_{\mathcal{D}}} + \mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{\Sigma}_{\mathbf{F}_{\mathcal{U}}}\mathbf{K}_{\mathbf{Z}_{\mathcal{D}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\mathbf{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_{\mathcal{U}}\widetilde{\mathbf{Z}}_{\mathcal{U}}}\right]$ Then we have

$$\langle \log p(\mathbf{Y}_{\mathcal{D}}|\mathbf{F}_{\mathcal{D}}) \rangle_{\mathcal{Q}} = -\frac{|\mathcal{D}|}{2}\ln 2\pi - \frac{|\mathcal{D}|}{2}\ln \sigma_n^2 - \frac{1}{2\sigma_n^2}\left(\mathbf{Y}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}\right)^{\top}\left(\mathbf{Y}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}\right) - \frac{1}{2\sigma_n^2}\mathrm{Tr}\left(\mathbf{\Sigma}_{\mathbf{F}_{\mathcal{D}}}\right) \tag{13}$$

Therefore, the lower bound can be written as

$$\mathcal{L} = -\frac{|\mathcal{D}|}{2}\ln 2\pi - \frac{|\mathcal{D}|}{2}\ln \sigma_n^2 - \frac{1}{2\sigma_n^2}\left(\mathbf{Y}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}\right)^{\top}\left(\mathbf{Y}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}\right) - \frac{1}{2\sigma_n^2}\mathrm{Tr}\left(\mathbf{\Sigma}_{\mathbf{F}_{\mathcal{D}}}\right) - \mathrm{KL}(q(\mathbf{F}_{\mathcal{U}})||p(\mathbf{F}_{\mathcal{U}})) - \mathrm{KL}(q(\mathbf{Z}_{\mathcal{U}})||p(\mathbf{Z}_{\mathcal{U}})) \tag{14}$$

where we would like to maximize the lower bound, if we look at a different perspective where we take the negative of $\mathcal{L}$ will render

$$-\mathcal{L} = \frac{1}{2\sigma_n^2}\left(\mathbf{Y}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}\right)^{\top}\left(\mathbf{Y}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathbf{F}_{\mathcal{D}}}\right) + \frac{1}{2\sigma_n^2}\mathrm{Tr}\left(\mathbf{\Sigma}_{\mathbf{F}_{\mathcal{D}}}\right) + \mathrm{KL}(q(\mathbf{F}_{\mathcal{U}})||p(\mathbf{F}_{\mathcal{U}})) + \mathrm{KL}(q(\mathbf{Z}_{\mathcal{U}})||p(\mathbf{Z}_{\mathcal{U}})) + \frac{|\mathcal{D}|}{2}\ln \sigma_n^2 + \mathrm{const} \tag{15}$$

Actually the negative lower bound above has some interesting properties, for the first item $\frac{1}{2\sigma_n^2}\left(\mathbf{Y}_\mathcal{D}-\boldsymbol{\mu}_{\mathbf{F}_\mathcal{D}}\right)^\top\left(\mathbf{Y}_\mathcal{D}-\boldsymbol{\mu}_{\mathbf{F}_\mathcal{D}}\right)$ denotes the loss function and the signal variance serves somehow as the controller, the smaller the signal variance be, the more we concern the loss between our model and the true observations.

The second item is the trace term $\frac{1}{2\sigma_n^2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{D}}\right)$, instead of arguing this one directly, we introduce the conditional entropy

$$\mathbb{H}\left[\mathbf{F}_\mathcal{D}|\mathbf{F}_\mathcal{U}\right]=\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{D}}\right| \tag{16}$$

From an information theoretic perspective, the conditional entropy represents the amount of additional information we need to specify the distribution of $\mathbf{F}_\mathcal{D}$ given that we already have the inducing variables $\mathbf{F}_\mathcal{U}$. In other words, this means that the smaller the entropy is, the less additional information we need to specify the true distribution of $\mathbf{F}_\mathcal{D}$ which means that the approximation is more accurate. Another issue is that the log determinant is upper bounded by trace: $\ln\left|\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{D}}\right| \leq\ Tr(\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{D}})$. So actually when we are minimizing the trace term we are also trying to find out the best inducing sets. And the expression for $\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{D}}$ is

$$\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{D}}=\mathbb{E}_{\mathbf{Z}_\mathcal{D}}\left[\mathbf{K}_{\mathbf{Z}_\mathcal{D}\mathbf{Z}_\mathcal{D}}-\mathbf{K}_{\mathbf{Z}_\mathcal{D}\widetilde{\mathbf{Z}}_\mathcal{U}}\boldsymbol{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_\mathcal{U}\widetilde{\mathbf{Z}}_\mathcal{U}}\mathbf{K}_{\widetilde{\mathbf{Z}}_\mathcal{U}\mathbf{Z}_\mathcal{D}}+\mathbf{K}_{\mathbf{Z}_\mathcal{D}\widetilde{\mathbf{Z}}_\mathcal{U}}\boldsymbol{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_\mathcal{U}\widetilde{\mathbf{Z}}_\mathcal{U}}\boldsymbol{\Sigma}_{\mathbf{F}_\mathcal{U}}\mathbf{K}_{\mathbf{Z}_\mathcal{D}\widetilde{\mathbf{Z}}_\mathcal{U}}\boldsymbol{\Sigma}^{-1}_{\widetilde{\mathbf{Z}}_\mathcal{U}\widetilde{\mathbf{Z}}_\mathcal{U}}\right] \tag{17}$$

where we take expectation with respect to variable $\mathbf{Z}_\mathcal{D}$ which kind of serves as averaging over the different $\mathbf{Z}_\mathcal{D}$. This could give us the benefits of the Bayesian averaging which could be robust to outliers (suppose that for certain input $\mathbf{Z}_i$ which is quite far from others inputs.). The distribution of $\mathbf{Z}_\mathcal{D}$ is governed by the true input $\mathbf{X}_\mathcal{D}$ by another GP.

And the third item corresponds to the KL divergences among inducing variables. Which we do not want our approximation to be way to far from our prior assumption.

As for the last item $\frac{|\mathcal{D}|}{2}\ln\sigma_n^2$, as we have argued previously, the smaller the signal variance σ_n^2 is, the more concern we are about the loss $\left(\mathbf{Y}_\mathcal{D}-\boldsymbol{\mu}_{\mathbf{F}_\mathcal{D}}\right)^\top\left(\mathbf{Y}_\mathcal{D}-\boldsymbol{\mu}_{\mathbf{F}_\mathcal{D}}\right)$, the more capable our model is. I do think this makes sense since if we examine

$$\mathbb{H}\left[\mathbf{Y}_\mathcal{D}|\mathbf{F}_\mathcal{D}\right]=\frac{1}{2}\ln|\sigma_n^2\mathbf{I}| \tag{18}$$

where entropy is small, the latent function $\mathbf{F}_\mathcal{D}$ is able to recover the observations.

6 Prediction

In terms of prediction, we actually would like to incorporate the existing various kinds of sparse GPR models, such as FITC, PITC, PIC and LMA framework. If we write down the exact predictive distribution, it looks like:

$$p(f_{\mathbf{x}^*}|\mathbf{Y}_\mathcal{D})\approx\int p(f_{\mathbf{x}^*}|\mathbf{Y}_{\mathcal{D}_B},\mathbf{F}_\mathcal{U},\mathbf{z}^*)q(\mathbf{F}_\mathcal{U})p(\mathbf{z}^*|\mathbf{H}_\mathcal{U},\mathbf{x}^*)q(\mathbf{H}_\mathcal{U})l \tag{19}$$

Based on the lower bound we discussed previously, it appears that to make it possible for various of SGPR models only if we change the conditional probability on the top layer.

7 Nested Variational Compression in Deep Gaussian Process Regression

This section we would like to talk about what [Hensman and Lawrence, 2014] has done, in his nested variational compression DGPRs, we could write the model as:

$$\mathbf{y}=h_l(h_{l-1}(\ldots h_1(\mathbf{X}))) \tag{20}$$

Then we could write the conditional probabilistic model as

$$\begin{array}{rcl} p(h_1|\mathbf{X}) & = & \mathcal{N}(\mathbf{0},\mathbf{K}_{h_1h_1}+\sigma_1^2\mathbf{I}) \\ p(h_2|h_1) & = & \mathcal{N}(\mathbf{0},\mathbf{K}_{h_2h_2}+\sigma_2^2\mathbf{I}) \\ & \vdots & \\ p(\mathbf{y}|h_{l-1}) & = & \mathcal{N}(\mathbf{0},\mathbf{K}_{h_lh_l}+\sigma_l^2\mathbf{I}) \end{array} \tag{21}$$

Actually the author just impose this signal variance σ_i^2 into each layer.

Another issue is that how [Hensman and Lawrence, 2014] construct the lower bound. Let us take a single layer for example, since we know a GPR model with inducing variables could be written as:

$$p(\mathbf{y})=\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) \tag{22}$$

then by applying this VI framework we can have this following inequality:

$$\log p(\mathbf{y}|\mathbf{u}) \geq \log \mathcal{N}(\mathbf{y}|\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \sigma^2\mathbf{I}) - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{\Sigma}) \quad \text{where} \quad \mathbf{\Sigma} = \mathbf{K_{ff}} - \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}} \tag{23}$$

where we define $\tilde{p}(\mathbf{y}|\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \sigma^2\mathbf{I})$. Then we could write the above inequality as:

$$p(\mathbf{y}|\mathbf{u}) \geq \tilde{p}(\mathbf{y}|\mathbf{u}) \times \exp(-\frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{\Sigma})) \tag{24}$$

By augmenting into deeper layers, we can apply this variational compression to achieve a bound on the conditional probability:

$$\begin{aligned}
p(\mathbf{y}|\mathbf{u}_l, \mathbf{h}_{l-1}) &\geq \tilde{p}(\mathbf{y}|\mathbf{u}_l, \mathbf{h}_{l-1}) \times \exp(-\frac{1}{2\sigma_l^2}\mathrm{Tr}(\mathbf{\Sigma}_l)) && \text{where} && \mathbf{\Sigma}_l = \mathbf{K}_{\mathbf{h}_l\mathbf{h}_l} - \mathbf{K}_{\mathbf{h}_l\mathbf{u}_l}\mathbf{K}_{\mathbf{u}_l\mathbf{u}_l}^{-1}\mathbf{K}_{\mathbf{u}_l\mathbf{h}_l} \\
&\vdots \\
p(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1) &\geq \tilde{p}(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1) \times \exp(-\frac{1}{2\sigma_2^2}\mathrm{Tr}(\mathbf{\Sigma}_2)) && \text{where} && \mathbf{\Sigma}_2 = \mathbf{K}_{\mathbf{h}_2\mathbf{h}_2} - \mathbf{K}_{\mathbf{h}_2\mathbf{u}_2}\mathbf{K}_{\mathbf{u}_2\mathbf{u}_2}^{-1}\mathbf{K}_{\mathbf{u}_2\mathbf{h}_2} \\
p(\mathbf{h}_1|\mathbf{u}_1) &\geq \tilde{p}(\mathbf{h}_1|\mathbf{u}_1) \times \exp(-\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1)) && \text{where} && \mathbf{\Sigma}_1 = \mathbf{K}_{\mathbf{h}_1\mathbf{h}_1} - \mathbf{K}_{\mathbf{h}_1\mathbf{u}_1}\mathbf{K}_{\mathbf{u}_1\mathbf{u}_1}^{-1}\mathbf{K}_{\mathbf{u}_1\mathbf{h}_1}
\end{aligned} \tag{25}$$

Then the final lower bound could be written as

$$\begin{aligned}
\log p(\mathbf{y}) = &-\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1) - \textstyle\sum_{i=2}^{l} \frac{1}{2\sigma_i^2}\left\langle \mathbf{K}_{\mathbf{h}_i\mathbf{h}_i} - \mathbf{K}_{\mathbf{h}_i\mathbf{u}_i}\mathbf{K}_{\mathbf{u}_i\mathbf{u}_i}\mathbf{K}_{\mathbf{u}_i\mathbf{h}_i}\right\rangle_{q(\mathbf{h}_i)} - \textstyle\sum_{i=1}^{l}\mathrm{KL}(q(\mathbf{u}_i)||p(\mathbf{u}_i)) \\
&-\frac{1}{2\sigma_l^2}\left\langle \left(\mathbf{y} - \mathbf{K}_{\mathbf{h}_l\mathbf{u}_l}\mathbf{K}_{\mathbf{u}_l\mathbf{u}_l}^{-1}\mathbf{u}\right)^\top \left(\mathbf{y} - \mathbf{K}_{\mathbf{h}_l\mathbf{u}_l}\mathbf{K}_{\mathbf{u}_l\mathbf{u}_l}^{-1}\mathbf{u}\right)\right\rangle_{q(\mathbf{h}_l)q(\mathbf{u}_l)}
\end{aligned} \tag{26}$$

So we can learn that the above method of [Hensman and Lawrence, 2014] is composing this lower bound by multiplying a set of inequality functions. Compared with our lower bound (after composing several layers) which I listed below:

$$\begin{aligned}
\log p(\mathbf{y}) = &-\textstyle\sum_{i=1}^{l}\mathrm{KL}(q(\mathbf{u}_i)||p(\mathbf{u}_i)) \\
&-\frac{1}{2\sigma_l^2}\left\langle \left(\mathbf{y} - \mathbf{K}_{\mathbf{h}_l\mathbf{u}_l}\mathbf{K}_{\mathbf{u}_l\mathbf{u}_l}^{-1}\mathbf{u}\right)^\top \left(\mathbf{y} - \mathbf{K}_{\mathbf{h}_l\mathbf{u}_l}\mathbf{K}_{\mathbf{u}_l\mathbf{u}_l}^{-1}\mathbf{u}\right)\right\rangle_{q(\mathbf{h}_l)q(\mathbf{u}_l)}
\end{aligned} \tag{27}$$

So the question remains about why these two lower bound are different, the difference lies in that we drop the noises connecting each layers hence making it possible to drop these regularization terms

$$-\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1) - \sum_{i=2}^{l}\frac{1}{2\sigma_i^2}\left\langle \mathbf{K}_{\mathbf{h}_i\mathbf{h}_i} - \mathbf{K}_{\mathbf{h}_i\mathbf{u}_i}\mathbf{K}_{\mathbf{u}_i\mathbf{u}_i}\mathbf{K}_{\mathbf{u}_i\mathbf{h}_i}\right\rangle_{q(\mathbf{h}_i)}$$

We can also explain this difference in lower bound by stating that they are using a double stage approximation scheme, here I would like to illustrate with a simple 2-layer case, following Eq.25 we have:

$$\begin{aligned}
p(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1) &\geq \tilde{p}(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1) \times \exp(-\frac{1}{2\sigma_2^2}\mathrm{Tr}(\mathbf{\Sigma}_2)) && \text{where} && \mathbf{\Sigma}_2 = \mathbf{K}_{\mathbf{h}_2\mathbf{h}_2} - \mathbf{K}_{\mathbf{h}_2\mathbf{u}_2}\mathbf{K}_{\mathbf{u}_2\mathbf{u}_2}^{-1}\mathbf{K}_{\mathbf{u}_2\mathbf{h}_2} \\
p(\mathbf{h}_1|\mathbf{u}_1) &\geq \tilde{p}(\mathbf{h}_1|\mathbf{u}_1) \times \exp(-\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1)) && \text{where} && \mathbf{\Sigma}_1 = \mathbf{K}_{\mathbf{h}_1\mathbf{h}_1} - \mathbf{K}_{\mathbf{h}_1\mathbf{u}_1}\mathbf{K}_{\mathbf{u}_1\mathbf{u}_1}^{-1}\mathbf{K}_{\mathbf{u}_1\mathbf{h}_1}
\end{aligned} \tag{28}$$

therefore, we have

$$\begin{aligned}
&\log \textstyle\int p(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1)p(\mathbf{h}_1|\mathbf{u}_1)\mathrm{d}\mathbf{h}_1 \geq \log \textstyle\int \tilde{p}(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1)\exp(-\frac{1}{2\sigma_2^2}\mathrm{Tr}(\mathbf{\Sigma}_2))\ \tilde{p}(\mathbf{h}_1|\mathbf{u}_1)\exp(-\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1))\mathrm{d}\mathbf{h}_1 \\
&\log \textstyle\int \tilde{p}(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1)\exp(-\frac{1}{2\sigma_2^2}\mathrm{Tr}(\mathbf{\Sigma}_2))\ \tilde{p}(\mathbf{h}_1|\mathbf{u}_1)\exp(-\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1))\mathrm{d}\mathbf{h}_1 \\
&= \textstyle\int \tilde{p}(\mathbf{h}_1|\mathbf{u}_1)\log \tilde{p}(\mathbf{h}_2|\mathbf{u}_2, \mathbf{h}_1)\mathrm{d}\mathbf{h}_1 + \left\langle -\frac{1}{2\sigma_2^2}\mathrm{Tr}(\mathbf{\Sigma}_2)\right\rangle_{\tilde{p}(\mathbf{h}_1|\mathbf{u}_1)} + -\frac{1}{2\sigma_1^2}\mathrm{Tr}(\mathbf{\Sigma}_1)
\end{aligned} \tag{29}$$

The first inequality in Eq. 29 comes from the approximation of VI; the second inequality comes from applying this Jensen inequality. Hence we can achieve a tighter lower bound than his.

References

[Bui et al., 2016] Bui, T. D., Hernández-Lobato, D., Li, Y., Hernández-Lobato, J. M., and Turner, R. E. (2016). Deep gaussian processes for regression using approximate expectation propagation. *arXiv preprint arXiv:1602.04133*.

[Damianou, 2015] Damianou, A. (2015). *Deep Gaussian processes and variational propagation of uncertainty*. PhD thesis, University of Sheffield.

[Damianou and Lawrence, 2013] Damianou, A. C. and Lawrence, N. D. (2013). Deep gaussian processes. In *Proc. AISTATS*, pages 207–215.

[Duvenaud, 2014] Duvenaud, D. (2014). *Automatic Model Construction with Gaussian Processes*. PhD thesis, Computational and Biological Learning Laboratory, University of Cambridge.

[Hensman and Lawrence, 2014] Hensman, J. and Lawrence, N. D. (2014). Nested variational compression in deep gaussian processes. *arXiv preprint arXiv:1412.1370*.

[Neal, 1995] Neal, R. M. (1995). *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto.

[Titsias, 2009] Titsias, M. K. (2009). Variational model selection for sparse Gaussian process regression. Technical report, School of Computer Science, University of Manchester.

[Titsias and Lawrence, 2010] Titsias, M. K. and Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In *Proc. AISTATS*, pages 844–851.