Function-Space Variational Inference in the context of Implicit Processes

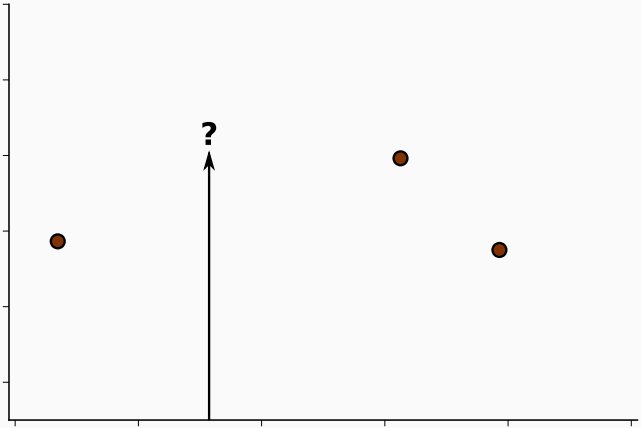
Machine Learning Group

Luis Antonio Ortega Andrés

December 1, 2022

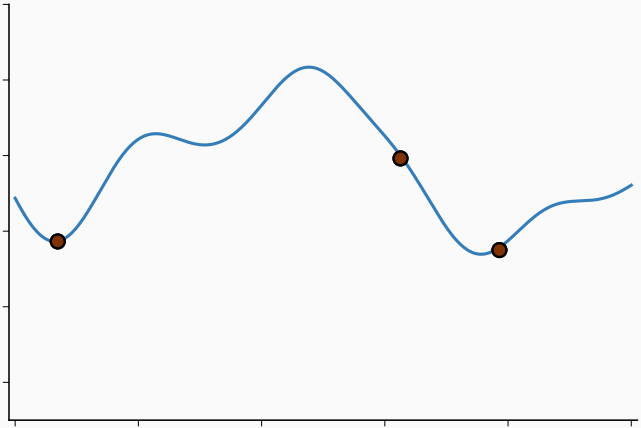
Autonomous University of Madrid

Motivation



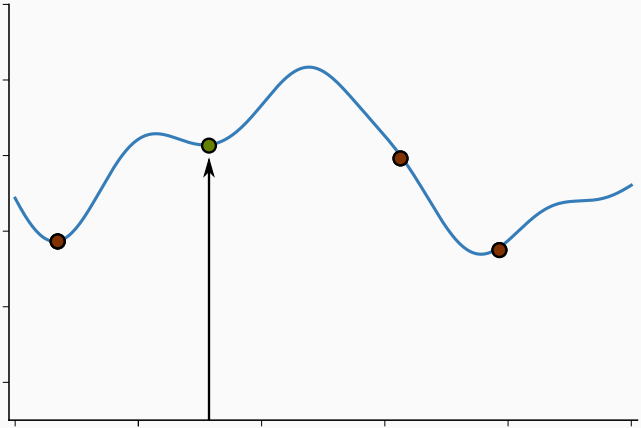
Objective: Make predictions over unknown points.

Motivation



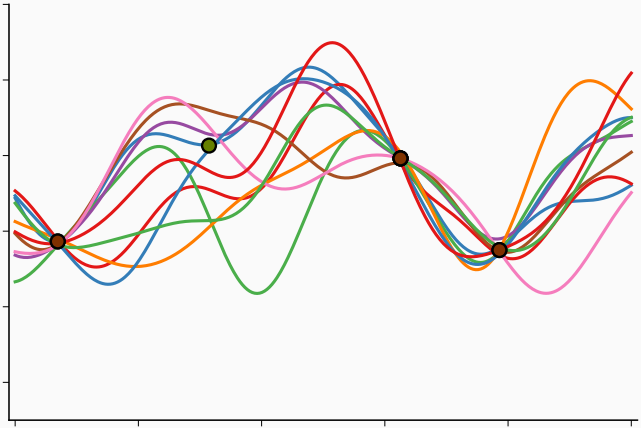
Procedure: Learn a function that explain the visible data.

Motivation



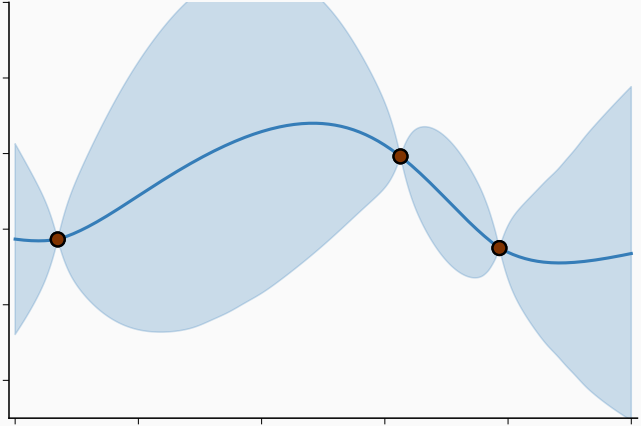
Outcome: The function can be used to predict.

Motivation



Question: How confident is the prediction?

Motivation



Answer: Bayesian approach.

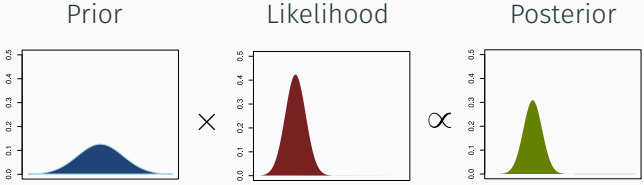
Overview

Bayesian Supervised Learning

Bayesian Supervised Learning

Objective: Learn an unknown function $f : \mathbb{R}^D \to \mathbb{R}^M$ given a set of observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \subset \mathbb{R}^D, \mathbf{y} = (y_1, \ldots, y_N) \subset \mathbb{R}^M$.

Approach. Consider a set of latent random variables $\mathbf{z}$ that model the generation of the dataset $P(\mathbf{y}|\mathbf{X}, \mathbf{z})$.

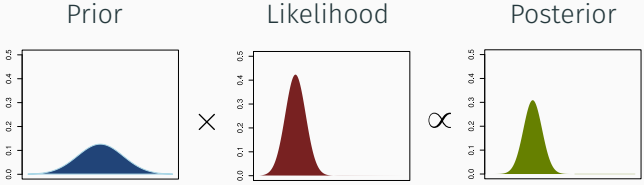


Problem. Prediction requires the posterior $P(\mathbf{z}|\mathbf{X}, \mathbf{y})$:

$$P(y_\star|\mathbf{x}_\star, \mathbf{X}, \mathbf{y}) = \int P(y_\star|\mathbf{x}_\star, \mathbf{z}) P(\mathbf{z}|\mathbf{X}, \mathbf{y})\, d\mathbf{z}\,,$$

which is usually intractable due to the integral

$$P(\mathbf{z}|\mathbf{X}, \mathbf{y}) = \frac{P(\mathbf{y}, \mathbf{z}|\mathbf{X})}{\int P(\mathbf{y}, \mathbf{z}|\mathbf{X})\, d\mathbf{z}}\,.$$



Variational Inference

Variational Inference

Idea. Approximate the posterior $P(\mathbf{z}|\mathbf{X}, \mathbf{y})$ with a simpler distribution $Q(\mathbf{z})$ and ensure that $KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z}|\mathbf{X}, \mathbf{y})\Big)$ is close to 0.

Formally,

$$\begin{aligned} Q^\star(\mathbf{z}) &= \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z}|\mathbf{X}, \mathbf{y})\Big) \\ &= \arg\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(\mathbf{z})}\Big[\log P(\mathbf{y}|\mathbf{X}, \mathbf{z})\Big] - KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z})\Big)\,. \end{aligned}$$

Variational Inference

Idea. Approximate the posterior $P(\mathbf{z}|\mathbf{X}, \mathbf{y})$ with a simpler distribution $Q(\mathbf{z})$ and ensure that $KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z}|\mathbf{X}, \mathbf{y})\Big)$ is close to 0.

Formally,

$$\begin{aligned} Q^\star(\mathbf{z}) &= \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z}|\mathbf{X}, \mathbf{y})\Big) \\ &= \arg\max_{Q \in \mathcal{Q}} \underbrace{\mathbb{E}_{Q(\mathbf{z})}\Big[\log P(\mathbf{y}|\mathbf{X}, \mathbf{z})\Big] - KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z})\Big)}_{\text{ELBO}} . \end{aligned}$$

Variational Inference

Idea. Approximate the posterior $P(\mathbf{z}|\mathbf{X}, \mathbf{y})$ with a simpler distribution $Q(\mathbf{z})$ and ensure that $KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z}|\mathbf{X}, \mathbf{y})\Big)$ is close to 0.

Formally,

$$
\begin{aligned}
Q^\star(\mathbf{z}) &= \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z}|\mathbf{X}, \mathbf{y})\Big) \\
&= \arg\max_{Q \in \mathcal{Q}} \underbrace{\mathbb{E}_{Q(\mathbf{z})}\Big[\log P(\mathbf{y}|\mathbf{X}, \mathbf{z})\Big]}_{\text{Data Fitting term}} - \underbrace{KL\Big(Q(\mathbf{z}) \mid P(\mathbf{z})\Big)}_{\text{Regularizer}} .
\end{aligned}
$$

Gaussian Processes

Gaussian Processes

A Gaussian process (GP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$ follows a Gaussian distribution

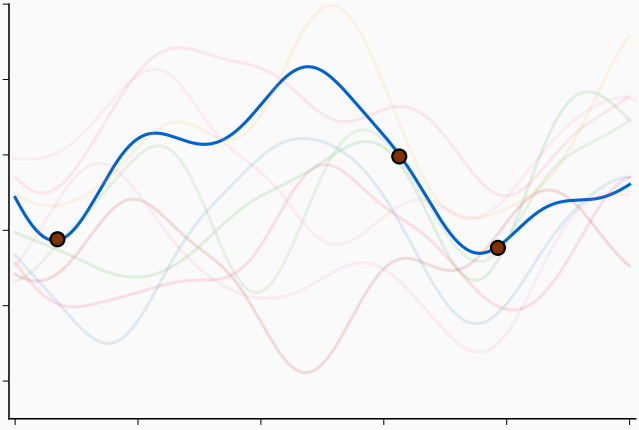
$$\mathbf{f} = f(\mathbf{X}) = \Big(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\Big) \sim \mathcal{N}\Big(m(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X})\Big)\,.$$

Gaussian processes place a **probability distribution over functions**.

Usually,

$$m(\mathbf{x}) = 0 \quad \text{and} \quad \kappa(\mathbf{x}_1, \mathbf{x}_2) = \sigma^2 \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{2l^2}\right)\,.$$

Hypothesis: The unknown function f is a (noisy) sample from a Gaussian process.



For a regression problem, using,

- the Gaussian process prior over functions,

$$P(\mathbf{f}) = \mathcal{N}\Big(m(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X})\Big)\,,$$

- a suitable likelihood

$$P(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})\,.$$

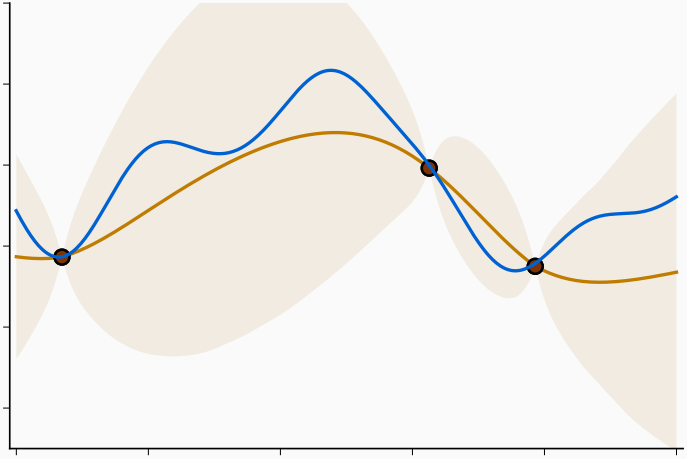
Predictive distribution is computable in closed form

$$P\big(f(\mathbf{x}^\star)|\mathbf{y}, \mathbf{X}\big) = \mathcal{N}(\mu^\star, \mathbf{\Sigma}^\star)\,,$$

with

$$\begin{aligned}\mu^\star &= m(\mathbf{x}^\star) + \kappa(\mathbf{x}^\star, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - m(\mathbf{X}))\,,\\ \mathbf{\Sigma}^\star &= \kappa(\mathbf{x}^\star, \mathbf{x}^\star) - \kappa(\mathbf{x}^\star, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})^{-1}\kappa(\mathbf{X}, \mathbf{x}^\star)\,.\end{aligned}$$

$$\mu^\star = m(\mathbf{x}^\star) + \kappa(\mathbf{x}^\star, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - m(\mathbf{X}))\,,$$
$$\mathbf{\Sigma}^\star = \kappa(\mathbf{x}^\star, \mathbf{x}^\star) - \kappa(\mathbf{x}^\star, \mathbf{X})(\kappa(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})^{-1}\kappa(\mathbf{X}, \mathbf{x}^\star)\,.$$



Hyper-parameters (kernel and likelihood) can be **optimized** using the marginal log likelihood:

$$\log P(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\det(\mathbf{K}+\sigma^2\mathbf{I}) - \frac{N}{2}\log 2\pi\,,$$

with $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$.

Limitation. Requires the computation of $(\mathbf{K}+\sigma^2\mathbf{I})^{-1}$ which is $\mathcal{O}(N^3)$.
Limitation. Mini-batches cannot be used for optimization.

Hyper-parameters (kernel and likelihood) can be **optimized** using the marginal log likelihood:

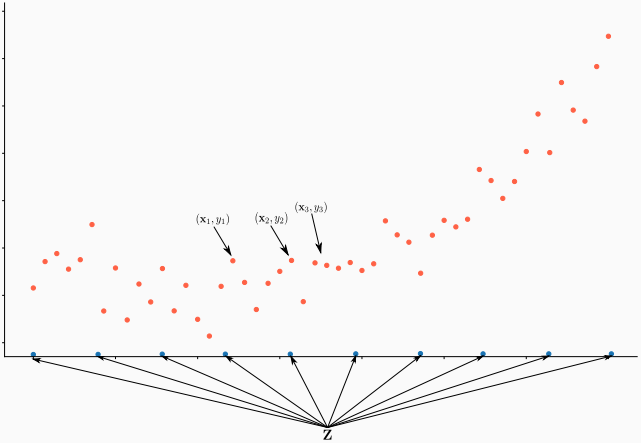
$$\log P(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\det(\mathbf{K}+\sigma^2\mathbf{I}) - \frac{N}{2}\log 2\pi\,,$$

with $\mathbf{K} = \kappa(\mathbf{X},\mathbf{X})$.

Limitation. Requires the computation of $(\mathbf{K}+\sigma^2\mathbf{I})^{-1}$ which is $\mathcal{O}(N^3)$.
Limitation. Mini-batches cannot be used for optimization.

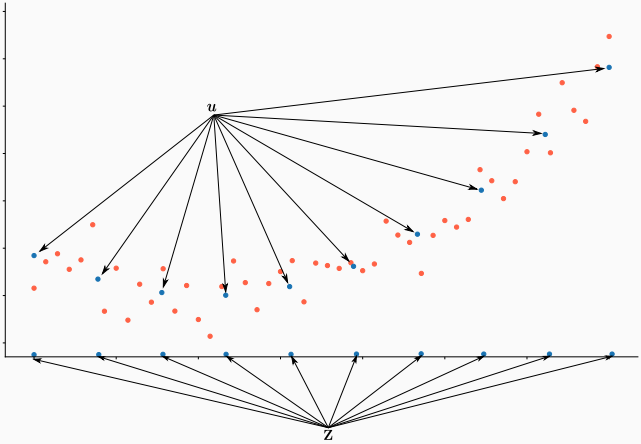
Sparse Gaussian Processes

Initialize a set of inducing locations $\mathbf{Z}$ with $|\mathbf{Z}| < |\mathbf{X}|$. For example $\mathbf{Z}$ can be the centers of applying KMeans to $\mathbf{X}$.



Sparse Gaussian Processes

Set a variational distribution over the inducing points $\boldsymbol{u} = f(\mathbf{Z})$, typically, $Q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$.

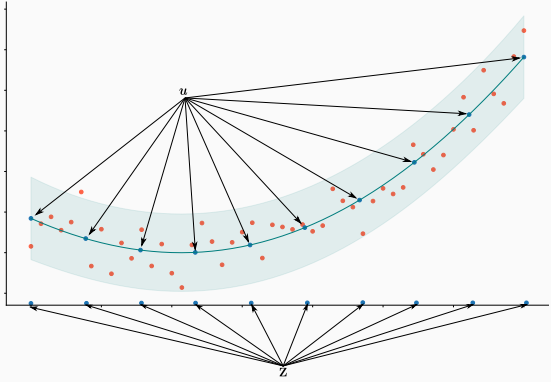


Sparse Gaussian Processes

The approximated posterior predictive distribution is,

$$Q(f(\mathbf{x}_\star)) = \int_{\boldsymbol{u}} P(f(\mathbf{x}_\star)|\boldsymbol{u})Q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

is Gaussian and suitable for making predictions.



Minimize the ELBO to optimize $Q(\boldsymbol{u}, \mathbf{f})$, with $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$,

$$\begin{aligned} Q^\star(\boldsymbol{u}, \mathbf{f}) &= \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\boldsymbol{u}, \mathbf{f}) \mid P(\boldsymbol{u}, \mathbf{f} | \mathbf{X}, \mathbf{y})\Big) \\ &= \arg\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(\mathbf{f})}\Big[\log P(\mathbf{y} | \mathbf{X}, \mathbf{f})\Big] - KL\Big(Q(\boldsymbol{u}) \mid P(\boldsymbol{u})\Big)\,. \end{aligned}$$

where

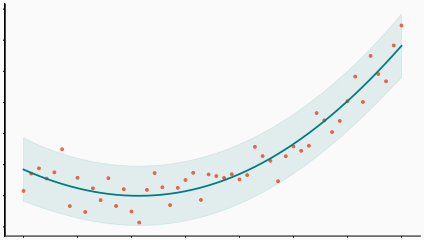
1. $P(\boldsymbol{u})$ is given by the Gaussian Process distribution evaluated at $\mathbf{Z}$.
2. The Kullback-Leibler divergence is between Gaussian distributions.
3. The data-fitting term can be computed for Gaussian likelihoods or approximated by quadrature.

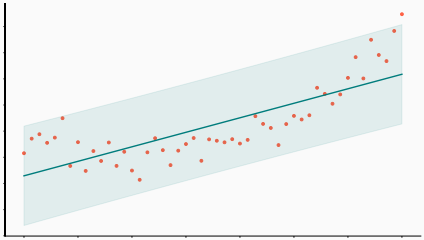
Implicit Processes

Implicit Processes

Motivation: Gaussian processes are limited by the parametric kernel family.

- Square exponential kernel: $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \sigma^2 \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{2l^2}\right)$.
- Linear kernel: $\kappa(\mathbf{x}_1, \mathbf{x}_2) = a\mathbf{x}_1^T\mathbf{x}_2 + b$.





The Gaussian process prior over functions is **too restrictive**.

An **implicit stochastic process**[1] (IP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$ is implicitly defined by the following generative process:

$$\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z}) \quad \textit{and} \quad f(\mathbf{x}_n) = g_\theta(\mathbf{x}_n, \mathbf{z}),\ \forall n = 1, \ldots, N\,.$$

[1] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.

An **implicit stochastic process**[1] (IP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$ is implicitly defined by the following generative process:

$$\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z}) \quad \textit{and} \quad f(\mathbf{x}_n) = g_\theta(\mathbf{x}_n, \mathbf{z}),\ \forall n = 1, \ldots, N\,.$$

Gaussian process

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \textit{and} \quad f(\mathbf{x}_n) = \boldsymbol{L}(\mathbf{x}_n)^T \mathbf{z},\ \forall n = 1, \ldots, N\,.$$

[1] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.

An **implicit stochastic process**[1] (IP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$ is implicitly defined by the following generative process:

$$\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z}) \quad \textit{and} \quad f(\mathbf{x}_n) = g_\theta(\mathbf{x}_n, \mathbf{z}),\ \forall n = 1, \ldots, N\,.$$

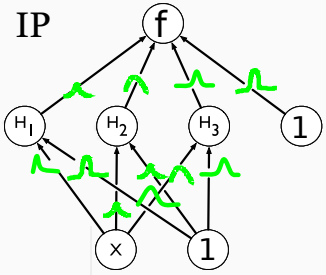
Bayesian Neural Networks.

$$(\mathbf{z}_1, \mathbf{z}_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$$\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$$

$$\boldsymbol{h} = r((\boldsymbol{\mu}_1 + \boldsymbol{\sigma}_1 \mathbf{z}_1)^T \mathbf{x}_n)\,.$$

$$g_\theta(\mathbf{x}_n, \mathbf{z}) = (\boldsymbol{\mu}_2 + \boldsymbol{\sigma}_2 \mathbf{z}_2)^T \boldsymbol{h}$$



[1] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.

Variational Inference on Implicit Processes

Problem statement:

1. The unknown target function **is a sample from an IP**, that is, an implicit distribution over stochastic processes $P(f(\cdot))$.
2. Given the set of observations $(\mathbf{X}, \mathbf{y})$, we aim to approximate the posterior distribution over functions $P(f(\cdot)|\mathbf{X}, \mathbf{y})$.

Approach: Use variational inference over the function-space distribution $Q(f(\cdot))$.

$$
\begin{aligned}
Q^\star(f(\cdot)) &= \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(f(\cdot)) \mid P(f(\cdot)|\mathbf{X}, \mathbf{y})\Big) \\
&= \arg\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(f)}\Big[\log P(\mathbf{y}|\mathbf{X}, f(\mathbf{X}))\Big] - KL\Big(Q(f(\cdot)) \mid P(f(\cdot))\Big)\,.
\end{aligned}
$$

Approach: Use variational inference over the function-space distribution

$$Q^\star(f(\cdot)) = \arg\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(f)}\Big[\log P(\mathbf{y}|\mathbf{X}, f(\mathbf{X}))\Big] - KL\Big(Q(f(\cdot)) \mid P(f(\cdot))\Big).$$

Difficulties:

1. The prior $P(f(\cdot))$ lacks a closed form.
2. The Kullback-Leibler divergence between stochastic processes is not well-defined.
3. The variational distribution $Q(f(\cdot))$ must allow to compute or approximate by samples the data-fitting term.

Existing Approaches

Variational Implicit Processes[2]: Approximates the distribution over functions using a **linear combination of samples**.

Sparse Implicit Processes[3]: Uses **inducing points** for scalability and approximates the KL using an **external discriminator** (a Neural Network).

Linearized approximation[4]: Approximates the distribution over functions using a **linearization** of the BNN over the parameters.

[2] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.
[3] Rodríguez-Santana, S., Zaldivar, B. & Hernandez-Lobato, D.. (2022). Function-space Inference with Sparse Implicit Processes.
[4] Rudner, T., Chen, Z., Whye Y. & Gal, Y.. (2022). Tractable Function-Space Variational Inference in Bayesian Neural Networks.

Variational Implicit Processes

Approximate $P(f(\cdot))$ with a GP $P_{\mathcal{GP}}(f(\cdot))$ **based on samples** $f_1(\cdot), \ldots, f_S(\cdot)$. Let

$$\hat{m}(\mathbf{x}) = \frac{1}{S}\sum_{s=1}^{S} f_s(\mathbf{x}), \quad \hat{\phi}(\mathbf{x}) = \frac{1}{\sqrt{S}}\Big(f_1(\mathbf{x}) - \hat{m}(\mathbf{x}), \ldots, f_S(\mathbf{x}) - \hat{m}(\mathbf{x})\Big)^T.$$

Then, setting a standard **Gaussian prior** $P(\mathbf{a}) = \mathcal{N}(\mathbf{a}|\mathbf{0}, \mathbf{I})$,

$$\hat{f}(\mathbf{x}) = \hat{m}(\mathbf{x}) + \mathbf{a}^T\hat{\phi}(\mathbf{x}) \implies P_{\mathcal{GP}}(\hat{f}(\mathbf{x})) = \mathcal{N}(\hat{m}(\mathbf{x}), \hat{\phi}(\mathbf{x})^T\hat{\phi}(\mathbf{x}))\,.$$

Using a variational distribution $Q(\mathbf{a}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ induces a variational distribution over functions

$$Q(\hat{f}(\mathbf{x})) = \int_{\boldsymbol{a}} P(\hat{f}(\mathbf{x})|\boldsymbol{a})Q(\boldsymbol{a}) = \mathcal{N}\Big(\hat{m}(\mathbf{x}) + \hat{\phi}(\mathbf{x})^T\mathbf{m}, \hat{\phi}(\mathbf{x})^T\mathbf{S}\hat{\phi}(\mathbf{x})\Big)\,.$$

Naming

$$\hat{\mathbf{f}} = (\hat{f}(\mathbf{x}_1), \ldots, \hat{f}(\mathbf{x}_N)) \,.$$

The ELBO is computed to minimized the KL divergence evaluated on $\hat{\mathbf{f}}$ and $\boldsymbol{a}$ rather than between stochastic processes

$$\cancel{Q^\star(f(\cdot)) = \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(f(\cdot)) \mid P(f(\cdot)|\mathbf{X}, \mathbf{y})\Big)}$$

$$Q^\star(\hat{\mathbf{f}}, \boldsymbol{a}) = \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\hat{\mathbf{f}}, \boldsymbol{a}) \mid P(\hat{\mathbf{f}}, \boldsymbol{a}|\mathbf{X}, \mathbf{y})\Big)$$

$$= \arg\max_{Q \in \mathcal{Q}} \mathbb{E}_{Q(\hat{\mathbf{f}})}\Big[\log P(\mathbf{y}|\mathbf{X}, \hat{\mathbf{f}})\Big] - KL\Big(Q(\boldsymbol{a}) \mid P(\boldsymbol{a})\Big) \,.$$

Sparse Implicit Processes

1. Considers an **inducing points** approach, leading to the ELBO
$$\mathcal{L} = \mathbb{E}_{Q(\mathbf{f})}\Big[\log P(\mathbf{y}|\mathbf{X},\mathbf{f})\Big] - KL\Big(Q(\boldsymbol{u}) \mid P(\boldsymbol{u})\Big)\,.$$
2. The variational distribution of the inducing points $Q(\boldsymbol{u})$ is an **IP**.
3. The Kullback-Leibler term is approximated using an **external discriminator** T (a neural network),
$$KL\Big(Q(\boldsymbol{u}) \mid P(\boldsymbol{u})\Big) \approx \mathbb{E}_{Q(\boldsymbol{u})}[T(\boldsymbol{u})]\,.$$
4. The data-fitting term is **approximated using Monte-Carlo** samples of $Q(\boldsymbol{u})$,
$$Q(\mathbf{f}) = \int_{\boldsymbol{u}} P(\mathbf{f}|\boldsymbol{u})Q(\boldsymbol{u}) \approx \frac{1}{S}\sum_{s=1}^{S} P(\mathbf{f}|\boldsymbol{u}_s)\,,$$
where $P(\mathbf{f}|\boldsymbol{u})$ is approximated as Gaussian with mean and covariances estimated empirically from the prior.

Linearized approximation

Seeing the stochastic function $f(\cdot, \boldsymbol{\theta})$ defined in terms of the stochastic parameters $\boldsymbol{\theta}$ according to a distribution $P_{\boldsymbol{\Theta}}$, with

$$P_{\boldsymbol{\Theta}} = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S}) \,.$$

The stochastic function can be approximated with a **Taylor approximation of order 1 over the parameter space, centered on $\boldsymbol{m}$,**

$$f(\cdot, \boldsymbol{\theta}) \approx \hat{f}(\cdot, \boldsymbol{\theta}) = f(\cdot, \boldsymbol{m}) + \mathcal{J}(\cdot, \boldsymbol{m})(\boldsymbol{\theta} - \boldsymbol{m}) \,,$$

where

$$\mathcal{J}(\cdot, \boldsymbol{m}) = \frac{\partial f(\cdot, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\boldsymbol{m}) \,.$$

As a result, $\hat{f}(\cdot, \boldsymbol{\theta})$ is a Gaussian process,

$$\hat{P}(\hat{f}(\cdot, \boldsymbol{\theta})) = \mathcal{N}(f(\cdot, \boldsymbol{m}), \mathcal{J}(\cdot, \boldsymbol{m}) \boldsymbol{S} \mathcal{J}(\cdot, \boldsymbol{m})^T) \,.$$

$$\mathcal{L} = \mathbb{E}_{Q(f))}\Big[\log P(\mathbf{y}|\mathbf{X}, f(\mathbf{X}))\Big] - KL\Big(Q(f(\cdot)) \mid P(f(\cdot))\Big)$$

Using a variational distribution over the parameters $Q_{\mathbf{\Theta}}$, **leads to a variational distribution over the stochastic functions** $Q(f)$, where samples can be easily taken.

The Kullback-Leibler divergence between stochastic processes can be defined using evaluations

$$KL\Big(Q(f(\cdot)) \mid P(f(\cdot))\Big) = \sup_{\mathbf{C}\in 2^{\mathcal{X}}} KL\Big(Q(f(\mathbf{C})) \mid P(f(\mathbf{C}))\Big)\,.$$

Therefore, **approximated empirically** using the linearized models on a set of **Context points** $\{\mathbf{C}_1, \ldots, \mathbf{C}_S\}$,

$$KL\Big(Q(f(\cdot)) \mid P(f(\cdot))\Big) \approx \max_{\mathbf{C}\in\{\mathbf{C}_1,\ldots,\mathbf{C}_S\}} KL\Big(\hat{Q}(\hat{f}(\mathbf{C})) \mid \hat{P}(\hat{f}(\mathbf{C}))\Big)\,,$$

which is a KL between Gaussian distributions.

Limitations

- **Variational implicit processes**. The linear approximation can be too strong,
$$f(\mathbf{x}) \approx \hat{m}(\mathbf{x}) + \boldsymbol{a}^T \hat{\phi}(\mathbf{x}) \,.$$
- **Sparse Implicit Processes**. Relies on an external discriminator T, increasing the training time due to a double-loop training.
$$KL\Big(Q(\boldsymbol{u}) \mid P(\boldsymbol{u})\Big) \approx \mathbb{E}_{Q(\boldsymbol{u})}[T(\boldsymbol{u})] \,.$$
- **Linearized approximation**. The set of **Context points** $\{\mathbf{C}_1, \ldots, \mathbf{C}_S\}$ must be defined by hand previously to any learning. With points both **in the training space and out of the training space** to ensure generalization.
$$KL\Big(Q(f(\cdot)) \mid P(f(\cdot))\Big) \approx \max_{\mathbf{C} \in \{\mathbf{C}_1, \ldots, \mathbf{C}_S\}} KL\Big(\hat{Q}(\hat{f}(\mathbf{C})) \mid \hat{P}(\hat{f}(\mathbf{C}))\Big) \,.$$

Sparse Linearized Implicit Processes

Linearized model with inducing points

We propose to use the linearized model among with the usage of inducing points to, simultaneusly,

1. **Avoid using Context points** to approximate the KL divergence between stochastic processes
2. **Avoid using a discriminator** for the KL divergence between IPs.

Features:

1. The variational distribution over the inducing points $Q(\boldsymbol{u})$ is Gaussian.
2. Both $P(\boldsymbol{u})$ and $P(\mathbf{f}|\boldsymbol{u})$ are approximated using the linearized model rather than samples from the IP.

How are $P(\boldsymbol{u})$ and $P(\mathbf{f}|\boldsymbol{u})$ approximated?

Consider the concatenation of the input features and the inducing locations $(\mathbf{X}, \mathbf{Z})$, and the linearized approximation of the prior evaluated on them,

$$\hat{f}\big((\mathbf{X}, \mathbf{Z}), \boldsymbol{\theta}\big) = f\big((\mathbf{X}, \mathbf{Z}), \boldsymbol{m}\big) + \mathcal{J}\big((\mathbf{X}, \mathbf{Z}), \boldsymbol{m}\big)(\boldsymbol{\theta} - \boldsymbol{m})\,.$$

Then, $\hat{f}\big((\mathbf{X}, \mathbf{Z}), \boldsymbol{\theta}\big)$ is a Gaussian process,

$$\hat{P}(\mathbf{f}, \boldsymbol{u}) = \mathcal{N}\left(\begin{pmatrix} f(\mathbf{X}, \boldsymbol{m}) \\ f(\mathbf{Z}, \boldsymbol{m}) \end{pmatrix}, \begin{pmatrix} \mathcal{J}(\mathbf{X}, \boldsymbol{m})\boldsymbol{S}\mathcal{J}(\mathbf{X}, \boldsymbol{m})^T & \mathcal{J}(\mathbf{X}, \boldsymbol{m})\boldsymbol{S}\mathcal{J}(\mathbf{Z}, \boldsymbol{m})^T \\ \mathcal{J}(\mathbf{Z}, \boldsymbol{m})\boldsymbol{S}\mathcal{J}(\mathbf{X}, \boldsymbol{m})^T & \mathcal{J}(\mathbf{Z}, \boldsymbol{m})\boldsymbol{S}\mathcal{J}(\mathbf{Z}, \boldsymbol{m})^T \end{pmatrix}\right)\,.$$

where $\hat{P}(\mathbf{f}|\boldsymbol{u})$ and $\hat{P}(\boldsymbol{u})$ can be easily computed.

The variational posterior distribution can be computed in closed form to be Gaussian

$$Q(\mathbf{f}) = \int_{\boldsymbol{u}} \hat{P}(\mathbf{f}|\boldsymbol{u})Q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,.$$

The ELBO can be easily computed for regression and approximated for classification

$$\mathcal{L} = \mathbb{E}_{Q(\mathbf{f})}\Big[\log P(\mathbf{y}|\mathbf{X}, \mathbf{f})\Big] - KL\Big(Q(\boldsymbol{u}) \mid \hat{P}(\boldsymbol{u})\Big)\,.$$

Experiments

How good is the Taylor approximation?

$$f(\cdot, \boldsymbol{\theta}) \approx \hat{f}(\cdot, \boldsymbol{\theta}) = f(\cdot, \boldsymbol{m}) + \mathcal{J}(\cdot, \boldsymbol{m})(\boldsymbol{\theta} - \boldsymbol{m})$$

Advantages:

1. The mean of the approximation is in the support of the function-space distribution.
2. Does not rely on taking samples of the prior, avoiding an hyperparameter that depends on the dimensionality of the data.

Disadvantages:

1. The approximation is degenerate when $\mathbb{E}_P[\boldsymbol{\theta}] = \boldsymbol{m} = \mathbf{0}$.

Assume that

$$f(\cdot,(\boldsymbol{w}_1,\boldsymbol{w}_2))=\boldsymbol{w}_2\tanh(\boldsymbol{w}_1\mathbf{x}),\quad \text{and}\quad \mathbb{E}_P[(\boldsymbol{w}_1,\boldsymbol{w}_2)]=(\boldsymbol{m}_1,\boldsymbol{m}_2)=(\mathbf{0},\mathbf{0})\,.$$

then

$$\mathcal{J}(\mathbf{x},\boldsymbol{m})=\frac{\partial f(\mathbf{x},(\boldsymbol{w}_1,\boldsymbol{w}_2))}{\partial(\boldsymbol{w}_1,\boldsymbol{w}_2)}(\boldsymbol{m}_1,\boldsymbol{m}_2)=\begin{pmatrix}\boldsymbol{m}_2(1-\tanh(\boldsymbol{m}_1\mathbf{x})^2)\mathbf{x}\\ \tanh(\boldsymbol{m}_1\mathbf{x})\end{pmatrix}=\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}$$

Meaning that

$$f(\cdot,\boldsymbol{\theta})\approx\hat{f}(\cdot,\boldsymbol{\theta})=f(\cdot,\boldsymbol{m})+\cancel{\mathcal{J}(\cdot,\boldsymbol{m})(\boldsymbol{\theta}-\boldsymbol{m})}\,.$$

However, this can be solved using random initialization of the mean values of the prior distribution $P_{\boldsymbol{\Theta}}$.

Is the Taylor GP close to the true GP?

We want to test if in cases where $P(f(\cdot))$ is a GP, the Taylor approximation GP is a good approximation.

Approach. Create a Bayesian Neural Network whose implicit distribution equals that of a Gaussian Process.

1. **Squared exponential Kernel**: Single hidden layer BNN with cos activation and infinite width. Gaussian weights and uniform biases.
2. **Gaussian c.d.f activation**: Single hidden layer BNN with Gaussian c.d.f activation.

GP approximation

Let

$$f(\mathbf{x}) = \frac{1}{\sqrt{H}} \boldsymbol{w}_2^T \phi(\boldsymbol{w}_1^T \mathbf{x} + \boldsymbol{b}_1) + b_2 \,,$$

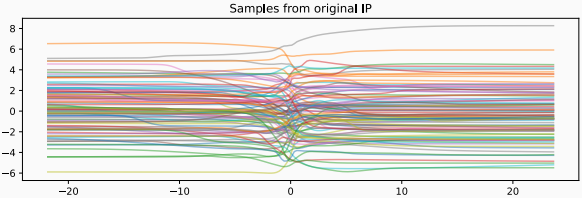
where the dimensionality of the parameter verctors is H

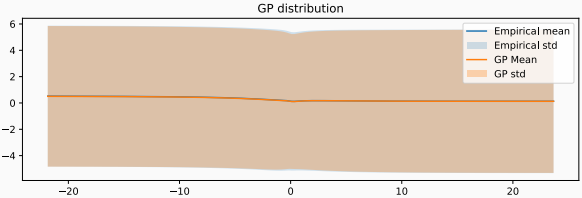
$$\begin{aligned} \boldsymbol{w}_1 &\sim \mathcal{N}(\boldsymbol{m}_{w_1}, \boldsymbol{\sigma}_{w_1} \boldsymbol{I}), \qquad & \boldsymbol{b}_1 &\sim \mathcal{N}(\boldsymbol{m}_{b_1}, \boldsymbol{\sigma}_{b_1} \boldsymbol{I}) \,, \\ \boldsymbol{w}_2 &\sim \mathcal{N}(\boldsymbol{m}_{w_2}, \boldsymbol{\sigma}_{w_2} \boldsymbol{I}), \qquad & \boldsymbol{b}_2 &\sim \mathcal{N}(\boldsymbol{m}_{b_2}, \boldsymbol{\sigma}_{b_2} \boldsymbol{I}) \,. \end{aligned}$$

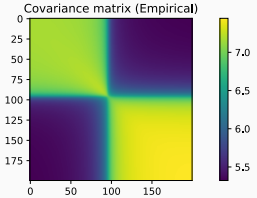
The distribution of $f(\mathbf{x})$ tends to a GP when $H \to \infty$. The mean and covariance of $f(\mathbf{x})$ can be computed using 1-dimensional quadrature.

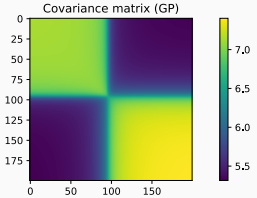
In practise, $H = 20$ is enough to approximate the GP.

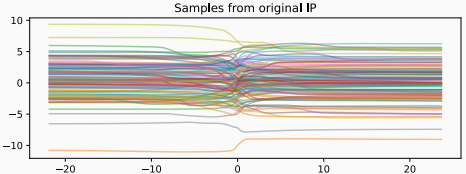
Showing that $H = 20$ is good enough

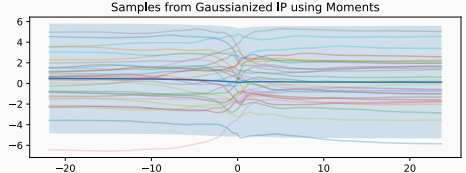


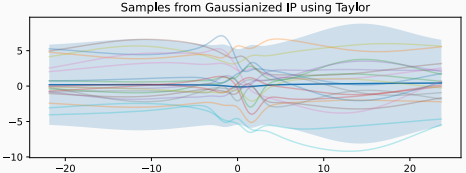


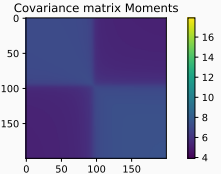


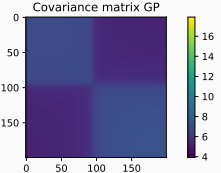


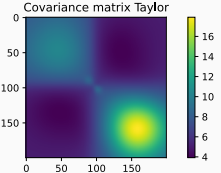

Samples from original IP
10
5
0
−5
−10
−20
−10
0
10
20


Samples from Gaussianized IP using Moments
6
4
2
0
−2
−4
−6
−20
−10
0
10
20


Samples from Gaussianized IP using Taylor
5
0
−5
−10
−20
−10
0
10
20


Covariance matrix Moments
0
50
100
150
0
50
100
150
16
14
12
10
8
6
4


Covariance matrix GP
0
50
100
150
0
50
100
150
16
14
12
10
8
6
4


Covariance matrix Taylor
0
50
100
150
0
50
100
150
16
14
12
10
8
6
4

Exact posterior distributions

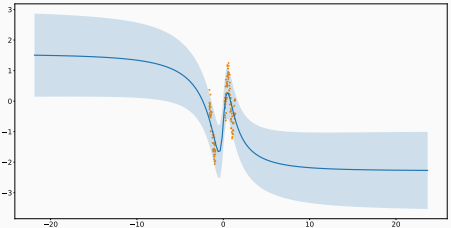
We are considering three main models based on the stochastic function

$$f(\mathbf{x}) = \frac{1}{\sqrt{H}} \boldsymbol{w}_2^T \phi(\boldsymbol{w}_1^T \mathbf{x} + \boldsymbol{b}_1) + b_2 \,.$$

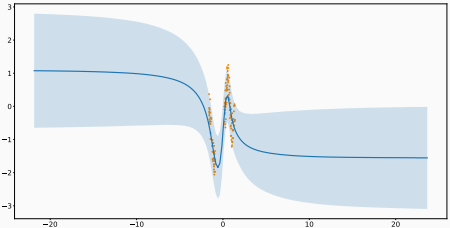
1. The **exact GP** of the stochastic function.
2. A GP where the mean and covariance are estimated using **samples** from the stochastic function.
3. A GP where the mean and covariance are the ones obtained from the **Taylor** approximation.

We are testing these three methods on a toy 1-D dataset where the exact GP posteriors can be computed.

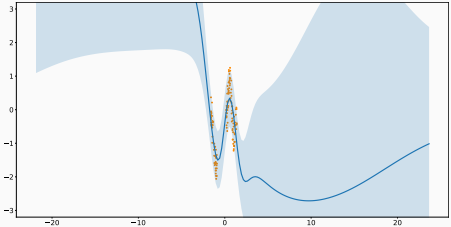
Theoretical GP



GP Samples



GP Taylor



Conclusions

Conclusions

- The Taylor approximation of order 1 is **not really close** to the exact distribution in the tested cases.
- **However**, preliminary results show that the predictive distribution close to the known points **is good enough** to encourage more research and testing.
- If the Taylor approximation results usable in practise, the proposed method would **avoid some of the problems of other function-space approaches**.
- The family of priors on which the Taylor approximation is good enough needs to be studied.

Conclusions

- The Taylor approximation of order 1 is **not really close** to the exact distribution in the tested cases.
- **However**, preliminary results show that the predictive distribution close to the known points **is good enough** to encourage more research and testing.
- If the Taylor approximation results usable in practise, the proposed method would **avoid some of the problems of other function-space approaches**.
- The family of priors on which the Taylor approximation is good enough needs to be studied.

Conclusions

- The Taylor approximation of order 1 is **not really close** to the exact distribution in the tested cases.
- **However**, preliminary results show that the predictive distribution close to the known points **is good enough** to encourage more research and testing.
- If the Taylor approximation results usable in practise, the proposed method would **avoid some of the problems of other function-space approaches**.
- The family of priors on which the Taylor approximation is good enough needs to be studied.

Conclusions

- The Taylor approximation of order 1 is **not really close** to the exact distribution in the tested cases.
- **However**, preliminary results show that the predictive distribution close to the known points **is good enough** to encourage more research and testing.
- If the Taylor approximation results usable in practise, the proposed method would **avoid some of the problems of other function-space approaches**.
- The family of priors on which the Taylor approximation is good enough needs to be studied.

Thank you for your attention