Input Dependent Sparse Gaussian Processes

Bahram Jafrasteh[1,2,*], Carlos Villacampa-Calvo[1,*], Daniel Hernández-Lobato[1]

[1]Computer Science Department, Universidad Autónoma de Madrid, Madrid, Spain

[2]Biomedical Research and Innovation Institute of Cádiz (INiBICA) Research Unit, Puerta del Mar University, Cádiz, Spain

[*]Equal contribution

July 2022

Paper at https://arxiv.org/pdf/2107.07281.pdf
Code at https://github.com/BahramJafrasteh/IDSGP

Gaussian Process

Having a set of data, we assume that $y_i = f(\mathbf{x_i}) + \epsilon_i$, with $f(\cdot)$ a latent function and ϵ_i Gaussian noise with variance σ^2, *i.e.*, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

A Gaussian process (GP) can be used as a prior for $f(\cdot)$.

Then, the posterior of f at a new point $\mathbf{x}^\star$ is Gaussian with mean and variance

$$\mu(\mathbf{x}^\star) = \mathbf{k}(\mathbf{x}^\star)^\mathsf{T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}\,,$$
$$\sigma^2(\mathbf{x}^\star) = k^\star - \mathbf{k}(\mathbf{x}^\star)^\mathsf{T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}(\mathbf{x}^\star)\,,$$

The cost of this approach is $O(N^3)$ since it needs the inversion of $\mathbf{K}$, a $N \times N$ matrix. This makes GPs unsuitable for large data sets.

Variational Sparse GPs

The most popular methods for Sparse GPs are using a new set of $M \ll N$ points , called the inducing points.

We focus on a widely used variational inference (VI) approach to approximate the posterior for f to improve the cost of Gaussian process.
In VI, the goal is to find an approximate posterior for $\mathbf{f}$ and $\mathbf{u}$, $q(\mathbf{f}, \mathbf{u})$, that resembles as much as possible the true posterior $p(\mathbf{f}, \mathbf{u}|\mathbf{y})$.
Then, the evidence lower bound (or ELBO) is:

$$\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{f})}[\log p(y_i|f_i)] - \mathrm{KL}[q(\mathbf{u})|p(\mathbf{u})]\,,$$

The expressive power of the VSGPs depends on the number of inducing points M and their correct location on the input space.

Input Dependent SGPs

- we consider a meta-point $\tilde{\mathbf{x}}$ that is used to determine the inducing points $\mathbf{Z}$ and the corresponding $\mathbf{u}$.
- Using posterior's factorization and Jensen's inequality we obtain the lower bound after some simplifications:

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}_q \left[\log \frac{p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{u}) p(\mathbf{u}|\tilde{\mathbf{x}}) p(\tilde{\mathbf{x}})}{q(\mathbf{f}, \mathbf{u}, \tilde{\mathbf{x}})} \right] \\
&= \mathbb{E}_q \left[\log \frac{p(\mathbf{y}|\mathbf{f}) \cancel{p(\mathbf{f}|\mathbf{u})} p(\mathbf{u}|\tilde{\mathbf{x}}) \cancel{p(\tilde{\mathbf{x}})}}{\cancel{p(\mathbf{f}|\mathbf{u})} q(\mathbf{u}|\tilde{\mathbf{x}}) \cancel{p(\tilde{\mathbf{x}})}} \right] \\
&= \sum_{i=1}^{N} \int p(\tilde{\mathbf{x}}) \left[p(f_i|\mathbf{u}) q(\mathbf{u}|\tilde{\mathbf{x}}) \log p(y_i|f_i) d\mathbf{f} d\mathbf{u} \right. \\
&\quad \left. - \frac{1}{N} \mathrm{KL}[q(\mathbf{u}|\tilde{\mathbf{x}}) | p(\mathbf{u}|\tilde{\mathbf{x}})] \right] d\tilde{\mathbf{x}} \,.
\end{aligned}
$$

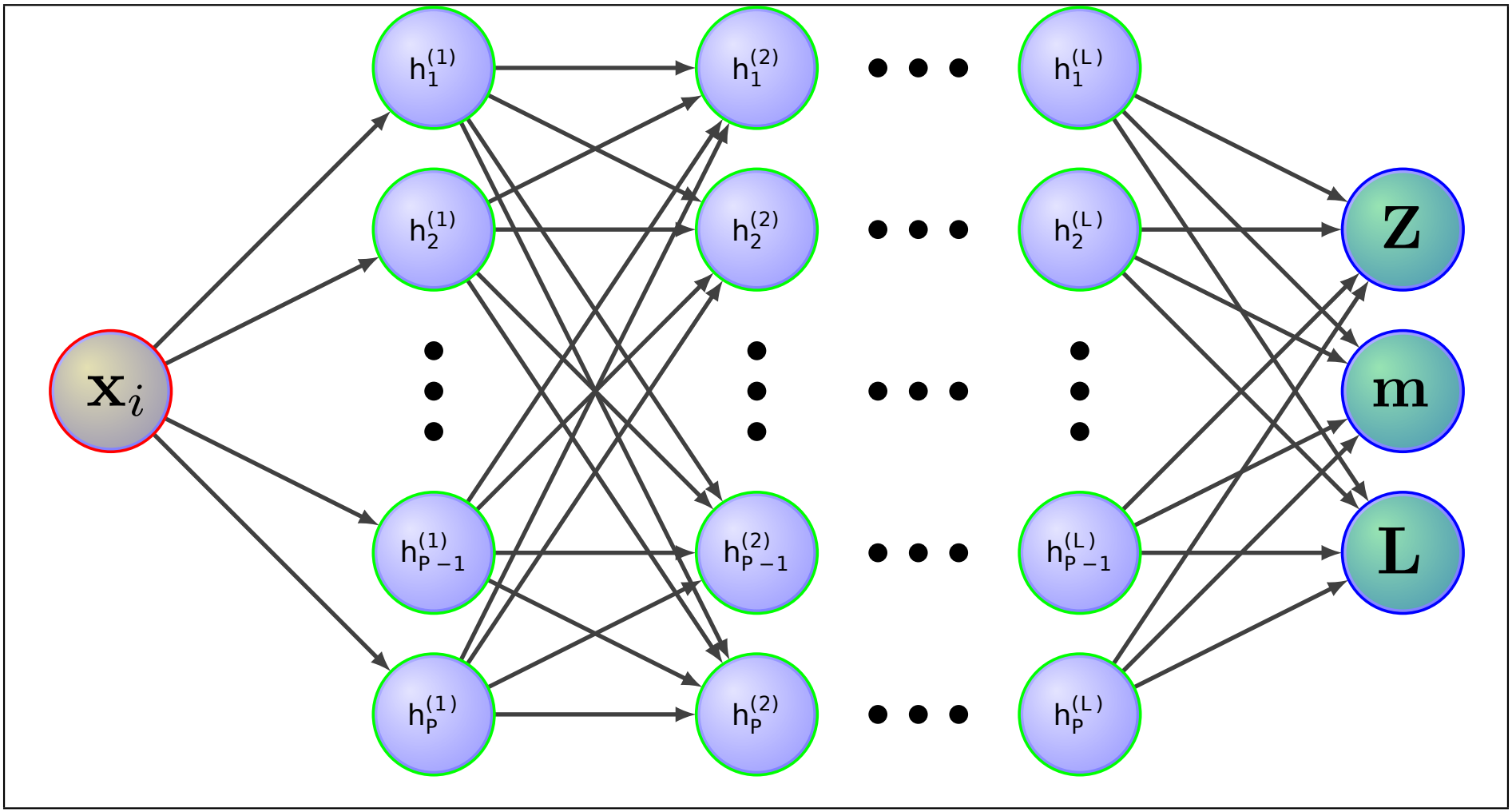
Input Dependent SGPs

- Assume that $p(\tilde{\mathbf{x}})$ is an implicit distribution. We can draw samples from it and approximate the expectation w.r.t $p(\tilde{\mathbf{x}})$. Thus, for a sample $\tilde{\mathbf{x}}_s$ from $p(\tilde{\mathbf{x}})$, ELBO is approximated as

$$\mathcal{L} \approx \sum_{i=1}^{N} \left[\mathbb{E}_{p(f_i|\mathbf{u})q(\mathbf{u}|\tilde{\mathbf{x}}_s)}[\log p(y_i|f_i)] - \frac{1}{N}\mathrm{KL}[q(\mathbf{u}|\tilde{\mathbf{x}}_s)|p(\mathbf{u}|\tilde{\mathbf{x}}_s)] \right] .$$

- Consider now that we use mini-batch-based training for optimization, and we set $\tilde{\mathbf{x}}_s = \mathbf{x}_i$.

Input Dependent SGPs (Amortization)



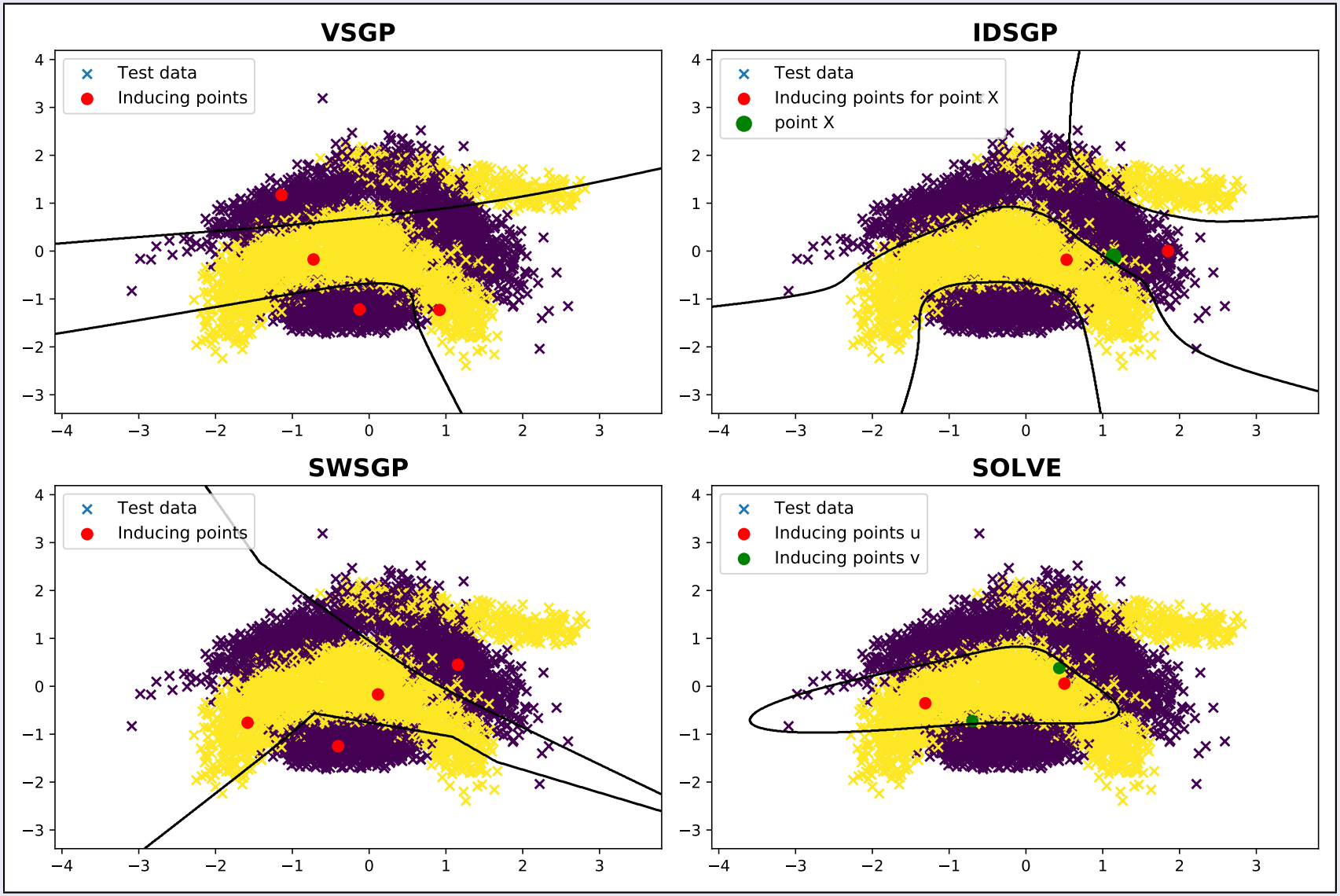
IDSGPs (Prediction)

The predictive distribution for $f(\mathbf{x}^\star)$ is Gaussian with mean and variance:

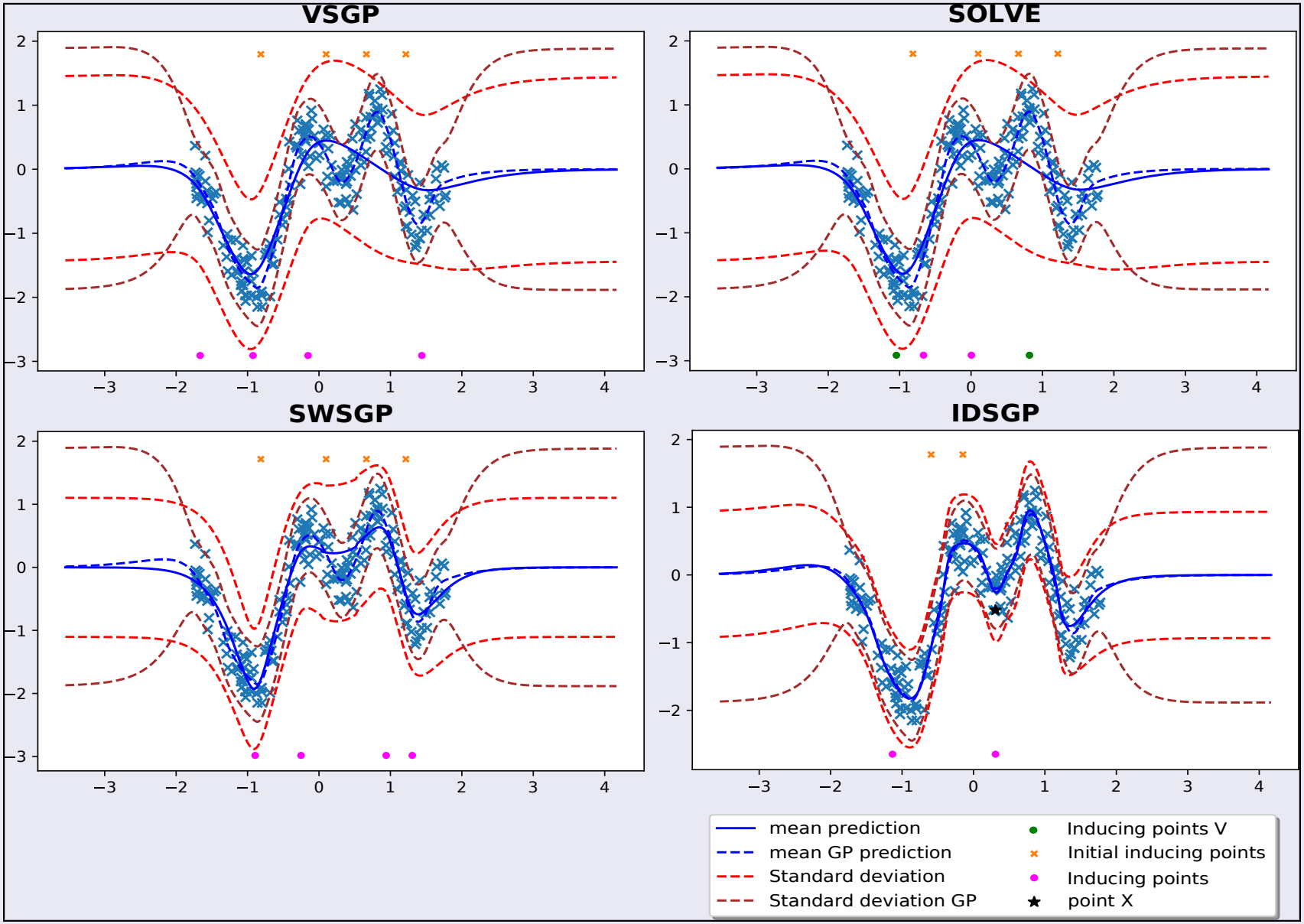
$$m^\star = \mathbf{k}_{\mathbf{x}^\star,\mathbf{Z}}\mathbf{K}_\mathbf{Z}^{-1}\mathbf{m}\,,$$
$$s^\star = k^\star + \mathbf{k}_{\mathbf{x}^\star,\mathbf{Z}}\mathbf{K}_\mathbf{Z}^{-1}\left(\mathbf{S} - \mathbf{K}_\mathbf{Z}\right)\mathbf{K}_\mathbf{Z}^{-1}\mathbf{k}_{\mathbf{x}^\star,\mathbf{Z}}^\mathsf{T}\,.$$

Given this distribution for $f(\mathbf{x}^\star)$, the probability distribution for $y^\star$ can be computed in closed-form in regression problems and with 1-dimensional quadrature in binary classification.

IDSGPs (Experiments)

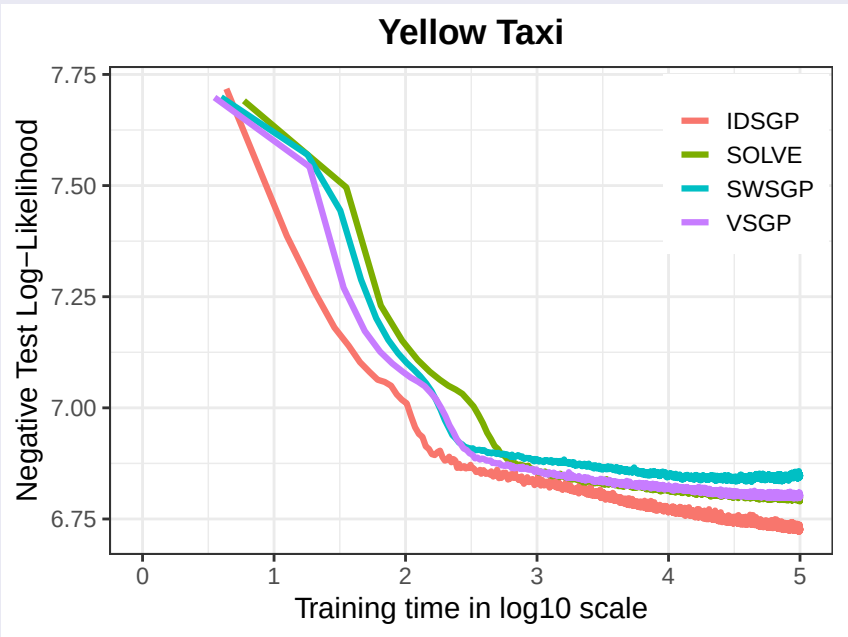


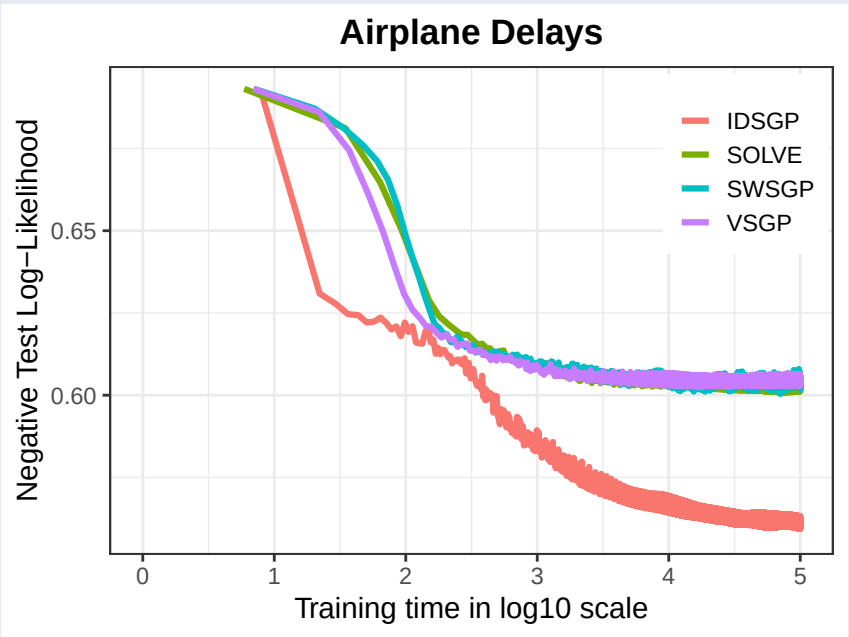
IDSGPs (Experiments)



IDSGPs (Large Scale)

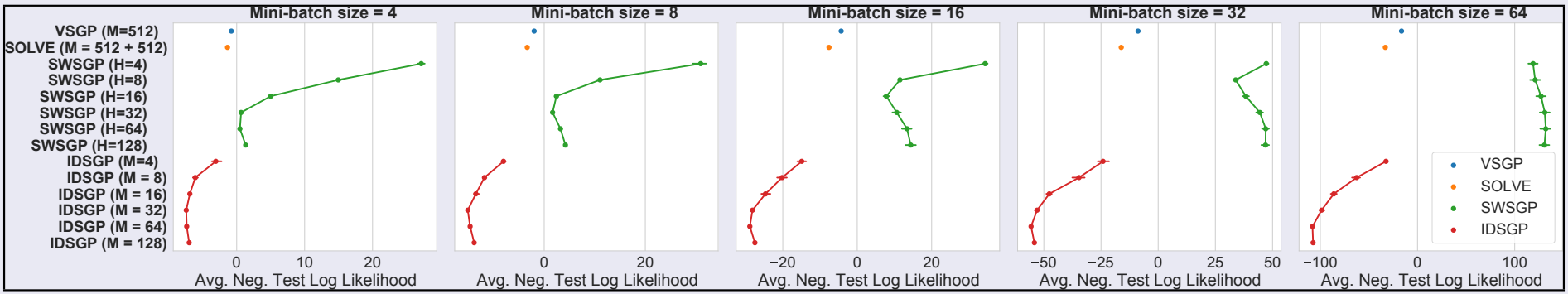
IDSGP performs best on each dataset. We believe this due to using a smaller number of inducing points, and also because of the extra flexibility of the NN that can specify an input-dependent location of the inducing points.





IDSGPs (Joint Predictive Covariances)

Increasing size of the mini-batch used for training and testing. Moreover, in the case of SWSGP and IDSGP, we show results for an increasing number of neighbors H and inducing points M.



These results confirm that IDSGP can provide accurate joint predictive distributions.

Conclusion

- IDSGP can improve the training time and the flexibility of sparse GP approximations.
- IDSGP keeps intact the GP prior on the latent function values associated to the training points.
- IDSGP uses a deep neural network (DNN) to output specific inducing points for each point
- The extra flexibility provided by the DNN allows to significantly reduce the number M of inducing points used in IDSGP.
- The scalability of IDSGP is also illustrated on massive datasets of up to 1 billion points.

Thank You!

References I