DEEP RECURRENT GAUSSIAN PROCESS WITH VARIATIONAL SPARSE SPECTRUM APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modeling sequential data has become more and more important in practice. Some applications are autonomous driving, virtual sensors and weather forecasting. To model such systems so called recurrent models are frequently used. In this paper we introduce several new Deep Recurrent Gaussian Process (DRGP) models based on the Sparse Spectrum Gaussian Process (SSGP) and the improved version called Variational Sparse Spectrum Gaussian Process (VSSGP). We follow the recurrent structure given by an existing DRGP based on a specific variational sparse Nyström approximation, the recurrent Gaussian Process (RGP). Therefore, we also variationally integrate out the input-space and hence can propagate uncertainty through the GP layers. Further, we combine the (V)SS approximations with a well known inducing-input regularization framework. This case naturally collapses to a tractable expression by calculating the integrals. For the simple extension of the (V)SS approximation an optimal variational distribution exists. Training is realized through optimizing the variational lower bounds. We improve over current state of the art methods in prediction accuracy for experimental data-sets used for their evaluation and introduce a new data-set for engine control, named Emission.

1 INTRODUCTION

Modeling sequential data for simulation tasks in the context of machine learning is hard for several reasons. Its internal structure poses the problem of modeling short term behavior and long term behavior together for different types of data variables, where the data variables itself might differ in the information gain in the chosen time frequency. Recurrent models (Hochreiter & Schmidhuber, 1997; Nelles, 2013; Pascanu et al., 2013) have proven to perform well on these tasks. They consist of output-data and input-data structured sequentially for shifted discrete time steps.
The general form of a *recurrent model* is given by

$$\mathbf{h}_i = \zeta(\mathbf{h}_{i-1}, \dots, \mathbf{h}_{i-H_\mathbf{h}}, \mathbf{x}_{i-1}, \dots, \mathbf{x}_{i-H_\mathbf{x}}) + \boldsymbol{\epsilon}_i^\mathbf{h}, \tag{1}$$

$$\mathbf{y}_i = \psi(\mathbf{h}_i, \dots, \mathbf{h}_{i-H_\mathbf{h}}) + \boldsymbol{\epsilon}_i^\mathbf{y}, \tag{2}$$

where $\mathbf{x}_i$ is an external input, $\mathbf{y}_i$ is an output observation, $\mathbf{h}_i$ is a latent hidden representation or state (details on dimensions and ranges will be specified in upcoming sections) at time $i = H_\mathbf{x}+1, \dots, N$, where $N \in \mathbb{N}$ is the number of data samples, $H_\mathbf{x}$, $H_\mathbf{h} \in \mathbb{N}$ are the chosen time horizons, ζ, ψ are non-linear functions modeling *transition* and *observation* and $\boldsymbol{\epsilon}_i^\mathbf{h}, \boldsymbol{\epsilon}_i^\mathbf{y}$ are transition and observation noise, which are adjusted for the specific problem.
In control and dynamical system identification previous work on Bayesian recurrent approaches for modeling sequential data usually make use of *(non-)linear auto-regressive with exogenous inputs models* ((N)ARX) and *state-space models* (SSM), for both see Nelles (2013). The general recurrent model given by the Equations (1), (2) represents both cases. This can be recognized by its general recurrent and hierarchical structure. This work deals with deep learning in a recurrent fashion for modeling sequential data in a Bayesian non-parametric approach by using *Gaussian Processes* (GP). To make a connection to the general recurrent model, the deep structure arises by defining ζ in Equation (1) in a deep manner (Pascanu et al., 2013, Section 3).
This paper proposes deep recurrent GP models based on (variational) Sparse Spectrum approximations, denoted by DRGP-(V)SS. For reproducibility of the experimental results, we provide the

code online[1]. Therefore, we follow the same deep recurrent structure as introduced in Mattos et al. (2016). To summarize the contributions of this paper:

- Extension of the sparse GP based on the SS approximation introduced by Lázaro-Gredilla et al. (2010) and the improved VSS approximation by Gal & Turner (2015) to DRGPs;
- Improving regularization properties of the variational bounds through the combination of the (V)SS approximation with the inducing-point (IP) regularization of Titsias & Lawrence (2010);
- Propagation of uncertainty through the hidden layers of our DGPs by variationally integrating out the input-space;
- The existence of an optimal variational distribution in the sense of a functional local optimum of the variational bounds of our DRGPs models is established.

The DRGP of Mattos et al. (2016) is limited to a small class of deterministic covariance functions, because the covariance functions variational expectation has to be analytically tractable. Using the (V)SS approximation instead, we can derive a valid approximation for every stationary covariance function, because the basis functions expectation is always tractable. We show that this approach improves over state of the art approaches in prediction accuracy on several cases of the experimental data-sets used in Mattos et al. (2016); Svensson et al. (2016); Al-Shedivat et al. (2017); Salimbeni & Deisenroth (2017); Dörr et al. (2018) in a simulation setting. For scalability, *Distributed Variational Inference* (DVI) (Gal et al., 2014) is recommended and can lower the complexity per core from $\mathcal{O}(NM^2Q_{\text{max}}L)$ down to $\mathcal{O}(M^3)$, for $N \leq M$, M the sparsity parameter, L the amount of GPs and Q_{max} is the maximum over all input dimensions used in our defined deep structure for ζ and ψ.

2 Related Work to Gaussian processes with SSM and DGPs

An *SSM with GPs* (GP-SSM) for transition and observation functions is used by Wang et al. (2005), where the uncertainty in the latent states is not accounted for, which can lead to overconfidence. Turner et al. (2010) solved this problem, but they have complicated approximate training and inference stages and the model is hard to scale. Frigola et al. (2014) used a GP for transition, while the observation is parametric. Svensson et al. (2016) used an approximation of the spectral representation by Bochner's theorem in a particular form and with a reduced rank structure for the transition function. They realize inference in a fully Bayesian approach over the amplitudes and the noise parameters. Dörr et al. (2018) introduced a GP-SSM with scalable training based on doubly stochastic variational inference for robust training. Our models extend the GP-SSM framework by defining the transition as a DRGP based on our newly derived (V)SS approximations in Sections 3.3, 3.4 where the latent (non-observed) output-data is learned as a hidden state. We refer to the report Föll et al. (2017)[2] for a detailed but preliminary formulation of our models and experiments.
Following Damianou & Lawrence (2013), a *Deep Gaussian Process* (DGP) is a model assuming

$$\mathbf{y}_i = \mathrm{f}^{(\mathrm{L}+1)}(\mathrm{f}^{(L)}(\mathrm{f}^{(L-1)}(\dots(\mathrm{f}^{(1)}(\mathbf{x}_i) + \boldsymbol{\epsilon}_i^{\mathbf{h}^{(1)}})\dots) + \boldsymbol{\epsilon}_i^{\mathbf{h}^{(L-1)}}) + \boldsymbol{\epsilon}_i^{\mathbf{h}^{(L)}}) + \boldsymbol{\epsilon}_i^{\mathbf{y}},$$

where the index $i = 1, \dots, N$ is *not* necessarily the time and where we define $\mathbf{h}_i^{(1)} \stackrel{\text{def}}{=} \mathrm{f}^{(1)}(\mathbf{x}_i) + \boldsymbol{\epsilon}_i^{\mathbf{h}^{(1)}}$, $\mathbf{h}_i^{(l+1)} \stackrel{\text{def}}{=} \mathrm{f}^{(l)}(\mathbf{h}_i^{(l)}) + \boldsymbol{\epsilon}_i^{\mathbf{h}^{(l)}}$, for $l = 2 \dots, L-1$, where $L \in \mathbb{N}$ is the number of hidden layers. The noise $\boldsymbol{\epsilon}_i^{\mathbf{h}^{(l)}}$, $\boldsymbol{\epsilon}_i^{\mathbf{y}}$ is assumed Gaussian and the functions $\mathrm{f}^{(l)}$ are modeled with GPs for $l = 1, \dots, L+1$. To obtain computational tractability, in most cases variational approximation and inference is used. Damianou & Lawrence (2013) introduced these kind of DGPs based on the sparse variational approximation following Titsias & Lawrence (2010). Based on this Dai et al. (2016) introduced a DGP with a variationally auto-encoded inference mechanism and which scales on larger data-sets. Cutajar et al. (2016) introduced a DGP for the so called *Random Fourier Features* (RFF) approach, where the variational weights for each GPs are optimized along with the hyperparameters. This approach does not variationally integrate out the latent inputs to carry through the uncertainty and no existence of an optimal distribution for the variational weights is proven, to reduce the amount of parameters to optimize in training. Further, Salimbeni & Deisenroth (2017) introduced a approximation framework for DGPs which is similar to the single GP of Hensman & Lawrence (2014) but

[1] DRGP-(V)SS code available from `http://github.com/RomanFoell/DRGP-VSS`.
[2] Available on the website `https://arxiv.org/`.

does not force independence between the GP layers and which scales to billions of data.
Two state of the art approaches for DRGPs have been introduced by Mattos et al. (2016), the RGP, which we call DRGP-Nyström, based on the variational sparse Nyström/inducing-point approximation introduced by Titsias (2009); Titsias & Lawrence (2010), as well as Al-Shedivat et al. (2017), which we call GP-LSTM, based on deep kernels via a *Long-short term memory* (LSTM) network, a special type of *Recurrent Neural Network* (RNN). DRGP-Nyström uses a recurrent construction, where the auto-regressive structure is not realized directly with the observed output-data, but with the GPs latent output-data and uses a Variational Inference (VI) framework named *Recurrent Variational Bayes* (REVARB). The structure acts like a standard RNN, where every parametric layer is a GP. So additionally uncertainty information can be carried through the hidden layers.
GP-LSTM is a combination of GPs and LSTMs. LSTMs have proven to perform well on modeling sequential data. LSTMs try to overcome vanishing gradients by placing a memory cell into each hidden unit. GP-LSTM uses special update rules for the hidden representations and the hyper parameters through a semi-stochastic optimization scheme. It combines a GP with the advantages of LSTMs by defining structured recurrent deep covariance functions, also called deep kernels, which reduces the time and memory complexities of the linear algebraic operations (Wilson et al., 2016).

3 Gaussian Process and Variational Sparse Spectrum GP

Loosely speaking, a GP can be seen as a Gaussian distribution over functions. We will first introduce GPs and GP regression and then recall the SSGP by (Lázaro-Gredilla et al., 2010) and its improved version VSSGP by (Gal & Turner, 2015). Based on these we derive new variational approximations.

3.1 Gaussian Process

A stochastic process f is a GP if and only if any finite collection of random variables $f_{\mathbf{X}} \stackrel{\text{def}}{=} [f_{\mathbf{x}_1}, \dots, f_{\mathbf{x}_N}]^T$ forms a Gaussian random vector (Rasmussen, 2006). A GP is completely defined by its mean function $m : \mathbb{R}^Q \to \mathbb{R}, \mathbf{x} \mapsto m(\mathbf{x})$ and covariance function $k : \mathbb{R}^Q \times \mathbb{R}^Q \to \mathbb{R}, (\mathbf{x}, \mathbf{x}') \mapsto k(\mathbf{x}, \mathbf{x}')$ (Kallenberg, 2006, Lemma 11.1), where

$$m(\mathbf{x}) \stackrel{\text{def}}{=} E\left[f(\mathbf{x})\right], \quad k(\mathbf{x}, \mathbf{x}') \stackrel{\text{def}}{=} \text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = E\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right],$$

and the GP will be written as $f \sim \mathcal{GP}(m, k)$.
Let $\mathbf{y} \stackrel{\text{def}}{=} [\mathrm{y}_1, \dots, \mathrm{y}_N]^T \in \mathbb{R}^N$ and we assume $\mathrm{y}_i = \mathrm{f}(\mathbf{x}_i) + \epsilon_i^{\mathrm{y}}$, where $\epsilon_i^{\mathrm{y}} \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$, for $i = 1, \dots, N$, and our aim is to model any set of function values $\mathbf{f} = [\mathrm{f}(\mathbf{x}_1), \dots, \mathrm{f}(\mathbf{x}_N)]^T \in \mathbb{R}^N$ at $\mathbf{X} \stackrel{\text{def}}{=} [\mathbf{x}_1, \dots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times Q}$ as samples from a random vector $f_{\mathbf{X}}$. Therefore, we assume the prior $f_{\mathbf{X}} | \boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, K_{NN})$, meaning that any set of function values $\mathbf{f}$ given $\mathbf{X}$ are jointly Gaussian distributed with mean $\mathbf{0} \in \mathbb{R}^N$ and a covariance matrix $K_{NN} \stackrel{\text{def}}{=} (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^N \in \mathbb{R}^{N \times N}$.
The predictive distribution $p_{f_{\mathbf{x}^*} | \boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y}}$ for a test point $\mathbf{x}^* \in \mathbb{R}^Q$, where $K_{N*} \stackrel{\text{def}}{=} (k(\mathbf{x}_i, \mathbf{x}^*))_{i=1}^N \in \mathbb{R}^N$, K_{**}, K_{*N} analogously, can be derived through the joint probability model and conditioning as

$$f_{\mathbf{x}^*} | \boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(K_{*N}(K_{NN} + \sigma^2_{\text{noise}} I_N)^{-1}\mathbf{y}, K_{**} - K_{*N}(K_{NN} + \sigma^2_{\text{noise}} I_N)^{-1} K_{N*}).$$

In preview of our experiments in Section 5 and the following sections we choose a specific covariance function, the *spectral mixture* (SM) covariance function (Wilson & Adams, 2013)

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2_{\text{power}} \left(\prod_{q=1}^Q \exp\left(-\frac{1}{2} \mathrm{l}_q^{-2} (\mathrm{x}_q - \mathrm{x}'_q)^2 \right) \right) \cos\left(2\pi \sum_{q=1}^Q \mathrm{p}_q^{-1} (\mathrm{x}_q - \mathrm{x}'_q) \right), \tag{3}$$

with length scales $\mathrm{l}_q, \mathrm{p}_q \in \mathbb{R}$, $q = 1, \dots, Q$. As $p_q \to \infty$, this corresponds to the *squared exponential* (SE) covariance function in the limit (Gal & Turner, 2015).

3.2 Variational Sparse Spectrum GP

We introduce the SSGP following Gal & Turner (2015). For a stationary covariance function k on $\mathbb{R}^Q \times \mathbb{R}^Q$ there exists a function $\rho : \mathbb{R}^Q \to \mathbb{R}, \boldsymbol{\tau} \mapsto \rho(\boldsymbol{\tau})$, such that $k(\mathbf{x}, \mathbf{x}') = \rho(\mathbf{x} - \mathbf{x}')$ for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^Q$. Bochner's theorem states that any stationary covariance function k can be represented as the Fourier transform of a positive finite measure $\boldsymbol{\mu}$ (Stein, 2012). Then $\rho(\boldsymbol{\tau})$, using $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$,

can be expressed via *Monte Carlo Approximation* (MCA) following Gal & Turner (2015) Section 2 as

$$\rho(\boldsymbol{\tau}) = \int_{\mathbb{R}^Q} e^{2\pi i \mathbf{z}^T \boldsymbol{\tau}} d\boldsymbol{\mu}(\mathbf{z}) \tag{4}$$
$$\approx \frac{\sigma^2_{\text{power}}}{M} \sum\nolimits_{m=1}^{M} 2\cos(2\pi \mathbf{z}_m^T(\mathbf{x}-\mathbf{u}_m) + \mathrm{b}_m)\cos(2\pi \mathbf{z}_m^T(\mathbf{x}'-\mathbf{u}_m) + \mathrm{b}_m) \stackrel{\text{def}}{=} \tilde{k}(\mathbf{x},\mathbf{x}')$$

We refer to $\mathbf{z}_m$ as the spectral points, b_m as the spectral phases and $\mathbf{u}_m$ as the pseudo-input/inducing points for $m = 1, \dots, M$. Choosing the probability density like in Gal & Turner (2015) Proposition 2, we approximate the SM covariance function with a scaling matrix $\mathfrak{L} \stackrel{\text{def}}{=} \text{diag}([2\pi \mathrm{l}_q]_{q=1}^Q) \in \mathbb{R}^{Q\times Q}$, a scaling vector $\mathbf{p} \stackrel{\text{def}}{=} [\mathrm{p}_1^{-1}, \dots, \mathrm{p}_Q^{-1}]^T \in \mathbb{R}^Q$ and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_M]^T \in \mathbb{R}^{Q\times M}$ via

$$\phi(\mathbf{x},\mathbf{Z}) \stackrel{\text{def}}{=} \sqrt{2\sigma^2_{\text{power}} M^{-1}} \left[\cos(2\pi(\mathfrak{L}^{-1}\mathbf{z}_1 + \mathbf{p})^T(\mathbf{x}-\mathbf{u}_1) + \mathrm{b}_1), \dots, \right. \tag{5}$$
$$\left. \cos(2\pi(\mathfrak{L}^{-1}\mathbf{z}_M + \mathbf{p})^T(\mathbf{x}-\mathbf{u}_M) + \mathrm{b}_M)\right]^T \in \mathbb{R}^M.$$

Here we sample $b \sim \text{Unif}\,[0, 2\pi]$, $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I_Q)$ and we set $\tilde{K}_{NN}^{(\text{SM})} \stackrel{\text{def}}{=} \Phi\Phi^T$ with $\Phi \stackrel{\text{def}}{=} [\phi(\mathbf{x}_1,\mathbf{Z}), \dots, \phi(\mathbf{x}_N,\mathbf{Z})]^T \in \mathbb{R}^{N\times M}$.

In Gal & Turner (2015) the SSGP was improved to VSSGP by variationally integrating out the spectral points and instead of optimizing the spectral points additionally optimizing the variational parameters. We follow the scheme of Gal & Turner (2015) Section 4 for the 1-dimensional output case $\mathbf{y} \in \mathbb{R}^{N\times 1}$. By replacing the covariance function with the sparse covariance function $\tilde{k}^{(\text{SM})}$ and setting the priors $p_{\boldsymbol{z}_m}, p_{\boldsymbol{a}}$ to

$$\boldsymbol{z}_m \sim \mathcal{N}(\mathbf{0}, I_Q), \quad \boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, I_M), \tag{6}$$

for $m = 1, \dots, M$, where we have $\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{X} \sim \mathcal{N}(\Phi\mathbf{a}, \sigma^2_{\text{noise}} I_N)$ with $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_M]^T \in \mathbb{R}^{Q\times M}$ (we do not define priors on $\mathbf{p}$, $\mathfrak{L}^{-1}$, $\mathbf{b} = [b_1, \dots, b_M]^T \in \mathbb{R}^M$), we can expand the *marginal likelihood* (ML) with Bishop (2006) Equations 2.113 - 2.115 to

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{a},\mathbf{Z},\mathbf{U},\mathbf{X})p(\mathbf{a})p(\mathbf{Z})d\mathbf{a}d\mathbf{Z}. \tag{7}$$

Now, to improve the SSGP to VSSGP, variational distributions $q_{\boldsymbol{z}_m}, q_{\boldsymbol{a}}$ are introduced in terms of

$$\boldsymbol{z}_m \sim \mathcal{N}(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m), \quad \boldsymbol{a} \sim \mathcal{N}(\mathbf{m}, \mathbf{s}),$$

with $\boldsymbol{\beta}_m \in \mathbb{R}^{Q\times Q}$ diagonal, for $m = 1, \dots, M$, and $\mathbf{s} \in \mathbb{R}^{M\times M}$ diagonal. From here on we use variational mean-field approximation to derive the approximate models with different lower bounds to the log ML introduced by Gal & Turner (2015):

$$\log(p(\mathbf{y}|\mathbf{X})) \geq \mathbf{E}[\log(p(\mathbf{Y}|\mathbf{a},\mathbf{Z},\mathbf{U},\mathbf{X}))]_{q_{\boldsymbol{a}} q_{\boldsymbol{Z}}} - \mathbf{KL}(q_{\boldsymbol{a}}||p_{\boldsymbol{a}}) - \mathbf{KL}(q_{\boldsymbol{Z}}||p_{\boldsymbol{Z}})). \tag{8}$$

As usual, $\mathbf{KL}$ defines the *Kullback-Leibler* (KL) divergence. By proving the existence of an optimal distribution $q_{\boldsymbol{a}}^{\text{opt}}$ for $\boldsymbol{a}$ Gal & Turner (2015) Proposition 3 in the sense of a functional local optimum of the right hand side of 8, where $\boldsymbol{a} \sim \mathcal{N}(A^{-1}\Psi_1^T\mathbf{y}, \sigma^2_{\text{noise}} A^{-1})$, with $A = \Psi_2 + \sigma^2_{\text{noise}} I_M$, $\Psi_1 = \mathbf{E}\,[\Phi]_{q_{\boldsymbol{Z}}} \in \mathbb{R}^{N\times M}$, $\Psi_2 = \mathbf{E}\left[\Phi^T\Phi\right]_{q_{\boldsymbol{Z}}} \in \mathbb{R}^{M\times M}$, we can derive the optimal bound case.

3.3 (V)SSGP WITH REGULARIZATION PROPERTIES VIA INDUCING POINTS (IP)

As a first contribution of this paper we combine two approximation schemes to two new methods (V)SSGP-IP. We want to point out that the (V)SSGP has not the same regularization properties as the GP of Titsias & Lawrence (2010) when optimizing the parameters $\boldsymbol{U}$ because the prior 6 of the weights $\mathbf{a}$ are defined generic via Bishop (2006) Equations 2.113 - 2.115. We now define them as

$$\boldsymbol{a}|\boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, K_{MM}), \quad \boldsymbol{y}|\boldsymbol{a},\boldsymbol{Z},\boldsymbol{U},\boldsymbol{X} \sim \mathcal{N}(\Phi K_{MM}^{-1}\mathbf{a}, K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}), \tag{9}$$

where K_{NN}, K_{NM}, K_{MM} are defined through the given covariance function in Equation (3).

Consequently, the resulting bound in 10 has an extra regularization property compared to the right hand side of 8, which is reflected in the different form of $A = \Psi_2 + \sigma^2_{\text{noise}} K_{MM}$, which involves K_{MM}, the choosen covariance matrix filled in with the pseudo-input points $\boldsymbol{U}$ and three extra terms

$$2^{-1}\log(|K_{MM}|), \quad (2\sigma^2_{\text{noise}})^{-1}\text{tr}(K_{NN}), \quad (2\sigma^2_{\text{noise}})^{-1}\text{tr}(K_{MM}^{-1}K_{MN}K_{NM}).$$

The optimal variational distribution for $\mathbf{a}$, which collapses by reversing Jensen's inequality, is similar to Titsias & Lawrence (2010) Section 3.1 and the resulting bounds (SSGP-IP without $\mathbf{KL}(q_{\boldsymbol{Z}}||p_{\boldsymbol{Z}})$ and $\Psi_1 = \Phi$, $\Psi_2 = \Phi^T\Phi$) can be calculated in the same way as Titsias & Lawrence (2010) Equation 14 in closed form by simply calculating the integral in Equation 7:

$$\log(p(\mathbf{y}|\mathbf{X})) \geq -\frac{(N-M)}{2}\log(\sigma^2_{\text{noise}}) - \frac{N}{2}\log(2\pi) - \frac{\mathbf{y}^T\mathbf{y}}{2\sigma^2_{\text{noise}}} + \frac{\log(|K_{MM}|\,|A^{-1}|)}{2} + \frac{\mathbf{y}^T\Psi_1 A^{-1}\Psi_1^T\mathbf{y}}{2\sigma^2_{\text{noise}}} - \frac{\text{tr}(K_{NN}) + \text{tr}(K_{MM}^{-1}K_{MN}K_{NM})}{2\sigma^2_{\text{noise}}} - \mathbf{KL}(q_{\boldsymbol{Z}}||p_{\boldsymbol{Z}}). \tag{10}$$

3.4 Variational approximation of the input-space for (V)SSGP(-IP)

As a second contribution of this paper, in this section we marginalize also the input-space. This is not straightforward as it is not clear whether for the (V)SS sparse covariance function these expressions even exist. To prevent misunderstanding, we will write from now on $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^T \in \mathbb{R}^{N\times Q}$ instead of $\mathbf{X}$, to highlight $\mathbf{H}$ is now a set of latent variables (later on the hidden states of the DGP). Therefore, we introduce priors and variational distributions

$$p_{\boldsymbol{H}} \stackrel{\text{def}}{=} \prod_{i=1}^N p_{\boldsymbol{h}_i}, \text{ where } \boldsymbol{h}_i \sim \mathcal{N}(\mathbf{0}, I_Q), \qquad q_{\boldsymbol{H}} \stackrel{\text{def}}{=} \prod_{i=1}^N q_{\boldsymbol{h}_i}, \text{ where } \boldsymbol{h}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\lambda}_i),$$

$\boldsymbol{\lambda}_i \in \mathbb{R}^{Q\times Q}$ diagonal, for $i = 1, \ldots, N$. As a consequence, for VSSGP-IP we overall derive statistics $\Psi_0 = \text{tr}\left(\mathbf{E}[K_{NN}]_{q_{\boldsymbol{H}}}\right) = N\sigma^2_{\text{power}}$, $\Psi_1 = \mathbf{E}\left[\Phi\right]_{q_{\boldsymbol{Z}}q_{\boldsymbol{H}}} \in \mathbb{R}^{N\times M}$, $\Psi_2 = \mathbf{E}\left[\Phi^T\Phi\right]_{q_{\boldsymbol{Z}}q_{\boldsymbol{H}}} \in \mathbb{R}^{M\times M}$ and $\Psi_{reg} = \mathbf{E}\left[K_{MN}K_{NM}\right]_{q_{\boldsymbol{H}}}$ as defined in the Appendix 6.3. The SSGP-IP model derives by being not variational over the spectral points.
This extra step will makes it possible to propagate uncertainty between the hidden layers of a DGP, as we gain an extra variance parameter for the inputs. For the IP case we get the lower bound:

$$\log(p(\mathbf{y}|\mathbf{H})) \geq -\frac{(N-M)}{2}\log(\sigma^2_{\text{noise}}) - \frac{N}{2}\log(2\pi) - \frac{\mathbf{y}^T\mathbf{y}}{2\sigma^2_{\text{noise}}} + \frac{\log(|K_{MM}|\,|A^{-1}|)}{2} + \frac{\mathbf{y}^T\Psi_1 A^{-1}\Psi_1^T\mathbf{y}}{2\sigma^2_{\text{noise}}} - \frac{\Psi_0 + \text{tr}(K_{MM}^{-1}\Psi_{reg})}{2\sigma^2_{\text{noise}}} - \mathbf{KL}(q_{\boldsymbol{Z}}||p_{\boldsymbol{Z}}) - \mathbf{KL}(q_{\boldsymbol{H}}||p_{\boldsymbol{H}}). \tag{11}$$

For the extension of the SSGP and VSSGP in the optimal bound case in 8 (we only focus on this case in the following), we again have $A = \Psi_2 + \sigma^2_{\text{noise}}I_M$ and eliminate $\frac{\log(|K_{MM}|)}{2}$, $\frac{\Psi_0 + \text{tr}(K_{MM}^{-1}\Psi_{reg})}{2\sigma^2_{\text{noise}}}$.

4 DRGP with Variational Sparse Spectrum Approximation

In this section we want to combine our newly derived (V)SSGP-(IP) approximations in 11, overall four cases SSGP, VSSGP, SSGP-IP, VSSG-IP, with the framework introduced in Mattos et al. (2016), to derive our models DRGP-SSGP, -VSSGP, -SSGP-IP, -VSSG-IP.

4.1 DRGP-(V)SS(-IP) Model Definition

Choosing the same recurrent structure as in Mattos et al. (2016), where now i represents the time, $\mathbf{y}_{H_{\mathbf{x}}+1:} = [y_{H_{\mathbf{x}}+1}, \ldots, y_N]^T \in \mathbb{R}^{\hat{N}}$(to simplify notation, we write $\mathbf{h}^{(L+1)}_{H_{\mathbf{x}}+1:} \stackrel{\text{def}}{=} \mathbf{y}_{H_{\mathbf{x}}+1:}$), we have

$$\begin{aligned}
\zeta: &\quad \text{h}^{(l)}_i = \text{f}^{(l)}(\hat{\mathbf{h}}^{(l)}_i) + \epsilon^{h^{(l)}}_i, &&\text{with prior} \quad f^{(l)}_{\hat{\mathbf{H}}^{(l)}}|\hat{\boldsymbol{H}}^{(l)} \sim \mathcal{N}(\mathbf{0}, K^{(l)}_{\hat{N}\hat{N}}), \qquad l = 1, \ldots, L\\
\psi: &\quad \text{y}_i = \text{f}^{(l)}(\hat{\mathbf{h}}^{(l)}_i) + \epsilon^y_i, &&\text{with prior} \quad f^{(l)}_{\hat{\mathbf{H}}^{(l)}}|\hat{\boldsymbol{H}}^{(l)} \sim \mathcal{N}(\mathbf{0}, K^{(l)}_{\hat{N}\hat{N}}), \qquad l = L+1
\end{aligned}$$

with ζ, ψ in Equation (1) and (2), $\epsilon^{\text{h}^{(l)}}_i \sim \mathcal{N}(0, (\sigma^{(l)}_{\text{noise}})^2)$, $\epsilon^{\text{y}}_i \sim \mathcal{N}(0, (\sigma^{(L+1)}_{\text{noise}})^2)$ and $\hat{N} = N - H_{\mathbf{x}}$, for $i = H_{\mathbf{x}} + 1, \ldots, N$. The matrix $K^{(l)}_{\hat{N}\hat{N}}$ represents a covariance matrix coming from our chosen $\tilde{k}$ from Equation 4 and 5. A set of input-data $\hat{\mathbf{H}}^{(l)} = [\hat{\mathbf{h}}^{(l)}_{H_{\mathbf{x}}+1}, \ldots, \hat{\mathbf{h}}^{(l)}_N]^T$ is specified as

$$\hat{\mathbf{h}}_i^{(l)} \stackrel{\text{def}}{=} \begin{cases} \left[\overline{\mathbf{h}}_{i-1}^{(1)}, \overline{\mathbf{x}}_{i-1}\right]^T \stackrel{\text{def}}{=} \left[\left[\mathrm{h}_{i-1}^{(1)}, \ldots, \mathrm{h}_{i-H_\mathrm{h}}^{(1)}\right], \left[\mathbf{x}_{i-1}, \ldots, \mathbf{x}_{i-H_\mathbf{x}}\right]\right]^T, & l = 1 \\ \left[\overline{\mathbf{h}}_{i-1}^{(l)}, \overline{\mathbf{h}}_{i}^{(l-1)}\right]^T \stackrel{\text{def}}{=} \left[\left[\mathrm{h}_{i-1}^{(l)}, \ldots, \mathrm{h}_{i-H_\mathrm{h}}^{(l)}\right], \left[\mathrm{h}_{i}^{(l-1)}, \ldots, \mathrm{h}_{i-H_\mathrm{h}+1}^{(l-1)}\right]\right]^T, & l = 2, \ldots, L \\ \overline{\mathbf{h}}_{i}^{(L)} \stackrel{\text{def}}{=} \left[\mathrm{h}_{i}^{(L)}, \ldots, \mathrm{h}_{i-H_\mathrm{h}+1}^{(L)}\right]^T, & l = L+1, \end{cases} \tag{12}$$

where $\hat{\mathbf{h}}_i^{(1)} \in \mathbb{R}^{H_\mathrm{h}+H_\mathbf{x}Q}$, $\hat{\mathbf{h}}_i^{(l)} \in \mathbb{R}^{2H_h}$ for $l = 2, \ldots, L$, $\hat{\mathbf{h}}_i^{(L+1)} \in \mathbb{R}^{H_\mathrm{h}}$, for $i = H_\mathbf{x}+1, \ldots, N$. For simplification we set $H \stackrel{\text{def}}{=} H_\mathbf{x} = H_\mathrm{h}$ in our experiments.
Now we use the new approximations in 11 to derive the new joint probability density

$$p_{\boldsymbol{y}_{H_\mathbf{x}+1:}\left[\boldsymbol{a}^{(l)}, \boldsymbol{Z}^{(l)}, \boldsymbol{h}^{(l)}, \boldsymbol{U}^{(l)}\right]_{l=1}^{L+1}|\boldsymbol{X}} = \prod_{l=1}^{L+1} p_{\boldsymbol{h}_{H_\mathbf{x}+1:}^{(l)}|\boldsymbol{a}^{(l)}, \boldsymbol{Z}^{(l)}, \hat{\boldsymbol{H}}^{(l)}, \boldsymbol{U}^{(l)}} p_{\boldsymbol{a}^{(l)}} p_{\boldsymbol{Z}^{(l)}} p_{\tilde{\boldsymbol{h}}^{(l)}},$$

with $p_{\tilde{\boldsymbol{h}}^{(L+1)}} = 1$, $\boldsymbol{h}^{(L+1)} = \{\}$ and $\tilde{\mathbf{h}}^{(l)} = [\mathrm{h}_{1+H_\mathbf{x}-H_\mathrm{h}}^{(l)}, \ldots, \mathrm{h}_{H_\mathrm{h}}^{(l)}]^T \in \mathbb{R}^{2H_\mathrm{h}-H_\mathbf{x}}$. Here the priors $p_{\boldsymbol{a}^{(l)}}, p_{\boldsymbol{z}_m^{(l)}}, p_{\tilde{\boldsymbol{h}}^{(l)}}$ are

$$\boldsymbol{a}^{(l)} \sim \mathcal{N}(\mathbf{0}, I_M), \quad \boldsymbol{z}_m^{(l)} \sim \mathcal{N}(\mathbf{0}, I_Q), \quad \tilde{\boldsymbol{h}}^{(l)} \sim \mathcal{N}(\mathbf{0}, I_{2H_\mathrm{h}-H_\mathbf{x}}),$$

the product of them is defined as P_{REVARB}, and the variational distributions $q_{\boldsymbol{a}^{(l)}}, q_{\boldsymbol{z}_m^{(l)}}, q_{h_i^{(l)}}$ are

$$\boldsymbol{a}^{(l)} \sim \mathcal{N}(\mathbf{m}^{(l)}, \mathbf{s}^{(l)}), \quad \boldsymbol{z}_m^{(l)} \sim \mathcal{N}(\boldsymbol{\alpha}_m^{(l)}, \boldsymbol{\beta}_m^{(l)}), \quad h_i^{(l)} \sim \mathcal{N}(\mu_i^{(l)}, \lambda_i^{(l)}),$$

the product of them is defined as Q_{REVARB}, and where $\boldsymbol{\beta}_m^{(l)} \in \mathbb{R}^{Q \times Q}$ is diagonal, for $i = 1 + H_\mathbf{x} - H_\mathrm{h}, \ldots, N$, $m = 1 \ldots, M$, $l = 1, \ldots, L+1$. For the setting (V)SSGP-IP, see Section 3.3, we choose no variational distribution for the weights $\mathbf{a}^{(l)}$ but for $l = 1, \ldots, L+1$ the prior assumptions

$$\boldsymbol{a}^{(l)}|\boldsymbol{U}^{(l)} \sim \mathcal{N}(\mathbf{0}, K_{MM}^{(l)}),$$
$$\boldsymbol{h}_{H_\mathbf{x}+1:}^{(l)}|\boldsymbol{a}^{(l)}, \boldsymbol{Z}^{(l)}, \hat{\boldsymbol{H}}^{(l)}, \boldsymbol{U}^{(l)} \sim \mathcal{N}(\Phi(K_{MM}^{(l)})^{-1}\mathbf{a}^{(l)}, K_{NN}^{(l)} - K_{NM}^{(l)}(K_{MM}^{(l)})^{-1}K_{MN}^{(l)}).$$

This defines our models for the cases DRGP-VSS, DRGP-VSS-IP. In the case where we are not variational over the spectral-points we derive the simplified versions DRGP-SS, DRGP-SS-IP.

4.2 DRGP-(V)SS(-IP) Variational Evidence Lower Bound

Using standard variational approximation techniques (Blei et al., 2017), the *Recurrent Variational Bayes* lower bound for the *(V)SS*(-IP) approximations, denoted as REVARB-(V)SS(-IP), is given by

$$\log(p(\mathbf{y}_{H_\mathbf{x}+1:}|\mathbf{X})) \geq$$
$$\mathbf{E}[\sum_{l=1}^{L+1} \log(p(\mathbf{h}_{H_\mathbf{x}+1:}^{(l)}|\mathbf{a}^{(l)}, \mathbf{Z}^{(l)}, \hat{\mathbf{H}}^{(l)}, \mathbf{U}^{(l)}))]_{Q_{\text{REVARB}}} - \mathbf{KL}(Q_{\text{REVARB}}||P_{\text{REVARB}}) \stackrel{\text{def}}{=} \boldsymbol{\mathcal{L}}_{\text{(V)SS(-IP)}}^{\text{REVARB}}. \tag{13}$$

Additionally, for the simple (V)SSGP approximations the optimal bound $\boldsymbol{\mathcal{L}}_{\text{(V)SS-opt}}^{\text{REVARB}}$ can be obtained immediately analogous to Gal & Turner (2015) Proposition 3 and the fact, that the bound decomposes into a sum of independent terms for $\boldsymbol{a}^{(l)}$. Maximizing the lower bounds is equivalent to minimizing the KL-divergence of Q_{REVARB} and the true posterior, therefore this is a way to optimize the approximated model parameter distribution with respect to the intractable, true model parameter posterior. Calculating $\boldsymbol{\mathcal{L}}_{\text{(V)SS-opt/IP}}^{\text{REVARB}}$ requires $\mathcal{O}(NM^2Q_{\max}L)$, where $Q_{\max} = \max_{l=1\ldots,L+1} Q^{(l)}$, $Q^{(l)} \stackrel{\text{def}}{=} \dim(\hat{\mathbf{h}}_i^{(l)})$ and $\hat{\mathbf{h}}_i^{(l)}$ from Equation (12). DVI can reduce the complexity to $\mathcal{O}(M^3)$ if the number of cores scales suitable with the number of training-data. A detailed derivation of the REVARB-(V)SS(-IP) lower bound can be found in the Appendix 6.3.4.

4.3 Model Predictions

After model optimization based on the lower bounds in Equation 13, model predictions for new $\hat{\mathbf{h}}_*^{(l)}$ in the REVARB-(V)SS(-IP) framework can be obtained based on the approximate, variational posterior distribution Q_{REVARB}. They are performed iteratively with approximate uncertainty propagation between each GP layer. We derive $q_{\boldsymbol{h}_*^{(l)}}$ from previous time-steps and it is per definition Gaussian with mean and variance derived from previous predictions for $l = 1, \ldots, L$. These kind of models propagate the uncertainty of the hidden GP layers' outputs, not of the observed output-data and are relevant for good model predictions. The detailed expressions for the mean and variance of the predictive distribution involved during the predictions can be found in the Appendix 6.3.5.

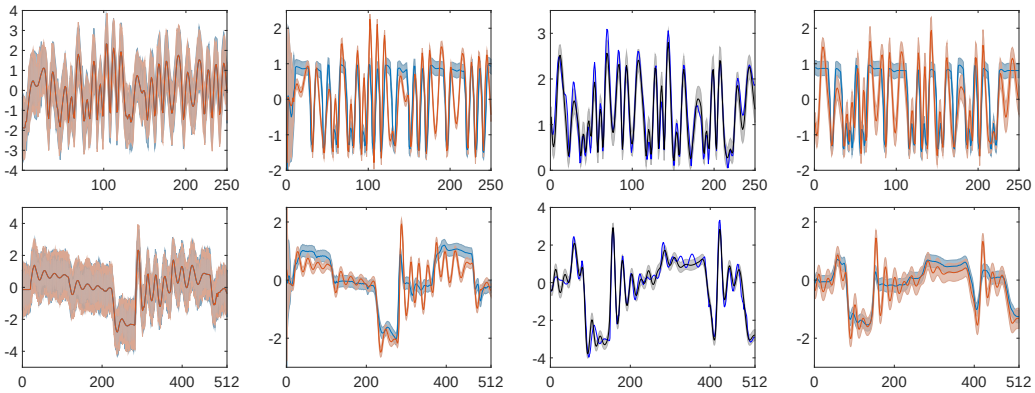

Figure 1: Simulation results visualized for the data-sets Drive (first row) and Actuator (second row) for the method DRGP-VSS. First column represents the initial hidden states, blue: first layer, and red: second layer. The second column represents the learned hidden states. The third column shows the simulation results, blue: real data, black: simulation, grey: ± 2 times Standard Deviation (SD) and the fourth column the predicted hidden states.

5 Experiments

In this section we want to compare our methods DRGP-SS, DRGP-VSS (optimal bound case) and DRGP-SS-IP, DRGP-VSS-IP, against other well known sparse GPs and the full GP with NARX structure, the GP-SSM of Svensson et al. (2016), the DRGP-Nyström of Mattos et al. (2016), the DGP-DS of Salimbeni & Deisenroth (2017) with NARX structure, the GP-LSTM of Al-Shedivat et al. (2017) and PR-SSM Dörr et al. (2018). The full GP is named GP-full, the FITC approximation of Snelson & Ghahramani (2006) is named GP-FITC, the DTC approximation of Williams & Seeger (2000) is named GP-DTC, the SSGP of Lázaro-Gredilla et al. (2010) is named GP-SS, the VSSGP of Gal & Turner (2015) is named GP-VSS. The setting in this system identification task is simulation. This means that no past measured output observations (but perhaps predicted output observations) are used to predict next output observations. To enable a fair comparison, all methods are given access to a predefined amount of data. Details about the methods, their configuration, as well as the benchmark data-sets can be found in the Appendix 6.1 and 6.2.

5.1 Implementation

Our methods DRGP-(V)SS(-IP) were implemented in Python, using the library Theano and Matlab R2016b. For the optimization/training we used Python, Theano. Theano allows us to take full advantage of the automatic differentiation to calculate the gradients. For simulation and visualization we used Matlab R2016b.
We further implemented in Matlab R2016b the methods DRGP-Nyström, GP-SS, GP-DTC, GP-FITC, GP-full and used these implementations for the experiments. For GP-VSS, GP-LSTM and DGP-DS we used the published code[3][4][5]. For GP-SSM the published code is only applicable for small range of state dimension and a small time horizon, so we just show the results from their paper. For PRR-SSM we also show the results from their paper.

5.2 Benchmark data-sets Model Learning and Comparison

In Figure 1 we show a comparison of the latent model states before and after training, the simulation results, as well as the simulated latent states for the data-sets Drive and Actuator for the model DRGP-VSS. The test RMSE results for all methods are summarized in Table 1. The results show that on most data-sets DRGP-(V)SS(-IP) improve slightly in comparison to other methods. In order

[3]GP-VSS code available from `https://github.com/yaringal/VSSGP`.
[4]GP-LSTM code available from `https://github.com/alshedivat/keras-gp`.
[5]DGP-GP code available from `https://github.com/ICL-SML/Doubly-Stochastic-DGP`.

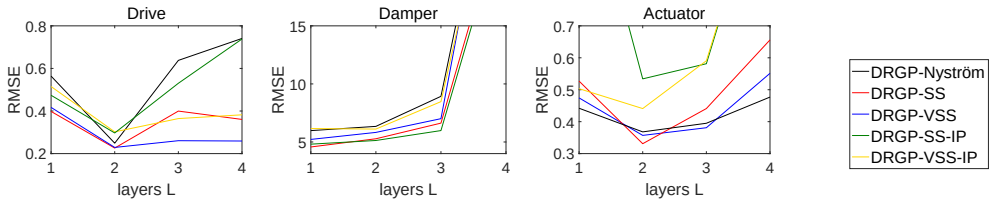

Figure 2: RMSE values for different number of layers L for the data-sets Drive, Damper, Actuator.

Table 1: Summary of RMSE values for the free simulation results on test data. Best values per data-set are bold. All values are calculated on the original data, unless the data-set Power Load, where the RMSE is shown for the normalized data. Here we have full recurrence for our methods and DRGP-Nyström and with auto-regressive part for standard sparse GPs and full GP. For non-rec we turned off the auto-regressive part in the first layer for our methods and DRGP-Nyström and also turned off the auto-regressive part for standard sparse GPs, full GP and DGP-DS.

methods-data	Emission	non-rec	Power Load	Damper	Actuator	non-rec	Ballbeam	Dryer	Drive	non-rec
DRGP-VSS	*0.104*	*0.062*	***0.457***	*5.825*	*0.357*	***0.388***	*0.084*	*0.109*	*0.229*	*0.268*
DRGP-VSS-IP	*0.119*	*0.064*	*0.544*	*6.112*	*0.441*	*0.546*	*0.071*	*0.107*	*0.302*	*0.293*
DRGP-SS	*0.108*	*0.062*	*0.497*	*5.277*	***0.329***	*0.563*	*0.081*	*0.108*	***0.226***	***0.253***
DRGP-SS-IP	*0.118*	*0.065*	*0.631*	***5.129***	*0.534*	*0.547*	*0.076*	*0.107*	*0.297*	*0.261*
DRGP-Nyström	0.109	0.059	0.493	6.344	0.368	0.415	0.082	0.109	0.249	0.289
DGP-DS	0.106	0.062	0.543	6.267	0.590	0.576	0.066	**0.085**	0.422	0.571
GP-LSTM	**0.096**	N/A	0.529	9.083	0.430	N/A	**0.062**	0.108	0.320	N/A
PR-SSM	N/A	N/A	N/A	N/A	0.502	N/A	0.073	0.140	0.492	N/A
GP-SSM	N/A	N/A	N/A	8.170	N/A	N/A	N/A	N/A	N/A	N/A
GP-VSS	0.130	0.058	0.514	6.554	0.449	0.767	0.120	0.112	0.401	0.549
GP-SS	0.128	0.060	0.539	6.730	0.439	0.777	0.077	0.106	0.358	0.556
GP-DTC	0.137	0.061	0.566	7.474	0.458	0.864	0.122	0.105	0.408	0.540
GP-FITC	0.126	**0.057**	0.536	6.754	0.433	0.860	0.084	0.108	0.403	0.539
GP-full	0.122	0.066	0.696	9.890	0.449	1.037	0.128	0.106	0.444	0.542

to evaluate the reproducing quality of our results, we provide a robustness test for our methods and DRGP-Nyström on the data-sets Drive and Damper. We run the optimization for different time horizons. For every method we visualized a boxplots with whiskers from minimum to maximum in Figure 4. For our models we derive good results compared with DRGP-Nyström on these data-sets, in particular for the setting of time horizons of Table 1 with $H = 10$. In Figure 2 the RMSE results for different layers L on the data-sets Drive, Damper and Actuator are shown. We can observe that different layers L are favored.

5.3 Conclusion

In this paper we introduced four new DRGPs based on the SS approximation introduced by Lázaro-Gredilla et al. (2010) and the improved VSS approximation by Gal & Turner (2015). We combined the (V)SS approximations with the variational inducing-point approximation from Titsias & Lawrence (2010), also integrated variationally over the input-space and established the existence of an optimal variational distribution of the weights $\boldsymbol{a}^{(l)}$ in the sense of a functional local optimum of the variational bounds for DRGP-(V)SS. We could show that our methods slightly improve on the data-sets used in this paper compared to the RGP from Mattos et al. (2016) and other state-of-the-art methods. Our sparse approximations are also practical for dimensionality reduction as shown in Titsias & Lawrence (2010) and can be further expanded to a deep version in this application (Damianou & Lawrence, 2013).

Acknowledgments

We would like to acknowledge support for this project from the ETAS GmbH.

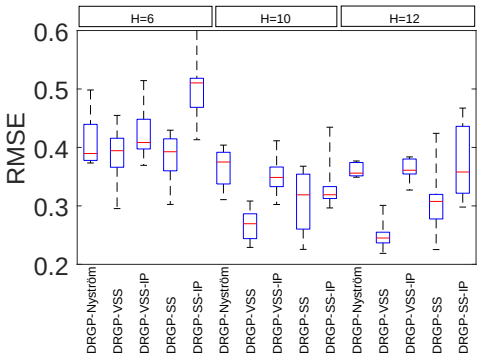

Figure 3: Data-set Drive, Boxplot, 10 runs.

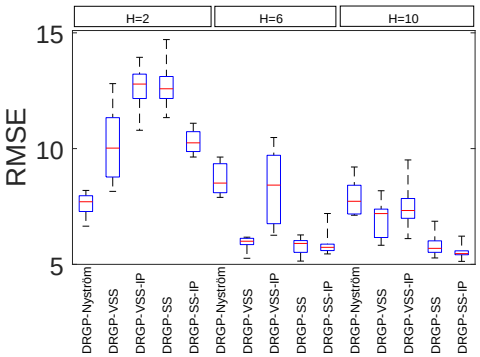

Figure 4: Data-set Damper, Boxplot, 10 runs.

REFERENCES

Maruan Al-Shedivat, Andrew Gordon Wilson, Yunus Saatchi, Zhiting Hu, and Eric P Xing. Learning scalable deep kernels with recurrent structure. *The Journal of Machine Learning Research*, 18(1): 2850–2886, 2017.

Christopher M Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Kurt Cutajar, Edwin V Bonilla, Pietro Michiardi, and Maurizio Filippone. Practical Learning of Deep Gaussian Processes via Random Fourier Features. *stat*, 1050:14, 2016.

Zhenwen Dai, Andreas Damianou, Javier González, and Neil D Lawrence. Variational Auto-encoded Deep Gaussian processes. *International Conference on Learning Representations*, 2016.

Andreas C Damianou and Neil D Lawrence. Deep Gaussian Processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

Andreas Dörr, Christian Daniel, Schiegg Martin, Duy Nguyen-Tuong, Stefan Schaal, Marc Toussaint, and Sebastian Trimpe. Probabilistic Recurrent State-Space Models. *arXiv preprint arXiv: 1801.10395*, 2018.

Roman Föll, Bernard Haasdonk, Markus Hanselmann, and Holger Ulmer. Deep Recurrent Gaussian Process with Variational Sparse Spectrum Approximation. *arXiv preprint arXiv:1711.00799v2*, 2017.

Roger Frigola, Yutian Chen, and Carl Edward Rasmussen. Variational Gaussian process state-space models. In *Advances in Neural Information Processing Systems*, pp. 3680–3688, 2014.

Yarin Gal and Richard Turner. Improving the Gaussian Process Sparse Spectrum Approximation by Representing Uncertainty in Frequency Inputs. In *International Conference on Machine Learning*, pp. 655–664, 2015.

Yarin Gal, Mark van der Wilk, and Carl Edward Rasmussen. Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Advances in Neural Information Processing Systems*, pp. 3257–3265, 2014.

James Hensman and Neil D Lawrence. Nested Variational Compression in Deep Gaussian Processes. *stat*, 1050:3, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881, 2010.

César Lincoln C Mattos, Zhenwen Dai, Andreas Damianou, Jeremy Forth, Guilherme A Barreto, and Neil D Lawrence. Recurrent Gaussian processes. *International Conference on Learning Representations*, 2016.

Oliver Nelles. *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer Science & Business Media, 2013.

Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013.

Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

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

Jonas Sjöberg, Qinghua Zhang, Lennart Ljung, Albert Benveniste, Bernard Delyon, Pierre-Yves Glorennec, Håkan Hjalmarsson, and Anatoli Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257, 2006.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Andreas Svensson, Arno Solin, Simo Särkkä, and Thomas Schön. Computationally efficient Bayesian learning of Gaussian process state space models. In *Artificial Intelligence and Statistics*, pp. 213–221, 2016.

Michalis K Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Artificial Intelligence and Statistics*, volume 5, pp. 567–574, 2009.

Michalis K Titsias and Neil D Lawrence. Bayesian Gaussian Process Latent Variable Model. In *Artificial Intelligence and Statistics*, volume 9, pp. 844–851, 2010.

Ryan D Turner, Marc Peter Deisenroth, and Carl Edward Rasmussen. State-Space Inference and Learning with Gaussian Processes. In *Artificial Intelligence and Statistics*, pp. 868–875, 2010.

Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *Neural Information Processing Systems*, volume 18, pp. 3, 2005.

Jiandong Wang, Akira Sano, Tongwen Chen, and Biao Huang. Identification of Hammerstein systems without explicit parameterisation of non-linearity. *International Journal of Control*, 82(5): 937–952, 2009.

Torbjörn Wigren. Input-output data sets for development and benchmarking in nonlinear identification. *Technical Reports from the department of Information Technology*, 20:2010–020, 2010.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pp. 661–667. MIT press, 2000.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pp. 1067–1075, 2013.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pp. 370–378, 2016.

6 APPENDIX

This additional material provides details about the derivations and configuration of the proposed DRGP-(V)SS(-IP) in the Sections 3-4. and elaborates on the methods and the employed data-sets in the experiments in Section 5.

6.1 DATA-SET DESCRIPTION

Table 2: Summary of the data-sets for the system identification tasks

parameters \ **data-sets**	N	N_{train}	N_{test}	input dimension	output dimension
Drive	500	250	250	1	1
Dryer	1000	500	500	1	1
Ballbeam	1000	500	500	1	1
Actuator	1024	512	512	2	1
Damper	3499	2000	1499	1	1
Power Load	9518	7139	2379	11	1
Emission	12500	10000	2500	6	1

In this section we introduce the data-sets we used in our experiments. We chose a large number of data-sets in training size going from 250 to 12500 data-points in order to show the performance for a wide range. We will begin with the smallest, the *Drive* data-set, which was first introduced by Wigren (2010). It is based on a system which has two electric motors that drive a pulley using a flexible belt. The input is the sum of voltages applied to the motors and the output is the speed of the belt. The data-set *Dryer*[6] describes a system, where air is fanned through a tube and heated at an inlet. The input is the voltage over the heating device (a mesh of resistor wires). The output is the air temperature measured by a thermocouple. The third data-set *Ballbeam*[6][7] describes a system, where the input is the angle of a beam and the output the position of a ball. *Actuator* is the name of the fourth data-set, which was described by Sjöberg et al. (1995) and which stems from an hydraulic actuator that controls a robot arm, where the input is the size of the actuator's valve opening and the output is its oil pressure. The *Damper* data-set, introduced by Wang et al. (2009), poses the problem of modeling the input–output behavior of a magneto-rheological fluid damper and is also used as a case study in the System Identification Toolbox of Mathworks Matlab. The data-set *Power Load*[8], used in Al-Shedivat et al. (2017), consists of data, where the power load should be predicted from the historical temperature data. This data-set was used for 1-step ahead prediction, where past measured output observations are used to predict current or next output observations, but we will use it here for free simulation. We down-sampled by starting with the first sample and choosing every 4th data-point, because the original data-set with a size of 38064 samples and a chosen time-horizon of 48 is too large for our implementation, which is not parallelized so far. The newly provided data-set *Emission*[9] contains an emission-level of nitrogen oxide from a driving car as output and as inputs the indicated torque, boost pressure, EGR (exhaust gas recirculation) rate, injection, rail pressure and speed. The numerical characteristics of all data-sets are summarized in Table 2. The separation of the data-sets Drive, Dryer, Ballbeam, Actuator, Damper, Power Load in training- and test-data was given by the papers we use for comparison. The separation of the Emission data-set was chosen by ourself.

[6]Received from `http://homes.esat.kuleuven.be/~tokka/daisydata.html`.

[7]Description can be found under `http://forums.ni.com/t5/NI-myRIO/myBall-Beam-Classic-Control-Experiment/ta-p/3498079`.

[8]Originally received from Global Energy Forecasting Kaggle competitions organized in 2012.

[9]Available from `http://github.com/RomanFoell/DRGP-VSS`.

6.2 NONLINEAR SYSTEM IDENTIFICATION

Table 3: Summary of the methods for the system identification tasks

method	references
DRGP-VSS	this one
DRGP-VSS-IP	this one
DRGP-SS	this one
DRGP-SS-IP	this one
DRGP-Nyström	(Mattos et al., 2016)
DGP-DS	(Salimbeni & Deisenroth, 2017)
GP-LSTM	(Al-Shedivat et al., 2017)
PR-SSM	(Dörr et al., 2018)
GP-SSM	(Svensson et al., 2016)
GP-VSS	(Gal & Turner, 2015)
GP-SS	(Lázaro-Gredilla et al., 2010)
GP-DTC	(Williams & Seeger, 2000)
GP-FITC	(Snelson & Ghahramani, 2006)
GP-full	(Rasmussen, 2006)

The methods for the system identification tasks and their references are summarized in Table 2.
For the data-sets Drive and Actuator we chose for our methods DRGP-(V)SS(-IP) the setting $L = 2$ hidden layers, $M = 100$ spectral-points and time-horizon $H_\mathrm{h} = H_\mathbf{x} = 10$, which was also used in Mattos et al. (2016), Al-Shedivat et al. (2017) and Dörr et al. (2018) for free simulation (using M pseudo-input points instead of spectral-points). For these two data-sets we filled the results from Mattos et al. (2016); Al-Shedivat et al. (2017) into Table 1. Further, for our methods DRGP-(V)SS(-IP) and DRGP-Nyström we chose on the data-sets Ballbeam and Dryer $L = 1$, $M = 100$ and $H_\mathrm{h} = H_\mathbf{x} = 10$. For the data-set Damper we chose $L = 2$, $M = 125$ and $H_\mathrm{h} = H_\mathbf{x} = 10$. For the data-set Power Load we chose $L = 1$, $M = 125$ and $H_\mathrm{h} = H_\mathbf{x} = 12$. For the data-set Emission we chose $L = 1$, $M = 125$ and $H_\mathrm{h} = H_\mathbf{x} = 10$.
The other sparse GP, the full GP and DGP-DS were trained with NARX structure $H_\mathbf{x} = H_\mathrm{y}$ with the same time horizon as for our DRGPs and with the same amount of pseudo-input points or spectral points.
For GP-LSTM we chose the same setting for the amount of hidden layers, pseudo-input points and time horizon as for our DRGPs. We used Adam optimizer with a learning rate of 0.01. As activation function we chose $\tanh$ and as setting free simulation. We tested with 8, 16, 32, 64, 128 hidden units (every hidden layer of a RNN is specified by a hidden unit parameter) for all training data-sets. For the data-sets with training size smaller or equal to 2000 we chose the version GP-LSTM in Al-Shedivat et al. (2017) and for the ones larger than 2000 the scalable version MSGP-LSTM. We did not pre-train the weights.
For DGP-DS we tested for $L = 1, \ldots, 3$ number of GPs. We used natural gradients for the last GP with gamma 0.1 and Adam optimizer for the others with a learning rate of 0.01. The batch size was chosen to be the training data-size.
All GPs which use the Nyström approximation were initialized for the pseudo-inputs points with a random subset of size M from the input-data and trained with SE covariance function. For the ones which use the (V)SS approximations, which includes our methods, we trained with a spectral-point initialization sampled from $\mathcal{N}(\mathbf{0}, I_Q)$, an initialization for the pseudo-input points with a random subset of size M from the input-data (or setting them all to zero; not in the IP case). For our methods DRGP-(V)SS(-IP) and GP-VSS we fixed the length scales $\mathrm{p}_q^{(l)} = \infty$, for all q, l. So all GPs with (V)SS approximations were also initialized as SE covariance function, see Equation 5.
For all methods we used automatic relevance determination, so each input dimension has its own length scale. For our methods DRGP-(V)SS(-IP) and DRGP-Nyström, the noise parameters and the hyperparameters were initialized by $\sigma_\text{noise}^{(l)} = 0.1$, $\sigma_\text{power}^{(l)} = 1$ and the length scales by either $\mathrm{l}_q^{(l)} = \sqrt{\max(\hat{\mathbf{H}}_q^{(l)}) - \min(\hat{\mathbf{H}}_q^{(l)})}$ or $\mathrm{l}_q^{(l)} = \max(\hat{\mathbf{H}}_q^{(l)}) - \min(\hat{\mathbf{H}}_q^{(l)})$, for all q, l, where $\hat{\mathbf{H}}_q^{(l)}$ is the data-vector containing the q-th input-dimension values of every input-data point $\hat{\mathbf{h}}_i^{(l)}$, for all i. Furthermore, we initialized the latent hidden GP states with the output-observation data $\mathbf{y}$.
The other standard GPs were also initialized with $\sigma_\text{noise} = 0.1$, $\sigma_\text{power} = 1$ and the same initialization for length scales with respect to the NARX structure input data as before.
For GP-LSTM we used the initialization for the weights provided by Keras, a Deep Learning library

for Theano and TensorFlow and $\sigma_{\text{noise}} = 0.1$, $\sigma_{\text{power}} = 1$ and for the length scale initialization we chose $\mathrm{l}_q = 1$ for all input-dimensions.
For DGP-DS we used the initialization $\sigma_{\text{noise}}^{(l)} = 0.1$ and $\sigma_{\text{power}}^{(l)} = 1$ and for the length scale initialization we chose $\mathrm{l}_q^{(l)} = 1$ for all l, q.
For all our implementations, see Section 5.1, we used the positive transformation $\mathrm{x}' = \log(1 + \exp(\mathrm{x}))^2$ for the calculation of the gradients in order for the parameters constrained to be positive and with L-BFGS optimizer, either from Matlab R2016b with fmincon, or Python 2.7.12 with scipy optimize.
All methods were trained on the normalized data $\frac{\cdots - \mu}{\sigma^2}$ (for every dimension independently), several times (same amount per data-set: the initializations are still not deterministic, e.g. for pseudo-inputs points and spectral points) with about 50 to 100 iterations and the best results in RMSE value on the test-data are shown in Table 1.
For our methods DRGP-(V)SS(-IP) and DRGP-Nyström we fixed $\sigma_{\text{noise}}^{(l)}$, $\sigma_{\text{power}}^{(l)}$ for all l (optional the spectral points/pseudo-input points for DRGP-(V)SS; for the IP case we never excluded the pseudo-input points because we would fall back to the DRGP-(V)SS case; for DRGP-Nyström we always included the pseudo-input points) during the first iterations to independently train the latent states. For all other GPs we also tested with fixed and not fixed $\sigma_{\text{noise}} = 0.1$, except GP-LSTM and DGP-DS. For DRGP-VSS(-IP) we fixed $\boldsymbol{\beta}_m^{(l)}$ for all m, l to small value around 0.001, as well as the spectral phases b_m for all m, l sampling from Unif $[0, 2\pi]$ (this seems to work better in practice). The limitations for $\boldsymbol{\beta}_m^{(l)}$ also holds for GP-VSS as well.
We want to signify at this point that setting $\mathbf{u}_m = \mathbf{0}$ for all $m = 1, \ldots, M$ worked sometimes better than choosing a subset from the input-data (not in the IP case). This seems to be different to Gal & Turner (2015), who pointed out: 'These are necessary to the approximation. Without these points (or equivalently, setting these to $\mathbf{0}$), the features would decay quickly for data points far from the origin (the fixed point $\mathbf{0}$).'
For GP-SSM we show the result of the data-set Damper from their paper in Table 1.
For PR-SSM we show the results from their paper in Table 1.
We show additional results for the data-sets Drive, Actuator and Emission with missing auto-regressive part for the first layer for our methods DRGP-(V)SS(-IP) and DRGP-Nyström in Table 1, named non-rec. For the sparse GP, the full GP and DGP-DS and the data-sets Drive, Actuator and Emission we show the results with missing auto-regressive part in Table 1, just modeling the data with exogenous inputs. Here we want to examine the effect of the auto-regressive part of the first layer for the DRGP models on the RMSE. GP-LSTM, GP-SSM and PR-SSM are not listed for this setting of recurrence (we used the original code for GP-LSTM, which is not adaptable for this recurrence setting).

6.3 Variational approximation and inference for DRGP-(V)SS(-IP)

In the Sections 6.3.1-6.3.3 we derive the statistics Ψ_0, Ψ_1, Ψ_2 and Ψ_{reg} for the four model cases DRGP-(V)SS(-IP). In Section 6.3.4 we derive the resulting variational lower bounds. In Section 6.3.5 we show the mean and variance expressions of the predictive distributions of our DRGP models.

In the following we use the abbreviations and formulas

- B.1, (Gal & Turner, 2015, see Section 4.1),
- B.2, (Rasmussen, 2006, see A.7.),
- $\frac{1}{2}(\cos(a - b + x - y) + \cos(a + b + x + y)) = \cos(x + a)\cos(y + b)$.
- JI, for Jensen's inequality.

6.3.1 DRGP-VSS(-IP), the statistics Ψ_1, Ψ_2

For the versions DRGP-VSS(-IP) the statistics are

$$
\begin{aligned}
(\Psi_1)_{nm} &= \mathbf{E}\left[\Phi_{nm}\right]_{q_{\boldsymbol{z}_m} q_{\boldsymbol{h}_n}} \\
&\overset{\text{B.1}}{=} \mathbf{E}\left[\sqrt{2\sigma_{\text{power}}^2 (M)^{-1}} e^{-\frac{1}{2}\bar{\mathbf{h}}_{nm}^T \boldsymbol{\beta}_m \bar{\mathbf{h}}_{nm}} \cos(\hat{\boldsymbol{\alpha}}_m^T (\mathbf{h}_n - \mathbf{u}_m) + \mathrm{b}_m)\right]_{q_{\boldsymbol{h}_n}} \\
&\overset{\text{B.2}}{=} \Sigma_m^1 \mathrm{Z}_{nm} e^{-\frac{1}{2}\hat{\boldsymbol{\alpha}}_m^T \mathrm{C}_{nm} \hat{\boldsymbol{\alpha}}_m} \cos(\hat{\boldsymbol{\alpha}}_m^T (\mathbf{c}_{nm} - \mathbf{u}_m) + \mathrm{b}_m),
\end{aligned}
$$

for $m, = 1, \ldots, M$, $n = 1, \ldots, N$ with

$$\begin{aligned}
\bar{\mathbf{h}}_{nm} &= 2\pi\mathfrak{L}^{-1}(\mathbf{h}_n - \mathbf{u}_m),\\
\hat{\boldsymbol{\alpha}}_m &= 2\pi(\mathfrak{L}^{-1}\boldsymbol{\alpha}_m + \mathbf{p}),\\
\mathbf{c}_{nm} &= \mathrm{C}_{nm}(\boldsymbol{\beta}_m(2\pi)^2\mathfrak{L}^{-2}\mathbf{u}_m + \boldsymbol{\lambda}_n^{-1}\boldsymbol{\mu}_n),\\
\mathrm{C}_{nm} &= (\boldsymbol{\beta}_m(2\pi)^2\mathfrak{L}^{-2} + \boldsymbol{\lambda}_n^{-1})^{-1},\\
\mathbf{v}_{nm} &= \mathbf{u}_m - \boldsymbol{\mu}_n,\\
\mathrm{V}_{nm} &= (2\pi)^{-2}\mathfrak{L}^2\boldsymbol{\beta}_m^{-1} + \boldsymbol{\lambda}_n,\\
\mathrm{Z}_{nm} &= \frac{1}{\sqrt{|\mathrm{V}_{nm}|}}e^{-\frac{1}{2}\mathbf{v}_{nm}^T\mathrm{V}_{nm}^{-1}\mathbf{v}_{nm}},\\
\Sigma_m^1 &= \sqrt{2\sigma_{\text{power}}^2\prod_{q=1}^{Q}\left(\frac{\mathrm{l}_q^2}{\beta_{m_q}}\right)(M)^{-1}},
\end{aligned}$$

and

$\Psi_2 = \sum_{n=1}^{N}(\Psi_2)^n$, where

$$\begin{aligned}
(\Psi_2)_{mm'}^n &= \mathbf{E}\left[\Phi_{nm}^T\Phi_{nm'}\right]_{q_{\boldsymbol{z}_m}q_{\boldsymbol{z}_{m'}}q_{\boldsymbol{h}_n}}\\
&\overset{\text{B.1}}{=} \mathbf{E}\Big[2\sigma_{\text{power}}^2(M)^{-1}e^{-\frac{1}{2}(\bar{\mathbf{h}}_{nm}^T\boldsymbol{\beta}_m\bar{\mathbf{h}}_{nm}+\bar{\mathbf{h}}_{nm'}^T\boldsymbol{\beta}_{m'}\bar{\mathbf{h}}_{nm'})}\\
&\cos(\hat{\boldsymbol{\alpha}}_m^T(\mathbf{h}_n - \mathbf{u}_m) + \mathrm{b}_m)\cos(\hat{\boldsymbol{\alpha}}_{m'}^T(\mathbf{h}_n - \mathbf{u}_{m'}) + \mathrm{b}_{m'})\Big]_{q_{\boldsymbol{h}_n}}\\
&\overset{\text{B.2}}{=} \Sigma_{mm'}^2\mathrm{Z}_{mm'}^n\Big(e^{-\frac{1}{2}\bar{\boldsymbol{\alpha}}_{mm'}^T\mathrm{D}_{mm'}^n\bar{\boldsymbol{\alpha}}_{mm'}}\cos(\bar{\boldsymbol{\alpha}}_{mm'}^T\mathbf{d}_{mm'}^n - \bar{\tau}_{mm'} + \bar{\mathrm{b}}_{mm'})\\
&+e^{-\frac{1}{2}\overset{+}{\boldsymbol{\alpha}}_{mm'}^T\mathrm{D}_{mm'}^n\overset{+}{\boldsymbol{\alpha}}_{mm'}}\cos(\overset{+}{\boldsymbol{\alpha}}_{mm'}^T\mathbf{d}_{mm'}^n - \overset{+}{\tau}_{mm'} + \overset{+}{\mathrm{b}}_{mm'})\Big).
\end{aligned}$$

for $m, m' = 1, \ldots, M$, $m \neq m'$, with

$$\begin{aligned}
\bar{\mathrm{b}}_{mm'} &= \mathrm{b}_m - \mathrm{b}_{m'},\\
\overset{+}{\mathrm{b}}_{mm'} &= \mathrm{b}_m + \mathrm{b}_{m'},\\
\bar{\tau}_{mm'} &= \hat{\boldsymbol{\alpha}}_m^T\mathbf{u}_m - \hat{\boldsymbol{\alpha}}_{m'}^T\mathbf{u}_{m'},\\
\overset{+}{\tau}_{mm'} &= \hat{\boldsymbol{\alpha}}_m^T\mathbf{u}_m + \hat{\boldsymbol{\alpha}}_{m'}^T\mathbf{u}_{m'},\\
\bar{\boldsymbol{\alpha}}_{mm'} &= \hat{\boldsymbol{\alpha}}_m - \hat{\boldsymbol{\alpha}}_{m'},\\
\overset{+}{\boldsymbol{\alpha}}_{mm'} &= \hat{\boldsymbol{\alpha}}_m + \hat{\boldsymbol{\alpha}}_{m'},\\
\mathbf{b}_{mm'} &= \mathrm{B}_{mm'}(2\pi)^2\mathfrak{L}^{-2}(\boldsymbol{\beta}_m\mathbf{u}_m + \boldsymbol{\beta}_{m'}\mathbf{u}_{m'}),\\
\boldsymbol{\beta}_{mm'} &= \boldsymbol{\beta}_m + \boldsymbol{\beta}_{m'},\\
\mathrm{B}_{mm'} &= (2\pi)^{-2}\mathfrak{L}^2\boldsymbol{\beta}_{mm'}^{-1},\\
\mathbf{d}_{mm'}^n &= \mathrm{D}_{mm'}^n(\mathrm{B}_{mm'}^{-1}\mathbf{b}_{mm'} + \boldsymbol{\lambda}_n^{-1}\boldsymbol{\mu}_n),\\
\mathrm{D}_{mm'}^n &= (\mathrm{B}_{mm'}^{-1} + \boldsymbol{\lambda}_n^{-1})^{-1},\\
\mathbf{w}_{mm'}^n &= \mathbf{b}_{mm'} - \boldsymbol{\mu}_n,\\
\mathrm{W}_{mm'}^n &= \mathrm{B}_{mm'} + \boldsymbol{\lambda}_n,\\
\mathbf{u}_{mm'} &= \mathbf{u}_m - \mathbf{u}_{m'},\\
\mathbf{U}_{mm'} &= (2\pi)^{-2}\mathfrak{L}^2(\boldsymbol{\beta}_m^{-1} + \boldsymbol{\beta}_{m'}^{-1}),\\
\mathrm{Z}_{mm'}^n &= \frac{1}{\sqrt{|\mathrm{W}_{mm'}^n|\,|\mathbf{U}_{mm'}|}}e^{-\frac{1}{2}(\mathbf{w}_{mm'}^{n\,T}\mathrm{W}_{mm'}^{n\,-1}\mathbf{w}_{mm'}^n+\mathbf{u}_{mm'}^T\mathbf{U}_{mm'}^{-1}\mathbf{u}_{mm'})},\\
\Sigma_{mm'}^2 &= \sigma_{\text{power}}^2\prod_{q=1}^{Q}\left(\frac{\mathrm{l}_q^2}{\sqrt{\beta_{m_q}\beta_{m'_q}}}\right)(M)^{-1},
\end{aligned}$$

and

$$
\begin{aligned}
(\Psi_2)^n_{mm} &= \mathbf{E}\left[\Phi_{nm}^T \Phi_{nm}\right]_{q_{\boldsymbol{z}_m} q_{\boldsymbol{h}_n}} \\
&\overset{\text{B.1}}{=} \mathbf{E}\left[2\sigma^2_{\text{power}}(M)^{-1}(\frac{1}{2} + \frac{1}{2} e^{-2\bar{\mathbf{h}}_{nm}^T \boldsymbol{\beta}_m \bar{\mathbf{h}}_{nm}} \cos(2(\hat{\boldsymbol{\alpha}}_m^T(\mathbf{h}_n - \mathbf{u}_m) + \mathrm{b}_m))\right]_{q_{\boldsymbol{h}_n}} \\
&\overset{\text{B.2}}{=} \sigma^2_{\text{power}}(M)^{-1}\left(1 + \tilde{\Sigma}_m^2 \tilde{\mathrm{Z}}_{nm} e^{-2\hat{\boldsymbol{\alpha}}_m^T \tilde{\mathrm{C}}_{nm} \hat{\boldsymbol{\alpha}}_m} \cos(2(\hat{\boldsymbol{\alpha}}_m^T(\tilde{\mathbf{c}}_{nm} - \mathbf{u}_m) + \mathrm{b}_m))\right),
\end{aligned}
$$

for $m, m' = 1, \ldots, M$, $m = m'$, with

$$
\tilde{\Sigma}_m^2 = \sqrt{\prod_{q=1}^{Q} \left(\frac{\mathrm{l}_q^2}{\beta_{m_q}}\right)} 2^{-Q}.
$$

6.3.2 DRGP-SS(-IP), THE STATISTICS Ψ_1, Ψ_2

For the versions DRGP-SS(-IP) the statistics are

$$
\begin{aligned}
(\Psi_1)_{nm} &= \mathbf{E}\left[\Phi_{nm}\right]_{q_{\boldsymbol{h}_n}} \\
&= \mathbf{E}\left[\sqrt{2\sigma^2_{\text{power}}(M)^{-1}} \cos(\hat{\mathbf{z}}_m^T(\mathbf{h}_n - \mathbf{u}_m) + \mathrm{b}_m)\right]_{q_{\boldsymbol{h}_n}} \\
&\overset{\text{B.2}}{=} \sqrt{2\sigma^2_{\text{power}}(M)^{-1}} e^{-\frac{1}{2}\hat{\mathbf{z}}_m^T \boldsymbol{\lambda}_n \hat{\mathbf{z}}_m} \cos(\hat{\mathbf{z}}_m^T(\boldsymbol{\mu}_n - \mathbf{u}_m) + \mathrm{b}_m),
\end{aligned}
$$

for $m = 1, \ldots, M$, $n = 1, \ldots, N$ with

$$
\hat{\mathbf{z}}_m = 2\pi(\mathfrak{L}^{-1}\mathbf{z}_m + \mathbf{p}),
$$

and

$\Psi_2 = \sum_{n=1}^{N} (\Psi_2)^n$, where

$$
\begin{aligned}
(\Psi_2)^n_{mm'} &= \mathbf{E}\left[\Phi_{nm}^T \Phi_{nm'}\right]_{q_{\boldsymbol{h}_n}} \\
&\overset{\text{B.1}}{=} \mathbf{E}\left[2\sigma^2_{\text{power}}(M)^{-1} \cos(\hat{\mathbf{z}}_m^T(\mathbf{h}_n - \mathbf{u}_m) + \mathrm{b}_m) \cos(\hat{\mathbf{z}}_{m'}^T(\mathbf{h}_n - \mathbf{u}_{m'}) + \mathrm{b}_{m'})\right]_{q_{\boldsymbol{h}_n}} \\
&= \sigma^2_{\text{power}}(M)^{-1}\Big(e^{-\frac{1}{2}\bar{\mathbf{z}}_{mm'}^T \boldsymbol{\lambda}_n \bar{\mathbf{z}}_{mm'}} \cos(\bar{\mathbf{z}}_{mm'}^T \boldsymbol{\mu}_n - \bar{\rho}_{mm'} + \bar{\mathrm{b}}_{mm'}) \\
&+ e^{-\frac{1}{2}\overset{+}{\mathbf{z}}_{mm'}^T \boldsymbol{\lambda}_n \overset{+}{\mathbf{z}}_{mm'}} \cos(\overset{+}{\mathbf{z}}_{mm'}^T \boldsymbol{\mu}_n - \overset{+}{\rho}_{mm'} + \overset{+}{\mathrm{b}}_{mm'}))\Big),
\end{aligned}
$$

for $m, m' = 1, \ldots, M$ with

$$
\begin{aligned}
\bar{\rho}_{mm'} &= \hat{\mathbf{z}}_m^T \mathbf{u}_m - \hat{\mathbf{z}}_{m'}^T \mathbf{u}_{m'}, \\
\overset{+}{\rho}_{mm'} &= \hat{\mathbf{z}}_m^T \mathbf{u}_m + \hat{\mathbf{z}}_{m'}^T \mathbf{u}_{m'}, \\
\bar{\mathbf{z}}_{mm'} &= \hat{\mathbf{z}}_m - \hat{\mathbf{z}}_{m'}, \\
\overset{+}{\mathbf{z}}_{mm'} &= \hat{\mathbf{z}}_m + \hat{\mathbf{z}}_{m'},
\end{aligned}
$$

for the other variables, see the defined variables in the DRGP-VSS case.

6.3.3 DRGP-(V)SS-IP, THE STATISTICS Ψ_{reg} AND Ψ_0

For $\Psi_{reg} = \mathbf{E}\left[K_{MN} K_{NM}\right]_{q_{\boldsymbol{H}}}$ see Ψ_2 in 6.3.1 but setting $b_m = 0$, $\alpha_m = 0$ and $\beta_m = 1$ for all $m = 1, \ldots, M$.

Ψ_0 naturally is given by $\Psi_0 = \mathrm{tr}\left(\mathbf{E}[K_{NN}]_{q_{\boldsymbol{H}}}\right) = N\sigma^2_{\text{power}}$ because of the chosen SE covariance function.

6.3.4 DRGP-(V)SS(-IP), LOWER BOUNDS

In this section we derive the different variational lower bounds for our models DRGP-(V)SS(-IP). We first show the bound $\mathcal{L}_{\text{(V)SS}}^{\text{REVARB}}$ without optimal variational distribution for $\boldsymbol{a}^{(l)}$. Then the bounds $\mathcal{L}_{\text{(V)SS-opt}}^{\text{REVARB}}$ with optimal variational distribution for $\boldsymbol{a}^{(l)}$ follows, as well as the $\mathcal{L}_{\text{(V)SS-IP}}^{\text{REVARB}}$ case.
We use the simplified notation $d\mathbf{A}d\mathbf{Z}d\mathbf{H} = d\mathbf{a}^{(1)} \dots d\mathbf{a}^{(L+1)} d\mathbf{Z}^{(1)} \dots d\mathbf{Z}^{(L+1)} d\mathbf{h}^{(1)} \dots d\mathbf{h}^{(L)}$.

$$
\begin{aligned}
&\log(p(\mathbf{y}_{H_\mathbf{x}+1:}|\mathbf{X})) \\
&= \log\left(\int p\left(\mathbf{y}_{H_\mathbf{x}+1:}, \left[\mathbf{a}^{(l)}, \boldsymbol{Z}^{(l)}, \mathbf{h}^{(l)}, \boldsymbol{U}^{(l)}\right]_{l=1}^{L+1} |\mathbf{X}\right) d\mathbf{A}d\mathbf{Z}d\mathbf{H}\right) \\
&= \log\left(\int \frac{Q_{\text{REVARB}}}{Q_{\text{REVARB}}} p\left(\mathbf{y}_{H_\mathbf{x}+1:}, \mathbf{a}^{(L+1)}, \boldsymbol{Z}^{(L+1)}, \boldsymbol{U}^{(L+1)}, \left[\mathbf{a}^{(l)}, \boldsymbol{Z}^{(l)}, \mathbf{h}^{(l)}, \boldsymbol{U}^{(l)}\right]_{l=1}^{L} |\mathbf{X}\right) d\mathbf{A}d\mathbf{Z}d\mathbf{H}\right) \\
&\overset{\text{JI}}{\geq} \int Q_{\text{REVARB}} \log\left(\frac{p\left(\mathbf{y}_{H_\mathbf{x}+1:}, \mathbf{a}^{(L+1)}, \boldsymbol{Z}^{(L+1)}, \boldsymbol{U}^{(L+1)}, \left[\mathbf{a}^{(l)}, \boldsymbol{Z}^{(l)}, \mathbf{h}^{(l)}, \boldsymbol{U}^{(l)}\right]_{l=1}^{L} |\mathbf{X}\right)}{Q_{\text{REVARB}}}\right) d\mathbf{A}d\mathbf{Z}d\mathbf{H} \\
&= \sum_{l=1}^{L+1} \int q(\mathbf{a}^{(l)}) q(\boldsymbol{Z}^{(l)}) q(\mathbf{h}^{(l)}) p(\mathbf{h}_{H_\mathbf{x}+1:}^{(l)} | \mathbf{a}^{(l)}, \boldsymbol{Z}^{(l)}, \hat{\mathbf{H}}^{(l)}, \boldsymbol{U}^{(l)}) d\mathbf{a}^{(l)} d\mathbf{Z}^{(l)} d\mathbf{h}^{(l)} \\
&- \sum_{l=1}^{L}\left(\frac{\hat{N}}{2} - \sum_{i=1+H_\mathbf{x}-H_\text{h}}^{N} \frac{\log(2\pi\lambda_i^{(l)})}{2} + \sum_{i=1+H_\mathbf{x}-H_\text{h}}^{H_\text{h}} \frac{\log(2\pi)}{2} + \frac{\left(\lambda_i^{(l)} + \left(\mu_i^{(l)}\right)^2\right)}{2}\right) \\
&- \sum_{l=1}^{L+1} \mathbf{KL}(q_{\boldsymbol{a}^{(l)}} || p_{\boldsymbol{a}^{(l)}}) - \mathbf{KL}(q_{\boldsymbol{Z}^{(l)}} || p_{\boldsymbol{Z}^{(l)}}) \\
&= -\frac{\hat{N}}{2} \sum_{l=1}^{L+1} \log(2\pi(\sigma_{\text{noise}}^{(l)})^2) - \frac{\mathbf{y}_{H_\mathbf{x}+1:}^T \mathbf{y}_{H_\mathbf{x}+1:}}{2\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} \\
&+ \frac{\mathbf{y}_{H_\mathbf{x}+1:}^T \Psi_1^{(\text{L+1})} \mathbf{m}^{(L+1)}}{\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} - \frac{\text{tr}\left(\Psi_2^{(\text{L+1})}(\mathbf{s}^{(L+1)} + \mathbf{m}^{(L+1)}(\mathbf{m}^{(L+1)})^T)\right)}{2\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} \\
&+ \sum_{l=1}^{L}\left(-\frac{1}{2\left(\sigma_{\text{noise}}^{(l)}\right)^2}\left(\left(\sum_{i=L+1}^{N} \lambda_i^{(l)}\right) + (\boldsymbol{\mu}^{(l)})_{H_\mathbf{x}+1:}^T \boldsymbol{\mu}_{H_\mathbf{x}+1:}^{(l)}\right)\right. \\
&+ \frac{\left((\boldsymbol{\mu}_{H_\mathbf{x}+1:}^{(l)})^T \Psi_1^{(l)} \mathbf{m}^{(l)}\right)}{\left(\sigma_{\text{noise}}^{(l)}\right)^2} - \frac{\text{tr}\left(\Psi_2^{(l)}(\mathbf{s}^{(l)} + \mathbf{m}^{(l)}(\mathbf{m}^{(l)})^T)\right)}{2\left(\sigma_{\text{noise}}^{(l)}\right)^2} \\
&\left.- \frac{\hat{N}}{2} + \sum_{i=1+H_\mathbf{x}-H_\text{h}}^{N} \frac{\log(2\pi\lambda_i^{(l)})}{2} - \sum_{i=1+H_\mathbf{x}-H_\text{h}}^{H_\text{h}} \frac{\log(2\pi)}{2} - \frac{\left(\lambda_i^{(l)} + \left(\mu_i^{(l)}\right)^2\right)}{2}\right) \\
&- \sum_{l=1}^{L+1} \mathbf{KL}(q_{\boldsymbol{a}^{(l)}} || p_{\boldsymbol{a}^{(l)}}) - \mathbf{KL}(q_{\boldsymbol{Z}^{(l)}} || p_{\boldsymbol{Z}^{(l)}}) \\
&\overset{\text{def}}{=} \mathcal{L}_{\text{VSS}}^{\text{REVARB}},
\end{aligned}
$$

where we derive $\mathcal{L}_{\text{SS}}^{\text{REVARB}}$ by being not variational over $\boldsymbol{Z}^{(l)}$. Using the optimal distribution for $\boldsymbol{a}^{(l)} \sim \mathcal{N}((A^{(l)})^{-1}(\Psi_1^{(l)})^T \boldsymbol{\mu}_{H_\mathbf{x}+1:}^{(l)}, (\sigma_{\text{noise}}^{(l)})^2 (A^{(l)})^{-1})$, with $A^{(l)} = \Psi_2^{(l)} + (\sigma_{\text{noise}}^{(l)})^2 I_M$, respective $\mathbf{y}_{H_\mathbf{x}+1:}$

for $L+1$ we obtain

$$
\begin{aligned}
&\geq -\frac{\hat{N}-M}{2}\sum_{l=1}^{L+1}\log\left(\left(\sigma_{\text{noise}}^{(l)}\right)^2\right) - \frac{\mathbf{y}_{H_{\mathbf{x}}+1:}^T\mathbf{y}_{H_{\mathbf{x}}+1:}}{2\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} \\
&+ \frac{\mathbf{y}_{H_{\mathbf{x}}+1:}^T\Psi_1^{(\text{L+1})}(A^{(L+1)})^{-1}(\Psi_1^{(\text{L+1})})^T\mathbf{y}_{H_{\mathbf{x}}+1:}}{2\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} + \frac{\log(|(A^{(L+1)})^{-1}|)}{2} \\
&+ \sum_{l=1}^{L}\left(-\frac{1}{2\left(\sigma_{\text{noise}}^{(l)}\right)^2}\left(\left(\sum_{i=L+1}^{N}\lambda_i^{(l)}\right) + (\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)})^T\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}\right)\right. \\
&+ \frac{(\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)})^T\Psi_1^{(l)}(A^{(l)})^{-1}(\Psi_1^{(l)})^T\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}}{2\left(\sigma_{\text{noise}}^{(l)}\right)^2} + \frac{\log(|(A^{(l)})^{-1}|)}{2} \\
&\left. -\frac{\hat{N}}{2} + \sum_{i=1+H_{\mathbf{x}}-H_{\text{h}}}^{N}\frac{\log(2\pi\lambda_i^{(l)})}{2} - \sum_{i=1+H_{\mathbf{x}}-H_{\text{h}}}^{H_{\text{h}}}\frac{\log(2\pi)}{2} - \frac{\left(\lambda_i^{(l)} + \left(\mu_i^{(l)}\right)^2\right)}{2}\right) \\
&- \frac{(L+1)\hat{N}}{2}\log(2\pi) - \sum_{l=1}^{L+1}\mathbf{KL}(q_{\boldsymbol{Z}^{(l)}}||p_{\boldsymbol{Z}^{(l)}}) \\
&\stackrel{\text{def}}{=} \boldsymbol{\mathcal{L}}_{\text{VSS-opt}}^{\text{REVARB}},
\end{aligned}
$$

where we derive $\boldsymbol{\mathcal{L}}_{\text{SS-opt}}^{\text{REVARB}}$ by being not variational over $\boldsymbol{Z}^{(l)}$.
The IP regularization case with $A^{(l)} = \Psi_2^{(l)} + (\sigma_{\text{noise}}^{(l)})^2 K_{MM}^{(l)}$ is given by:

$$
\begin{aligned}
&\geq -\frac{\hat{N}-M}{2}\sum_{l=1}^{L+1}\log\left(\left(\sigma_{\text{noise}}^{(l)}\right)^2\right) - \frac{\mathbf{y}_{H_{\mathbf{x}}+1:}^T\mathbf{y}_{H_{\mathbf{x}}+1:}}{2\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} + \frac{\log(|K_{MM}^{(L+1)}|)}{2} \\
&+ \frac{\mathbf{y}_{H_{\mathbf{x}}+1:}^T\Psi_1^{(\text{L+1})}(A^{(L+1)})^{-1}(\Psi_1^{(\text{L+1})})^T\mathbf{y}_{H_{\mathbf{x}}+1:}}{2\left(\sigma_{\text{noise}}^{(\text{L+1})}\right)^2} + \frac{\log(|(A^{(L+1)})^{-1}|)}{2} \\
&- \frac{\Psi_0^{(L+1)} + \text{tr}((K_{MM}^{(L+1)})^{-1}\Psi_{reg}^{(L+1)})}{2\sigma_{\text{noise}}^{(\text{L+1})}} \\
&- \sum_{l=1}^{L}\left(-\frac{1}{2\left(\sigma_{\text{noise}}^{(l)}\right)^2}\left(\left(\sum_{i=L+1}^{N}\lambda_i^{(l)}\right) + (\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)})^T\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}\right) + \frac{\log(|K_{MM}^{(l)}|)}{2}\right. \\
&+ \frac{(\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)})^T\Psi_1^{(l)}(A^{(l)})^{-1}(\Psi_1^{(l)})^T\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}}{2\left(\sigma_{\text{noise}}^{(l)}\right)^2} + \frac{\log(|(A^{(l)})^{-1}|)}{2} \\
&- \frac{\Psi_0^{(l)} + \text{tr}((K_{MM}^{(l)})^{-1}\Psi_{reg}^{(l)})}{2\sigma_{\text{noise}}^{(\text{l})}} - \frac{\hat{N}}{2} + \sum_{i=1+H_{\mathbf{x}}-H_{\text{h}}}^{N}\frac{\log(2\pi\lambda_i^{(l)})}{2} \\
&\left. - \sum_{i=1+H_{\mathbf{x}}-H_{\text{h}}}^{H_{\text{h}}}\frac{\log(2\pi)}{2} - \frac{\left(\lambda_i^{(l)} + \left(\mu_i^{(l)}\right)^2\right)}{2}\right) - \frac{(L+1)\hat{N}}{2}\log(2\pi) - \sum_{l=1}^{L+1}\mathbf{KL}(q_{\boldsymbol{Z}^{(l)}}||p_{\boldsymbol{Z}^{(l)}}) \\
&\stackrel{\text{def}}{=} \boldsymbol{\mathcal{L}}_{\text{VSS-IP}}^{\text{REVARB}},
\end{aligned}
$$

where again we derive $\boldsymbol{\mathcal{L}}_{\text{SS-IP}}^{\text{REVARB}}$ by being not variational over $\boldsymbol{Z}^{(l)}$.

6.3.5 PREDICTIONS

Predictions for each layer l and new $\hat{\mathbf{h}}_*^{(l)}$ are performed in the simple DRGP-(V)SS case with

$$\mathbf{E}\left[f^{*(l)}\right]_{q_{f^{*(l)}}} = \Psi_{1*}^{(l)}\mathbf{m}^{(l)},$$
$$\mathbf{V}\left[f^{*(l)}\right]_{q_{f^{*(l)}}} = \left(\mathbf{m}^{(l)}\right)^T\left(\Psi_{2*}^{(l)} - \left(\Psi_{1*}^{(l)}\right)^T\Psi_{1*}^{(l)}\right)\mathbf{m}^{(l)} + \mathrm{tr}\left(\mathbf{s}^{(l)}\Psi_{2*}^{(l)}\right),$$

where for the optimal distribution case for $\boldsymbol{a}^{(l)}$ we have with $A^{(l)} = \Psi_2^{(l)} + (\sigma_{\text{noise}}^{(l)})^2 I_M$

$$\mathbf{m}^{(l)} \overset{\text{opt}}{=} \left(A^{(l)}\right)^{-1}\left(\Psi_1^{(l)}\right)^T\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}, \quad \mathbf{s}^{(l)} \overset{\text{opt}}{=} (\sigma_{\text{noise}}^{(l)})^2\left(A^{(l)}\right)^{-1},$$

for $1,\dots,L$, and fully analog for $l = L+1$ by replacing $\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}$ with $\mathbf{y}_{H_{\mathbf{x}}+1:}$.
In the DRGP-(V)SS-IP case we make predictions for each layer l and new $\hat{\mathbf{h}}_*^{(l)}$ through

$$\mathbf{E}\left[f^{*(l)}\right]_{q_{f^{*(l)}}} = \Psi_{1*}^{(l)}\boldsymbol{\Lambda}^{(l)},$$
$$\mathbf{V}\left[f^{*(l)}\right]_{q_{f^{*(l)}}} = \left(\boldsymbol{\Lambda}^{(l)}\right)^T\left(\Psi_{2*}^{(l)} - \left(\Psi_{1*}^{(l)}\right)^T\Psi_{1*}^{(l)}\right)\boldsymbol{\Lambda}^{(l)}$$
$$+ \Psi_0 - \mathrm{tr}\left((K_{MM}^{(l)})^{-1}\Psi_{reg*}^{(l)} - (\sigma_{\text{noise}}^{(l)})^2\left(A^{(l)}\right)^{-1}\Psi_{2*}^{(l)}\right),$$

where $A^{(l)} = \Psi_2^{(l)} + (\sigma_{\text{noise}}^{(l)})^2 K_{MM}^{(l)}$ and

$$\boldsymbol{\Lambda}^{(l)} \overset{\text{opt}}{=} \left(A^{(l)}\right)^{-1}\left(\Psi_1^{(l)}\right)^T\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)},$$

for $1,\dots,L$, and fully analog for $l = L+1$ by replacing $\boldsymbol{\mu}_{H_{\mathbf{x}}+1:}^{(l)}$ with $\mathbf{y}_{H_{\mathbf{x}}+1:}$.