4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–17

Matrix Inversion free variational inference in Conditional Student's T Processes

Anonymous Authors
Anonymous Institution

1. Introduction

Gaussian processes (GP) are data-efficient Bayesian non-parametric models that offer calibrated uncertainty quantification and are robust to overfitting, recently finding applicability in data-scarce domains (Timonen et al., 2019; Wang et al., 2020) or where the uncertainty is of utmost importance (Chen et al., 2014) .Their drawback resides in the computational complexity of inverting the covariance matrix, which is cubic in computation and quadratic in memory. This has motivated research on sparse GP (SGP) methods (Seeger et al., 2003; Quinonero-Candela and Rasmussen, 2005). Titsias (2009) has addressed this problem by leaving the prior distribution of the GP unchanged, with sparsity being enforced in the posterior through inducing points learnt by variational inference. Hensman et al. (2013) proposed an inducing point framework scalable to large data, obtaining posterior formulas conditioned on these artificial points. However, this scales supralinearly with regards to inducing point numbers, resulting in $\mathcal{O}(M^2N + M^3)$ computation and $\mathcal{O}(MN + M^2)$ storage complexity, where M is the number of inducing points. Therefore, a major obstacle towards the wider adoption of GP in large scale datasets is related to the computational cost of matrix inversion and log determinants. With this in mind, in van der Wilk et al. (2020) the authors propose a lower bound that can be computed without computationally expensive matrix operations such as inversion. Similar in scope, we propose to learn variational approximations of the covariance matrix, implicitly also over inverses, of inducing and training points, thereby proposing a computationally efficient approximate posterior over covariance matrices within the probabilistic framework of Student's T Processes (STP) (Shah et al., 2014). Compared to van der Wilk et al. (2020), we propose a variational lower bound and show that it works on large scale datasets, a task which was not tackled in the latter work due to training instability.

2. Generative process of conditional sparse Student's T Processes

We follow the hierarchical Bayesian construction in prior space introduced in Shah et al. (2014) for the non-sparse scenario:

$$\Sigma_{ff} \sim W_n^{-1}\left(v_p + n + 1, v_p K_{ff}\right) \tag{1}$$

$$f \mid \Sigma_{ff} \sim \mathcal{N}\left(0, \Sigma_{ff}\right) \tag{2}$$

$$y \mid f \sim \mathcal{N}\left(f, \sigma^2 \mathbb{I}\right) \tag{3}$$

© A. Authors.

In Shah et al. (2014), the authors show that the latent function follows a $MVT(v_p, 0, K_{ff})$, where MVT stands for multivariate t-distribution, which is a generalization to random vectors of the Student's t-distribution. For completeness, we provide in Appendix A a more detailed proof of this statement.

We adapt the generative process to the sparse scenario. In order to achieve prior conditional matching in the Inverse Wishart prior over covariance matrices we add another layer in our Bayesian hierarchical generative process, namely:

$$T \sim W_m\left(v,\ \frac{1}{v}K_{uu}^{-1}\right) \tag{4}$$

$$\Sigma_{fu,fu} \mid T \sim W_{n+m}^{-1}\left(v_{f,p}+n,\ v_{f,p}\begin{pmatrix} T^{-1} & K_{uf} \\ K_{fu} & \overline{K_{ff\cdot u}} + K_{fu}TK_{uf}\end{pmatrix}\right) \tag{5}$$

$$U \mid \Sigma_{uu} \sim \mathcal{N}\left(0, \Sigma_{uu}\right) \tag{6}$$

$$f \mid U, \Sigma_{fu,fu}, T \sim \mathcal{N}\left(\Sigma_{fu}\Sigma_{uu}^{-1}U, \Sigma_{ff\cdot u}\right) \tag{7}$$

$$y \mid f \sim \mathcal{N}\left(f, \sigma^2\mathbb{I}\right) \tag{8}$$

where $v_{f,p} = v_p + m + 1$ for notation purposes. Details and motivation behind the change of scaling for mean covariance are given in Appendix B. We denote $\overline{K_{ff\cdot u}} = K_{ff} + K_{fu}TK_{uu}TK_{uf} - 2K_{fu}TK_{uf} \geq K_{ff\cdot u}$, with equality if and only if $T = K_{uu}^{-1}$ (Davies, 2015). As a consequence of Theorem 3 in Bodnar and Okhrin (2008), we can further express $\Sigma_{ff\cdot u} \mid T \sim W_n^{-1}\left(v_{f,p}+n,\ v_{f,p}\overline{K_{ff\cdot u}}\right)$ and $\Sigma_{uu}^{-1}\Sigma_{uf} \mid \Sigma_{ff\cdot u}, T \sim \mathcal{MN}\left(TK_{uf},\ \Sigma_{ff\cdot u} \otimes \frac{T}{v_{f,p}}\right)$, which are present in equation (7), where $\otimes$ denotes the Kronecker product and $\mathcal{MN}$ denotes the matrix variate distribution. Lastly, we notice that f no longer follows a $MVT(v_p, 0, K_{ff})$ distribution. Instead we have $f \mid T \sim MVT(v_p, 0, \overline{K_{ff\cdot u}} + K_{fu}TK_{uf})$, which converges to $MVT(v_p, 0, K_{ff})$ as $v \to \infty$ and $T \approx K_{uu}^{-1}$.

2.1. Variational Posteriors over Inverse-Wishart Processes

To perform inference over our hierarchical Bayesian model, we need a way of defining a valid variational mean covariance matrix for the posterior Inverse-Wishart Process over the space of possibly infinitely large covariance matrices: $q\left(\Sigma_{fu,fu}\right) = W_{n+m}^{-1}\left(v_{f,q}+n, v_{f,q}\tilde{K_{fu,fu}}\right)$ where $\tilde{K_{fu,fu}} \succ 0$ positive definite matrix defined as $\tilde{K_{fu,fu}} = \begin{pmatrix} T^{-1} & K_{uf} \\ K_{fu} & \overline{K_{ff\cdot u}} + K_{fu}TK_{uf}\end{pmatrix}$. To see this is true, we make use of the following proposition.

Proposition 1 (Proposition 2.1 in Gallier et al. (2010)) *For any symmetric matrix M of the form: $M = \begin{pmatrix} A & B \\ B^T & C\end{pmatrix}$ and if C is invertible then $M \succ 0 \iff A \succ 0$ and $A - BC^{-1}B^T \succ 0$.*

We note that $T^{-1} \succ 0$ and $\overline{K_{ff\cdot u}} \succ 0$ by construction, thereby $\tilde{K_{fu,fu}} \succ 0$, so that it satisfies the condition for defining mean covariance matrices for Inverse Wishart distributions.

3. Variational lower bound of conditional inverse free sparse Student's T Processes

We defer the full derivation of the ELBO to Appendix C.

$$= \mathbb{E}_{Q(F,U,\Sigma_{fu,fu},T)} \log p(Y|F,U,\Sigma_{fu,fu},T) - \int q(F,U,\Sigma_{fu,fu},T) \log \frac{q(F,U,\Sigma_{fu,fu},T)}{p(F,U,\Sigma_{fu,fu},T)} dF\ dU\ d\Sigma_{fu,fu}\ dT \tag{9}$$

where we define constituent terms of the factorized approximate posterior as follows:

$$q(T) = W_m\left(v_q,\ \frac{1}{v_q} L_{\tilde{K}_{uu}^{-1}} L_{\tilde{K}_{uu}^{-1}}^T\right) \tag{10}$$

$$q(\Sigma_{fu,fu} \mid T) = W_{n+m}^{-1}\left(v_{f,q}+n,\ v_{f,q}\begin{pmatrix} T^{-1} & K_{uf} \\ K_{fu} & \overline{K_{ff\cdot u}} + K_{fu} T K_{uf}\end{pmatrix}\right) \tag{11}$$

$$q(U) = \mathcal{N}\left(m, L_S L_S^T\right) \tag{12}$$

$$q(f \mid U, \Sigma_{fu,fu}) = \mathcal{N}\left(\Sigma_{fu}\Sigma_{uu}^{-1}U, \Sigma_{ff\cdot u}\right) \tag{13}$$

Expectation over data fit term in SVI setting We can re-express expectations over individual terms as follows:

$$\sum_{k=1}^{n} E_{q(\Sigma_{f_k u, f_k u})}\left[\int \log p(Y_k|F_k) q(F_k; \tilde{U}(\Sigma_{f_k u, f_k u}), \tilde{\Sigma}(\Sigma_{f_k u, f_k u}))\ dF_k\right]$$

where $\tilde{U}(\Sigma_{f_k u, f_k u}) = \Sigma_{f_k u}\Sigma_{uu}^{-1} m$ and $\tilde{\Sigma}(\Sigma_{f_k u, f_k u}) = \Sigma_{f_k f_k} - \Sigma_{f_k u}\Sigma_{uu}^{-1}\left[\Sigma_{uu} - S\right]\Sigma_{uu}^{-1}\Sigma_{u f_k}$ define distributions over scalars. Hence, the aforementioned multivariate formulas can now be rewritten as:

$$\Sigma_{f_k f_k \cdot u} \mid T \sim IG\left(\frac{1}{2}\left(v_{f,q}+1\right),\ \frac{v_{f,q}}{2}\overline{K_{f_k f_k \cdot u}}\right) \tag{14}$$

$$\Sigma_{uu}^{-1}\Sigma_{u f_k} \mid \Sigma_{f_k f_k \cdot u}, T \sim \mathcal{N}\left(T K_{u f_k},\ \frac{\Sigma_{f_k f_k \cdot u}}{v_{f,q}} \otimes T\right) \tag{15}$$

where IG represents in the inverse gamma distribution.

Kullback-Leibler Divergences In the KL term present in our lower bound we can separate the fraction terms that with respect to integrants that they exclusively depend on:

$$\mathbb{E}_{q(F,U,\Sigma_{fu,fu},T)} \log \frac{q(F|U,\Sigma_{fu,fu},T)}{p(F|U,\Sigma_{fu,fu},T)} + \mathbb{E}_{q(\Sigma_{uu}^{-1},T)} \int q(U) \log \frac{q(U)}{p(U|\Sigma_{uu}^{-1})} dU + \mathbb{E}_{q(T)}\mathbb{E}_{q(\Sigma_{fu,fu}|T)} \log \frac{q(\Sigma_{fu,fu} \mid T)}{p(\Sigma_{fu,fu} \mid T)}) + \mathbb{E}_{q(T)} \log \frac{q(T)}{p(T)} \tag{16}$$

where the first and third terms cancel out due to conditional prior matching.

KL-divergence between inducing points The term can be written as:

$$\mathbb{E}_{q(\Sigma_{uu}^{-1},T)} \int q(U) \log \frac{q(U)}{p(U|\Sigma_{uu}^{-1})} dU \tag{17}$$

$$\mathbb{E}_{q(T)} \mathbb{E}_{q(\Sigma_{uu}^{-1}|T)} \frac{1}{2} \left[-\log |\Sigma_{uu}^{-1}| - \log |S| - d + \mathbf{Tr}(\Sigma_{uu}^{-1} S) + m^T \Sigma_{uu}^{-1} m \right] \tag{18}$$

We remind ourselves that we can obtain samples from Σ_{uu}^{-1} via following Bayesian hierarchical construction:

$$q\left(T\right) = W_m\left(v_q,\ \frac{1}{v_q} \tilde{L_{K_{uu}^{-1}}} \tilde{L_{K_{uu}^{-1}}}^T\right); \qquad q\left(\Sigma_{uu}^{-1} \mid T\right) = W_m\left(v_{f,q},\ \frac{1}{v_{f,q}} T\right) \tag{19}$$

No matrix inversion is required for this divergence term. Moreover, T is obtained via the Bartlett decomposition $\tilde{L_{K_{uu}^{-1}}} A_T A_T^T \tilde{L^T_{K_{uu}^{-1}}}$, where A_T is a lower triangular matrix defined in Appendix D. Then, we can apply again the Bartlett decomposition to obtain samples from Σ_{uu}^{-1} by noticing that $\tilde{L_{K_{uu}^{-1}}} A_T$ represents a lower triangular matrix. Finally, samples are obtained via $\tilde{L_{K_{uu}^{-1}}} A_T A_{\Sigma_{uu}^{-1}} A_{\Sigma_{uu}^{-1}}^T A_T^T \tilde{L^T_{K_{uu}^{-1}}}$ Hence, Σ_{uu}^{-1} is a product of lower triangular matrices $L_{K_{uu}^{-1}} A_T A_{\Sigma_{uu}^{-1}}$, with the determinant $\log |\Sigma_{uu}^{-1}|$ having an analytic formula.

KL-divergence between inverse inducing points' covariance matrices We now focus on $KL\left[q(T) \mid p(T)\right]$, which has an analytic formula:

$$-\frac{v_p}{2}\left[\log \mid v_p K_{uu} \mid + \log \mid \frac{1}{v_q} \tilde{K_{uu}^{-1}} \mid\right] + \frac{v_q}{2}\left[Tr\left[\frac{v_p}{v_q} K_{uu} \tilde{K_{uu}^{-1}}\right] - m\right] + \log \frac{\Gamma_m(v_p/2)}{\Gamma_m(v_q/2)} + \frac{v_q - v_p}{2} \psi_m(\frac{v_q}{2}) \tag{20}$$

where ψ_m denotes the multivariate digamma function (the derivative of the log of the multivariate gamma function). From the above equation we can notice that no matrix inverses are required in the computation of the KL divergence. Directly computing $\log \mid K_{uu} \mid$ would incur an $\mathcal{O}\left(m^3\right)$ computational cost. To bypass it, we use the following proposition.

Proposition 2 *For a p.s.d. matrix $K \in M_{n\times n}$, x being a conjugate gradient solution to $Kx = g$, where $g \in M_{n\times m}$ are independent standard normal samples and $\tilde{K} \succ 0$ we have the following lower bound on the log determinant of K:*

$$log \mid K \mid \geq n + \frac{1}{m} \sum_{i=1}^{m} -g_i^T \tilde{K} x_i + log \mid \tilde{K} \mid \tag{21}$$

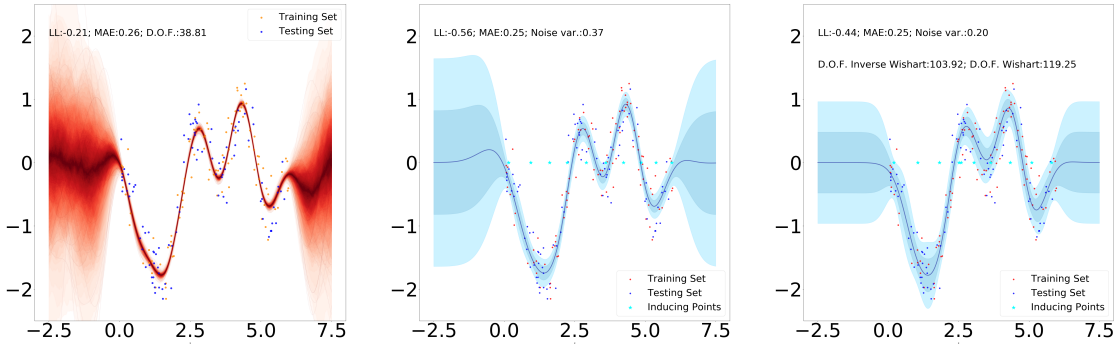
with the bound being tight if and only if $\tilde{K} \approx K$.

A proof of this proposition can be found in Appendix E. Whereas in current work we use standard preconditioned conjugate gradients, one could use the Unbiased LInear Systems SolvEr (ULISSE) Filippone and Engler (2015), a randomly truncated CG run, to compute unbiased estimate of x_i.

Remark 3 *Translating Proposition 2 to our case, we have $log \mid K_{uu} \mid \geq n + \frac{1}{m} \sum_{i=1}^{m} -g_i^T T^{-1} x_i - log \mid T \mid$. In the computation of x_i we precondition with T. Hence if $T \approx K_{uu}^{-1}$ the CG routine will finish in one step. We empirically prove on UCI datasets that this method is close to $\mathcal{O}\left(m^2\right)$ (Appendix F.4).*

4. Experiments

Convergence to Matrix Inversion based counterparts We are interested to find out whether our proposed model (denoted as IF-CSTP) is capable to recover the testing time prediction behaviour of Student's T Processes (STP) (Shah et al., 2014) and of SVGP (Hensman et al., 2013).



(a) Student's T Process (b) SVGP (c) IF-CSTP

Figure 1: Predictive mean and distributional variance for models trained on "snelson" dataset. Likelihood variance is not added.

IF-STP is capable of almost recovering the testing time predictive behaviour of both STP and SVGP on the "snelson" dataset (Snelson and Ghahramani, 2006) (Figure 1).

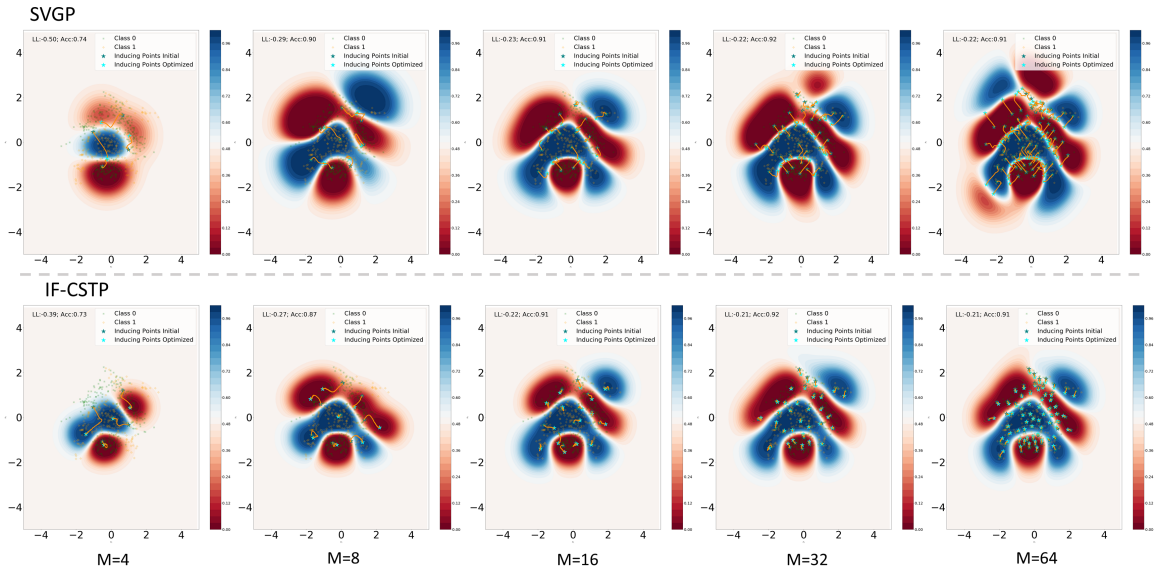


Figure 2: Effects on predictive mean of increasing number of inducing points for the "banana" dataset. With dark orange we plot the optimization trajectory of Z.

Increasing the number of inducing points We train IF-CSTP and SVGP for varying number of inducing points on 'banana" dataset (Figure 2). Whereas SVGP brings the inducing points' locations Z exactly on the classes' demarcation lines, IF-CSTP does not exhibit similar behaviour.

Initial results on large-scale datasets We evaluate our proposed model on selected UCI datasets (Figure 3). The gap between ELBOs is predominantly caused by the KL-divergence pertaining to T (eqn. 20). IF-CSTP's testing time behaviour almost approaches SVGP. However, in some cases, such as on "Protein" for 250 inducing points, the gap widens.

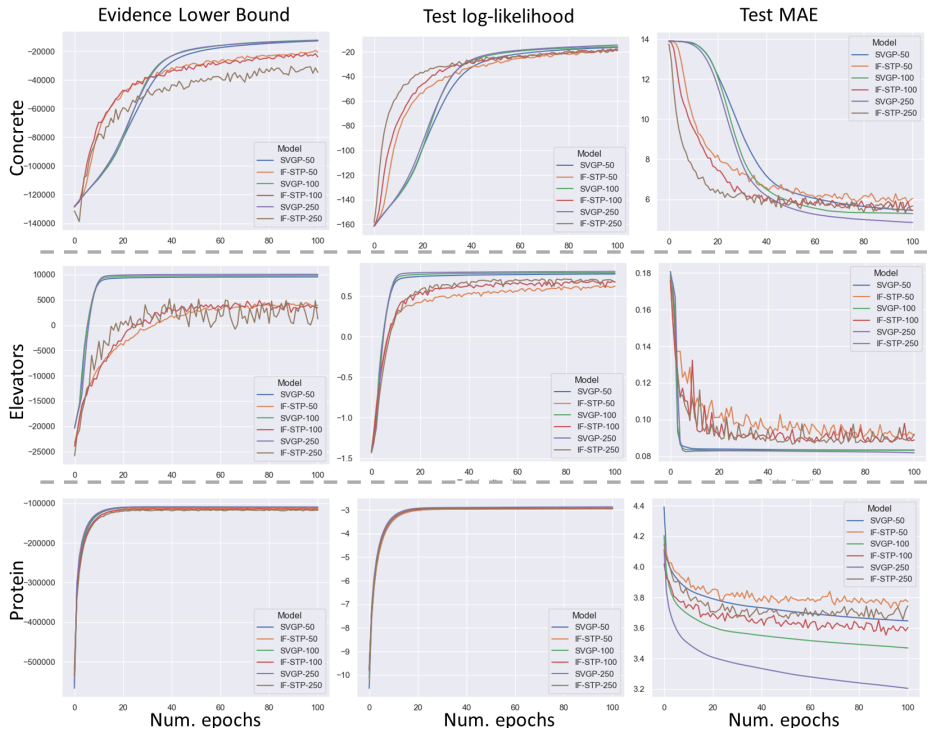


Figure 3: ELBO, Testing log-likelihood and mean absolute error values during optimization for varying inducing points numbers.

5. Discussion

Using our matrix inversion-free variational lower bound we showed similar behaviour to matrix inversion-dependent counterparts on toy tasks. However, scaling our proposed method to work on large-scale data still remains a research avenue as there is a tendency of over-stability of Z for large numbers of inducing points due to large penalty stemming from $KL\left[q(T) \mid p(T)\right]$. Lastly, low tolerance for the CG-based lower bound on log determinant does not guarantee that our objective remains a lower bound. Nevertheless, we have shown empirically that this does not constitute a problem in practice.

References

Taras Bodnar and Yarema Okhrin. Properties of the singular, inverse and generalized inverse partitioned wishart distributions. *Journal of Multivariate Analysis*, 99(10):2389–2405, 2008.

Tongtong Chen, Bin Dai, Ruili Wang, and Daxue Liu. Gaussian-process-based real-time ground segmentation for autonomous land vehicles. *Journal of Intelligent & Robotic Systems*, 76(3):563–582, 2014.

Alexander James Davies. *Effective implementation of Gaussian process regression for machine learning*. PhD thesis, University of Cambridge, 2015.

Maurizio Filippone and Raphael Engler. Enabling scalable stochastic gradient-based inference for gaussian processes by employing the unbiased linear system solver (ulisse). In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1015–1024, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/filippone15.html.

Jean Gallier et al. The schur complement and symmetric positive semidefinite (and definite) matrices. *Penn Engineering*, pages 1–12, 2010.

Didier Girard. *Un algorithme simple et rapide pour la validation croisée généralisée sur des problèmes de grande taille*. 1987.

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

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005.

Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse gaussian process regression. Technical report, 2003.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. In *Artificial intelligence and statistics*, pages 877–885, 2014.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

Juho Timonen, Henrik Mannerström, Aki Vehtari, and Harri Lähdesmäki. lgpr: An interpretable nonparametric method for inferring covariate effects from longitudinal data. *arXiv preprint arXiv:1912.03549*, 2019.

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

Mark van der Wilk, ST John, Artem Artemev, and James Hensman. Variational gaussian process models without matrix inverses. In Cheng Zhang, Francisco Ruiz, Thang Bui, Adji Bousso Dieng, and Dawen Liang, editors, *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–9. PMLR, 08 Dec 2020. URL http://proceedings.mlr.press//v118//wilk20a.html.

Dennis Wang, James Hensman, Ginte Kutkaite, Tzen S Toh, Ana Galhoz, Jonathan R Dry, Julio Saez-Rodriguez, Mathew J Garnett, Michael P Menden, Frank Dondelinger, et al. A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *Elife*, 9:e60352, 2020.

Appendix A. LATENT FUNCTIONS F ARE STUDENT-T DISTRIBUTED

In Shah et al. (2014), the authors provide a derivation showing that by marginalizing out Σ_{zz} one obtains a multivariate t-distribution. We provide here a more detailed derivation taking into consideration that we have the general result:

$$\int p(y|0,\Sigma)p(\Sigma|\mathbf{v},\psi)d\ \Sigma = \frac{|\psi|^{\mathbf{v}/2}\gamma_p(\frac{\mathbf{v}+n}{2})}{\pi^{np/2}|\psi+A|^{\frac{\mathbf{v}+n}{2}}\gamma_p(\frac{\mathbf{v}}{2})} \tag{22}$$

where $A = XX^T$. Adapting this result to our scenario we get that:

$$p(K_{zz}) = \frac{|\mathbf{v}K_{zz}|^{\mathbf{v}/2}\gamma_p(\frac{\mathbf{v}+m}{2})}{\pi^{m^2/2}|vK_{zz}+ZZ^T|^{\frac{\mathbf{v}+m}{2}}\gamma_p(\frac{\mathbf{v}}{2})} \tag{23}$$

Focusing on the determinant terms we can rearrange:

$$\frac{|\mathbf{v}K_{zz}|^{\frac{\mathbf{v}}{2}}}{|vK_{zz}+ZZ^T|^{\frac{\mathbf{v}+m}{2}}} = \frac{|vK_{zz}|^{-\frac{m}{2}}}{|\mathcal{I}+\frac{1}{\mathbf{v}}K_{zz}^{-1}ZZ^T|^{\frac{\mathbf{v}+m}{2}}} \tag{24}$$

We remind the matrix determinant lemma : $|A+uv^T| = (\mathcal{I}+v^TA^{-1}u)|A|$. We apply it to the denonimator of the previous equation:

$$|\mathcal{I}+\frac{1}{\mathbf{v}}K_{zz}^{-1}ZZ^T|^{\frac{\mathbf{v}+m}{2}} = \left[1+\frac{1}{\mathbf{v}}ZK_{zz}^{-1}Z^T\right]^{\frac{\mathbf{v}+m}{2}} \tag{25}$$

Plugging this into equation 23 we get that:

$$\frac{\gamma_p(\frac{\mathbf{v}+m}{2})}{\pi^{m^2/2}\gamma_p(\frac{\mathbf{v}}{2})v^{\frac{m^2}{2}}|K_{zz}|^{\frac{m}{1}2}}\left[1+\frac{1}{\mathbf{v}}Z^TK_{zz}^{-1}Z\right]^{-\frac{\mathbf{v}+m}{2}} \tag{26}$$

which we can now recognize as a $MVT(\mathbf{v},0,K_{zz})$, where MVT stands for multivariate t-distribution, which is a generalization to random vectors of the Student's t-distribution. The p.d.f. of this distribution is defined as $\frac{\Gamma(\frac{\mathbf{v}+m}{2})}{\Gamma(\frac{\mathbf{v}}{2})\mathbf{v}^{m/2}\pi^{m/2}|\Sigma|^{1/2})}\left[1+\frac{1}{\mathbf{v}}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]^{-(\mathbf{v}+m)/2}$, where μ is the location, Σ is a positive-definite $m*m$ matrix and $\mathbf{v}$ represent the degrees of freedom. For $\mathbf{v}\geq 1$ the mean is μ and for $\mathbf{v}\geq 2$ the variance is defined as $\frac{\mathbf{v}}{\mathbf{v}-2}\Sigma$. Therefore, by starting from a Inverse-Wishart distributed prior over the covariance of a GP evaluated at X, we have implicitly imposed a Student-t distribution over our latent function.

Appendix B. Properties of different mean covariance matrix scalings

Under the parametrization introduced in Shah et al. (2014), respectively:

$$\Sigma_{fu,fu} \sim W_{n+m}^{-1}\left(v_p+n+1, v_pK_{fu,fu}\right) \tag{27}$$

which implies the following probabilistic formulas and their expectations:

$$\Sigma_{uu} \sim W_m^{-1}\left(v_p + m + 1,\ v_p K_{uu}\right);\ \mathbb{E}\left[\Sigma_{uu}\right] = K_{uu} \tag{28}$$

$$\Sigma_{uu}^{-1} \sim W_m\left(v_p + m + 1,\ \frac{1}{v_p} K_{uu}^{-1}\right);\ \mathbb{E}\left[\Sigma_{uu}^{-1}\right] = \frac{m + 1 + v_p}{v_p} K_{uu}^{-1} \tag{29}$$

$$\Sigma_{f_k f_k \cdot u} \sim IG\left(\frac{1}{2}\left(v_p + m + 2\right),\ \frac{v_p}{2} K_{f_k f_k \cdot u}\right);\ \mathbb{E}\left[\Sigma_{f_k f_k \cdot u}\right] = \frac{v_p}{v_p + m} K_{f_k f_k \cdot u} \tag{30}$$

We can notice that for increasingly larger number of inducing points, the expectations over Σ_{uu}^{-1} and $\Sigma_{ff \cdot u}$ get increasingly upscaled, respectively downscaled with respect to their true solution. For $v_p \to \infty$, the expectations will converge towards the real solution and their variance will collapse to zero.

Under the parametrization introduced in section 2:

$$\Sigma_{fu,fu} \sim W_{n+m}^{-1}\left(v_p + m + n + 1, \left(v_p + m + 1\right) K_{fu,fu}\right) \tag{31}$$

we have:

$$\Sigma_{uu} \sim W_m^{-1}\left(v_p + m + 1,\ \left(v_p + m + 1\right) K_{uu}\right);\ \mathbb{E}\left[\Sigma_{uu}\right] = \frac{v_p + m + 1}{v_p} K_{uu} \tag{32}$$

$$\Sigma_{uu}^{-1} \sim W_m\left(v_p + m + 1,\ \frac{1}{v_p + m + 1} K_{uu}\right);\ \mathbb{E}\left[\Sigma_{uu}^{-1}\right] = K_{uu}^{-1} \tag{33}$$

$$\Sigma_{f_k f_k \cdot u} \sim IG\left(\frac{1}{2}\left(v_p + m + 2\right),\ \frac{v_p + m + 1}{2} K_{f_k f_k \cdot u}\right);\ \mathbb{E}\left[\Sigma_{f_k f_k \cdot u}\right] = \frac{v_p + m + 1}{v_p + m} K_{\tilde{f_k f_k} \cdot u} \tag{34}$$

In stark contrast to the parameterization introduced in Shah et al. (2014), we now have an upscaling of expectations over covariance matrices, which only reverts to the real solution for very large values of v_p. However, in the interest of obtaining predictions at testing time from our sparse Student's T Process, we are only interested in obtaining unbiased samples from Σ_{uu}^{-1} and $\Sigma_{f_k f_k \cdot u}$, for which we can notice that our scaling factor is more appropiate.

Appendix C. Derivation of evidence lower bound

We consider our training data $D = \{X_i, Y_i\}_{i=1,\cdots,n}$. Our goal is to approximate the true posterior distribution $p(F, U, \Sigma_{fu,fu}, T|D)$ by utilising the following approximate posterior $q(F, U, \Sigma_{fu,fu}, T) = p(F|U, \Sigma_{fu,fu}, T)q(U)q(\Sigma_{fu,fu} \mid T)q(T)$. Our minimization goal can be expressed as:

$$= \int q(F, U, \Sigma_{fu,fu}, T) \log \frac{q(F, U, \Sigma_{fu,fu}, T)}{p(F, U, \Sigma_{fu,fu}, T|D)} dF\ dU\ d\Sigma_{fu,fu}\ dT \tag{35}$$

$$= \int q(F, U, \Sigma_{fu,fu}, T) \left[\log \frac{q(F, U, \Sigma_{fu,fu}, T)}{p(F, U, \Sigma_{fu,fu}, T, D)} + \log p(D)\right] dF\ dU\ d\Sigma_{fu,fu}\ dT \tag{36}$$

$$= \mathbb{E}_{q(F,U,\Sigma_{fu,fu},T)} \left[\log \frac{q(F, U, \Sigma_{fu,fu}, T)}{p(D|F, U, \Sigma_{fu,fu}, T)p(F, U, \Sigma_{fu,fu}, T)} + \log p(D)\right] \tag{37}$$

which after some rearrangements becomes:

$$\log p(D) - \Big[\mathbb{E}_{q(F,U,\Sigma_{fu,fu},T)} \log p(D|F,U,\Sigma_{fu,fu},T) - KL\left[q(F,U,\Sigma_{fu,fu},T) \mid p(F,U,\Sigma_{fu,fu},T)\right]\Big] \tag{38}$$

We can now re-express this derivation as:

$$\begin{aligned}\log p(D) - KL\left[q(F,U,\Sigma_{fu,fu},T \mid p(F,U,\Sigma_{fu,fu},T))\right] = \\ \mathbb{E}_{q(F,U,\Sigma_{fu,fu},T)} \log p(D|F,U,\Sigma_{fu,fu},T) - KL\left[q(F,U,\Sigma_{fu,fu},T) \mid p(F,U,\Sigma_{fu,fu},T)\right]\end{aligned} \tag{39}$$

Hence, minimizing the intractable $KL\left[q(F,U,\Sigma_{fu,fu},T) \mid p(F,U,\Sigma_{fu,fu},T|D)\right]$ translates to maximizing a tractable lower bound on the log marginal likelihood, respectively the r.h.s. of the previous equation, which can be expressed as follows:

$$\begin{aligned}= \mathbb{E}_{Q(F,U,\Sigma_{fu,fu},T)} \log p(Y|F,U,\Sigma_{fu,fu},T) - \\ \int q(F,U,\Sigma_{fu,fu},T) \log \frac{q(F,U,\Sigma_{fu,fu},T)}{p(F,U,\Sigma_{fu,fu},T)} dF\ dU\ d\Sigma_{fu,fu}\ dT\end{aligned} \tag{40}$$

Appendix D. Sampling Wishart Distributions

For this operation we use the Bartlett decomposition of the matrix $R \sim W_p(v, \Sigma)$, which can be written in the following factorized manner $K_{zz}^{-1} = LAA^TL^T$, where A has the following matrix form $A = \begin{pmatrix} c_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ n_{p1} & \dots & c_p \end{pmatrix}$, where $n_{ij} \sim \mathcal{N}(0,1)$ and $c_i^2 \sim \chi^2_{v-i+1}$ is the chi-squared distribution and $LL^T = \Sigma$.

Appendix E. Proof of Proposition 1

We start with a lower bound on $\log \mid K \mid$, which can be expressed as:

$$log \mid K \mid \geq Tr\left[\mathbb{I} - K^{-1}\right] \tag{41}$$

Hutchinson (1989) introduced a stochastic estimation method for computing matrix traces. It is defined as follows:

$$H_m(A) = \frac{1}{m} \sum_{i=1}^{m} g_i^T A g_i \tag{42}$$

which was shown to converge towards $Tr(A)$ as $m \to \infty$, where $A \in M_{n\times n}$. Earlier work by Girard (1987) suggests taking $g_i \in M_{n\times 1}$ with individual elements of the column matrix stemming from standard normal, whereas Hutchinson (1989) advocated for taking the individual elements as Rademacher random variables, which translates into equally probable $\{-1, 1\}$ splits.

We can rewrite our objective as follows $\log \mid K \mid = \log \mid K\tilde{K}^{-1} \mid + \log \mid \tilde{K} \mid$ for any arbitrary p.s.d. matrix $\tilde{K}$. We are interested in obtaining reliable estimates of the following

bound without performing the matrix inversion operation:

$$log \mid K\tilde{K}^{-1} \mid + log \mid \tilde{K} \mid \geq Tr\left[\mathbb{I}\right] - Tr\left[\left(K\tilde{K}^{-1}\right)^{-1}\right] + log \mid \tilde{K} \mid \tag{43}$$

$$\geq Tr\left[\mathbb{I}\right] - \frac{1}{m}\sum_{i=1}^{m} g_i^T \left(K\tilde{K}^{-1}\right)^{-1} g_i + log \mid \tilde{K_{fu,fu}} \mid \tag{44}$$

Rather than obtaining lower/upper bounds on the part in involving matrix inversion, we can reliably estimate the solution via conjugate gradients. The solution will converge in one iteration if the condition number is close to 1, which will happen if $\tilde{K}^{-1}$ is an accurate estimate of K^{-1}.

We briefly summarize the theory behind conjugate gradients and how to compute them. We are interested in solution to $Kx = y$, where the optimal solution $x^{opt} = K^{-1}y$, where $K \in M_{n\times n}$ and $y \in M_{n\times 1}$. Conjugate gradients start with an initial solution x_0, which progressively gets refined and is guaranteed to converge in the worst case scenario in n iterations. We provide the algorithm in pseudocode 1.

Algorithm 1: Preconditioned Conjugate Gradient routine

Input: $K \in M_{n\times n}, y \in M_{n\times 1}$

Preconditioning matrix: $M \in M_{n\times n}$

Initial estimate: $x_0 \in M_{n\times 1}$

Compute first search direction and basis

$r_0 = Kx_0 - y$

$z_0 = M^{-1}r_0$

$p_0 = z_0$

1. For $i = 1$ to n
   (a) **Alpha:** $\alpha_i = \frac{r_i^T r_i}{p_i^T K p_i}$
   (b) **Solution update:** $x_{i+1} = x_i + \alpha_i p_i$
   (c) if i=n; return x_i
   (d) $z_{i+1} = M^{-1}r_{i-1}$
   (e) Beta: $\beta_i = \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$
   (f) Search direction update: $p_{i+1} = z_{i+1} - \beta_i p_i$

For the purposes of our problem at hand, we have $y = g_i$ and the CG-based solution $x_i \approx K^{-1}g_i$. Then, the lower bound translates to:

$$log \mid K \mid \geq Tr\left[\mathbb{I}\right] - \frac{1}{m}\sum_{i=1}^{m} g_i^T \tilde{K} x_i + log \mid \tilde{K} \mid \tag{45}$$

The tightness of the lower bound in the case $\tilde{K} \approx K$ can be shown as follows:

$$log \mid K \mid \geq Tr\left[\mathbb{I}\right] - \frac{1}{m}\sum_{i=1}^{m} g_i^T \tilde{K} x_i + log \mid \tilde{K} \mid \tag{46}$$

$$log \mid K \mid \geq Tr\left[\mathbb{I}\right] - \frac{1}{m}\sum_{i=1}^{m} g_i^T \tilde{K} K^{-1} g_i + log \mid \tilde{K} \mid \tag{47}$$

$$log \mid K \mid \geq Tr\left[\mathbb{I}\right] - Tr\left[\mathbb{I}\right] + log \mid \tilde{K} \mid \tag{48}$$

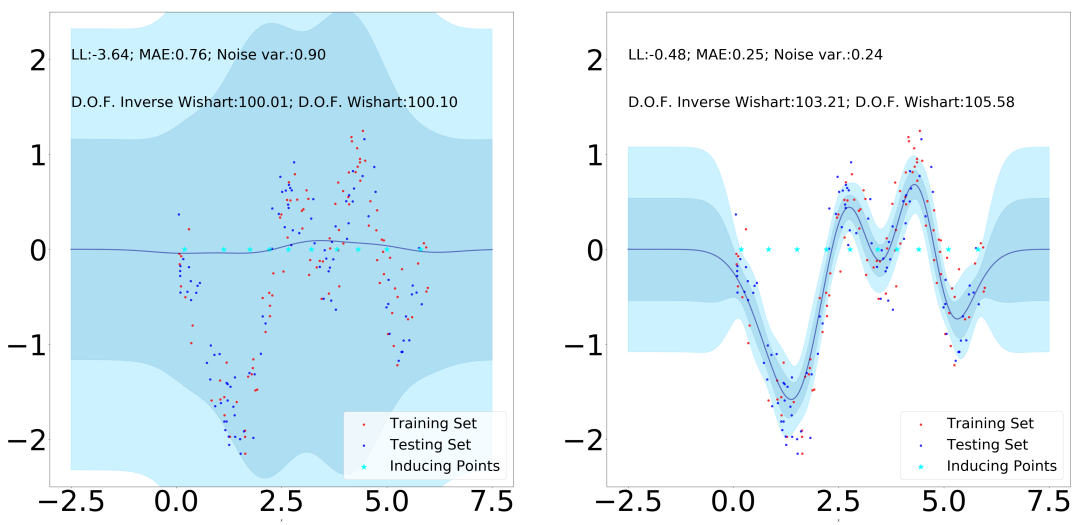
$$log \mid K \mid = log \mid \tilde{K} \mid \tag{49}$$

where we used the fact that $x_i = K^{-1} g_i$ since the condition number is 1 in this scenario.

Appendix F. Additional results

F.1. Lower bound recovers true approximate posteriors despite erroneous initialization

We aim to quantify the influence of "accurate" initializations of variational parameters on the converged solution. Whereas one would expect to initialize $L_{\Sigma_{uu}^{-1}}$ with their respective values given initial estimates of Z and kernel hyperparameters θ, we instead initialize them with identity matrices to determine if the bound can recover optimal posteriors. The lower



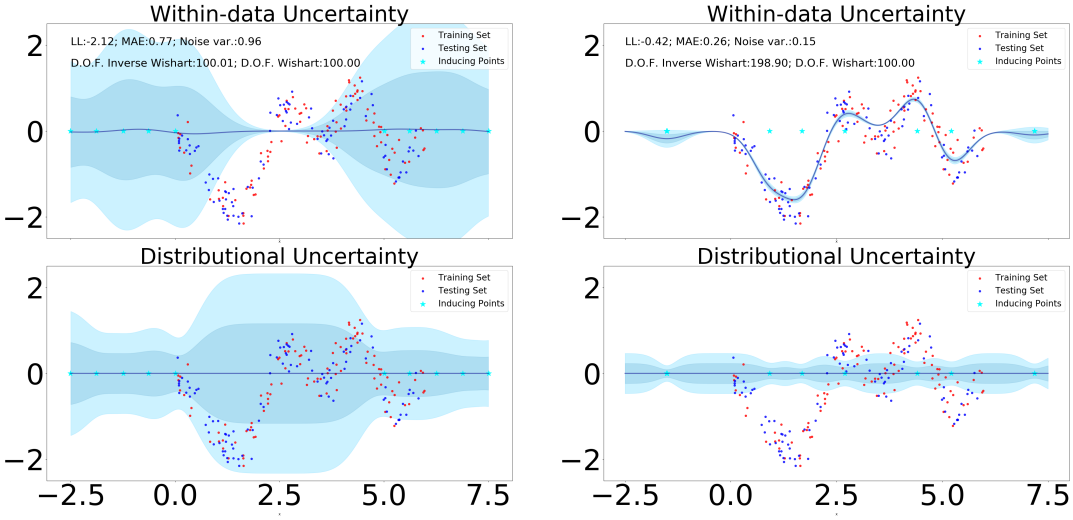
(a) Initial (b) Optimized

Figure 4: Predictive mean and variance plots at initialization and after optimization. Variational parameters are badly initialized.

bound manages to recover an accurate estimate of $\tilde{K_{fu,fu}} \approx K_{fu,fu}$.

The influence of the initial values of v_q is of paramount importance as it can make any change to $K_{fu,fu}$ impossible to take, possibly leading to scenarios where the optimization prefers to keep Z and kernel hyperparamters fixed. We explore this scenario and see if

the inducing points' locations Z will get optimized to ideal locations despite starting from values outside the training set range. From Figure 5 we can easily notice that the optimized Z values are brought back to locations similar to ones obtained from running k-means.



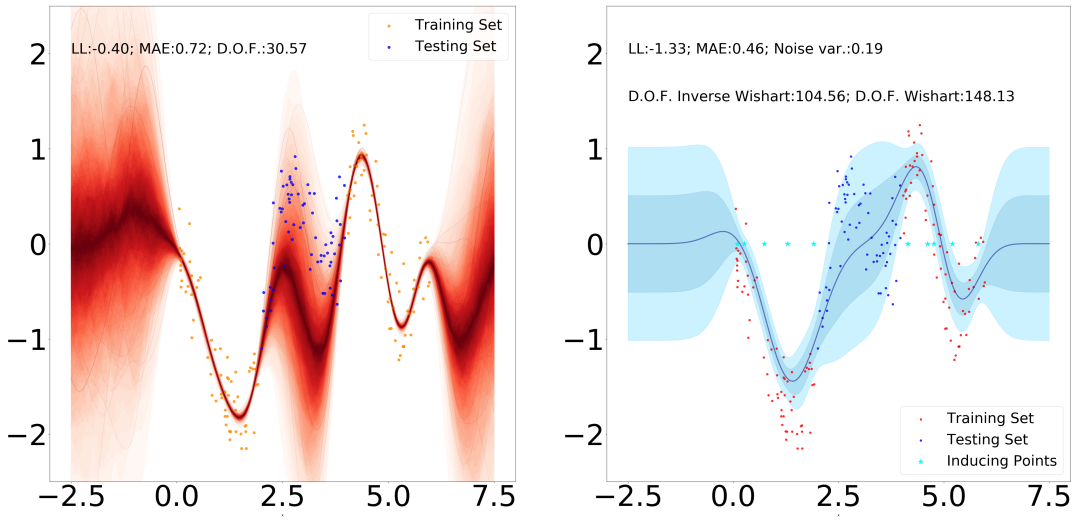
(a) Initial (b) Optimized

Figure 5: Predictive mean and variance plots at initialization and after optimization, split in parametric and non-parametric components. Inducing points' location is badly initialized.

F.2. Reliability of uncertainty estimates

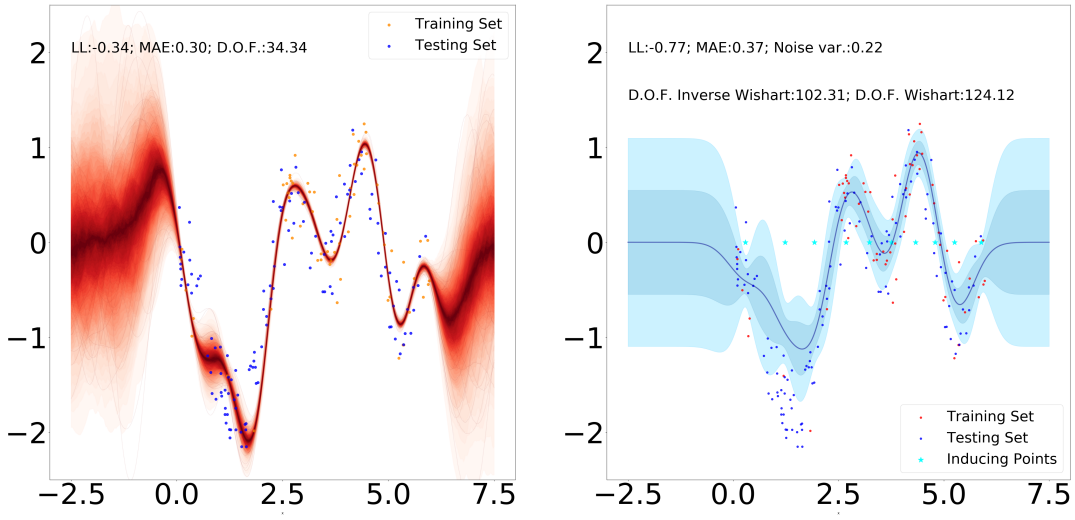
We are interested to stress test our proposed model in three different scenarios, which will individually point towards biases or discrepancies in the final estimates for $\Sigma_{ff\cdot u}$ and $\Sigma_{fu}\Sigma_{uu}^{-1}$. For this we use the "snelson" dataset, with the training set taken to comprise the intervals between 0.0 and 2.0, respectively 4.0 and 6.5. Thereby, in an ideal scenario we would expect our model to offer high uncertainty estimates between 2.0 and 4.0. From Figure 6 we can notice that indeed IF-STP does not suffer from pathologies.

Subsampling the "snelson" dataset allows us to discern if our method is capable to increase its uncertainty. For this we only subsample data between 0.0 and 2.5. Naturally, one should expect an increase in uncertainty just in that interval, with the remainder of the data range reverting back to levels seen for the full "snelson" dataset (Figure 7).



(a) Student's T Process (b) IF-CSTP

Figure 6: Predictive mean and distributional variance for models trained on "snelson with gap" dataset. Likelihood variance is not added.



(a) Student's T Process (b) IF-CSTP

Figure 7: Predictive mean and distributional variance for models trained on locally subsampled "snelson" dataset. Likelihood variance is not added.

F.3. Additional figures for "snelson" and "banana"

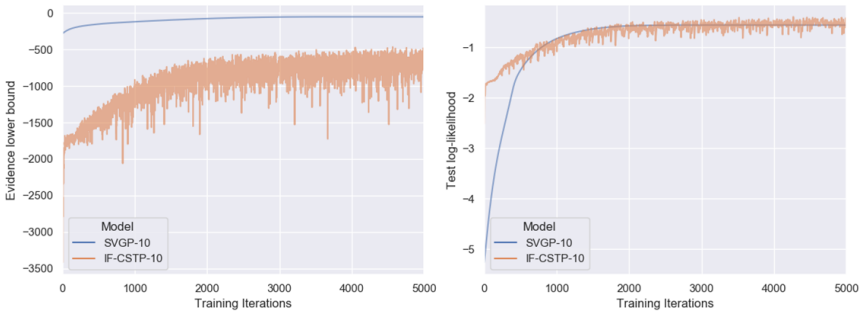


Figure 8: ELBO and Testing log-likelihood values during optimization for 10 inducing points on "snelson".

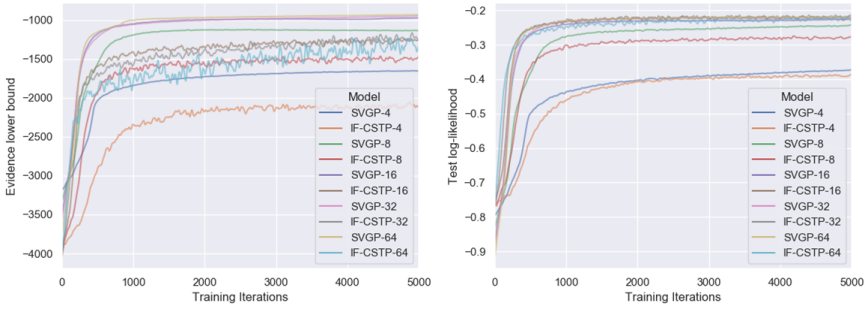


Figure 9: ELBO and Testing log-likelihood values during optimization for varying inducing points on "banana".

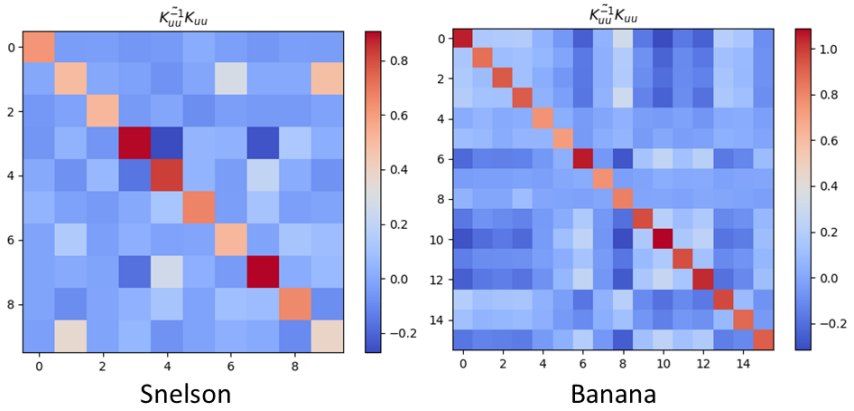


Figure 10: In optimal scenario, TK_{uu} should equate an identity matrix.

F.4. Conjugate Gradient steps during optimization

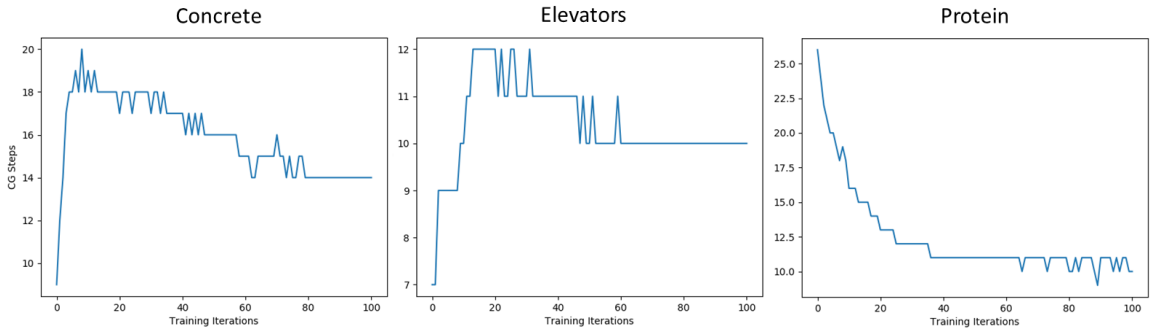


Figure 11: Number of CG steps needed to satisfy a tight tolerance for IF-CSTP trained on various datasets with 250 inducing points.