Supplement: Tighter Bounds on the Log Marginal Likelihood of Gaussian Process Regression using Conjugate Gradients

Artem Artemev [*1 2] **David R. Burt** [*3] **Mark van der Wilk** [1]

1. Additional bounds on the Log marginal likelihood of Gaussian process regression

In this section, we discuss additional bounds on the log marginal likelihood of GPR regression that can be computed efficiently. We focus on the case when we have access to a rank-m-plus-diagonal approximate $\mathrm{Q} \prec \mathrm{K}$, where we write $\mathrm{A} \prec \mathrm{B}$ if (and only if) A and B and $\mathrm{A} - \mathrm{B}$ is PSD (i.e. $\prec$ is the Loewner partial order on PSD matrices). We also assume we can efficiently compute the trace of K. We can then think of finding the optimal lower bound on $\log p_Y(\mathbf{y};\theta)$ as a constrained optimisation problem:

$$\log p_Y(\mathbf{y};\theta) = c - \frac{1}{2}\mathbf{y}^\mathsf{T}\mathrm{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathrm{K}| \geq c + \inf_{\substack{\mathrm{A}\succ\mathrm{Q}\\ \mathrm{tr}(\mathrm{A})=t}} \left(-\frac{1}{2}\mathbf{y}^\mathsf{T}\mathrm{A}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathrm{A}|\right). \tag{1}$$

We then apply an element-wise bound,

$$\log p_Y(\mathbf{y};\theta) \geq c + \inf_{\substack{\mathrm{A}\succ\mathrm{Q}\\ \mathrm{tr}(A)=t}} \left(-\frac{1}{2}\mathbf{y}^\mathsf{T}A^{-1}\mathbf{y} - \frac{1}{2}\log|A|\right) \geq c + \inf_{\substack{\mathrm{A}\succ\mathrm{Q}\\ \mathrm{tr}(\mathrm{A})=t}} -\frac{1}{2}\mathbf{y}^\mathsf{T}\mathrm{A}^{-1}\mathbf{y} + \inf_{\substack{\mathrm{A}\succ\mathrm{Q}\\ \mathrm{tr}(\mathrm{A})=t}} -\frac{1}{2}\log|\mathrm{A}| \tag{2}$$

We now consider each term separately,

1.1. Quadratic term

Since $\mathrm{A} \prec \mathrm{Q}$, we may write $\mathrm{A} = \mathrm{Q} + \mathrm{EE}$, where E is the PSD square root of $\mathrm{A} - \mathrm{Q}$. Applying Woodbury's Lemma,

$$\mathbf{y}^\mathsf{T}\mathrm{A}^{-1}\mathbf{y} = \mathbf{y}^\mathsf{T}\mathrm{Q}^{-1}\mathbf{y} - \mathbf{y}^\mathsf{T}\mathrm{Q}^{-1}\mathrm{E}(\mathrm{I} + \mathrm{E}\mathrm{Q}^{-1}\mathrm{E})^{-1}\mathrm{E}\mathrm{Q}^{-1}\mathbf{y}. \tag{3}$$

The second term is non-negative, but can be 0; in particular if $\mathrm{K} = \mathrm{Q} + t\mathbf{z}\mathbf{z}^\mathsf{T}$, where $\mathbf{z}$ is a unit vector orthogonal to $\mathrm{Q}^{-1}\mathbf{y}$. Hence,

$$\inf_{\substack{\mathrm{A}\succ\mathrm{Q}\\ \mathrm{tr}(\mathrm{A})=t}} -\frac{1}{2}\mathbf{y}^\mathsf{T}\mathrm{A}^{-1}\mathbf{y} = -\frac{1}{2}\mathbf{y}^\mathsf{T}\mathrm{Q}^{-1}\mathbf{y}, \tag{4}$$

which does not lead to any improvement in the quadratic term.

1.2. Log-determinant term

We recall the following property of PSD matrices,

Proposition 1 (Horn & Johnson, 2012, Corollary 7.7.4)**.** *Let* $\mathrm{A}_1, \mathrm{A}_2 \in \mathbb{R}^{k\times k}$ *such that* $\mathrm{A}_1 - \mathrm{A}_2$ *is PSD. Let* $\lambda_i(\mathrm{A}_j), 1 \leq i \leq k, j \in \{1, 2\}$ *denote the* i^{th} *largest eigenvalue of* A_j*. Then,* $\lambda_i(\mathrm{A}_1) \geq \lambda_i(\mathrm{A}_2)$ *for* $1 \leq i \leq k$*.*

We write

$$\log|\mathrm{A}| = \sum_{i=1}^{n}\log(a_i) = \sum\log(\ell_i + e_i), \tag{5}$$

where a_i are the eigenvalues of A and ℓ_i are the eigenvalues of Q, both sorted in descending order, and $e_i = a_i - \ell_i$. We can then translate the constraints on K as $e_i \geq 0$ and $\sum_i(e_i + \ell_i) = t$. The ℓ_i can be computed in $O(nm^2)$ by noting that $n - m$ of them are σ^2, and the remaining ones are the eigenvalues of $MM^\mathsf{T} + \sigma^2 I$, where $M = \mathrm{K}_{\mathrm{uu}}^{-1/2}\mathrm{K}_{\mathrm{uf}}$.

We define $t' = t - \mathrm{tr}(\mathrm{Q})$ and consider

$$\sup_{\substack{\mathrm{A} \succ \mathrm{Q} \\ \mathrm{tr(A)}=t}} \log|\mathrm{A}| = \sup_{\substack{e_i > 0 \\ \sum e_i = t'}} \log(e_i + \ell_i). \tag{6}$$

This coincides with a problem in information theory related to power-allocation over channels. It is a convex optimization which can be solved using Lagrange multiplier and the KKT conditions. The solution of this equation is,

$$e_i = \max(0, t'/\nu - t'\ell_i), \tag{7}$$

where ν is chosen such that, $\sum_{i=1}^{n} \max(0, 1/\nu - \ell_i) = 1$. See Example 5.2 in Boyd & Vandenberghe (2004) for details of this maximization. The Lagrange multiplier can be found in $O(n)$ as it is the solution to a piecewise linear problem.

In preliminary experiments, we found using this bound with hyperparameter optimisation did not yield significant gains over the simpler AM-GM based bound, and hence used the latter in the main experiments.

1.2.1. Lower bounding the log determinant

In some cases upper bounds on the log marginal likelihood are of interest (Titsias, 2014; Kim & Teh, 2018). We therefore turn to the problem of lower bounding the log determinant of $\mathbf{K}$. Improvements in this bound can be used in conjunction with either of the bounds on the quadratic term given in Titsias (2014) or Kim & Teh (2018). We have for any A satisfying the constraints,

$$\log(\mathrm{A}) = \sum \log\left(e_i + \ell_i\right) = \log|\mathrm{Q}| + \sum \log(1 + \frac{e_i}{\ell_i}). \tag{8}$$

Rewriting the second term on the right hand side as the log of a product, expanding the product and using that $e_i \geq 0$,

$$\sum_{i=1}^{n} \log\left(1 + \frac{e_i}{\ell_i}\right) = \log \prod_{i=1}^{n} \left(1 + \frac{e_i}{\ell_i}\right) \geq \log\left(1 + \sum_{i=1}^{n} \frac{e_i}{\ell_i}\right). \tag{9}$$

Using that $\ell_i \leq \ell_1$,

$$\log|\mathbf{K}| \geq \inf_{\substack{\mathrm{A} \succ \mathrm{Q} \\ \mathrm{tr(A)}=\mathrm{tr(K-Q)}}} \log|\mathrm{A}| \geq \log|\mathrm{Q}| + \log\left(1 + \frac{\mathrm{tr(K-Q)}}{\ell_1}\right). \tag{10}$$

We now show that this is the greatest upper bound given the constraints on A by constructing an A satisfying this bound. Consider $\mathrm{A} = \mathrm{Q} + \mathrm{tr(K-Q)}\mathbf{w}\mathbf{w}^{\mathsf{T}}$, where $\mathbf{w}$ is the eigenvector of Q corresponding to ℓ_1. Then all of the eigenvalues of A coincide with the eigenvalues of Q, except the largest, which is $\ell_1 + \mathrm{tr(K-Q)}$. Rearranging the formula for the log determinant, we see that this implies the second inequality in eq. (10) is in fact an equality.

2. Additional Experiments

We performed Iterative GP experiments with all combinations of Adam learning rates $(0.1, 0.01)$ and minimum likelihood noise constraints ($\sigma^2_{min} = 1\mathrm{e}{-4}$, $\sigma^2_{min} = 1\mathrm{e}{-6}$) for `poletele`, `kin40k`, `elevators`, `bike` and `protein` datasets. The experiments displayed instability and sensitivity to some settings (fig. 6, fig. 8, and fig. 9). In particular, we found that the setting with $\sigma^2_{min} = 1\mathrm{e}{-6}$ and Adam learning rate 0.1 (Wang et al. (2019) uses $\sigma^2_{min} = 1\mathrm{e}{-4}$ and Adam learning rate 0.1), causes severe predictive performance degradation in `poletele`, `bike`, and `kin40k` datasets after 2000 iterations. Lowering the learning rate to 0.01 facilitates smoother, but much slower convergence. However, datasets such as `poletele` (fig. 9) and `protein` (fig. 10) still exhibit a decrease in the predictive performance during training.

We also used Adam optimiser with 0.1 and 0.01 learning rates to train CGLB with 2048 inducing points on `poletele`, `bike`, and `elevators` datasets. We investigate how performance characteristics change with Adam optimiser, and we explore whether Adam optimiser can cause instabilities similar to Iterative GP in CGLB predictive performance. Experiments show that the CGLB model trained with Adam optimizer and 0.1 learning rate on `poletele` (fig. 11) and `elevators` (fig. 12) datasets achieves performance on par with results for the CGLB trained with L-BFGS. However, CGLB with Adam performs much worse on `bike` dataset than it did with L-BFGS. None of those experiments exhibit significant degradation in performance over the training time analogous to the behaviour of Iterative GP.

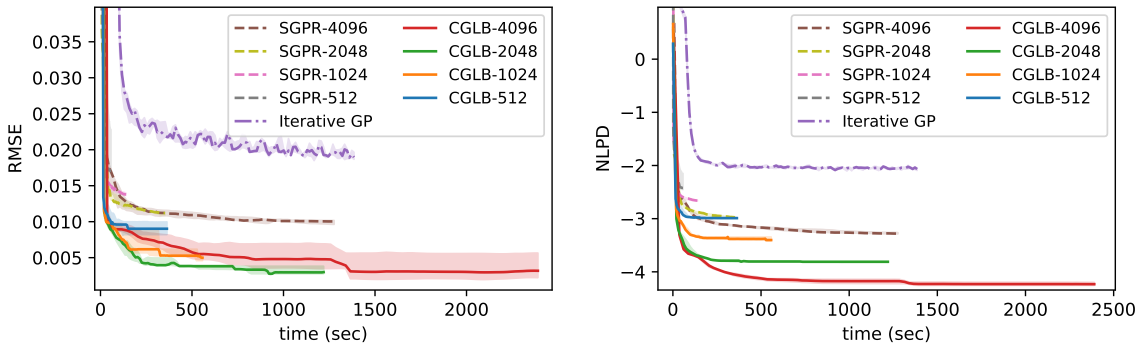

Figure 1. Test root mean square error (RMSE) and negative log predictive density (NLPD) metrics of CGGP, SGPR and Iterative GP models computed on `bike` dataset. The shaded area is IQR region, and the line is a median over five experiment trials with different dataset splits.

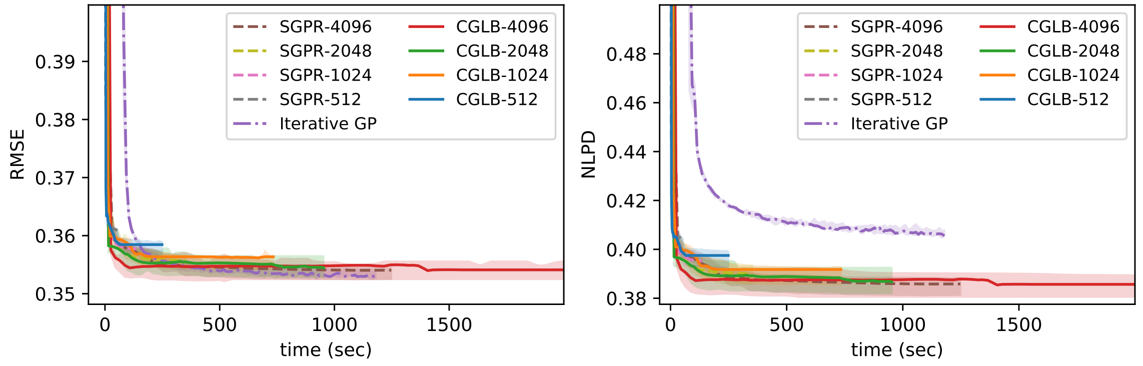

Figure 2. Test root mean square error (RMSE) and negative log predictive density (NLPD) metrics of CGLB, SGPR and Iterative GP models computed on `elevators` dataset. The shaded area is IQR region, and the line is a median over five experiment trials with different dataset splits.

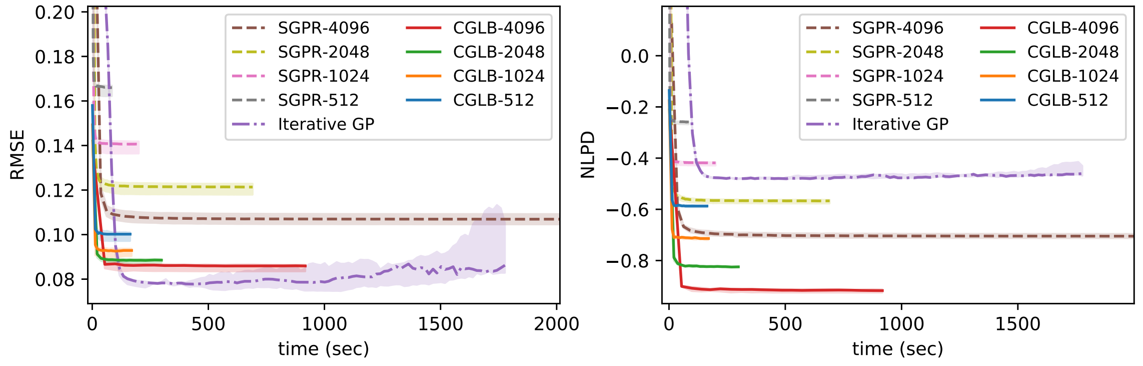

Figure 3. Test root mean square error (RMSE) and negative log predictive density (NLPD) metrics of CGLB, SGPR and Iterative GP models computed on `kin40k` dataset. The shaded area is IQR region, and the line is a median over five experiment trials with different dataset splits.

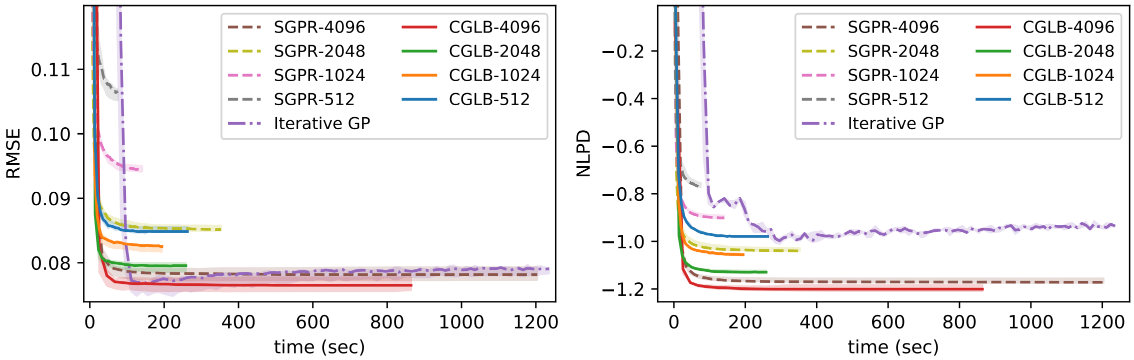

Figure 4. Test root mean square error (RMSE) and negative log predictive density (NLPD) metrics of CGLB, SGPR and Iterative GP models computed on `poletele` dataset. The shaded area is IQR region, and the line is a median over five experiment trials with different dataset splits.

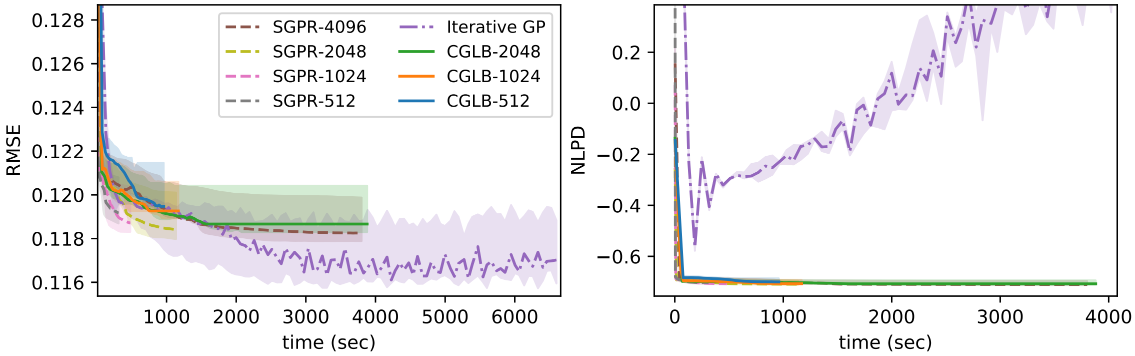

Figure 5. Test root mean square error (RMSE) and negative log predictive density (NLPD) metrics of CGLB, SGPR and Iterative GP models computed on `keggundirected` dataset. The shaded area is IQR region, and the line is a median over five experiment trials with different dataset splits.

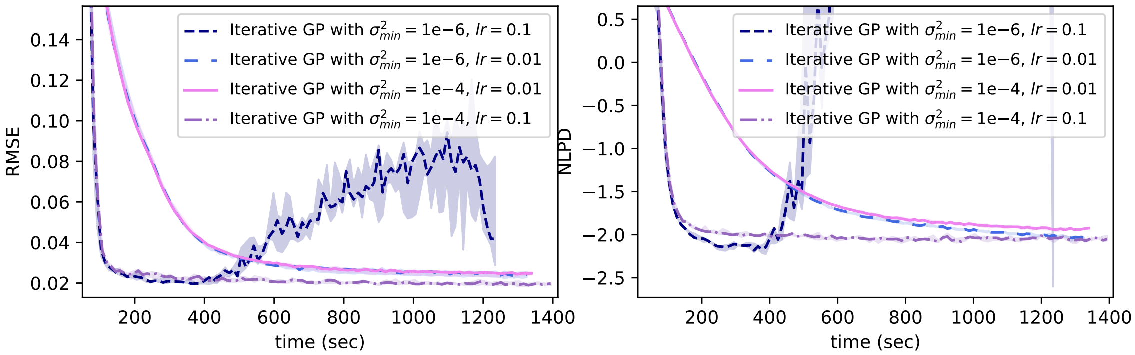

Figure 6. Test RMSE and NLPD for the Iterative GP method with learning rates 0.01 and 0.1 and minimum likelihood noise constrained at $1e{-}4$ and $1e{-}6$ on the `bike` dataset.

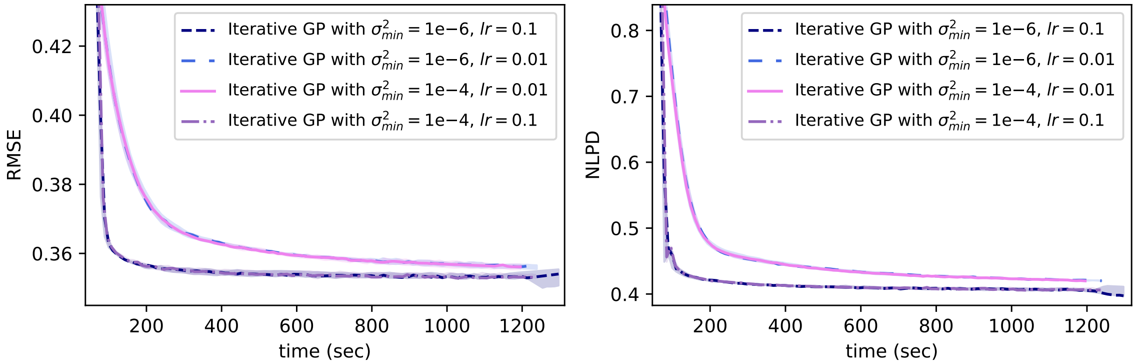

Figure 7. Test RMSE and NLPD for the Iterative GP method with learning rates 0.01 and 0.1 and minimum likelihood noise constrained at $1e{-4}$ and $1e{-6}$ on the `elevators` dataset.

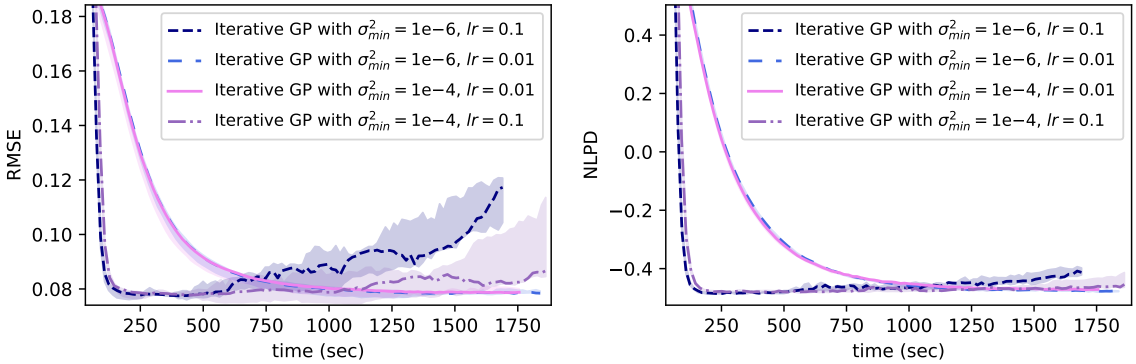

Figure 8. Test RMSE and NLPD for the Iterative GP method with learning rates 0.01 and 0.1 and minimum likelihood noise constrained at $1e{-4}$ and $1e{-6}$ on the `kin40k` dataset.

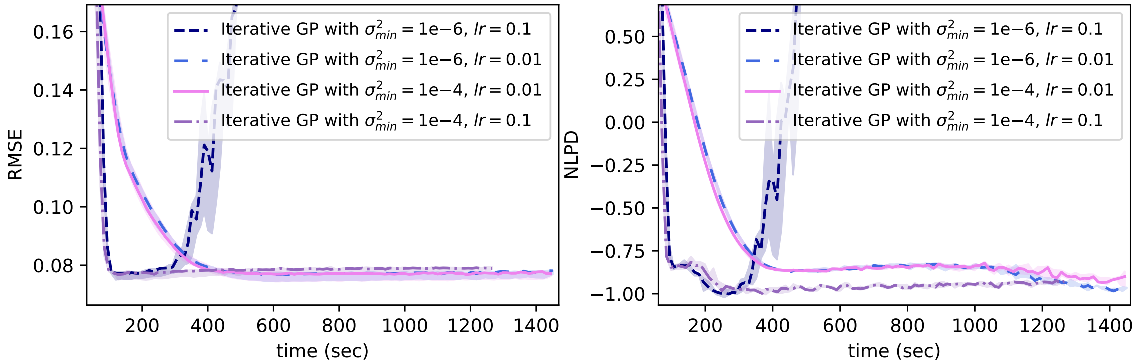

Figure 9. Test RMSE and NLPD for the Iterative GP method with learning rates 0.01 and 0.1 and minimum likelihood noise constrained at $1e{-4}$ and $1e{-6}$ on the `poletele` dataset.

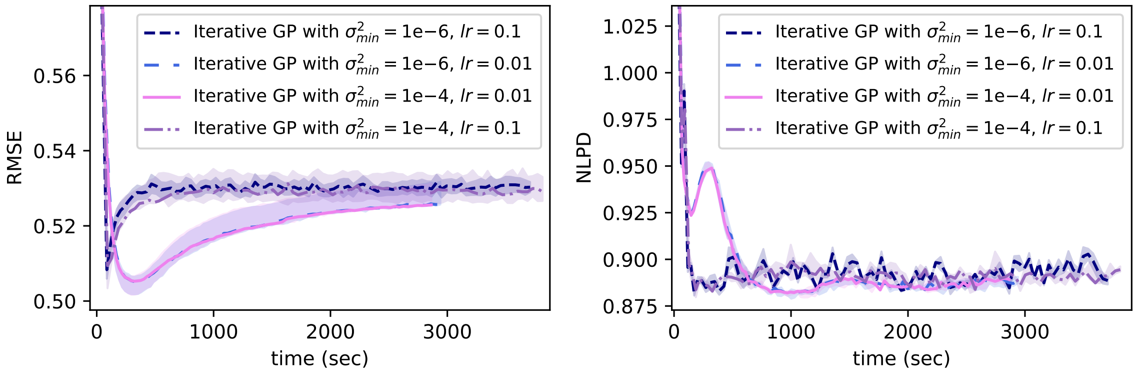

Figure 10. Test RMSE and NLPD for the Iterative GP method with learning rates 0.01 and 0.1 and minimum likelihood noise constrained at 1e−4 and 1e−6 on the `protein` dataset.

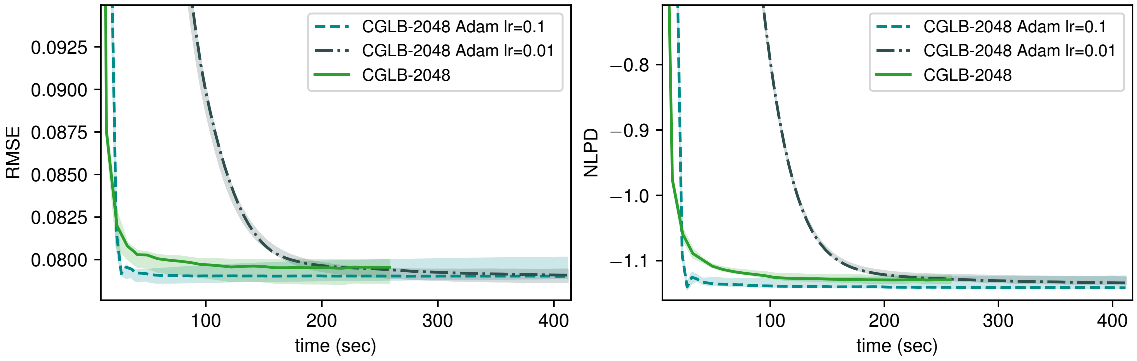

Figure 11. Test RMSE and NLPD for CGLB with 2048 inducing points which trained with L-BFGS and Adam with 0.1 and 0.01 learning rates on `poletele` dataset.

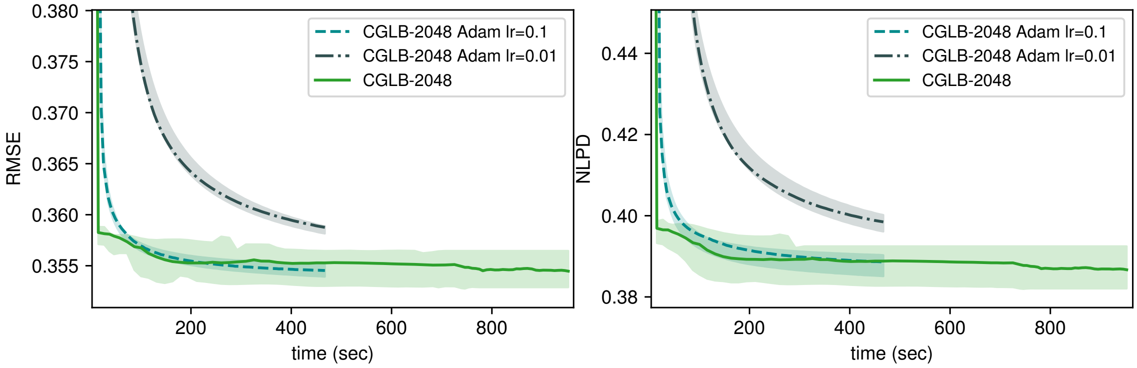

Figure 12. Test RMSE and NLPD for CGLB with 2048 inducing points which trained with L-BFGS and Adam with 0.1 and 0.01 learning rates on `elevators` dataset.

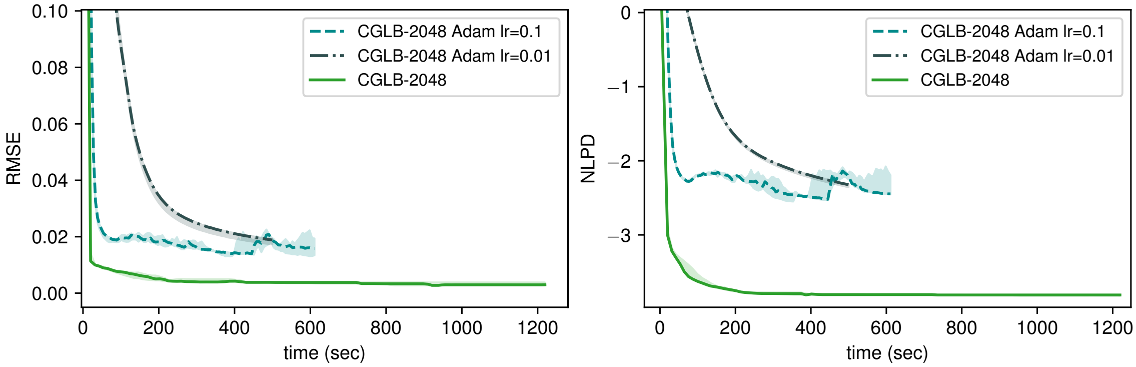

Figure 13. Test RMSE and NLPD for CGLB with 2048 inducing points which trained with L-BFGS and Adam with 0.1 and 0.01 learning rates on `bike` dataset.

References

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge University Press, 2012.

Kim, H. and Teh, Y. W. Scaling up the automatic statistician: Scalable structure discovery using Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 575–584, 2018.

Titsias, M. K. Variational inference for Gaussian and determinantal point processes. In *Workshop on Advances in Variational Inference (NIPS)*, 2014.

Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., and Wilson, A. G. Exact Gaussian processes on a million data points. In *Advances in Neural Information Processing Systems*, 2019.