Memory-Based Dual Gaussian Processes for Sequential Learning

Paul Chang Prakhar Verma Ti John Arno Solin Emtiyaz Khan

Aalto University · FCAI Finnish Center for Artificial Intelligence · RIKEN · AIP

TL;DR

- We devise a sparse GP method for inference, learning, and maintaining a representation in the sequential data setting.
- Key to our method is the use of what we call the dual parameters, which allow for easy assimilation of new data, a data-focused learning objective, and the derivation of a new score for picking important points.
- We show that our method can perform well on various sequential tasks that involve a sparse GP, such as streaming, building batches in Bayesian optimization/active learning, and continual learning.

Dual Form of the SVGP Solution

The approximate posterior process $q_{\mathbf{u}}(f_i)$ can be expressed as follows:

$$\mathbb{E}_{q_{\mathbf{u}}(f(\cdot))}[f_i] = \mathbf{k}_{\mathbf{z}i}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \boldsymbol{\alpha}_{\mathbf{u}},$$
$$\mathrm{Var}_{q_{\mathbf{u}}(f(\cdot))}[f_i] = \kappa_{ii} - \mathbf{k}_{\mathbf{z}i}^\top \left[\mathbf{K}_{\mathbf{zz}}^{-1} - (\mathbf{K}_{\mathbf{zz}} + \boldsymbol{B}_{\mathbf{u}})^{-1}\right] \mathbf{k}_{\mathbf{z}i}.$$

The dual parameters are $\boldsymbol{\alpha}_{\mathbf{u}}$ and $\boldsymbol{B}_{\mathbf{u}}$ which can be derived as

$$\boldsymbol{\alpha}_{\mathbf{u}} = \sum_{i \in \mathcal{D}} \mathbf{k}_{\mathbf{z}i}\, \mathbb{E}_{q_{\mathbf{u}}(f(\cdot))}\left[\nabla_{f_i} \log p(y_i \mid f_i)\right],$$
$$\boldsymbol{B}_{\mathbf{u}} = \sum_{i \in \mathcal{D}} \mathbf{k}_{\mathbf{z}i}\, \mathbb{E}_{q_{\mathbf{u}}(f(\cdot))}\left[-\nabla^2_{f_i} \log p(y_i \mid f_i)\right] \mathbf{k}_{\mathbf{z}i}^\top.$$

Sequential Inference and Learning

In the sequential setting, we no longer have access to previous data $\mathcal{D}_{\text{old}}$ once we receive new data $\mathcal{D}_{\text{new}}$. In the dual parameter form, updating our variational parameters becomes

$$\boldsymbol{\alpha}_{\mathbf{u}} = \boldsymbol{\alpha}_{\mathbf{u}}^{\text{old}} + \boldsymbol{\alpha}_{\mathbf{u}}^{\text{new}} \quad \text{and} \quad \boldsymbol{B}_{\mathbf{u}} = \boldsymbol{B}_{\mathbf{u}}^{\text{old}} + \boldsymbol{B}_{\mathbf{u}}^{\text{new}}.$$

To find $\boldsymbol{\alpha}_{\mathbf{u}}^{\text{new}}$ and $\boldsymbol{B}_{\mathbf{u}}^{\text{new}}$, we run an adjusted natural gradient algorithm detailed in [?]. The dual parameters can be viewed as a compact representation of the data, meaning our posterior approximation can be decomposed as follows:

$$q(\mathbf{u}) \propto p_{\boldsymbol{\theta}}(\mathbf{u}) \mathrm{N}(\mathbf{u} \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{\Sigma}}),$$

where $\tilde{\mathbf{y}} = \mathbf{K}_{\mathbf{zz}} \boldsymbol{B}_{\mathbf{u}} \boldsymbol{\alpha}_{\mathbf{u}}$ and $\tilde{\boldsymbol{\Sigma}} = \mathbf{K}_{\mathbf{zz}} \boldsymbol{B}_{\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{zz}}$.

Using this formulation, we can derive the dual-based sequential ELBO for hyperparameters, which crucially includes a memory term:

$$\mathcal{L}^{\boldsymbol{\theta}}_{\text{seq}} = \log \mathcal{Z}(\boldsymbol{\theta}) + \sum_{i \in \mathcal{D}_{\text{new}}} \mathbb{E}_{q_{\mathbf{u}}(f_i;\boldsymbol{\theta})}[\log p(y_i \mid f_i)] - \mathbb{E}_{q_{\mathbf{u}}(f_i;\boldsymbol{\theta})}[\log \mathrm{N}(\mathbf{u} \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{\Sigma}})]$$
$$+ \frac{n_{\text{old}}}{n_{\mathcal{M}}} \sum_{i \in \mathcal{M}} \mathbb{E}_{q(f_i;\boldsymbol{\theta})}[\log p(y_i \mid f_i)].$$

The memory term lets us trade off compute against how close the sequential solution will be to the offline batch solution.

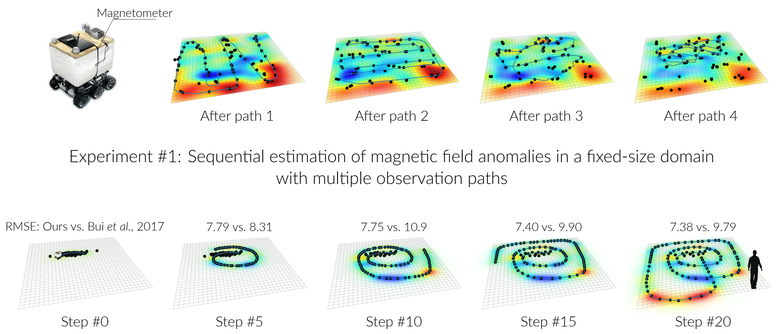


Figure 1: Sequential estimation of anomalies by the proposed method in the local ambient magnetic field strength (10μT 90μT) as mapped by a small wheeled robot. Two experiments for learning the hyperparameters and representation: (a) Complete trajectories are received and the model does not have access to the previous trajectories. (b) Data is received during the exploration of the space requiring the method to spread out a fixed number of inducing points.

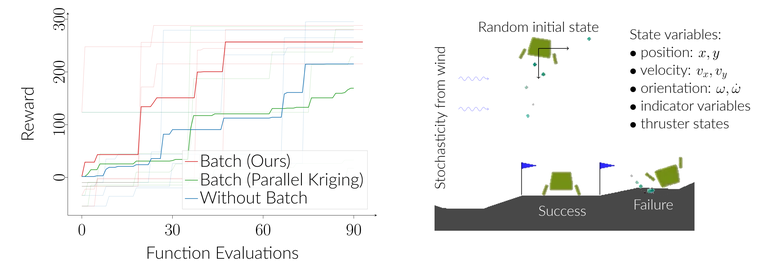
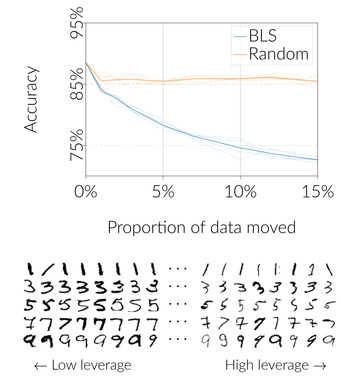
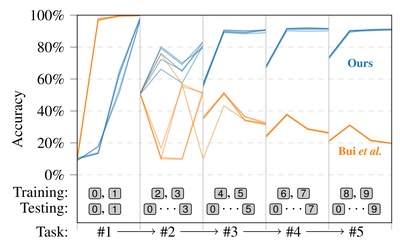


Figure 2: Bayesian optimization on a lunar landing setup. The goal is to successfully land on the surface between the flags. Compared against the non-batch solution and a batch solution using parallel Kriging, the proposed approach achieves higher reward in less function evaluations.





Accuracy on *split MNIST* over training with different random seeds. The training starts with 0 vs. 1 and each task introduces new digits while testing on all classes thus far. The overall accuracy (mean over tasks) drops when introducing a new task, but recovers and does not suffer from forgetting older tasks.

Representation: Memory Selection and $\mathbf{Z}$

The selection of memory and the placement of the inducing points $\mathbf{Z}$ define the representation of the data. We need a memory that reflects previous data for learning $\boldsymbol{\theta}$ and inducing points that provide support throughout $\mathcal{D}_{\text{old}} \cup \mathcal{D}_{\text{new}}$.

For memory selection, we derive a new score called the Bayesian leverage score (BLS),

$$h_i^{\text{BLS}} = \mathbf{a}_i^\top \left[\mathbf{A}\mathbf{W}\mathbf{A}^\top + \mathbf{K}_{\mathbf{zz}})^{-1}\right]^{-1} \mathbf{a}_i,$$

where $\mathbf{A}$ is the projection matrix whose rows are $\mathbf{a}_i = \mathbf{k}_{\mathbf{z},i}^\top \mathbf{K}_{\mathbf{zz}}^{-1}$, and $\mathbf{W} = \mathrm{diag}([-\nabla^2_{f_i} \log p(y_i \mid f_i)])$. The BLS generalises the ridge leverage score which only applies to Gaussian likelihoods, where $\mathbf{W} = \mathrm{diag}([\frac{1}{\sigma^2}])$.

- The BLS can be seen as example difficulty and we sample memory points proportional to the BLS score.
- For selection of $\mathbf{Z}$, we use a sequential pivoted Cholesky algorithm which returns M points from the kernel evaluated on $[\mathbf{Z}_{\text{old}}, \mathbf{X}_{\text{new}}]$.

Experiments

Real world robot estimation of magnetic field

- We consider the task of online mapping of local anomalies in the ambient magnetic field.
- We assign a GP prior $\mathcal{GP}(0, \sigma_0^2 + \kappa^{\text{Matérn}}_{\sigma^2,\ell}(\mathbf{x}, \mathbf{x}'))$ to the magnetic field strength over the space under the presence of Gaussian measurement noise.

Building batches in Bayesian optimization / Active learning

- The *lunar lander problem* is a challenging optimization problem due to environment stochasticity.
- Our method supports an acquisition function that is a product of the Expected Improvement of the regression model and the predictive mean of the classification model. In both the models, we make a batch of fantasy points.

Continual learning split-MNIST

- The BLS score helps maintain a compact and efficient representation by ranking data examples.
- Our method handles non-stationary sequential tasks such as split-MNIST by combining representation learning with a better-conditioned learning objective and efficient variational updates.

References

[1] V. Adam, P. Chang, M. E. Khan, and A. Solin, "Dual parameterization of sparse variational Gaussian processes," *Advances in Neural Information Processing Systems*, 2021.

paul.chang@aalto.fi