Bayesian Optimization using Deep Gaussian Processes

Sergio Pascual Diaz, Thang Bui, Richard Turner
Department of Engineering University of Cambridge

Objectives

My thesis focuses on the use Deep Gaussian Processes for Bayesian Optimization with Expected Improvement, where the main objectives are:

- Assess the quality of Expectation Propagation as inference for regression of DGP samples.
- Test performance of AEP-EI DGPs on benchmark Bayesian Optimization problems.

Bayesian Optimization

Interested in finding minima of black-box functions where evaluation is expensive, noisy and gradient information is not available. Two components:

- **Regression** of objective function using standard *Gaussian Processes* (GPs) to yield predictive distribution $P(y|\mathbf{x}, \mathcal{D}_n) = \mathcal{N}(y|\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$.
- **Data collection strategy** by maximizing *Expected Improvement* (EI) acquisition function $\alpha_{EI}(\mathbf{x}) = \mathbf{E}_{P(y|\mathbf{x},\mathcal{D}_n)}(\max(0, \mu_{min} - y))$.

Although robust and analytically tractable, the expressiveness of standard GPs is limited by the choice of kernel $\mathcal{K}(\cdot, \cdot)$ (e.g.: SE, Matern).

Limitations of GP Regression

In Bayesian Optimization problems often length-scales vary as a function of space requiring:

- Covariance function design, which only guarantees local smoothing with a fixed length-scale.
- Sophisticated kernel design in higher dimensions.
- Input and output warping [1].
- Leverage multiple correlated outputs.

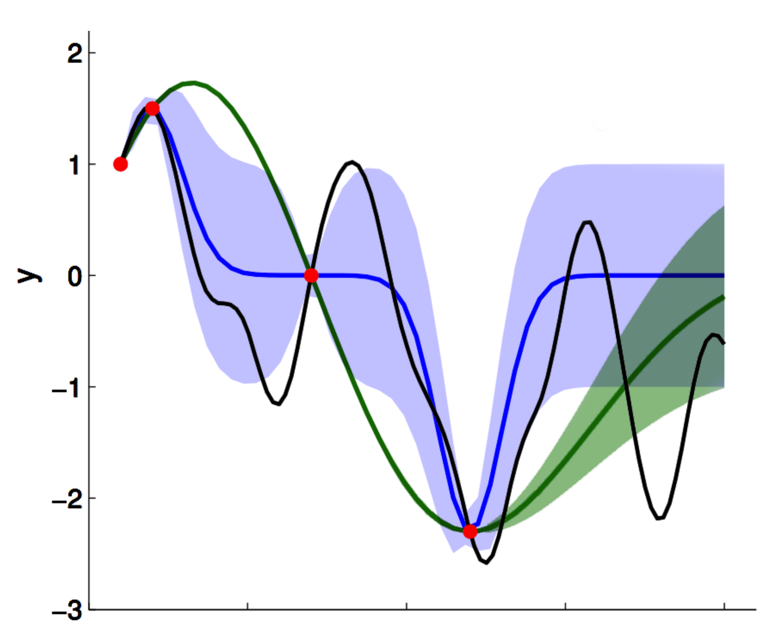
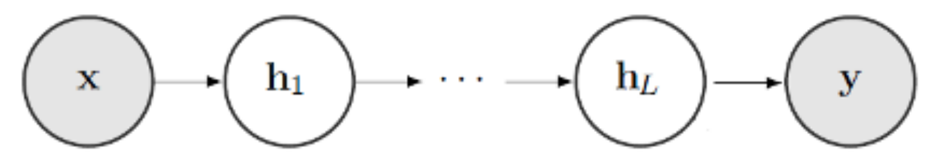


Figure: GP regression with two different kernel functions

Deep Gaussian Processes



Deep Gaussian Processes (DGPs) [2] are a multi-layer hierarchical generalisation of standard GPs. Given inputs and observation pairs $(\mathbf{x}_n, y_n)_{n=1}^N$:

$$p(f_l|\theta_l) = \mathcal{GP}(f_l; \mathbf{0}, \mathbf{K}_l) \quad (1)$$

$$p(\mathbf{h}_l|f_l, \mathbf{h}_{l-1}, \sigma_l^2) = \prod_n \mathcal{N}(h_{l,n}; f_l(h_{l-1,n}), \sigma_l^2) \quad (2)$$

$$p(\mathbf{y}|f_L, \mathbf{h}_{L-1}, \sigma_L^2) = \prod_n \mathcal{N}(y_n; f_L(h_{L-1,n}), \sigma_L^2) \quad (3)$$

with hidden layers $\mathbf{h}_l$ for $l = 1, \ldots, L$.

Technical Challenges

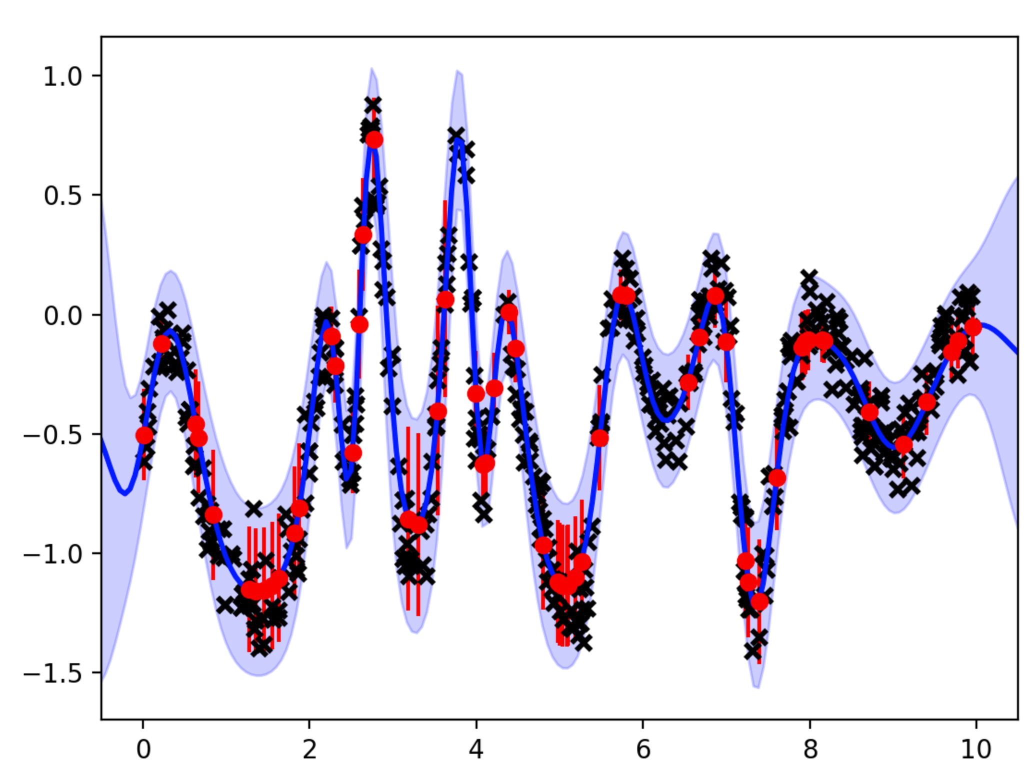
In practice, the DGP objective, the marginal likelihood $p(y|\mathbf{x}, \alpha)$, is intractable and regression has a high computational cost, $\mathcal{O}(LN^3)$. The following approximation scheme was proposed in [3]:

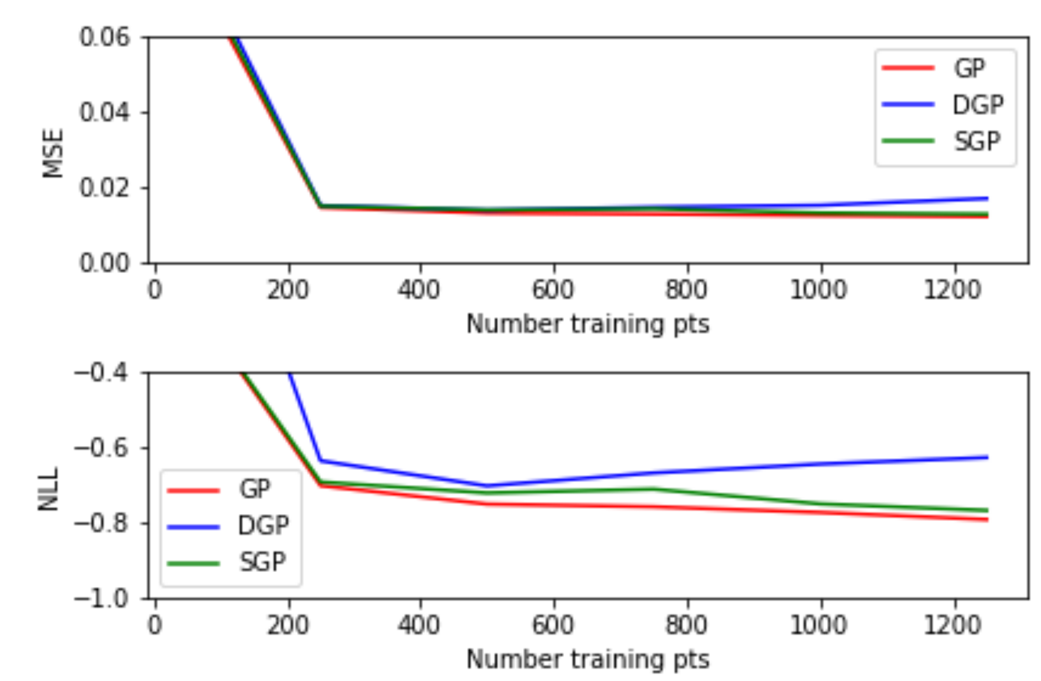
- *FITC* based pseudo-point sparse approximations, lowering the cost to $\mathcal{O}(LNM^2)$ where M is the number of pseudo points induced.
- Approximate inference for objective using approximate *Expectation Propagation* (AEP).
- *Probabilistic Backpropagation Algorithm* for training.

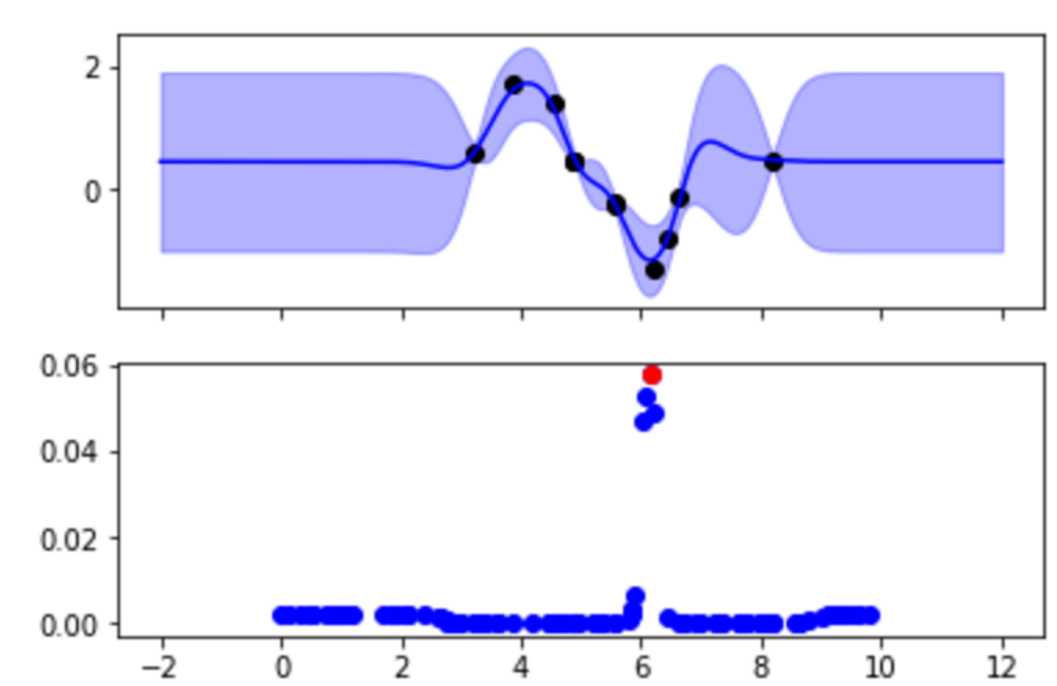
Experimental Results

Perform regression on multiple one-hidden-layer one-dimensional DGP samples to assess quality of Approximate EP inference scheme:

- AEP-DGP MSE and NLL scores compared to Full standard GPs and Sparse GPs. (Top-right)
- AEP-DGP fit to test data from DGP sample. (Bottom-left)
- Bayesian Optimization example for a DGP sample, using EI optimized by max pooling. (Bottom-right)







Alternative Methods

- *Bayesian Neural Networks* [4] require Monte Carlo methods making EI intractable.
- Rich and complex *Covariance Functions* require bayesian handling of hyperparameters to prevent overfitting.

Discussion

- DGPs present a flexible, more analytic generative model, making EI tractable.
- Hierarchical structure of DGPs allows for automatic input and output warping as well as non-parametric kernel design.
- DGPs are able to find correlations between multiple outputs and estimate expensive objectives based on cheaper ones.

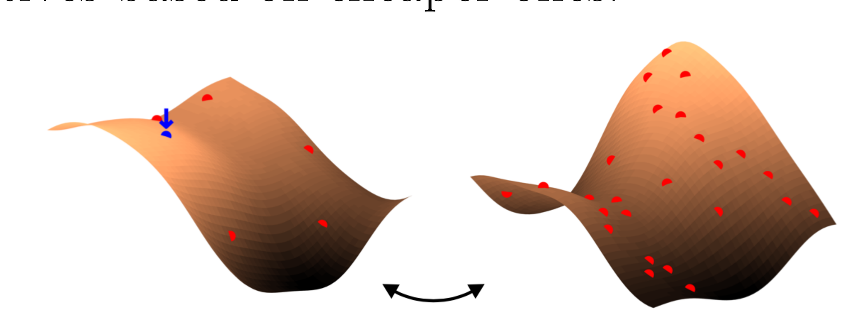


Figure: Expensive surface (right) and cheap surface (left) with multiple evaluations.

Future Work

- Investigate Sequential Monte Carlo based inference for DGPs in small-size Bayesian Optimization regression problems.

References

[1] Nando de Freitas Bobak Shahriari, Ryan P. Adams. Taking the human out of the loop: A review of bayesian optimization (2016).

[2] Neil D. Lawrence Andreas C. Damianou. Deep gaussian processes (2014).

[3] Yingzhen Li JosÃl Miguel HernÃąndez-Lobato Richard E. Turner Thang D. Bui, Daniel HernÃąndez-Lobato. Deep gaussian processes for regression using approximate expectation propagation (2016).

[4] Falkner Frank Hutter Jost Tobias, Klein Stefan. Bayesian optimization with robust bnns.