UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Designing Neural Network Hardware Accelerators Using Deep Gaussian Processes

M. Havasi,
Supervisor: Dr. J. M. Hernandez-Lobato

June, 2017

Introduction

In this project, Bayesian optimization methods are employed in a hardware design setting where simulation of a given hardware configuration can be computationally very expensive. An additional twist is that there are multiple objective functions, the power consumption and the accuracy, to balance against each other.

Deep Gaussian Processes

Deep Gaussian processes (DGPs) are multi-layer hierarchical generalizations of Gaussian processes (GPs) and are formally equivalent to neural networks with multiple, infinitely wide hidden layers.

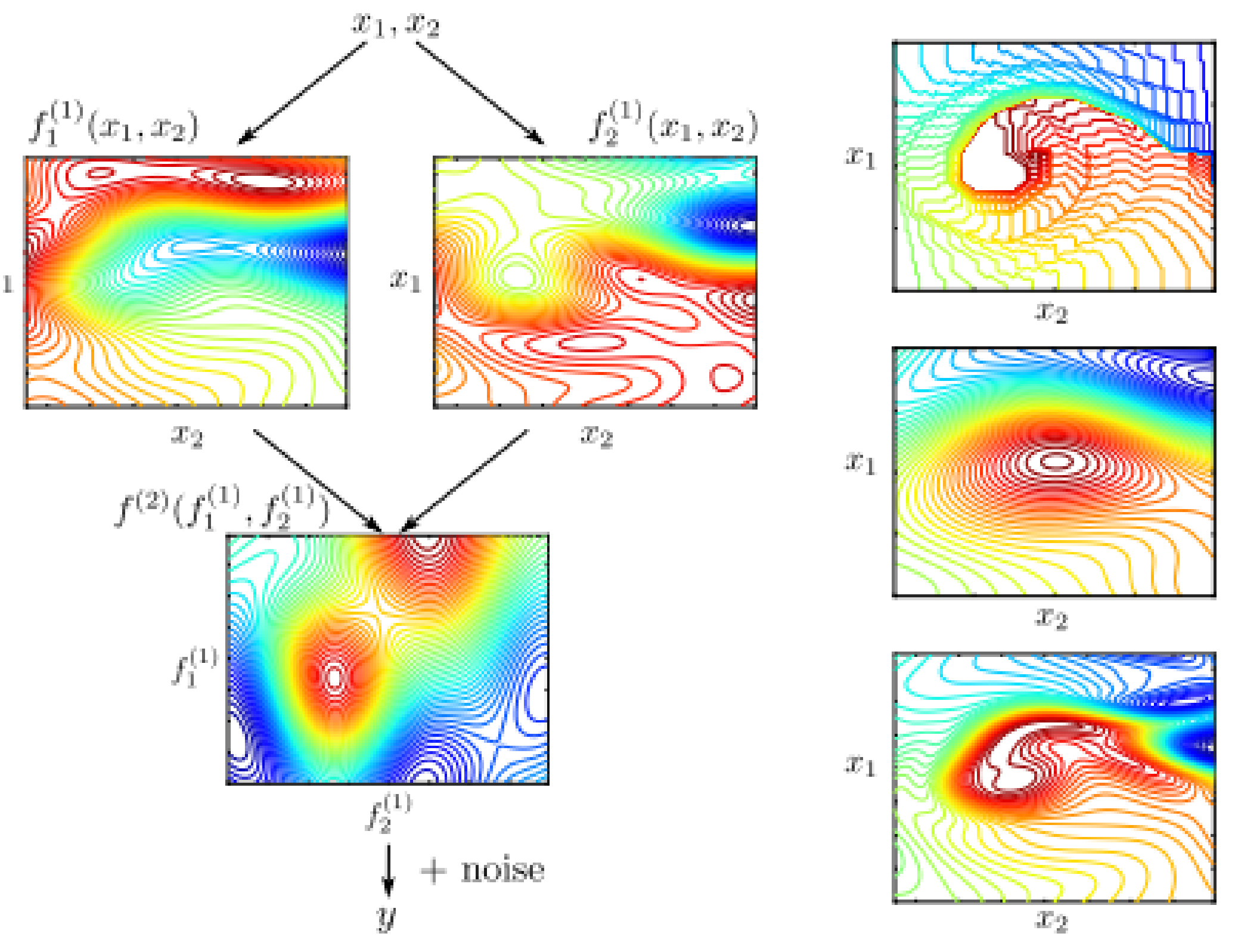
$$p(f_l|\theta_l) = \mathcal{GP}(f_l; \mathbf{0}, \boldsymbol{K}_l)$$

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

Approximate Inference is attained using Expectation Propagation.

$$log(\boldsymbol{y}|\alpha) \approx \mathcal{F}(\alpha) = \phi(\theta) - \phi(\theta_{prior}) + \sum_{n=1}^{N} log\tilde{Z}_n$$

$$log\tilde{Z}_n = logZ_n + \phi(\theta^{\backslash n}) - \phi(\theta)$$



Multiobjective Optimization

The goal is to find the set of optimal configurations that are not outperformed in both objective functions. These points are called the Pareto front.
We are using an acquisition function to determine where the next evaluation will take place. The aim of this heuristic function is to determine which evaluations will lead to the best Pareto front approximation.

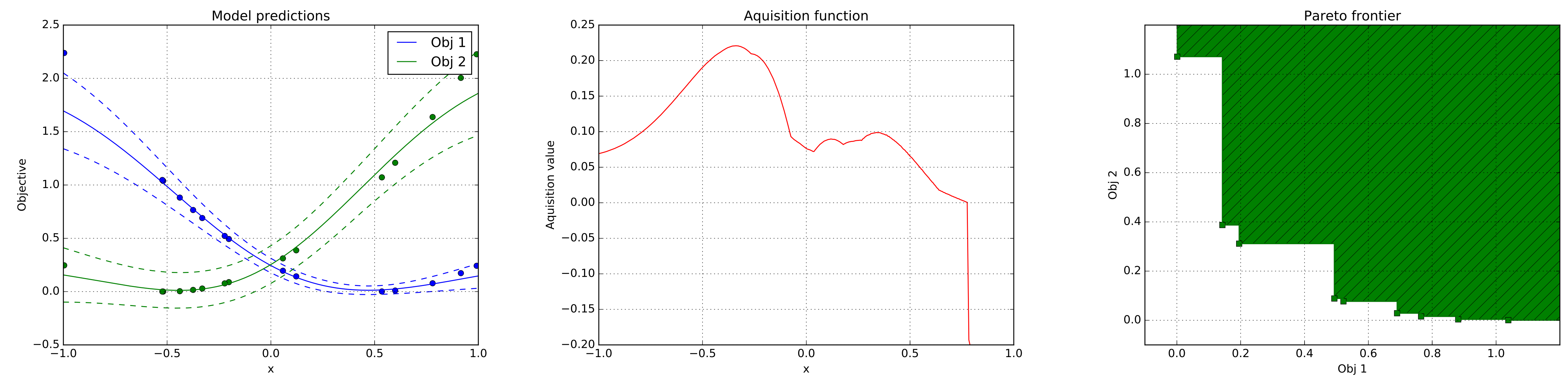
Acquisition function

Our aquisition function is S-Metric Selection-based Efficient Global Optimization (SMSego) which aims to maximize the hypervolume of the Pareto front.
It uses the lower confidence bound (LCB) for predicted values $\tilde{y}$ and uncertainties $\tilde{s}$:

$$\tilde{y}_{pot} = \tilde{y} - \alpha\tilde{s}$$

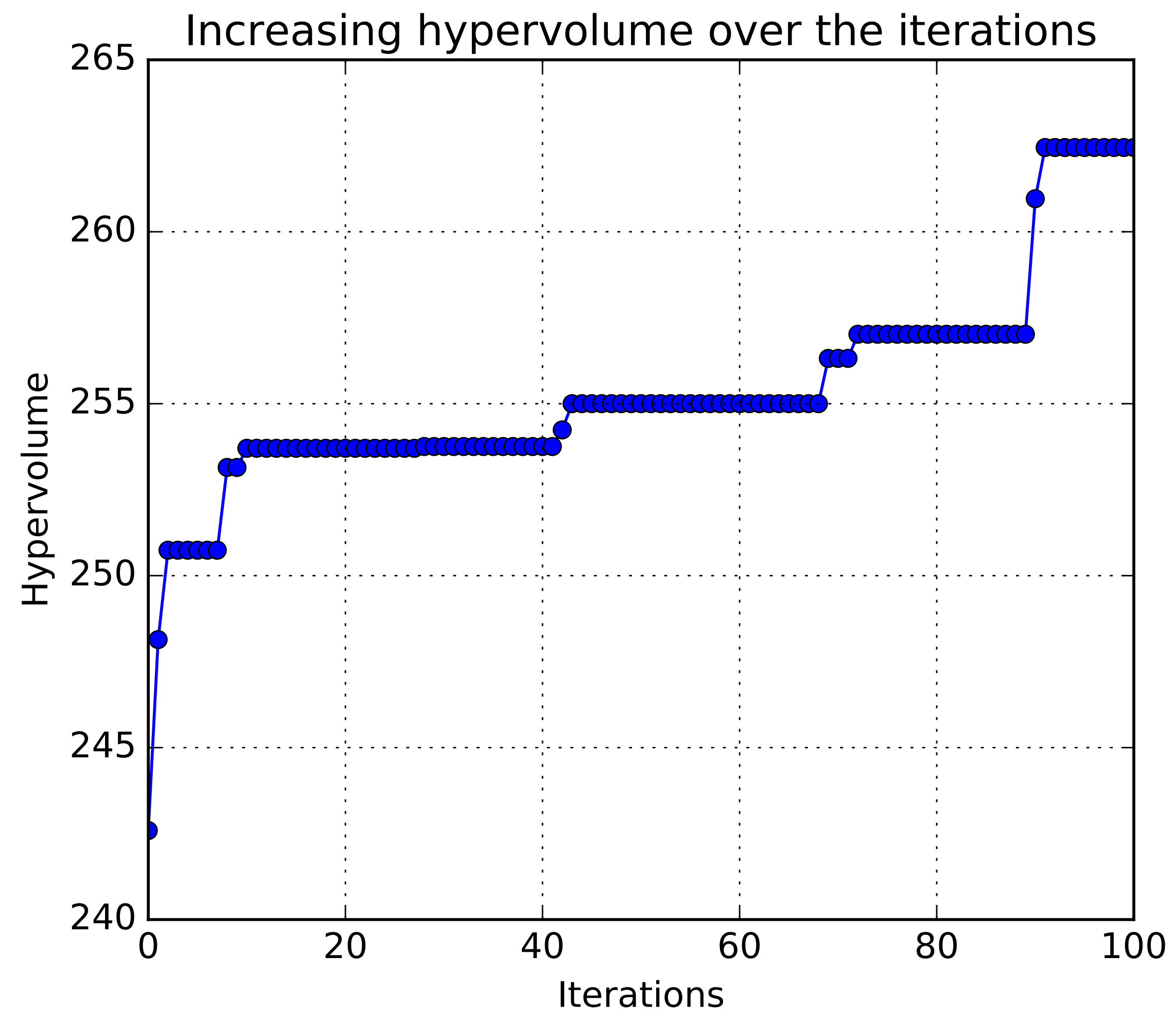
If the potential is dominated by a Pareto point then a penalty is applied along the non-ϵ-dominated dimensions.

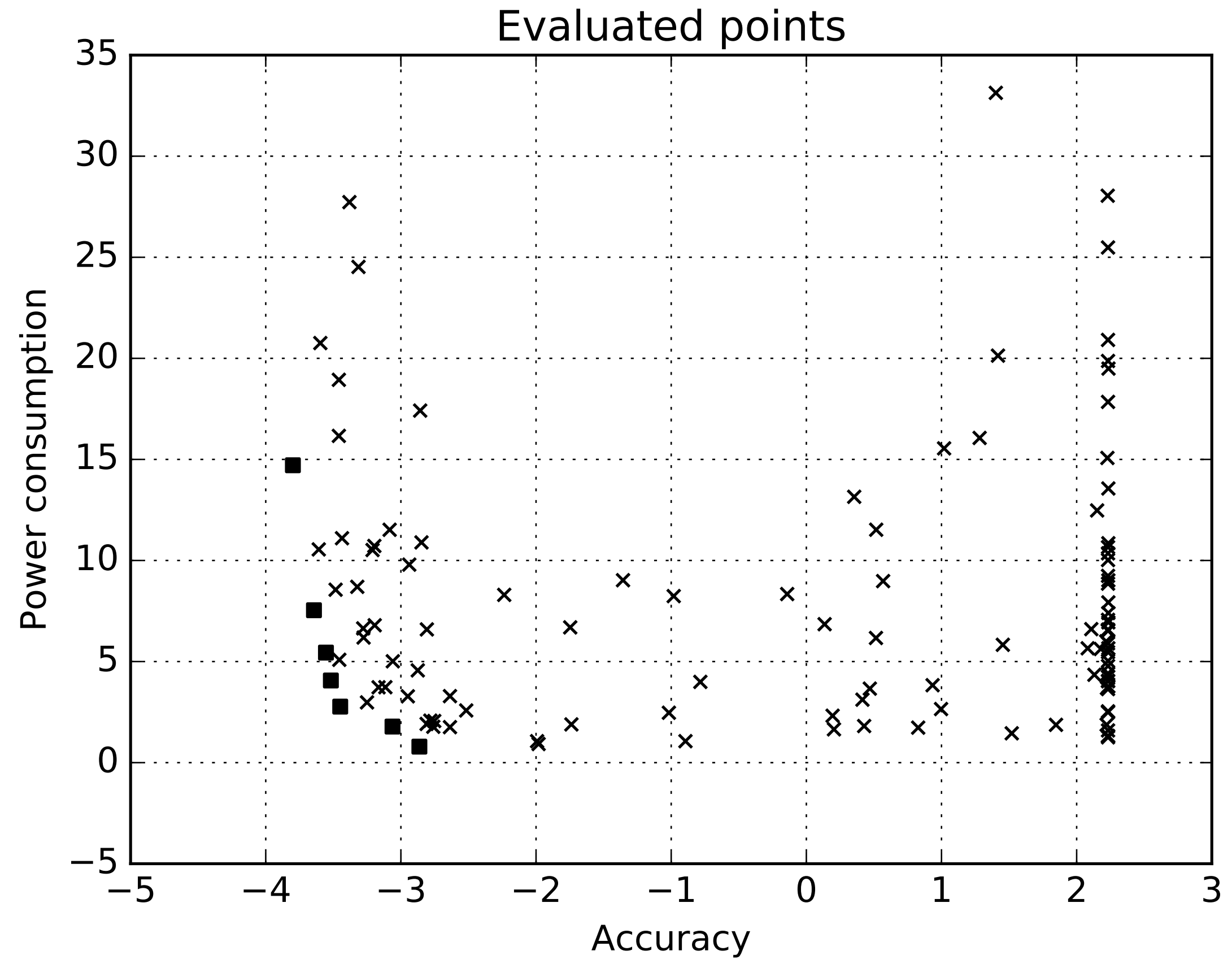
2D Example



Initial Results

This is an example run on real data draw from neural network hardware accelerators. The inputs were 13 dimensional configurations and the two objectives were power consumption and prediction accuracy.





Applications

The methodology is applicable in any setting where evaluating the objective function is costly and there are multiple objectives to balance against each other. There is a wide range of applications such as:

- Optimizing hardware configurations
- Optimizing neural-network hyperparameters

Further Work

During the remaining time of this project, we will focus on collecting more data. We want to establish baselines using existing methods and evaluate the relative performance.
We are also considering extensions that will speed up the otherwise expensive ($O(NLK^2)$) runtime.

References

[1] Bui T. D., Hernández-Lobato J. M., Li Y., Hernández-Lobato D. and Turner R. E. Deep *Gaussian Processes for Regression using Approximate Expectation Propagation*, In ICML, 2016.

[2] W. Ponweiser, T. Wagner, D. Biermann, and M. Vincze, *Multiobjective optimization on a limited budget of evaluations using model-assisted s-metric selection* in Proc. Parallel Problem Solving from Nature (PPSN X) , Dortmund, Germany, Sept. 2008, pp. 784âĂŞ794.

[3] Hernández-Lobato J. M., Gelbart M. A., Reagen B., Adolf R., Hernández-Lobato D., What- mough P., Brooks D., Wei G.-Y. and Adams R. P. *Designing Neural Network Hardware Accelerators with Decoupled Objective Evaluations*, In NIPS Workshop on Bayesian Opti- mization, Barcelona, Spain, 2016.

[4] Reagen B., Whatmough P. Adolf R., Rama S., Lee H., Lee S., Hernández-Lobato J. M., Wei G. Y. and Brooks D. *Minerva: Enabling Low-Power, High-Accuracy Deep Neural Network Accelerators*, In ISCA, 2016.