Bayesian Optimization with Neural Network Prior Mean Models

Tobias Boltz, Jose L. Martinez, Connie Xu, Kathryn R. L. Baker, Ryan Roussel, Daniel Ratner, Brahim Mustapha, Auralee L. Edelen

6 March 2024

SLAC NATIONAL ACCELERATOR LABORATORY
Argonne NATIONAL LABORATORY
UKRI Science and Technology Facilities Council
ISIS Neutron and Muon Source
Stanford University
U.S. DEPARTMENT OF ENERGY

Motivation

Why use Bayesian Optimization?

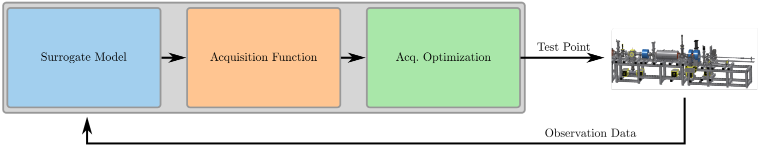
- Sample-efficiency: objectives are commonly expensive to evaluate at particle accelerators
- Flexibility and ease of use: successful applications at several facilities
- Possibility to include prior information: $p(A|B) \propto p(B|A)\, p(A)$
- Often some prior knowledge is available, but not used: beam dynamics principles, historical data, physics simulations, data from previous optimization runs etc.
- Large amounts of prior data are difficult to incorporate directly into GPs
- Convergence time can be a major limitation for high dimensional problems

How to make use of the available information?

- Prior mean models can improve sample-efficiency $\Rightarrow$ scale to higher dimensions
- NNs are flexible and scale well with the size of the available training data set

$\Rightarrow$ **Combine sample-efficiency of BO with computational scaling of NNs!**

Preliminaries



- Posterior mean for standard BO[1]:

$$\mu_* = K(X_*, X)[K(X, X) + \sigma_\epsilon^2 I]^{-1}\mathbf{y}$$

- Non-constant prior adds extra term[1]:

$$\mu_* = \mathbf{m}(X_*) + K(X_*, X)[K(X, X) + \sigma_\epsilon^2 I]^{-1}(\mathbf{y} - \mathbf{m}(X))$$

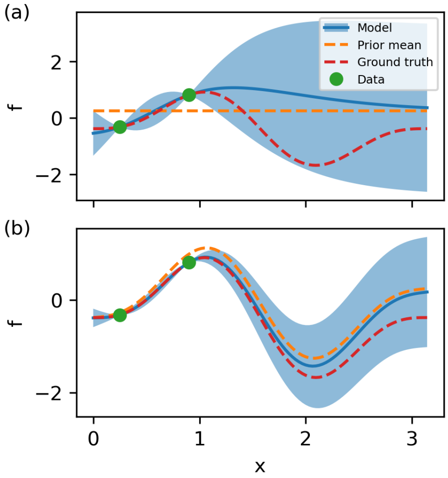
$\Rightarrow$ GP model is trained to predict the difference to the prior mean function $\mathbf{m}(X_*)$

$\Rightarrow$ Beneficial if the difference is small and/or easier to learn

[1] C.E. Rasmussen and C.K.I. Williams, MIT Press (2006)

Non-Constant Prior Mean Functions

- Posterior reverts to prior mean in the absence of local data
- Inaccurate predictions are updated with available data samples
- Good prior mean functions lead to better model predictions if no local data is available



BO with NN Prior Mean Model

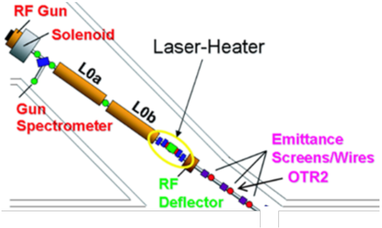
- LCLS Injector Surrogate Model:

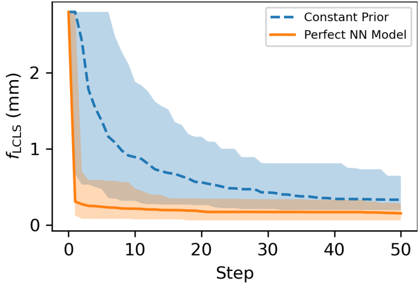
  9 layer NN trained on simulation data generated with IMPACT-T

- Minimize beam size of a round beam:

$$f_{\mathrm{LCLS}}(\boldsymbol{x}) = \sqrt{\sigma_x^2 + \sigma_y^2} + |\sigma_x - \sigma_y|$$

- Using a perfect prior mean model, the optimization problem is solved within a few steps



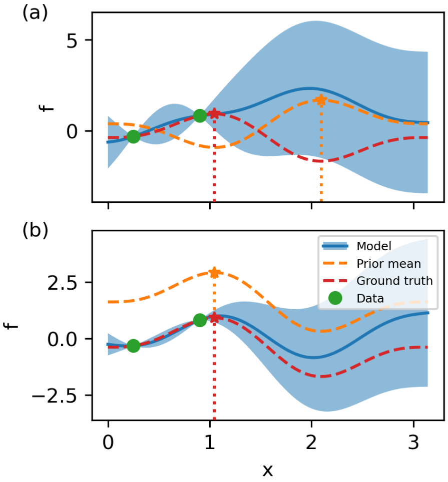


Metrics for Prior Mean Models

- NN models are commonly trained on absolute error metrics like MSE/MAE
- Low MSE/MAE may not translate to good predictions in the context of BO (a)
- Correlation can be a better metric as it captures the shape of the function (b)

$\Rightarrow$ Use combination of correlation and MAE to describe the model

$\Rightarrow$ Ideal metric remains an open question

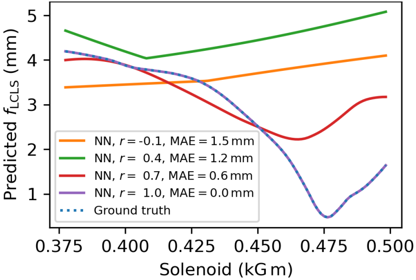


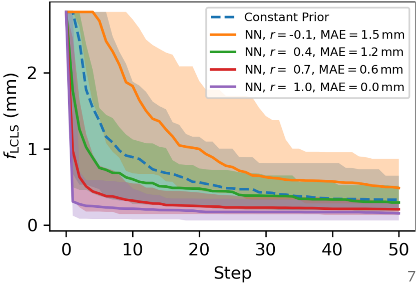
Simulations with LCLS Injector Surrogate Model

- Trained models with different levels of accuracy to test impact on BO
- Models with strong correlation improve initial performance and lead to better convergence
- Low or negative correlation can reduce performance below standard BO

⇒ Initial performance can be improved significantly

⇒ Better models lead to better performance

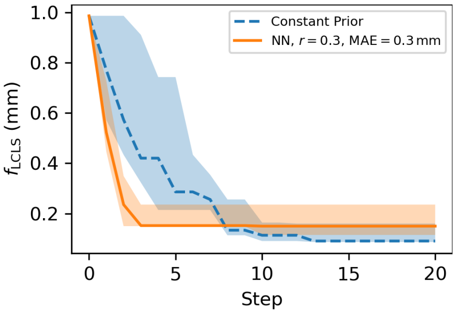




Experimental Results at LCLS

- Prior mean model consistently leads to better initial performance
- BO with constant prior mean eventually converges to better values

  $\Longrightarrow$ Probably due to the low model correlation
- Model calibration with additional linear layers for inputs and outputs

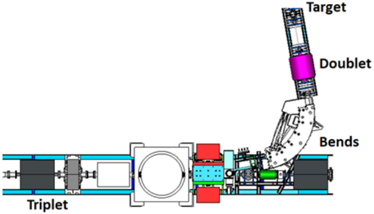


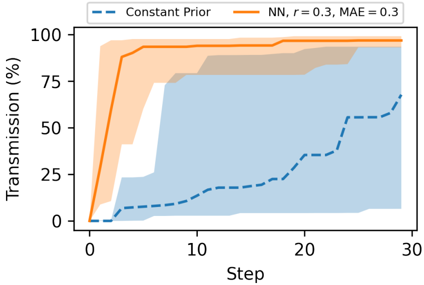
Science and Technology Facilities Council
ISIS Neutron and Muon Source

Experimental Results at ATLAS

- Optimize beam transmission while preserving overall beam quality
- Trained NN model on 3k samples from a previous experiment with a ^{14}N beam
- BO with NN prior model to optimize transmission for ^{16}O beam

$\Rightarrow$ Successful transfer learning!



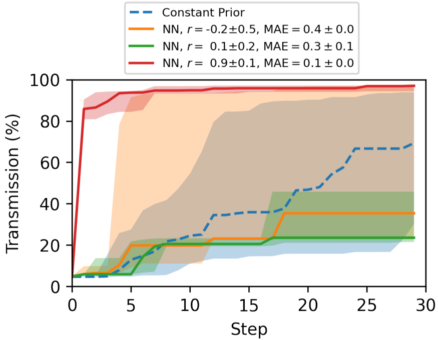


Experimental Results at ATLAS

- Trained models with different levels of accuracy to test impact on BO

- Experimental results also show BO performance depends on model quality

$\Rightarrow$ Performance with the same model can vary depending on which parts of the domain are sampled during a run



Low-Quality Prior Mean Models

- Obtaining models with high accuracy can be challenging in practice
- Convergence can suffer under the biased search with an inaccurate prior mean model
- Improve robustness by weighting NN model against a constant prior mean:

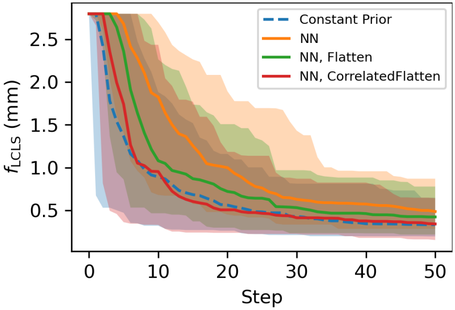
$$m'(\mathbf{x}) = w\, m(\mathbf{x}) + (1 - w)\,\text{const.}$$

$\Rightarrow$ "Flatten" prior mean as more steps are taken

- Weighting based on correlation:

$$w = \text{clip}(r - w_0, 0, 1)$$

$\Rightarrow$ Standard BO performance can be recovered



Summary

- NN priors are a flexible way to incorporate prior knowledge from different sources

  $\Rightarrow$ Enables incorporating large data sets into GPs
- Prior mean models can improve BO performance dramatically

  $\Rightarrow$ Successful demonstration at LCLS injector

  $\Rightarrow$ Successful demonstration at ATLAS (including transfer learning across different beam types!)
- Model accuracy and calibration are crucial (**see Eric's talk on Friday!**)
- Performance can be recovered if model quality is low

Outlook

- Application to constrained optimization
- Improved sample-efficiency allows scaling BO to high-dimensional problems

https://arxiv.org/abs/2403.03225

Questions?

Appendix

Calibration of LCLS Injector Model

- Calibration approach: linear transformation of individual inputs and outputs

$$y' = y_{\text{scale}}\,\text{model}\left(x_{\text{scale}}\,x + x_{\text{offset}}\right) + y_{\text{offset}}$$

- Linear approach helps to retain interpretability
- Regularization helps to get conservative estimates of the calibration parameters

Model	Correlation r[1]	MAE (mm)[1]
uncalibrated	0.56 ± 0.37	1.00 ± 0.31
low reg. ($w = 10^{-4}$)[2]	0.29 ± 0.18	2.13 ± 0.90
medium reg. ($w = 10^{-3}$)[3]	0.35 ± 0.21	0.54 ± 0.24
high reg. ($w = 10^{-2}$)[2]	0.20 ± 0.19	0.78 ± 0.27

Science and Technology Facilities Council
ISIS Neutron and Muon Source

[1] evaluated on 385 samples from different BO runs
[2] trained on 834 samples from previous BO runs
[3] trained on larger set of archived data with 36k samples