Mapping Gaussian Processes to Bayesian Neural Networks

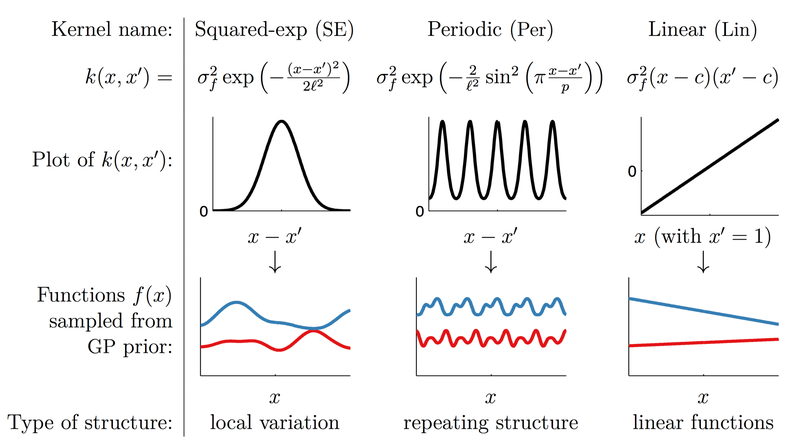
Daniel Flam-Shepherd [1], James Requeima [2 3], David Duvenaud [1]
1 University of Toronto, 2 University of Cambridge, 3 Invenia Labs

Priors in Function Space

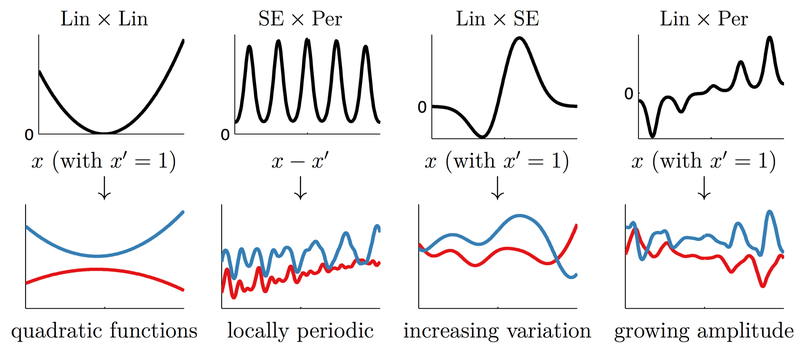
- **Baysian Neural Network Priors** are specified in parameter space. The implications of these priors in function space are hard to interpereate.
- How to we encorporate prior knowledge about function properties in our prior?

Gaussian Processes

Gaussian Processes have a elegant mechanism for incorporating prior beliefs about the underlying function - the mean and covariance functions.



Kernels can be combined using addition and multiplication to construct kernels with desired properties.



Mapping the Prior

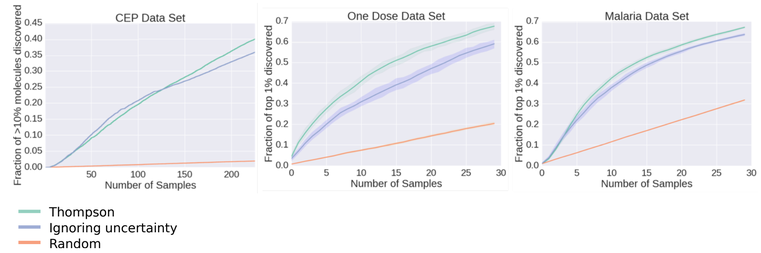
We apprioximate the **KL divergence** between the Gaussian process $p_{\mathsf{GP}}(\boldsymbol{f})$ and the BNN prior over functions $p_{\mathsf{BNN}}(\boldsymbol{f})$).

$$\begin{aligned}\mathcal{L}_{p(\boldsymbol{X})}(\boldsymbol{\phi}) &= \mathbb{E}_{p(\boldsymbol{X})}[\mathcal{KL}[p_{\mathsf{BNN}}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{\phi}) \mid p_{\mathsf{GP}}(\boldsymbol{f}(\boldsymbol{X}))]] \\ &= \mathbb{E}_{p(\boldsymbol{X})}[-\mathcal{H}[p_{\mathsf{BNN}}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{\phi})] \\ &\qquad -\mathbb{E}_{p_{\mathsf{BNN}}(\boldsymbol{f}|\boldsymbol{\phi})}[\log p_{\mathsf{GP}}(\boldsymbol{f}(\boldsymbol{X}))]]\end{aligned}$$

The second term in this expectation can be approximated using Monte Carlo. The entropy term can be approximated

1.

Results: Bayesian Neural Networks



Data sets:

- CEP: Harvard Clean Energy Project data, 2.3M molecules.
- One-dose: percentage cell growth relative to control, 27,000 molecules.
- Malaria: drug concentration giving half max response, 19,000 molecules.

Batch sizes: 500 (CEP) and 200 (Malaria and One-dose).

Comparison with ϵ-greedy sampling

Average rank and standard errors:

Method	Rank
$\epsilon = 0.01$	3.42±0.28
$\epsilon = 0.025$	3.02±0.25
$\epsilon = 0.05$	2.86±0.23
$\epsilon = 0.075$	3.20±0.26
Thompson	**2.51±0.20**

Conclusions

We have proposed a batch BO method that

- runs in a parallel and distributed manner.
- can handle large batch sizes and large molecule libraries.
- is comparable to non-scalable approaches (parallel EI) in small problems with GPs.
- outperforms other alternative scalable approaches in large scale settings with Bayesian neural networks.