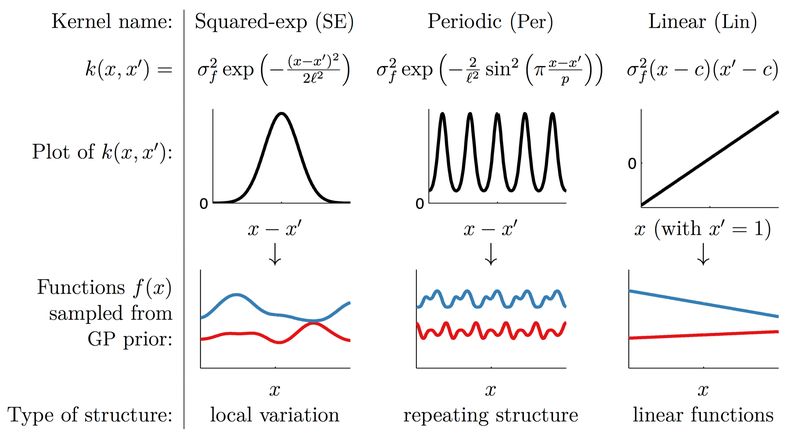
Mapping Gaussian Processes to Bayesian Neural Networks

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 are Interpretable

- **Baysian Neural Network Priors** are specified in parameter space. The implications of these priors in function space are hard to interpret.
- How to we incorporate prior knowledge about function properties in our prior?
- **Gaussian Processes** can elegantly incorporate prior beliefs about functions through the mean and covariance functions.



- Can we specify BNN priors using GP machinery?

Minimizing Divergence in Function Space

How can we find a prior on weights $p(w)$ that produces functions similar to a GP prior?

$$\boldsymbol{\phi}^* = \underset{\boldsymbol{\phi}}{\text{argmin}}\ \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\mathbb{KL}[p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X})|\boldsymbol{\phi})|p_{\text{GP}}(\boldsymbol{f}(\mathbf{X})]$$

$$\mathbb{KL}[p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X})|\boldsymbol{\phi})||p_{\text{GP}}(\boldsymbol{f}(\mathbf{X})] = -\mathbb{H}[p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X})|\boldsymbol{\phi})] - \mathbb{E}_{p_{\text{BNN}}(\boldsymbol{f}|\boldsymbol{\phi})}[\log p_{\text{GP}}(\boldsymbol{f}(\mathbf{X}))]$$

We can estimate the likelihood with simple Monte Carlo, but how can we estimate the entropy of p_{BNN}?

Estimating the Entropy of BNN Priors

1) Moment matching:

Approximate $p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X})|\boldsymbol{\phi}) \approx \mathbb{N}(\boldsymbol{f}|\boldsymbol{\mu}_{\boldsymbol{f}(\mathbf{X})},\ \boldsymbol{\Sigma}_{\boldsymbol{f}(\mathbf{X})})$ using empirical mean and covariance. In this case,

$$\mathbb{H}[p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X})|\boldsymbol{\phi})] \approx \frac{1}{2}\log|2\pi e\boldsymbol{\Sigma}_{\boldsymbol{f}(\mathbf{X})}|$$

Reasonable since $p_{\text{BNN}}(\boldsymbol{f}|\boldsymbol{\phi}) \to p_{\text{GP}}(\boldsymbol{f})$ over $\mathbf{X} \sim p(\mathbf{X})$.

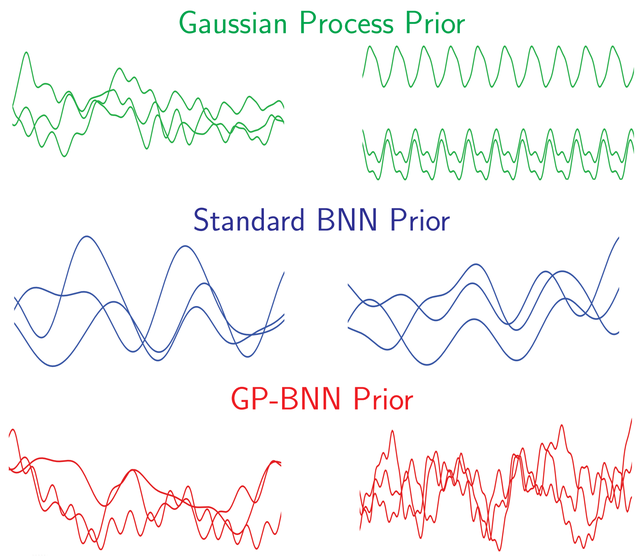
2) Nested Variational Bound:

Introduce noise $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\epsilon}$, and approximate $\log p(\mathbf{y}(\mathbf{X})|\boldsymbol{\phi})$ with a variational lower bound $\mathcal{L}(\boldsymbol{\lambda}, \mathbf{y}^{(s)}, \boldsymbol{X})$.

$$-\mathbb{H}[p_{\text{BNN}}(\mathbf{y}(\mathbf{X})|\boldsymbol{\phi})] \approx \frac{1}{S}\sum_{s=1}^{S}\log p_{\text{BNN}}(\mathbf{y}^{(s)}(\mathbf{X})|\boldsymbol{\phi}) \approx \frac{1}{S}\sum_{s=1}^{S}\max_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}, \mathbf{y}^{(s)}, \boldsymbol{X})$$

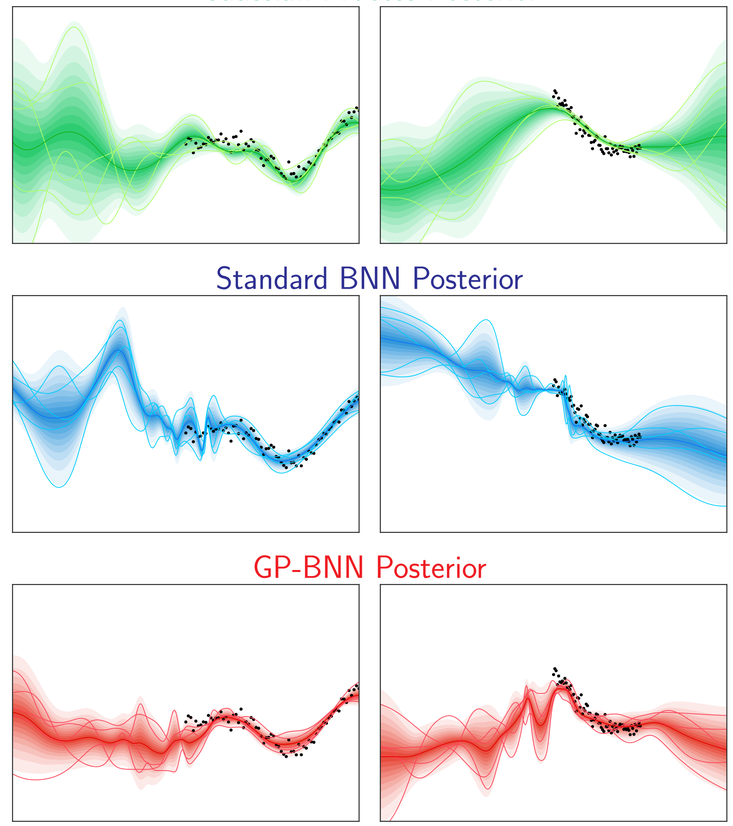
3) Early stopping to avoid mode collapse.

Samples from Approximate Priors



Posterior Results using GP Priors

We test our model (blue) on 2 different toy problems. All BNNs are 2 layer with rbf activation functions. The GP priors used RBF kernels.



Manifesto

- Specifying properties of functions should be the first decision, and computational architecture follows.
- Priors on functions are more interpretable than priors on parameters.