Mapping Gaussian Process priors to Bayesian Neural Networks

Daniel Flam-Shepherd, James Requeima and David Duvenaud

NIPS Workshop on Bayesian Deep Learning

The main objective

Minimize the KL divergence of prior $p_{\mathrm{GP}}(\boldsymbol{f}) \equiv \mathcal{GP}(\boldsymbol{f}|\mathbf{0}, \mathbf{K})$ to prior $p_{\mathrm{BNN}}(\boldsymbol{f}|\boldsymbol{\phi})$ in function space! Note $\boldsymbol{\phi}$ are variational parameters of $q(\mathbf{w}|\boldsymbol{\phi})$

$$\mathcal{K}(\boldsymbol{\phi}) = \mathbb{KL}[p_{\mathrm{BNN}}(\boldsymbol{f}|\boldsymbol{\phi}) \mid p_{\mathrm{GP}}(\boldsymbol{f})] \tag{1}$$

$$= -\mathbb{H}[p_{\mathrm{BNN}}(\boldsymbol{f}|\boldsymbol{\phi})] - \mathbb{E}_{p_{\mathrm{BNN}}(\boldsymbol{f}|\boldsymbol{\phi})}[\log p_{\mathrm{GP}}(\boldsymbol{f})] \tag{2}$$

But this is infinite! Approximate it by taking expectations over $p(\mathbf{X})$:

$$\mathcal{L}_{\mathbf{X}}(\boldsymbol{\phi}) \equiv \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}[\mathcal{K}(\boldsymbol{\phi})]$$

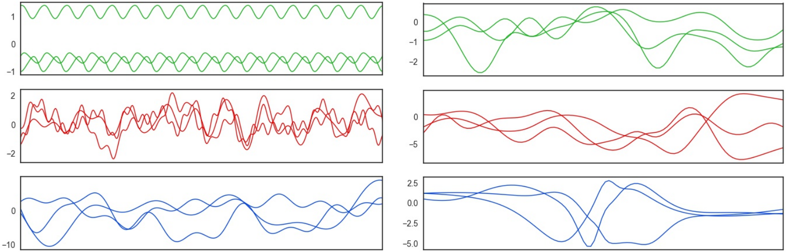
Assume normality of $p_{\mathrm{BNN}}(\boldsymbol{f}|\boldsymbol{\phi})$ and use MC estimation:

$$\mathcal{L}_{\mathbf{X}}(\boldsymbol{\phi}) \approx -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{f}}| - \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}[\log p_{\mathrm{GP}}(\boldsymbol{f}^{(s)}(\mathbf{X}))] \tag{3}$$

Now find $\boldsymbol{\phi}^* = \operatorname*{argmin}_{\boldsymbol{\phi}} \mathcal{L}_{\mathbf{X}}(\boldsymbol{\phi})$

Two Experiments with different kernels

Red are samples of $p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X})|\boldsymbol{\phi}^*)$ (optimized ones)! Green are samples from the GP prior and Blue are non-optimized samples of $p_{\text{BNN}}(\boldsymbol{f}(\mathbf{X}))$.



$k(x, x') = e^{-2\sin^2(\pi|x-x'|)/25}$

(a)

$k(x, x') = e^{-(x-x')^2/2}$

(b)

Check out our poster!!