In a similar setting where the last layer is linear that gets input $phi(x)$ we can write the predictive distribution in terms of a kernel. For a derivation, see . This kernel gives us how far away a point in the pool is from the dataset. Hence, points where the kernel function is low will have higher epistemic uncertainty.

Therefore, we try to learn the deep kernel that best describes the predictive distribution. Hence, goal is to minimize the KL divergence $\mathrm{KL}(p(y^*|x^*, X, Y)||q_\Theta(y^*|x^*))$ w.r.t. Θ where q is the model containing ϕ and the kernel k.

The kernel distribution that is learned is

$$\mathcal{N}(\sigma^{-2}\mathbf{Y}^T\big(\sigma^{-2}\mathbf{K}+s^{-2}I\big)^{-1}\mathbf{k}(X,x_*), s^2k(x_*,x_*)-s^2k(x_*,X)\big(\sigma^2s^{-2}I+\mathbf{K}\big)^{-1}k(X,x_*))$$

Derivation: weight-space posterior to kernel-space predictive

We have the predictive distribution for y_* at x_* in terms of posterior covariance Σ', posterior mean μ' as

$$p(y_* \mid x_*, X, Y) = \mathcal{N}\big(y_*;\ {\mu'}^T\phi(x_*),\ \sigma^2 + \phi(x_*)^T\Sigma'\phi(x_*)\big).$$

To write this in terms of a kernel k, write

$$ {\mu'}^T\phi(x_*) = \big(\sigma^{-2}\Sigma'\mathbf{\Phi}^T\mathbf{Y}\big)^T\phi(x_*) = \sigma^{-2}\mathbf{Y}^T\big(\Sigma'\mathbf{\Phi}^T\big)\phi(x_*). \tag{1}$$

Then apply the Woodbury identity to get

$$\Sigma'\mathbf{\Phi}^T = \big(s^{-2}I + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}\big)^{-1}\mathbf{\Phi}^T = s^2\mathbf{\Phi}^T\big(\sigma^2I + s^2\mathbf{K}\big)^{-1}. \tag{2}$$

Substituting (2) into (1) and using $\mathbf{k}(X, x_*) = \mathbf{\Phi}\phi(x_*)$ yields

$${\mu'}^T\phi(x_*) = \sigma^{-2}\mathbf{Y}^T\big[s^2\mathbf{\Phi}^T(\sigma^2I+s^2\mathbf{K})^{-1}\big]\phi(x_*) = \sigma^{-2}s^2\mathbf{Y}^T(\sigma^2I+s^2\mathbf{K})^{-1}\mathbf{k}(X,x_*).$$

So we obtain the kernel-space predictive mean

$${\mu'}^T\phi(x_*) = \sigma^{-2}\mathbf{Y}^T\big(\sigma^{-2}\mathbf{K} + s^{-2}I\big)^{-1}\mathbf{k}(X,x_*)$$

For the predictive variance start from

$$\phi(x_*)^T\Sigma'\phi(x_*).$$

Using the Woodbury-derived expansion for Σ' (one convenient form obtained from standard Woodbury algebra) is

$$\Sigma' = s^2I - s^4\mathbf{\Phi}^T\big(\sigma^2I + s^2\mathbf{K}\big)^{-1}\mathbf{\Phi}.$$

Hence

$$\begin{aligned}\phi(x_*)^T\Sigma'\phi(x_*) &= s^2\phi(x_*)^T\phi(x_*) - s^4\phi(x_*)^T\mathbf{\Phi}^T\big(\sigma^2I + s^2\mathbf{K}\big)^{-1}\mathbf{\Phi}\phi(x_*)\\ &= s^2k(x_*,x_*) - s^4\mathbf{k}(x_*,X)\big(\sigma^2I + s^2\mathbf{K}\big)^{-1}\mathbf{k}(X,x_*),\end{aligned}$$

where $\mathbf{k}(x_*, X) = \mathbf{k}(x_*, X)^\top = \mathbf{k}(x_*, X)^\top$ is just $\mathbf{k}(x_*, X) = \mathbf{k}(X, x_*)^T$. Pulling out an overall factor s^2 and rescaling the inverse gives the kernel-form

$$\phi(x_*)^T \Sigma' \phi(x_*) = s^2 k(x_*, x_*) - s^2 k(x_*, X)\big(\sigma^2 s^{-2} I + \mathbf{K}\big)^{-1} k(X, x_*) \; .$$

Further, find the derivation of the negative log-likelihood below.

When training the model, one tries to get as close to the predictive distribution. Therefore, the goal is to minimize the KL divergence $\mathrm{KL}(p(y^*|x^*, X, Y)||q_\Theta(y^*|x^*))$ w.r.t. Θ. Writing out

$$\mathrm{KL}(p(y^*|x^*, X, Y)||q_\Theta(y^*|x^*)) = \int p(y^*|x^*, X, Y)(\log p(y^*|x^*, X, Y) - \log q_\Theta(y^*|x^*))dy^*.$$

W.r.t. Θ, $\int p(y^*|x^*, X, Y) \log p(y^*|x^*, X, Y)dy^*$ is constant. Therefore, it is enough to minimize $\mathrm{NLL} = -\int p(y^*|x^*, X, Y) \log q_\Theta(y^*|x^*)dy^*$. For a given train set D this means minimizing the negative log likelihood $-\sum_{x,y \in D} \log(q_\Theta(y|x))$ for Θ as the real distribution is still independent of Θ.