Expressive yet Tractable Bayesian Deep Learning via Subnetwork Inference

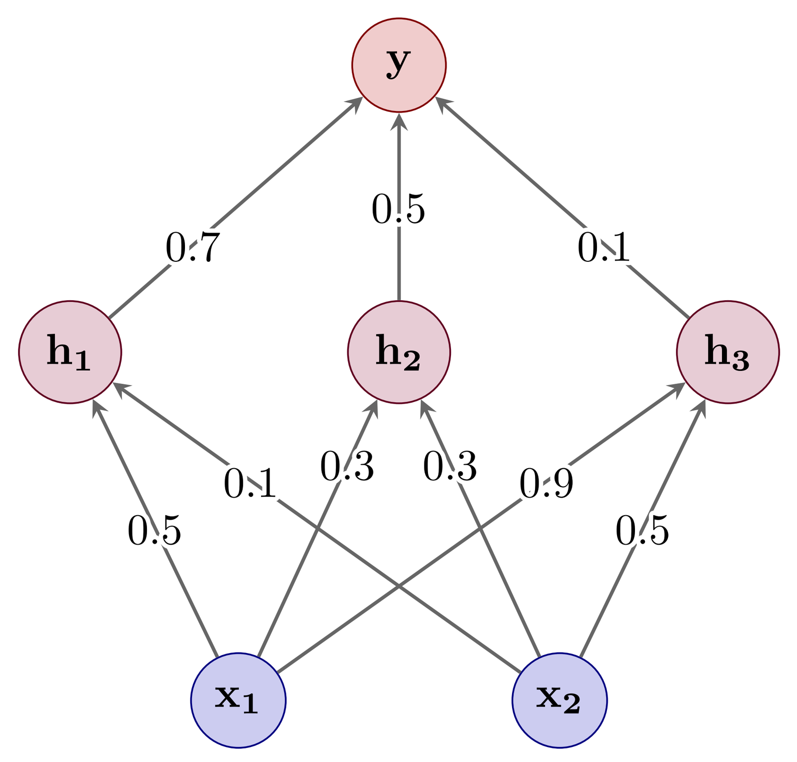
Erik Daxberger, Eric Nalisnick, James Urquhart Allingham, Javier Antorán, José Miguel Hernández-Lobato

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF AMSTERDAM

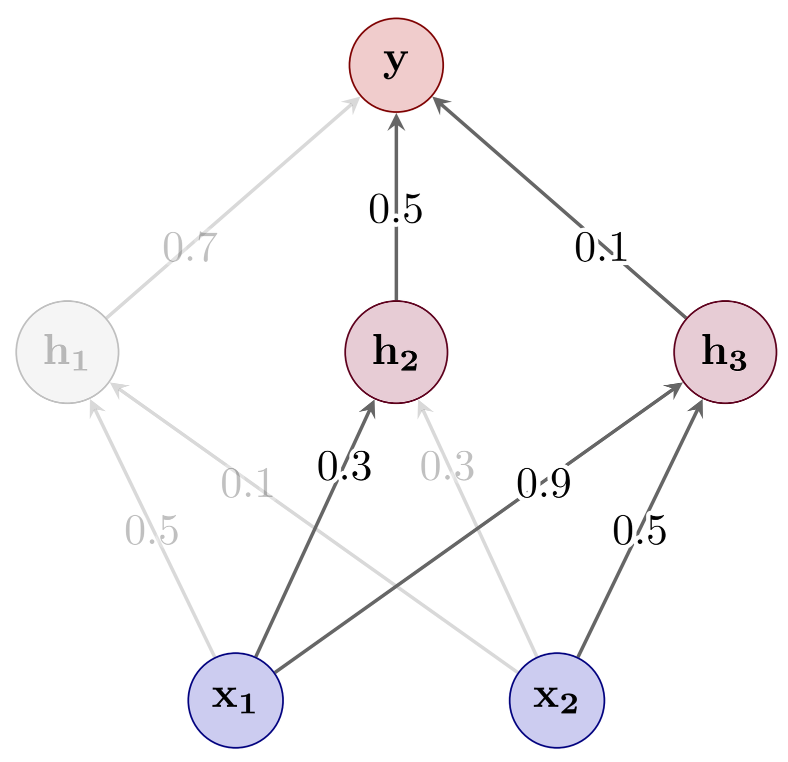
We propose a Bayesian deep learning method that does ***expressive inference*** over a carefully chosen ***subnetwork*** within a neural network, and show that this works better than doing crude inference over the full network.

1 MAP Estimation



Use SGD to obtain a **point estimate** over the weights: $\boldsymbol{W}_{MAP} = \arg\max_{\boldsymbol{W}} [\log p(\mathbf{y}|\boldsymbol{X}, \boldsymbol{W}) + \log p(\boldsymbol{W})]$

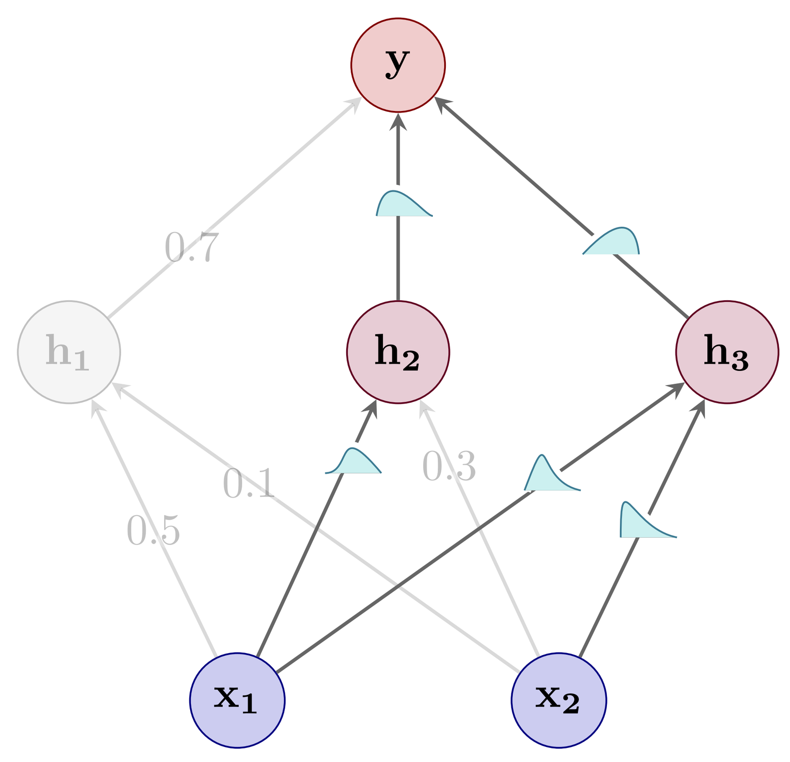
2 Subnetwork Selection



Find the subnetwork whose posterior is **closest to the full network posterior** in terms of Wasserstein distance:

1) Estimate a **factorized Gaussian** posterior over all weights
2) Subnetwork = weights with **largest marginal variances**

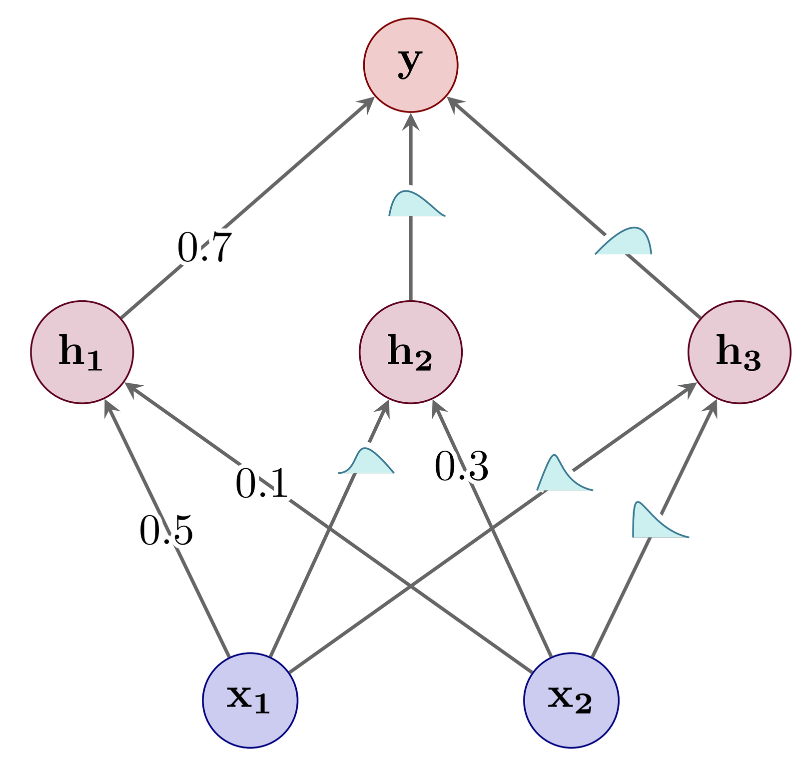
3 Bayesian Inference



Use the linearized **Laplace approximation** to infer a **full-covariance** Gaussian posterior over the subnet. All other weights are fixed to their **MAP estimates.**

$$p(\mathbf{W}|\boldsymbol{y}, \boldsymbol{X}) \approx \mathcal{N}(\mathbf{W}_S; \boldsymbol{W}^S_{MAP}, \widetilde{H}^{-1}) \prod_r \delta(\mathrm{w}_r - \mathrm{w}_r^*)$$

4 Prediction



Make predictions using the **full network** of mixed Bayesian/deterministic weights

$$p(\mathbf{y}^*|\boldsymbol{X}^*, \boldsymbol{y}, \boldsymbol{X}) \approx \int_{\mathbf{W}} p(\mathbf{y}^*|\boldsymbol{X}^*, \boldsymbol{W}) \, \mathcal{N}(\mathbf{W}_S; \boldsymbol{W}^S_{MAP}, \widetilde{H}^{-1}) \prod_r \delta(\mathrm{w}_r - \mathrm{w}_r^*) \, d\mathbf{W}$$