Concrete Dropout

MLSALT4 Paper Replication Exercise

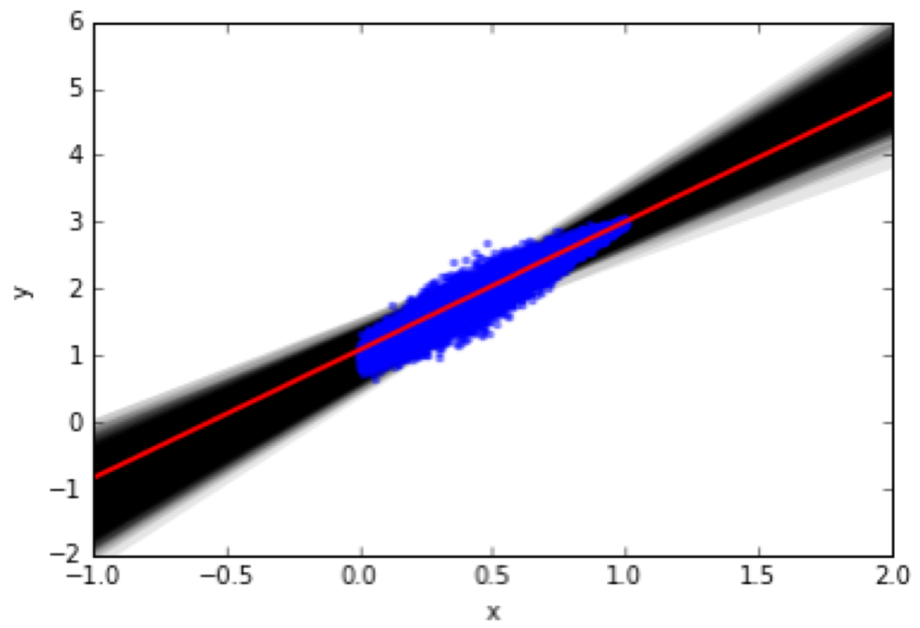
L. Chai, F. Ding, L. Foglianti Spadini, P. L. Tan

UNIVERSITY OF CAMBRIDGE

What is uncertainty?

Not knowing which line is the true model creates **epistemic uncertainty** in **y**

Inherent unpredictable noise creates **aleatoric uncertainty** in **y**



Dropout can measure uncertainty

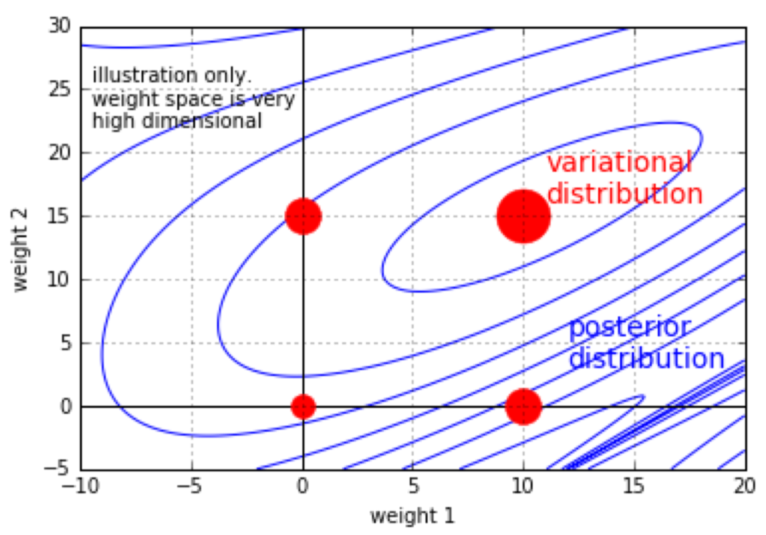
Gal and Ghahramani (2015) reinterpreted dropout regularisation as variational inference in BNNs

Posterior distribution $p(\omega|D)$ is approximated as

$$q_\theta(\omega) = \sum_{z \sim Bernoulli} p(z)\delta(\omega = W)$$

where

$$W = M \cdot diag(z)$$



Variational distribution is a hypercube of delta peaks in weight space. It is parameterised by furthest corner of cube from origin, M, and dropout probability, p. An optimal p is a proxy measure of **epistemic uncertainty**.

Optimal variational distribution found by minimising

$$\mathcal{L}(M,p) = KL[q_{M,p}(\omega)||p(\omega|D)]$$

$$= \underbrace{KL[q_{M,p}(\omega)||p(\omega)]}_{\textbf{Regulariser}} - \underbrace{\mathbb{E}_{q(\omega)}[\log p(D|\omega)] + \log p(D)}_{\textbf{MLE Loss}}$$

In a single layer of size K, for a Gaussian prior with variance l^{-2}, the regulariser is approximated as:

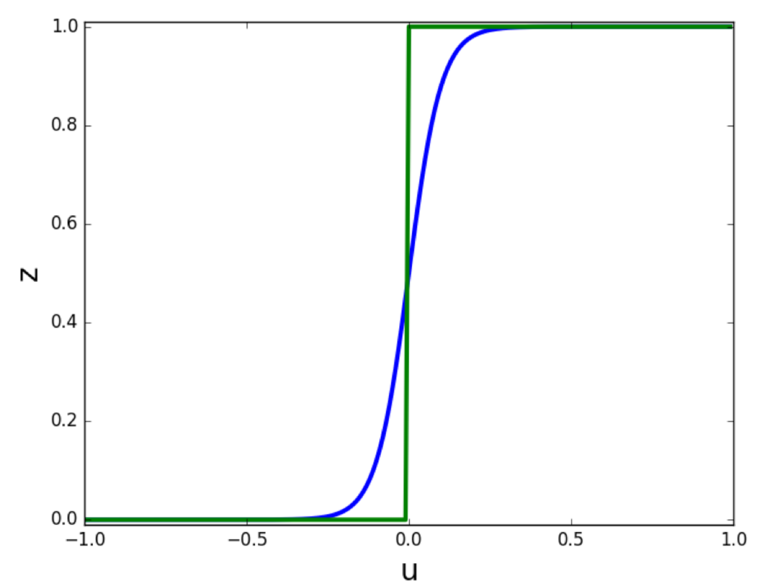
$$KL[q_{M,p}(\omega)||p(\omega)] \approx \frac{l^2(1-p)}{2}||M||^2 - KH(p)$$

Problems with tuning dropout p

Grid search over dropout probability is expensive:

- Wastes computing resources and experimental time
- Exponential increase in number of dropout configurations with number of NN layers
- In RL, dropout p should decrease as more data becomes available

Learning dropout probabilities



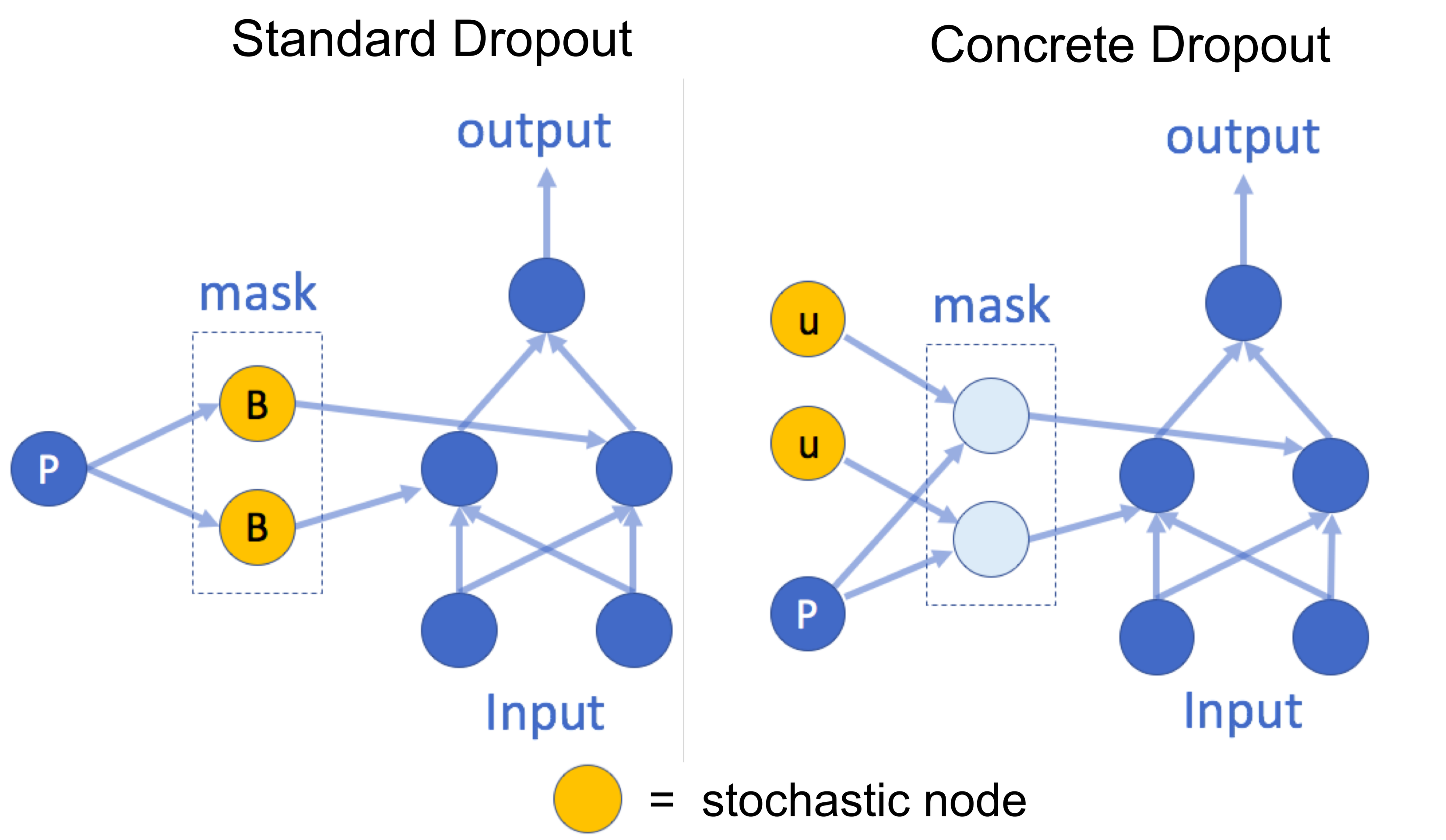
The **concrete distribution** is a relaxation of a categorical distribution onto the simplex.

In 1D, $z{\sim}Bernoulli(1-p)$ becomes

$$z \sim Sigmoid\left(\frac{\log\left(\frac{p}{1-p}\right)+\log\left(\frac{u}{1-u}\right)}{T}\right)$$

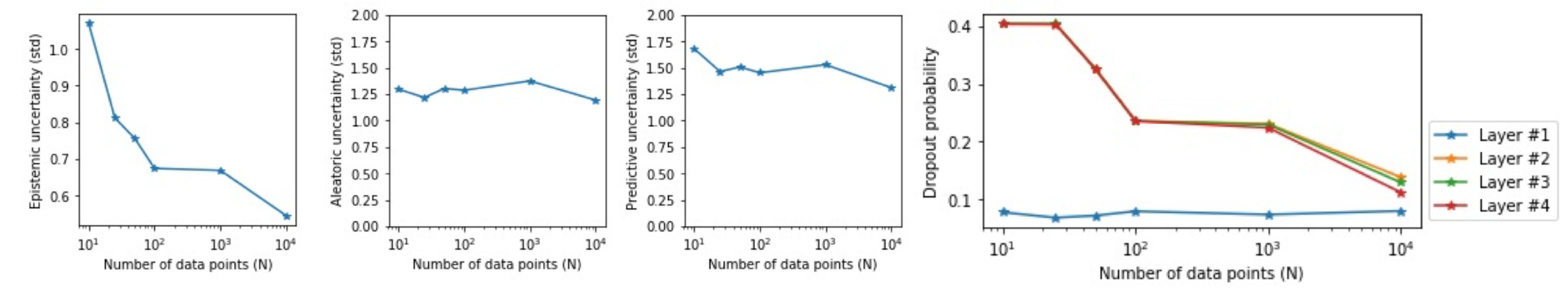
where $u \sim Uniform\,(0,1)$

- Dropout mask is now a **smooth deterministic function** of p and u
- Stochasticity is moved from the mask to uniform noise
- Gradient can freely flow through the network during backprop
- This is also called the "Reparameterisation Trick"

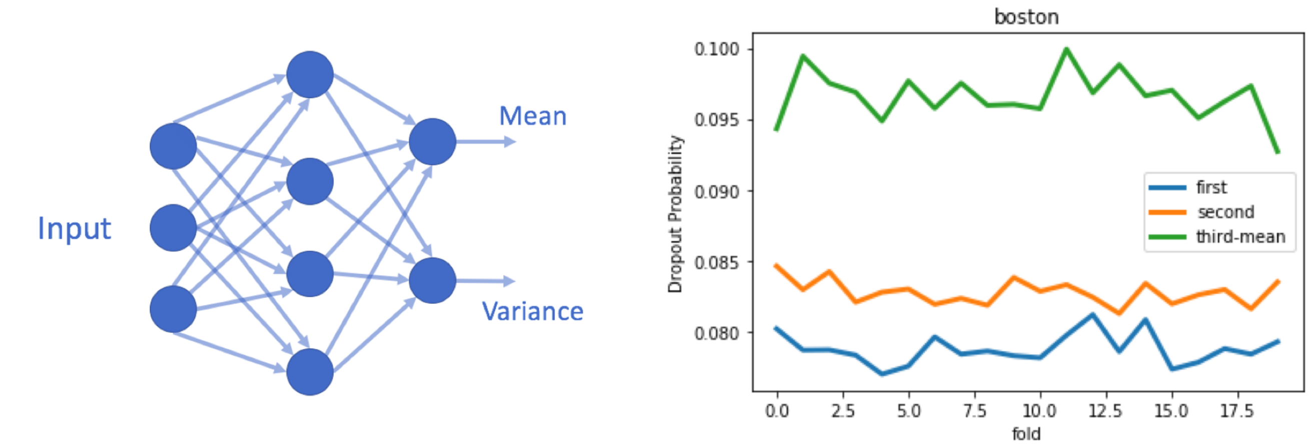


Experimental results

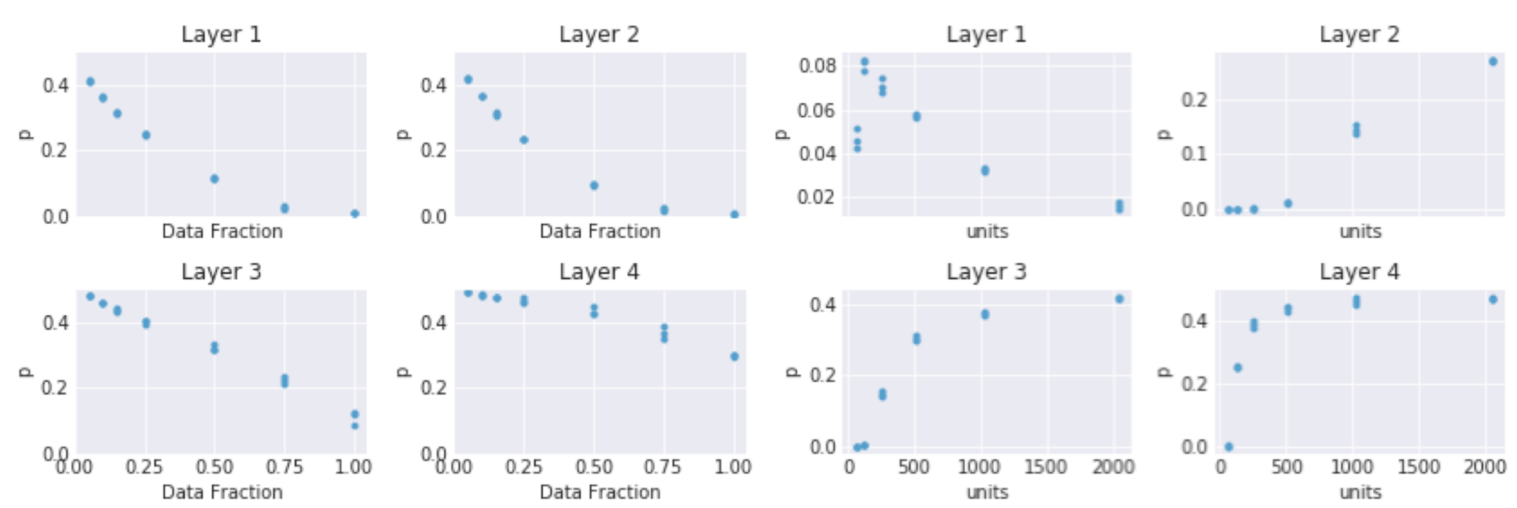
- Synthetic Data
  - 1-D linear regression model: $y = 2x + 8 + \epsilon$
  - Dropout probability decreases with more data points



- UCI Datasets
  - Dropout probability increases with depth



- MNIST
  - Dropout probability as a function of training set size (left; 3x512 MLP) and number of hidden units (right)



Proposed extensions

- Dropout can **reduce overfitting in RNNs**
- We propose applying concrete dropout to LSTMs/GRUs
- Possible architectures include input layer dropout, recurrent layer dropout, and combining the two

Gal, Y., Hron, J., & Kendall, A. (2017). Concrete dropout. In *Advances in Neural Information Processing Systems* (pp. 3584-3593).

Gal, Y., & Ghahramani, Z. (2015). Dropout as a Bayesian approximation: Insights and applications. In *Deep Learning Workshop, ICML* (Vol. 1, p. 2).