Stabilising priors for robust Bayesian deep learning

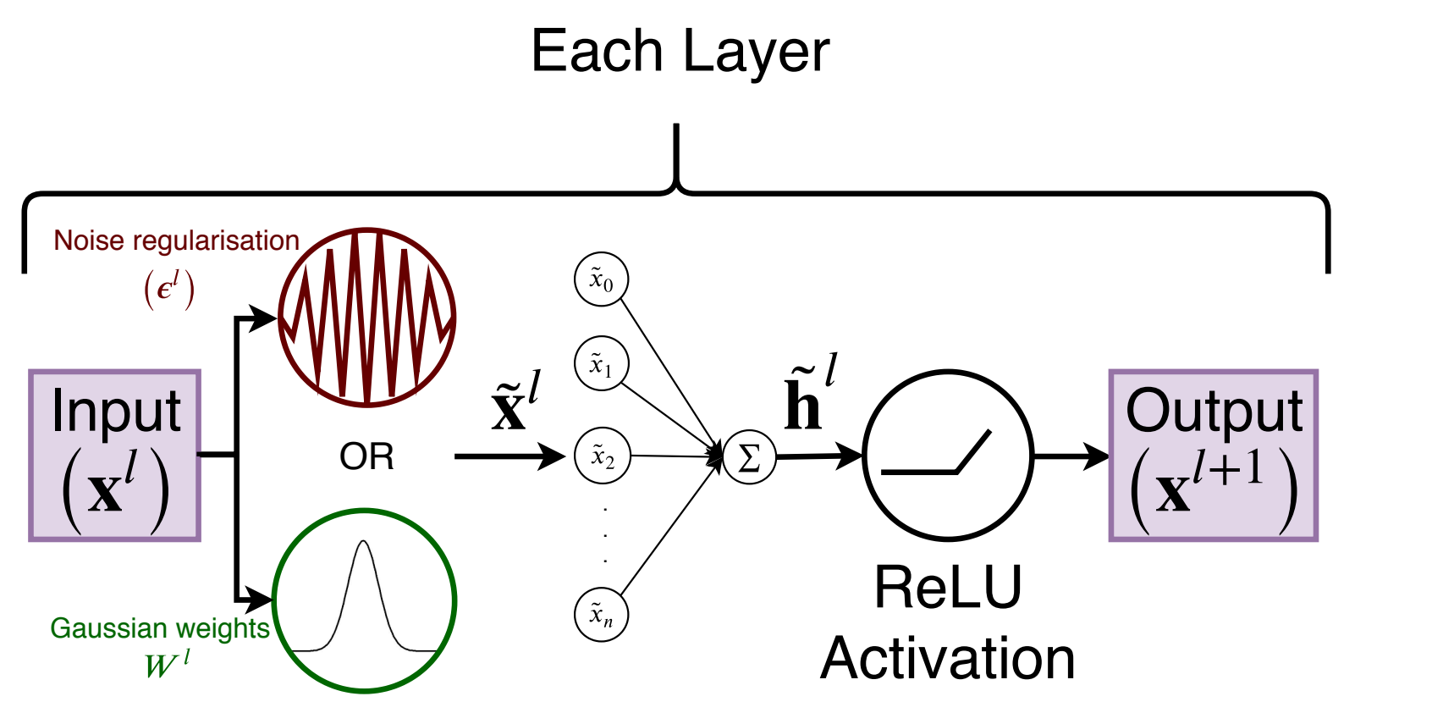
Felix McGregor* Arnu Pretorius† Johan du Preez* Steve Kroon†

*E&E Engineering, †Computer Science, Stellenbosch University, South Africa

Contributions

- We extend noisy signal propagation theory developed by Pretorius et al. (2018), to describe signal propagation in Bayesian neural networks (BNNs) in order to gain a theoretical understanding of the challenges of training deep stochastic networks.
- Signal propagation analysis of BNNs leads us to propose a *self-stabilising prior* where prior hyperparameters are derived to be optimal in the sense that they promote stable signal propagation.
- We derive a novel ELBO that allows the prior to affect the network during the forward pass.

1. Signal propagation in Bayesian Neural Networks (BNNs)



We study signal propagation of an input $\mathbf{x}^0 \in \mathbb{R}^{D_0}$ to a deep fully connected BNN with ReLU activations, where posterior distributions over weights are fully factorised Gaussian distributions $p(w_{ij}) \sim \mathcal{N}(\mu_q, \sigma_q^2)$ $W^l \in \mathbb{R}^{D_l \times D_{l-1}}$ sampled at each layer $l = 1, \ldots, L$.

We use the mean field assumption and Central Limit Theorem to approximate the distribution of the pre-activations at each layer l, in the large width limit, with a Gaussian. In expectation of the network parameters in a ReLU network, and assuming the input to each layer has zero mean, the variance ν^l governing the signal propagation dynamics of a BNN at a pre-activation at layer l becomes

$$\nu^l_j = \left[\left(1-\frac{1}{\pi}\right)(\mu^l_{\tilde{q}_j})^2 + (\sigma^l_{\tilde{q}_j})^2\right]\frac{\nu^{l-1}}{2}. \qquad (1)$$

2. Reformulating the ELBO

The prior usually impacts the ELBO through an additive KL term, affecting the weights at update time. These updates only take place after a completed forward pass, having no effect on the signal propagation dynamics of the network. We combine the prior and approximating posterior as $\tilde{q}_{\{\alpha,\phi\}}(W) = p_\alpha(W)q_\phi(W)/Z$, where Z is the normalising constant. This ensures that the sampled weights of the network are being influenced by the current prior during the forward pass. Then, we can construct a lower bound

$$\mathcal{L}_{\tilde{q}} = \mathbb{E}_{p(\epsilon)}\left[\log p(\mathbf{y}|\mathbf{x}, \mathbf{b}, W = \xi(\epsilon, \alpha, \phi))\right] - \mathrm{KL}(\tilde{q}_{\{\alpha,\phi\}}(W)||p_\alpha(W)).$$

3. Self-stabilising priors

To stabilise signal propagation we want to *preserve the variance during the forward pass*, or set $\nu^l_j = \nu^{l-1}_j$ where $j \in \{1, \ldots, D_{l-1}\}$.

- Setting variances equal defines the condition for BNN pre-activations with stable signal propagation given by

$$(1-1/\pi)(\mu^l_{\tilde{q}_j})^2 + (\sigma^l_{\tilde{q}_j})^2 = 2.$$

- Optimal prior parameters to stabilise a signal are then given by

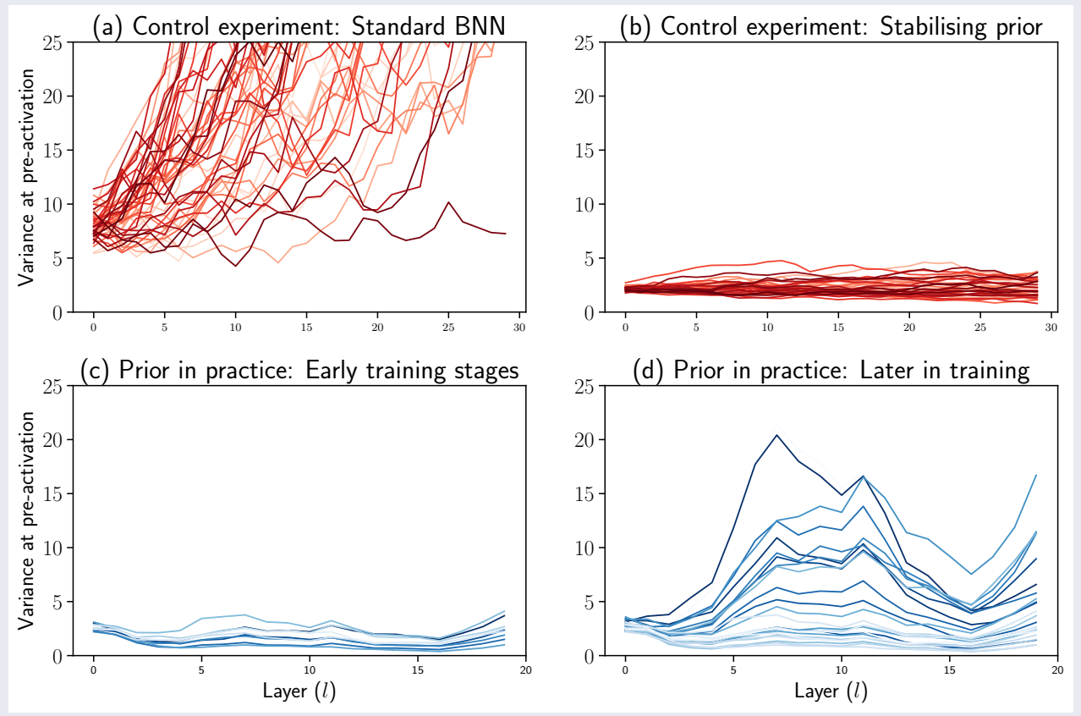
$$\mu^l_{p_{ij}} = \mu^l_{q_{ij}}, \quad \sigma^l_{p_j} = \sqrt{\frac{(\sigma^l_{\tilde{q}_j})^2\gamma}{(\sigma^l_{\tilde{q}_j})^2 - \gamma}},$$

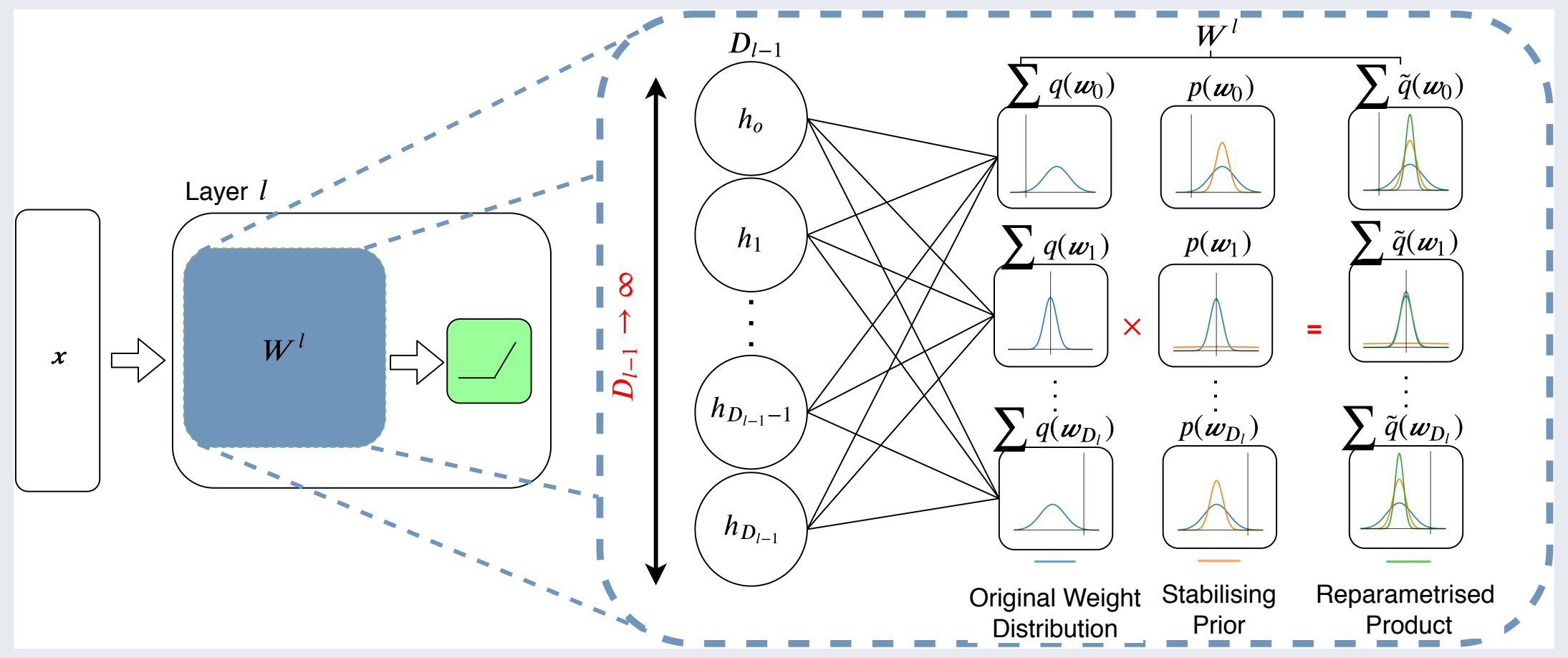
where $\gamma = |2 - (1-1/\pi)(\mu^l_{\tilde{q}_j})^2|$.

Sampling $w \sim \tilde{q}(W)$ at each forward pass, while setting the prior $p(w_{ij}) = \mathcal{N}(\mu^l_q, |(\sigma^l_q)^2\gamma/((\sigma^l_q)^2 - \gamma)|/D_{l-1})$, enables the network to simultaneously update our current posterior as well as promote stable signal propagation.

Intuition and Discussion

- In a controlled setting we can achieve perfectly stable signal propagation. In practice our assumptions hold for the early stages of training.
- The larger the mean and variance of the incoming weights, the more likely it is to destroy the signal. The prior becomes active when the second moment of the distribution is large which then urges the weights to sample closer to their means.





4. Large scale experiments

BNNs become untrainable at a certain depth. We observe our stabilising prior makes it possible to train deeper BNNs and in more noisy conditions.

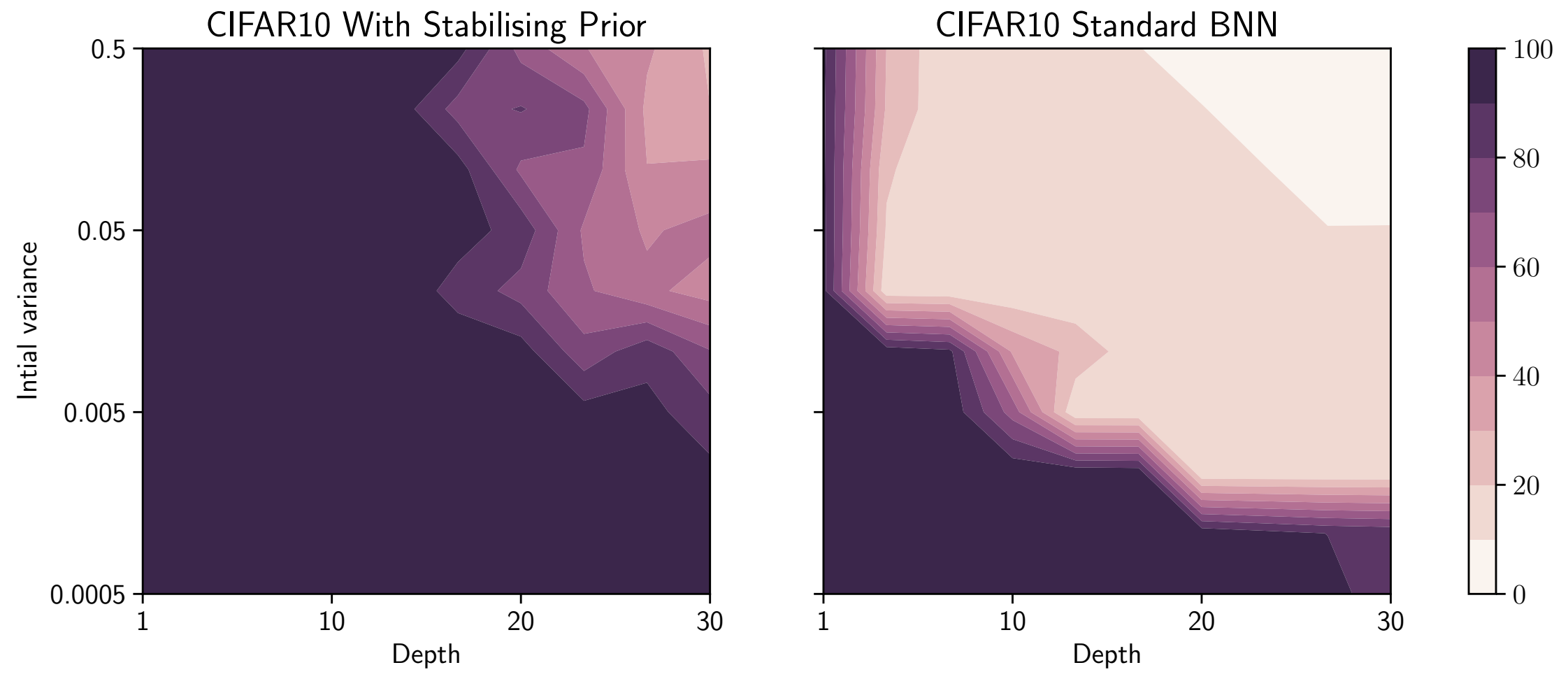
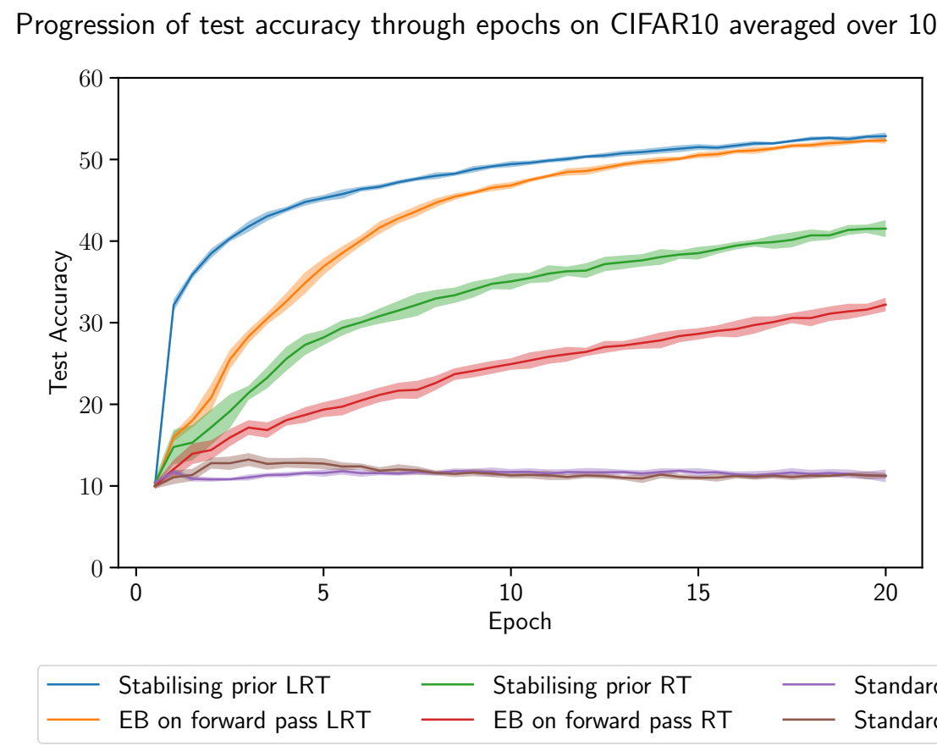


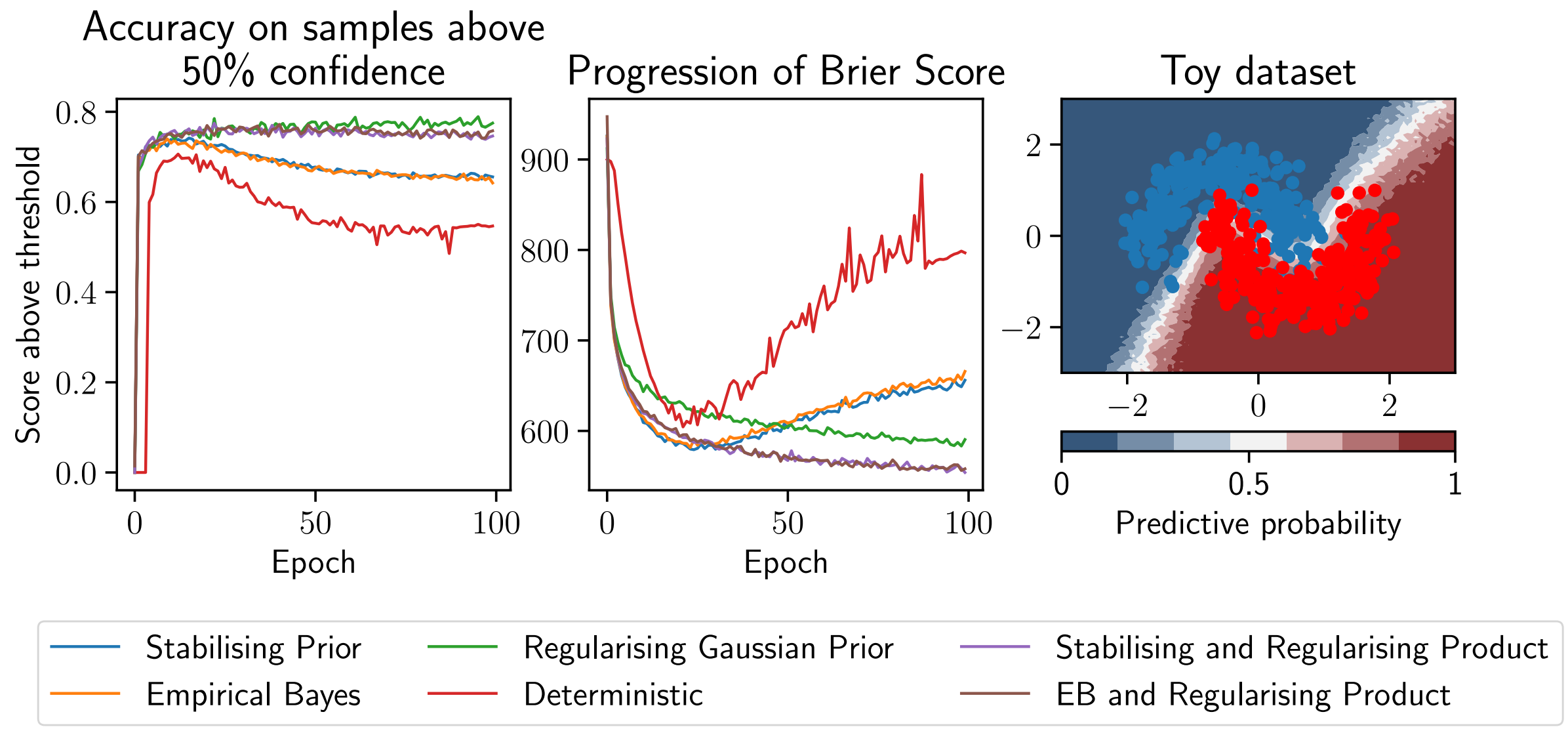
Figure 1: Classification accuracy grid of ReLU networks trained on CIFAR-10 with varying depths and initial variance conditions.

5. Improved convergence and effect on calibration

In general, we also observe that our prior increases the training speed.



As with EB and the stabilising prior, iteratively updating priors may adapt to the dataset and overfit. We combine a regularising and stabilising prior which trains faster and results in a well calibrated model with better Briers cores than any solitary prior.



Takeaways

- We examined BNN signal propagation dynamics, and used this result in combination with a novel ELBO to derive a self-stabilising prior. The prior essentially incorporates knowledge of model architecture and activation function, derived from how signals propagate in the network in the infinite width limit to promote stable signal propagation.
- Stabilising priors offer an attractive alternative prior for neural networks where designing meaningful priors is invariably obscure. They exhibit improved convergence and are often essential in deep BNNs to allow any training to occur i.e. to enable the signal to reach the outputs exhibit improved convergence.

Acknowledgements

We would like to thank the Technology and Human Resource for Industry Programme (THRIP) as well as Centre for Artificial Intelligence Research (CAIR) for financial support. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] A. Pretorius, E. Van Biljon, S. Kroon, and H. Kamper, Critical initialisation for deep signal propagation in noisy rectifier neural networks. NeurIPS, 2018.

[2] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. NeurIPS, 2015.

[3] M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. ICML, 2014.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

Code: https://github.com/felixmcgregor/Bayesian-Neural-Networks-with-self-stabilising-priors