Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations

Sebastian Farquhar, Lewis Smith, Yarin Gal {first.last@cs.ox.ac.uk}

Is mean-field a bad variational approximation?

- People often assume that the commonly used **mean-field approximation** (independent weights) is bad in variational inference for Bayesian Neural Networks.
- We challenge this. Instead, in **deeper models the mean-field approximation is fine.**

“diagonal approximation is ~~no good~~ fine in deep models because of the ~~strong~~ weak posterior correlations in at least one mode of the parameters.”

– David MacKay, 1992

- This challenges a foundational assumption that **motivates much research** (e.g., Louizos and Welling 2016; Sun et al. 2017; Oh et al. 2019).

Implications: Solve Difficulties of Scaling MFVI. Don’t Be Fancy.

- In bigger networks, MFVI gets better. So we must **resolve problems scaling MFVI**. E.g., gradient variance, computational cost.
- We **may not need complex posterior approximations**.
- **Architecture matters** for Bayesian approximations. Can’t evaluate methods meant for deep models in shallow models.

Empirically: Mean-field Assumption Is Better In Deep Models

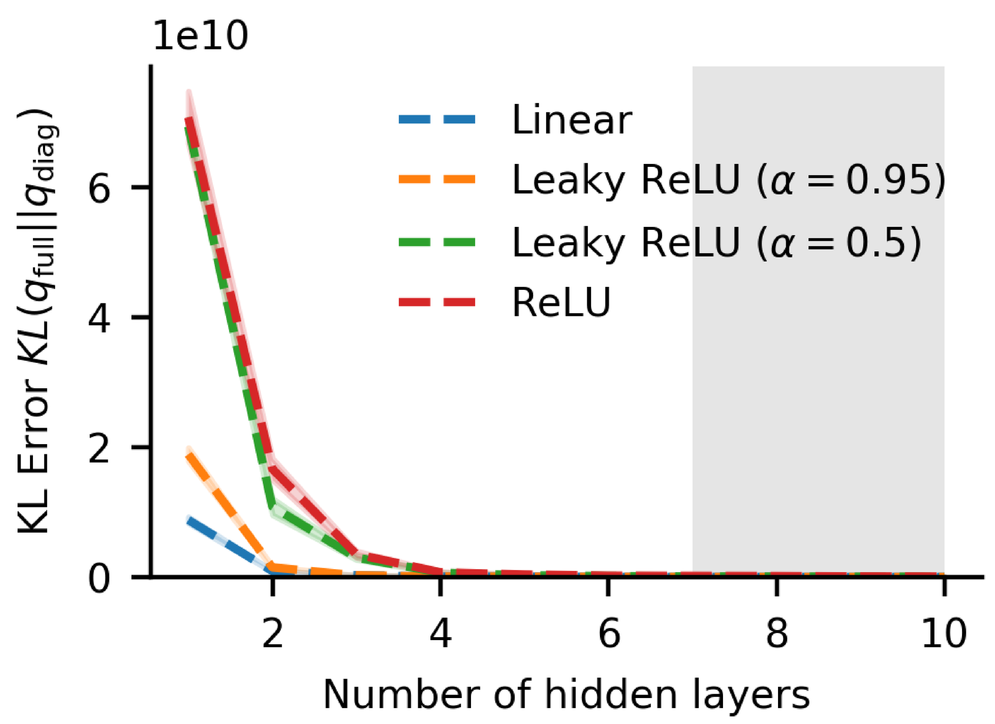


Fig 1. As depth increases, for fixed num. params., we lose less information by using a diagonal approx. instead of full-covariance to HMC posterior samples.

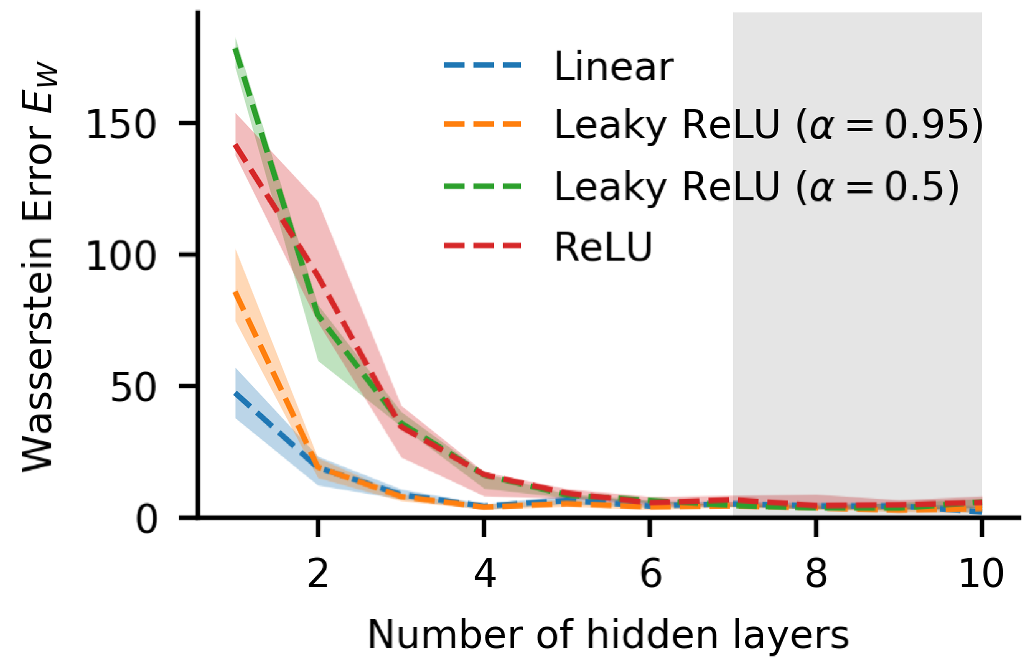


Fig 2. Similarly, the Wasserstein-2 distance between the diagonal approx. posterior and HMC samples falls in deeper models.

Arch.	Method	Covariance	Acc.	NLL	ECE
ResNet -18	VOGN‡	Diagonal	67.4%	1.37	0.029
ResNet -18	Noisy K-FAC††	MVG	66.4%	1.44	0.080
DenseNet -161	SWAG -Diag†	Diagonal	78.6%	0.86	0.046
DenseNet -161	SWAG†	Low-rank	78.6%	0.83	0.020
ResNet -152	SWAG -Diag†	Diagonal	80.0%	0.86	0.057
ResNet -152	SWAG†	Low-rank	79.1%	0.82	0.028

Fig 3. On Imagenet, in deep models there is no clear advantage to complex covariance approximations over mean-field (diagonal).

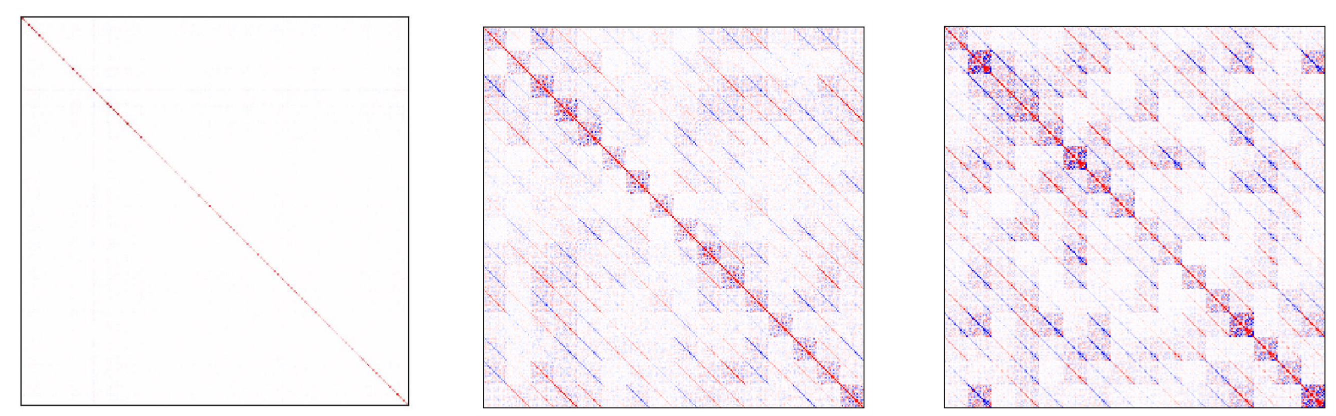
Theoretically: Depth ‘Simulates’ Correlation and True Posterior Has At Least One ‘Approximately Mean-field’ Mode

We introduce a new tool: ‘**local product matrices**’ for analysis piecewise linear models via linear ones. In linear case, we can ‘flatten’ a deep model with matrix multiplication and look at correlations in the ‘product matrix’. This lets us analyse the **function output distribution** induced by weights. Diagonal weights induce complex covariance in product.

$$\mathbf{o} = \begin{pmatrix} W^{(2)}_{11} & W^{(2)}_{12} \\ W^{(2)}_{21} & W^{(2)}_{22} \end{pmatrix} \mathbf{x}$$
$$= \begin{pmatrix} B_{11}A_{11} + B_{12}A_{21} & B_{11}A_{12} + B_{12}A_{22} \\ B_{21}A_{11} + B_{22}A_{21} & B_{21}A_{12} + B_{22}A_{22} \end{pmatrix} \mathbf{x}$$

- **Prop 1:** 3+ layers gives off-diagonal covariance anywhere in product matrix.
- **Prop 2:** MVG/K-FAC is special case of 3 mean-field layers.

Fig 4. Visualizing covariance of product matrix of K mean-field Gaussian layers trained on FashionMNIST. Starts diagonal. Develops covariance.



(a) 1-layer. **(b)** 5-layers. **(c)** 10-layers.

- **Prop 3:** Extends Prop 1 to ‘local product matrix’ for piecewise non-linearities like Leaky ReLU.
- **Prop 4:** A mean-field network can approximate true posterior density arbitrarily closely.

This research was supported by The Alan Turing Institute and the EPSRC grant number EP/P00881X/1 at the Centre for Doctoral Training in Autonomous Intelligent Machines and Systems and Cyber Security at the University of Oxford.