IBM Research

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

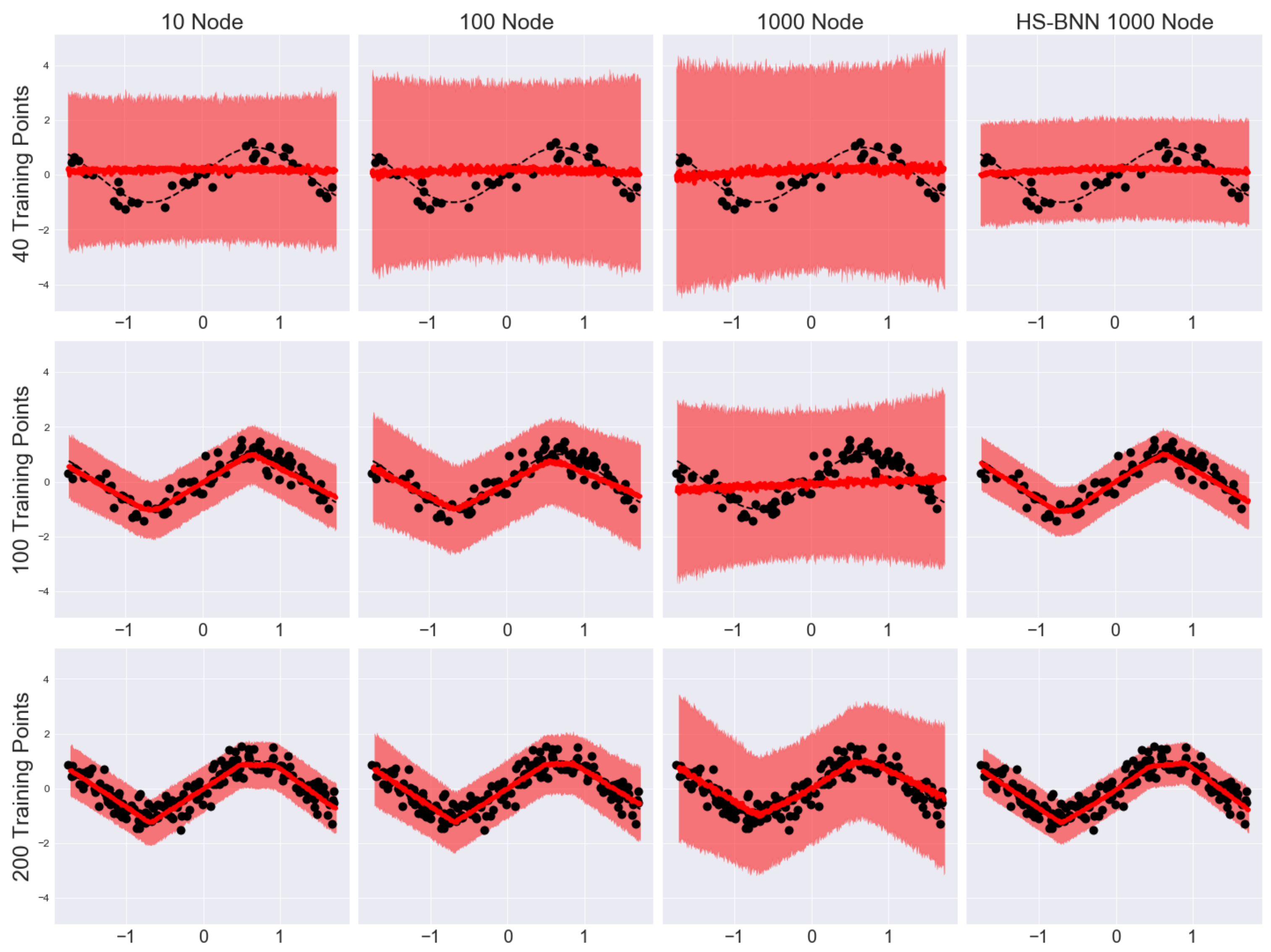
Model Selection in Bayesian Neural Networks via Horseshoe Priors

Soumya Ghosh
IBM Research, Cambridge

Finale Doshi-Velez
Harvard University

Model Selection in BNNs

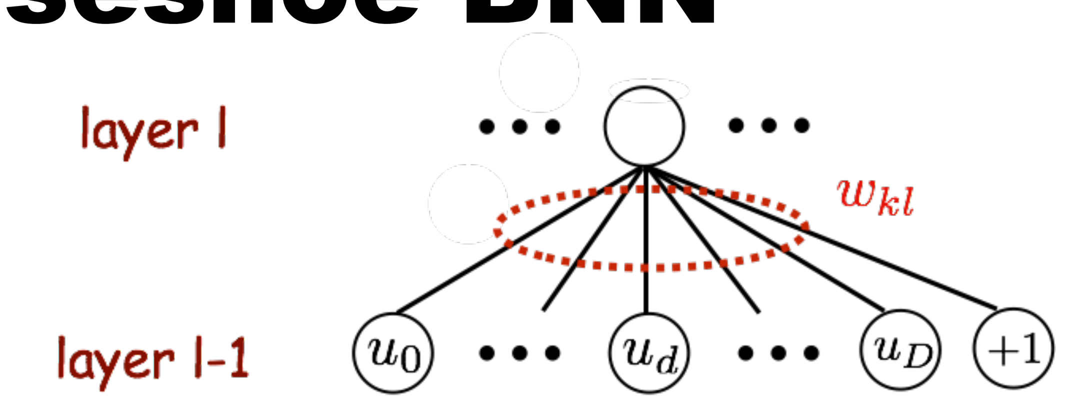
- *Bayesian NNs with large capacity & insufficient data can underfit, have large predictive variances.*



** BNNs have unit normal prior on weights, all models have Gaussian output noise:* $\mathcal{N}(y \mid f(x; \mathcal{W}), \gamma^{-1})$
** Thirty random inits, highest ELBO solution is visualized.*

- *We develop BNNs with group Horseshoe priors to prune away additional capacity.*
- *Utilize alternate parameterizations necessary for effective inference.*
- *Develop variants that nearly halve training time and storage requirements*

Horseshoe BNN



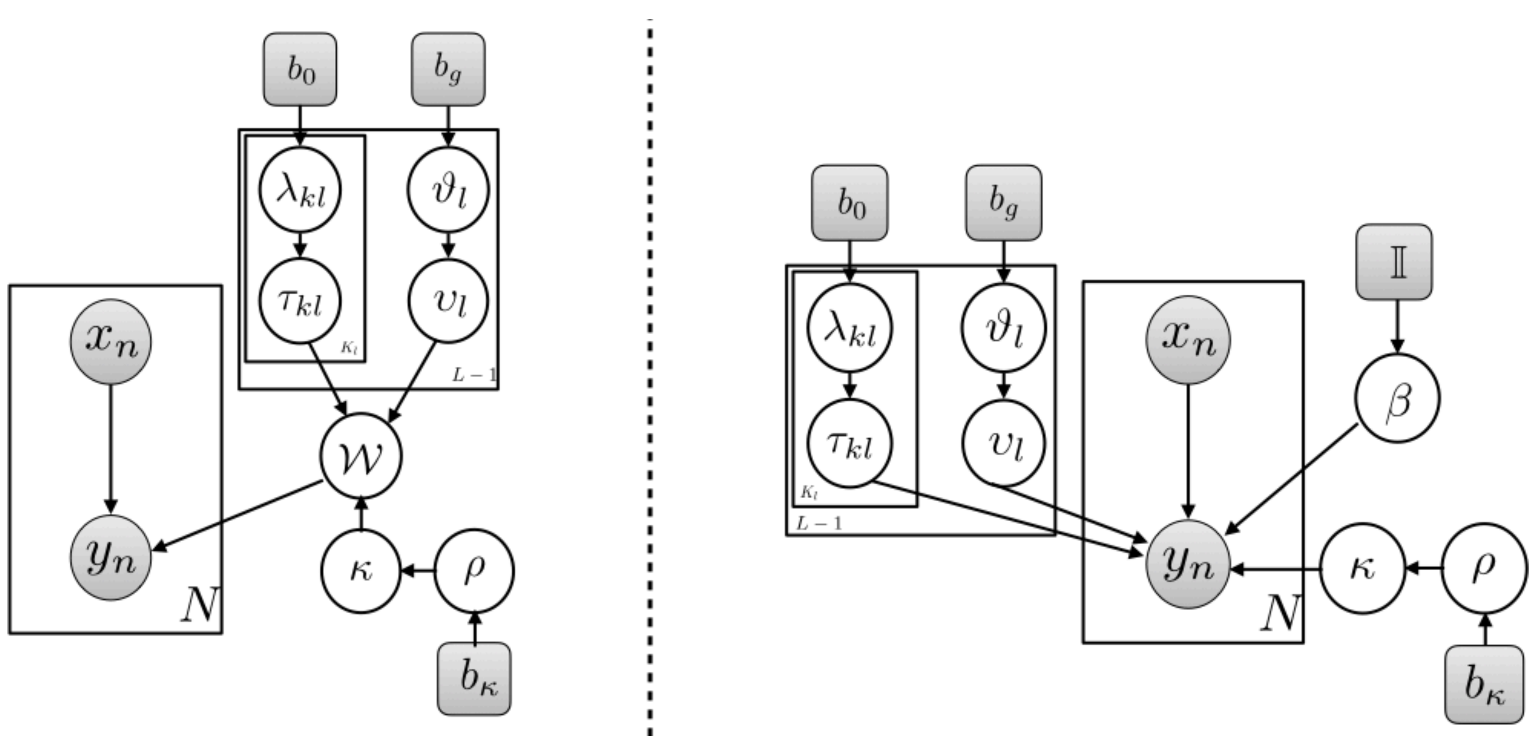
$$w_{kl} \mid \tau_{kl}, \upsilon_l \sim \mathcal{N}(0, (\tau_{kl}^2 \upsilon_l^2)\mathbb{I}),$$
$$\tau_{kl} \sim C^+(0, b_0), \quad \upsilon_l \sim C^+(0, b_g).$$

Inverse Gamma Parameterization

$$a \sim C^+(0, b) \iff a^2 \mid \lambda \sim \text{Inv-Gamma}(\frac{1}{2}, \frac{1}{\lambda});$$
$$\lambda \sim \text{Inv-Gamma}(\frac{1}{2}, \frac{1}{b^2}),$$

Non-centered Parameterization

$$\beta_{kl} \sim \mathcal{N}(0, \mathbb{I}), \quad w_{kl} = \tau_{kl} \upsilon_l \beta_{kl},$$



Inference

- *Black box variational inference with reparameterization gradients.*
- *Factorized approximation in the reparameterized space*
- *Two variants:*

$$q(\beta_{ij,l}) = \mathcal{N}(\mu_{ij,l}, \sigma_{ij,l}^2) \qquad q(\beta_{ij,l}) = \mathcal{N}(\mu_{ij,l}, 1)$$

full | *tied* — *less memory; faster training;*

- *Factorized approximation in non-centered space, couples weights and scales,*

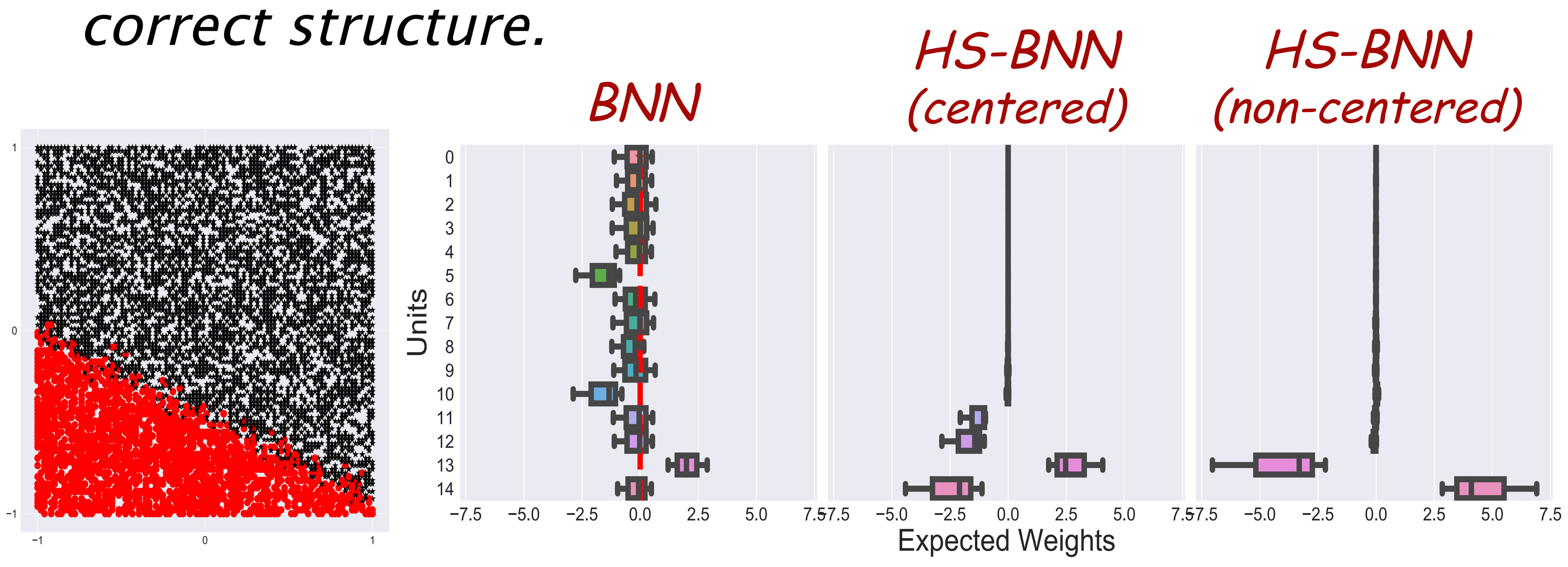
$$q(w_{kl} \mid \tau_{kl}, \upsilon_l) = \mathcal{N}(\tau_{kl}\upsilon_l\mu_{kl}, (\tau_{kl}\upsilon_l)^2\Psi)$$

- *Learning alternates between gradient updates and fixed point updates.*

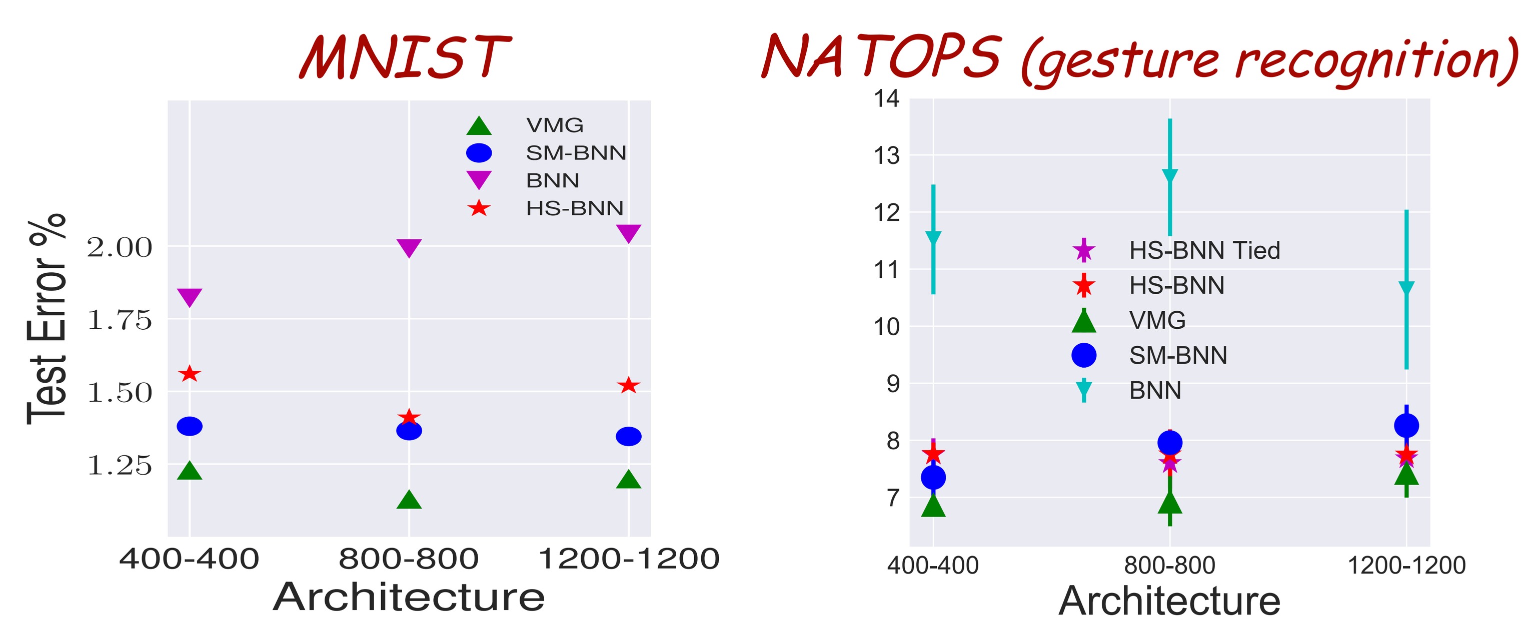
Results

Model Selection

- *Linearly separable data generated from a 2-2-1 network with known weights.*
- *Non centered HS-BNN (2-15-1, 2-100-1) recovers the correct structure.*



Predictive Performance



Faster Training

- *Variational parameter tying leads to faster training and lower storage requirements.*