IBM Research

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Structured Variational Learning of Bayesian Neural Networks with Horseshoe Priors

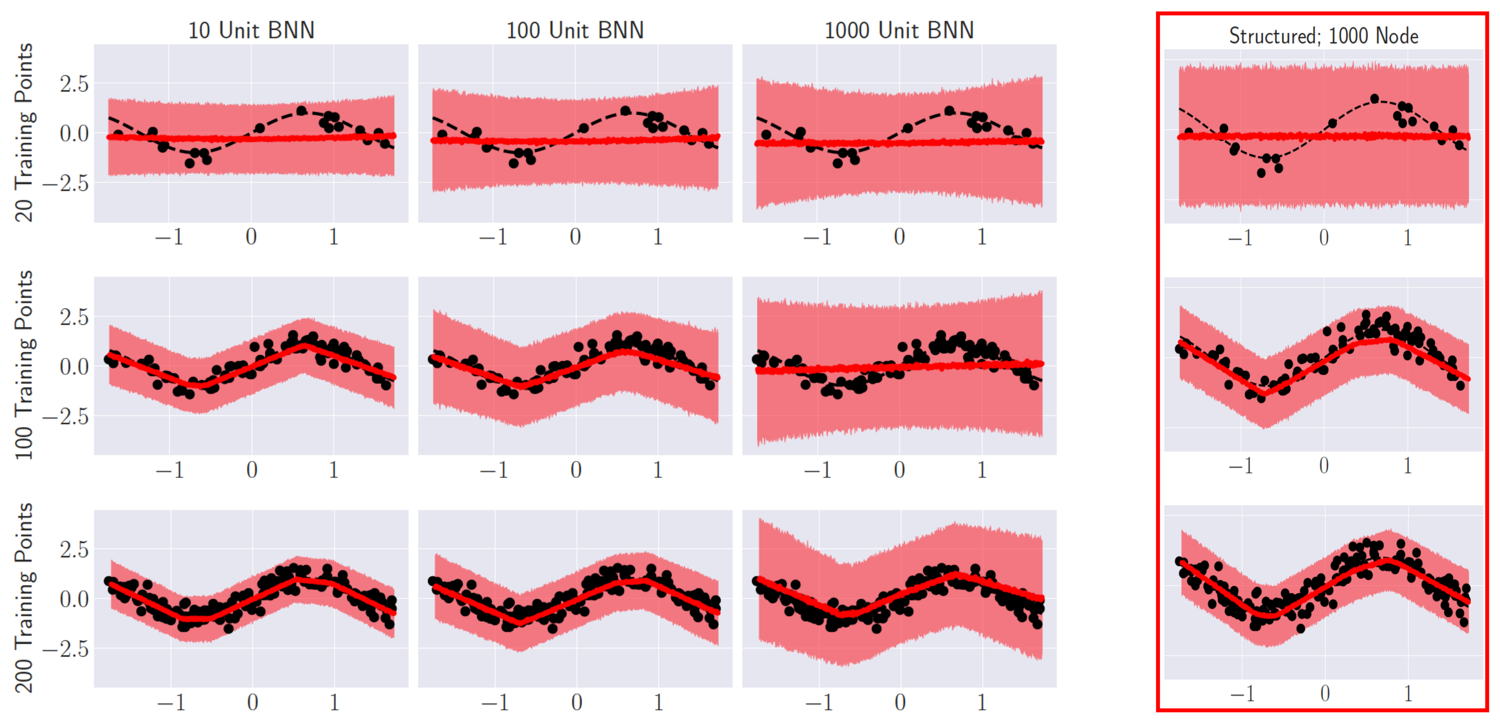
Soumya Ghosh — *MIT-IBM Watson AI Lab & IBM Research*

Jiayu Yao — *Harvard University*

Finale Doshi-Velez — *Harvard University*

Model Selection in BNNs

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



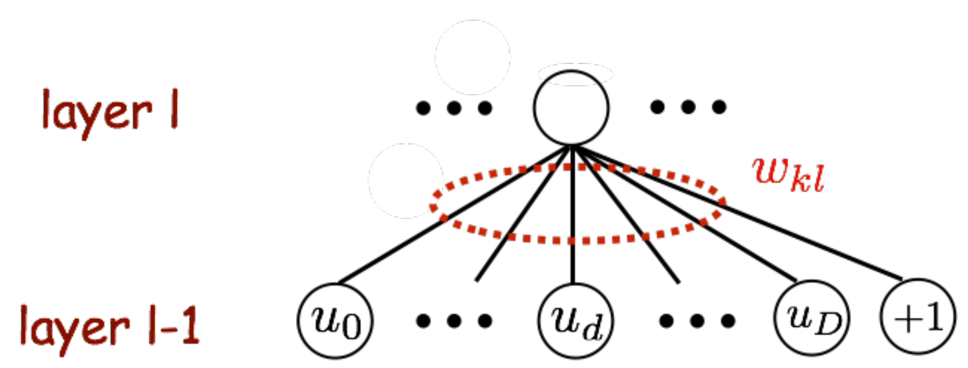
BNNs have unit normal prior on weights, all models have Gaussian output noise:

$$\mathcal{N}(y \mid f(x;\mathcal{W}), \gamma^{-1})$$

Thirty random inits, highest ELBO solution is visualized.

- We develop BNNs with *regularized* group Horseshoe priors to prune away additional capacity.
- Develop structured mean field inference, that provides stronger shrinkage.

All weights incident onto a node share a common scale:



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

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

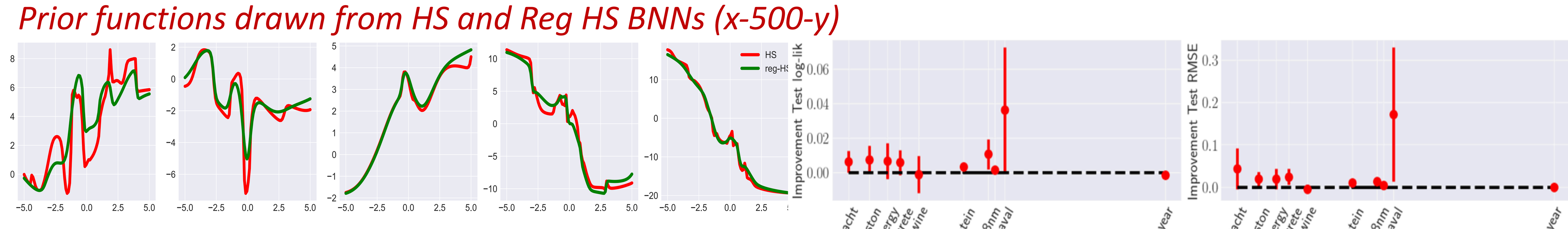
Regularized Horseshoe BNN

$$p(w_{kl} \mid c, \tau_{kl}, \upsilon_l) \propto \mathcal{N}(0, (\tau_{kl}^2 \upsilon_l^2)\mathbb{I})\mathcal{N}(0, c^2\mathbb{I})$$
$$c^2 \sim \text{Inv-Gamma}(c_a, c_b)$$

Non-centered Parameterization

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

Prior functions drawn from HS and Reg HS BNNs (x-500-y)



Regularized Horseshoe prefers smoother functions and results in better performance on smaller datasets

Inference

- Stochastic gradient Variational inference with reparameterization gradients.
- Approximations in the reparameterized space

$$q(\nu_l \mid \phi_{\nu_l})q(\beta_l \mid \phi_{\beta_l}) = \prod_{i,j}\mathcal{N}(\beta_{ij,l} \mid \mu_{ij,l}, \sigma^2_{ij,l})\prod_k q(\nu_{kl}, \phi_{\nu_{kl}})$$

Factorized

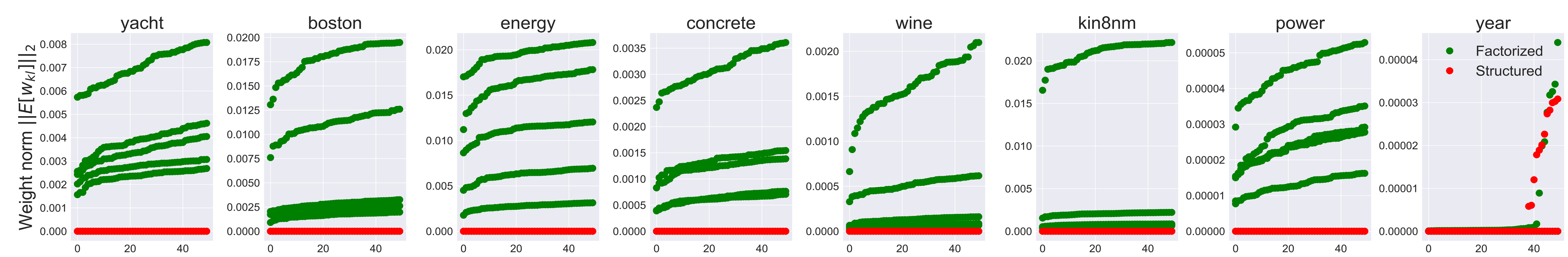
$$q(\beta_l, \nu_l \mid \phi_{B_l}) = \mathcal{MN}(B_l \mid M_l, U_l, V_l)$$

Structured
retains scale <-> weight structure; stronger shrinkage;

- Low variance gradients available through local re-parameterization, since $q(\beta_l \mid \nu_l, \phi_{\beta_l}) = \mathcal{MN}(M_{\beta_l|\nu_l}, U_{\beta_l|\nu_l}, V)$
- Learning alternates between gradient steps & fixed point updates.

Results

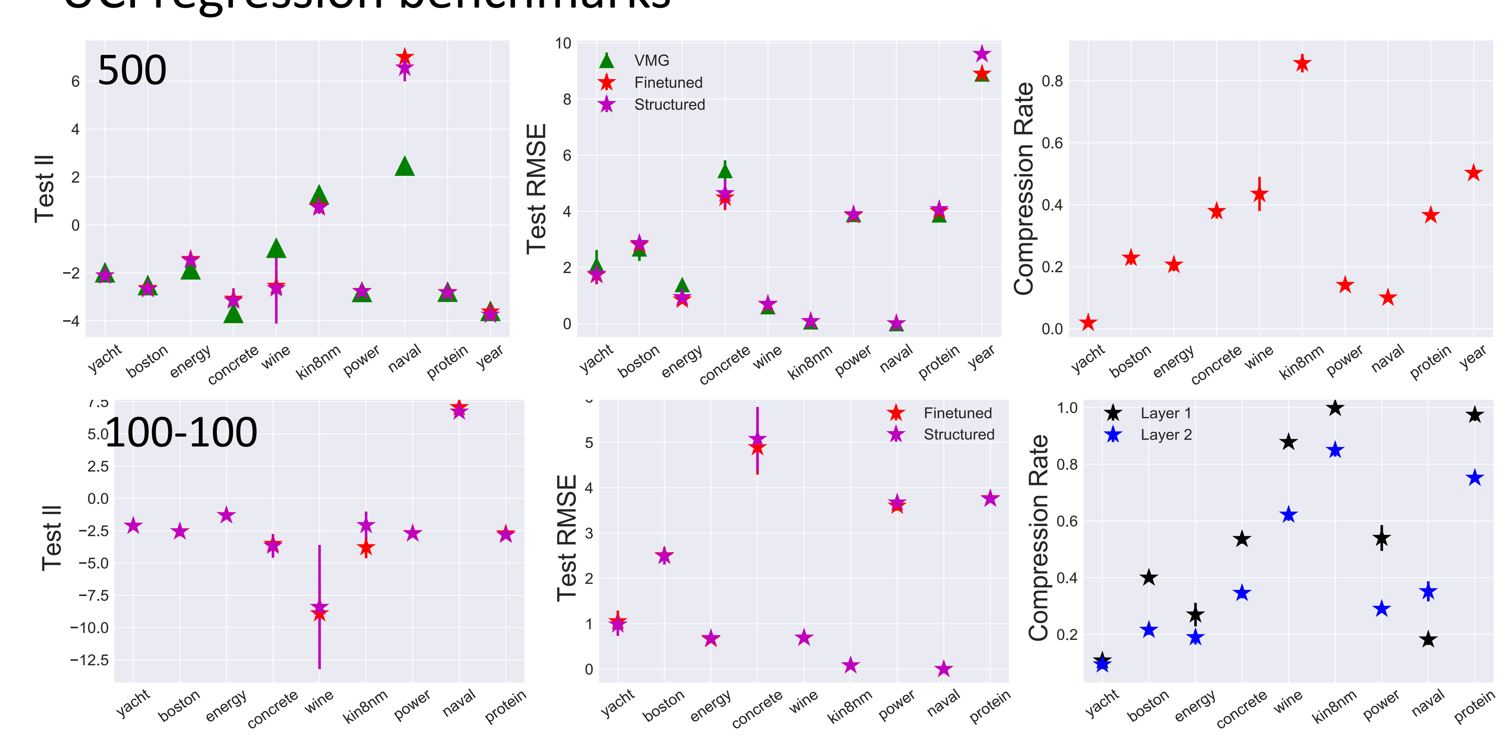
Structured vs Factorized approximations



Structured approximation consistently provides stronger shrinkage

Predictive Performance

UCI regression benchmarks



Similar in performance to state-of-the-art, but uses smaller networks.

Pruning rule uses the entire variational posterior: $q(\tau_{kl}\upsilon_l < \delta) > p_0$

RL experiments

	2D Map	
	Test RMSE	Avg. Reward
BNN x-500-y	0.187	975.386
BNN x-100-100-y	0.089	966.716
Structured x-500-y	**0.058**	**995.416**
Structured x-100-100-y	0.061	992.893
	Acrobot	
BNN x-500-y	0.924	-156.573
BNN x-100-100-y	0.710	-23.419
Structured x-500-y	**0.558**	-108.443
Structured x-100-100-y	0.656	**-17.530**

Outperforms on small data both for regression and reinforcement learning