IBM Research

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

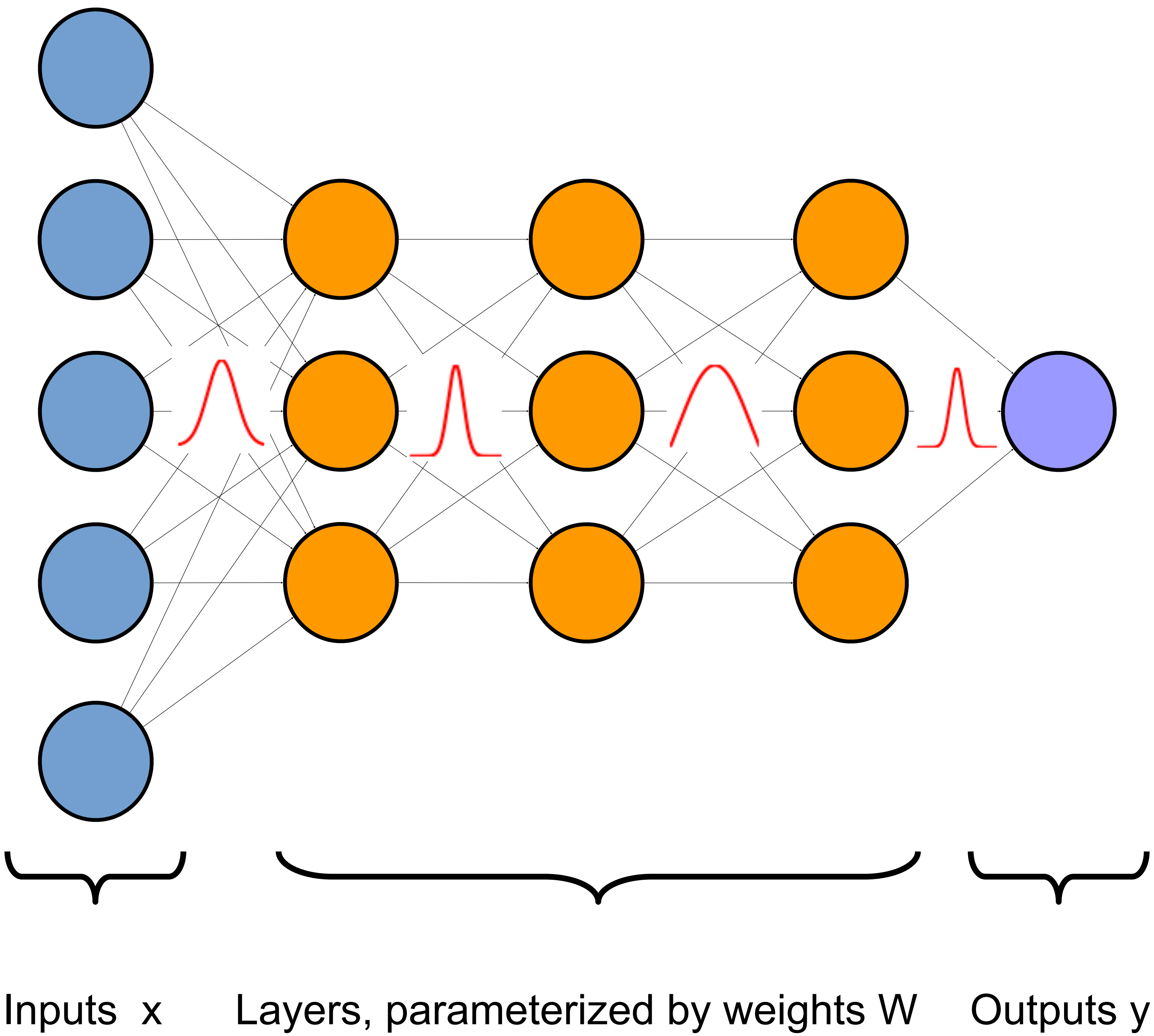
Structured Variational Learning of Bayesian Neural Networks with Horseshoe Priors

Soumya Ghosh
IBM Research
MIT-IBM Watson AI lab

Jiayu Yao
Harvard

Finale Doshi-Velez
Harvard

Bayesian Neural Networks (BNNs)



$$y = f(x, \mathcal{W})$$

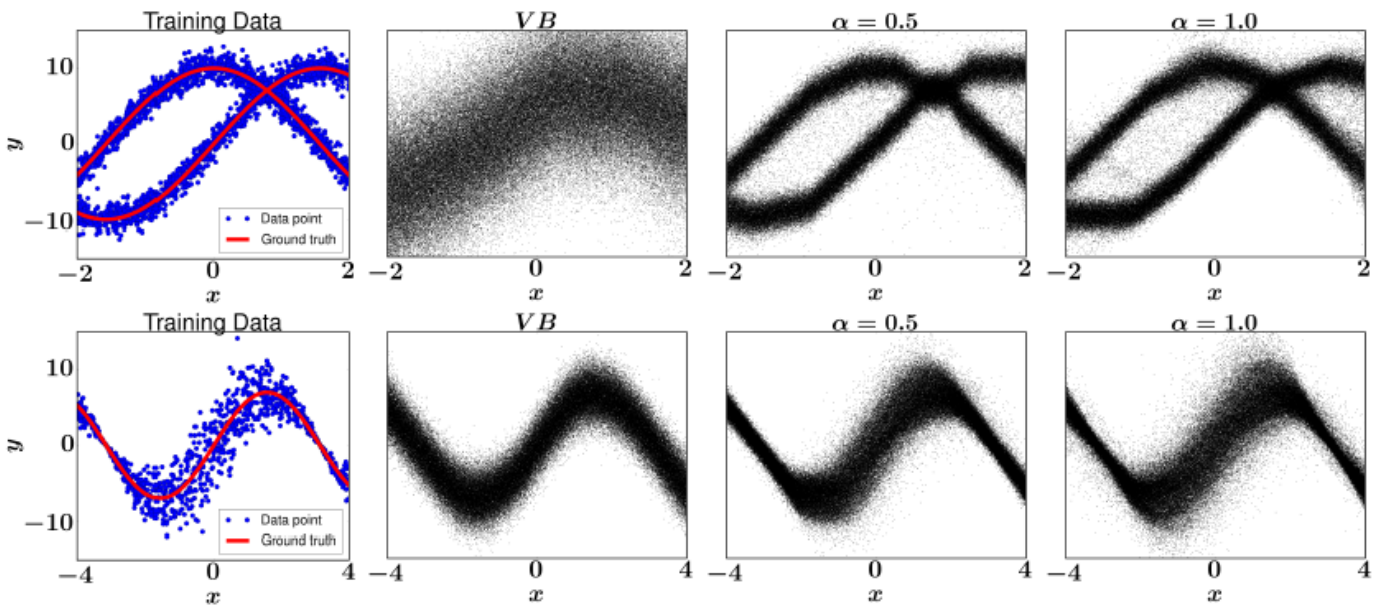
Being Bayesian:

$$p(\mathcal{W} \mid \lambda) \to p(\mathcal{W} \mid y, x, \lambda)$$

$$\downarrow$$

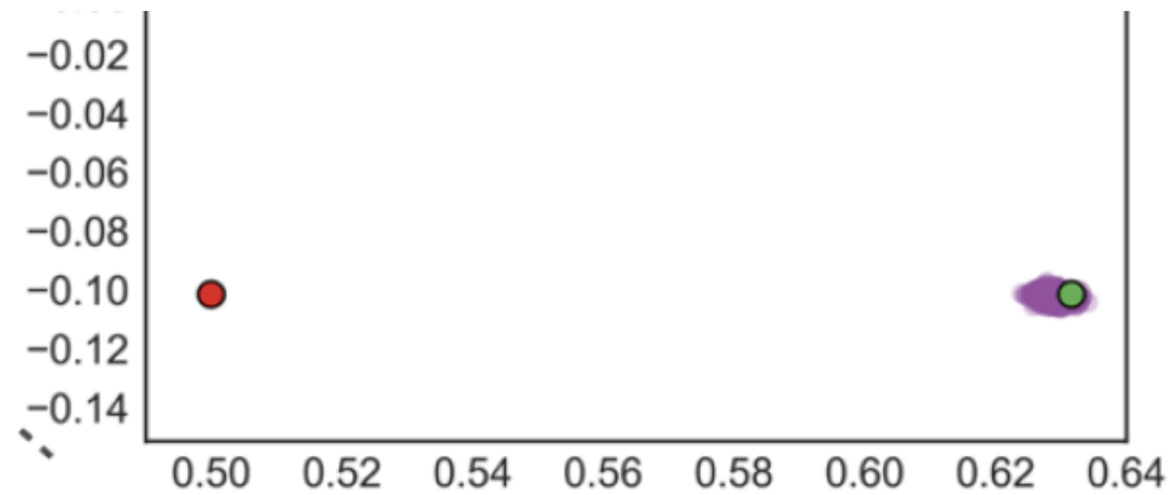
$$p(y_* \mid x_*, y, x, \lambda)$$

Why do we like BNNs?



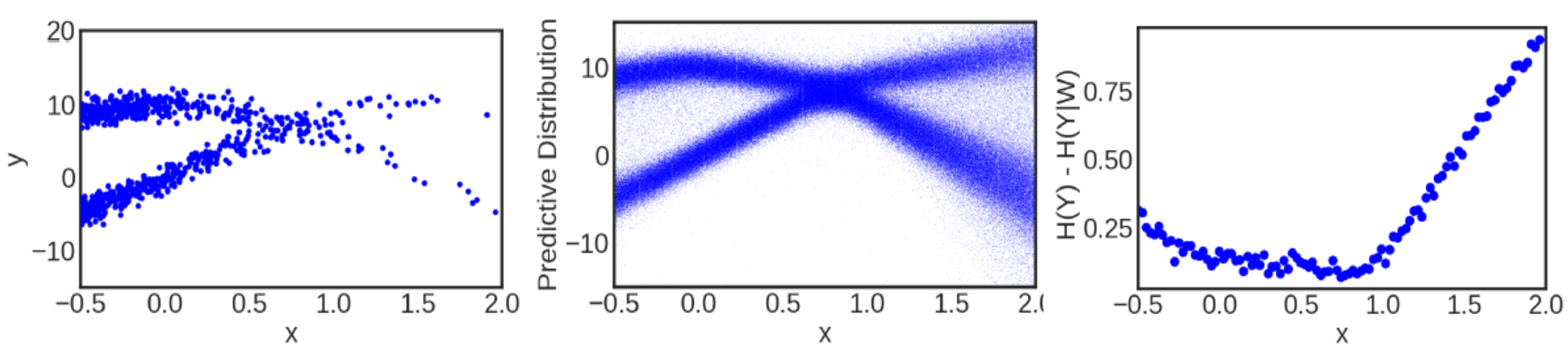
Model stochastic functions

Depweg et al., ICLR 2017



Model uncertainty in deterministic functions

Gal et al., 2016, Killian et al., NIPS 2017



Predictive uncertainties for active learning, sequential decision making

Hernández-Lobato et al., ICML 2015, Gal et al., ICML 2017, Joshi et al., CVPR 2017, Zhang et al., AISTATS 2018, Depweg et al., ICML 2018

Predictive Uncertainties?

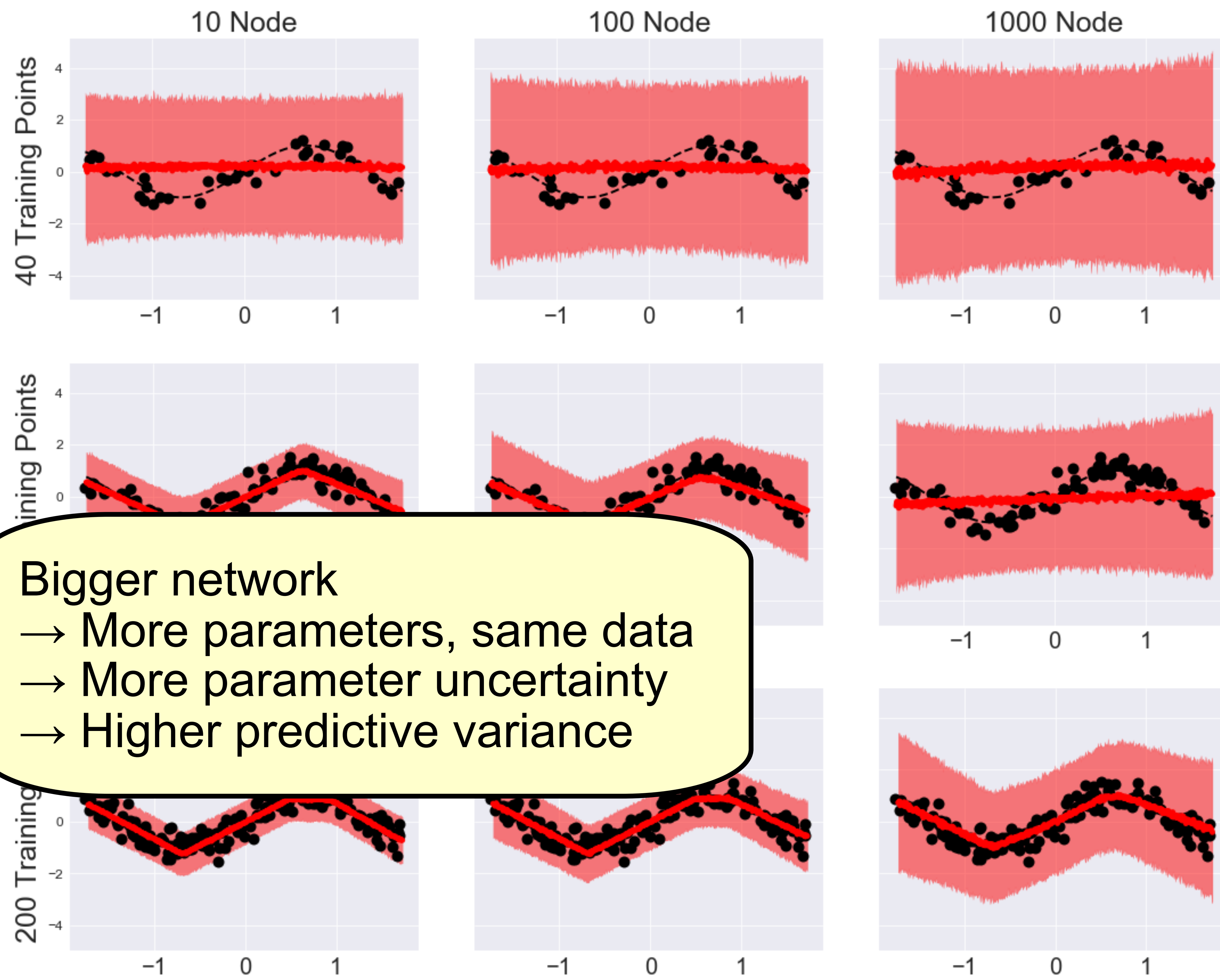
Single layer network, with prior:

$$\mathcal{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

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

(Same results across many initialization strategies)

What is happening?

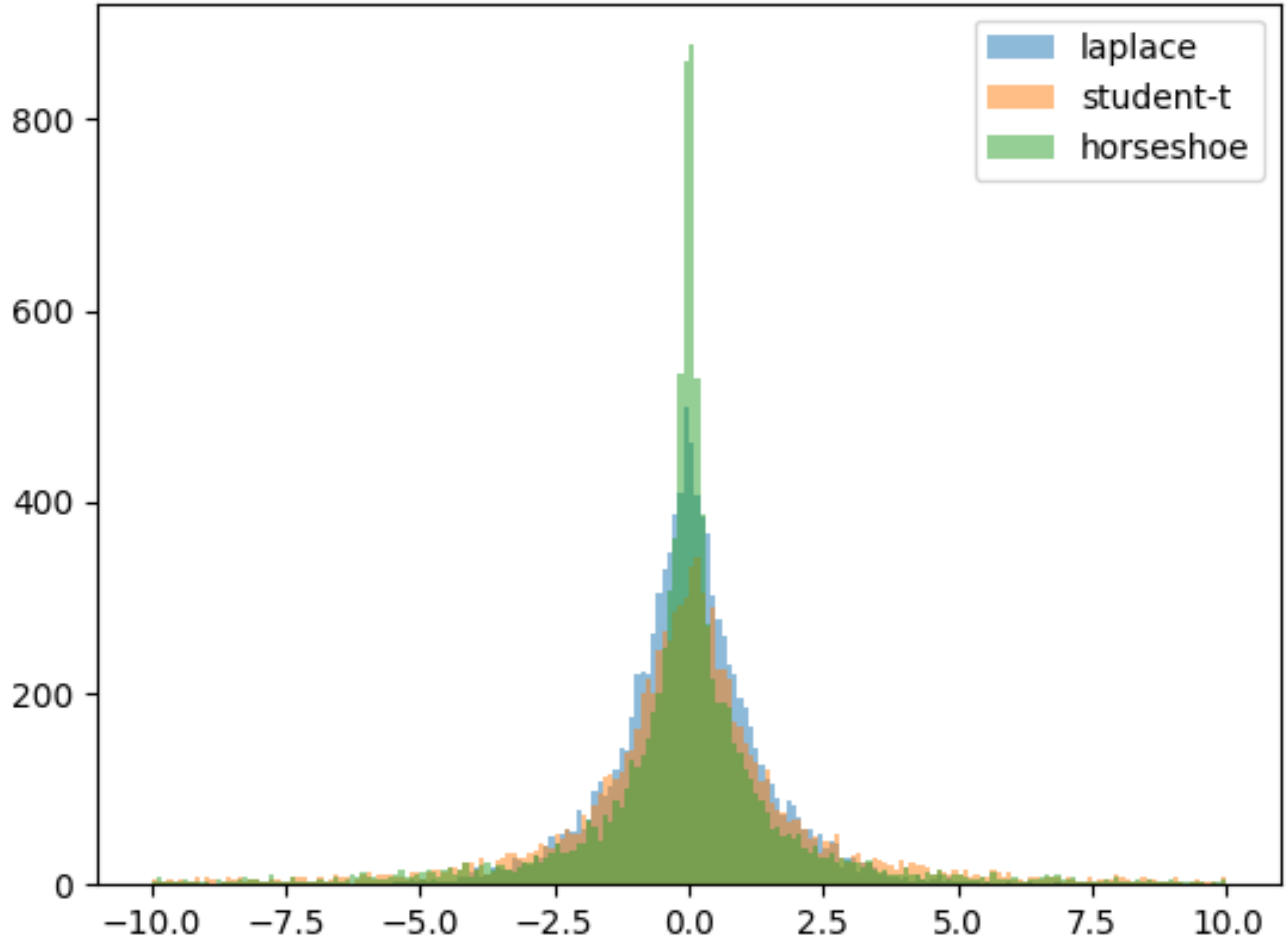


Bigger network
→ More parameters, same data
→ More parameter uncertainty
→ Higher predictive variance

Horseshoe Priors for Model selection

The horseshoe prior is a scale mixture of normals:

$$w_k \sim \mathcal{N}(0, \tau_k^2 v^2)$$
$$\tau_k \sim C^+(0, 1)$$



Group Horseshoe Priors for BNNs

- ~~Horseshoe BNN:~~ Regularized Horseshoe BNN

  For each layer l, draw a global scale: $v_l \sim C^+(0, b_g)$

  For node k in layer l:

  - Draw a local scale for the node: $\tau_{kl} \sim C^+(0, b_0)$
  - For each incident weight: $w_{kk',l} \sim \mathcal{N}(0, \tau_{kl}^2 v_l^2)$

- Inference:

  Stochastic gradient variational Bayes with ~~naive~~ ~~*fully factorized*~~ *structured* variational approximations.

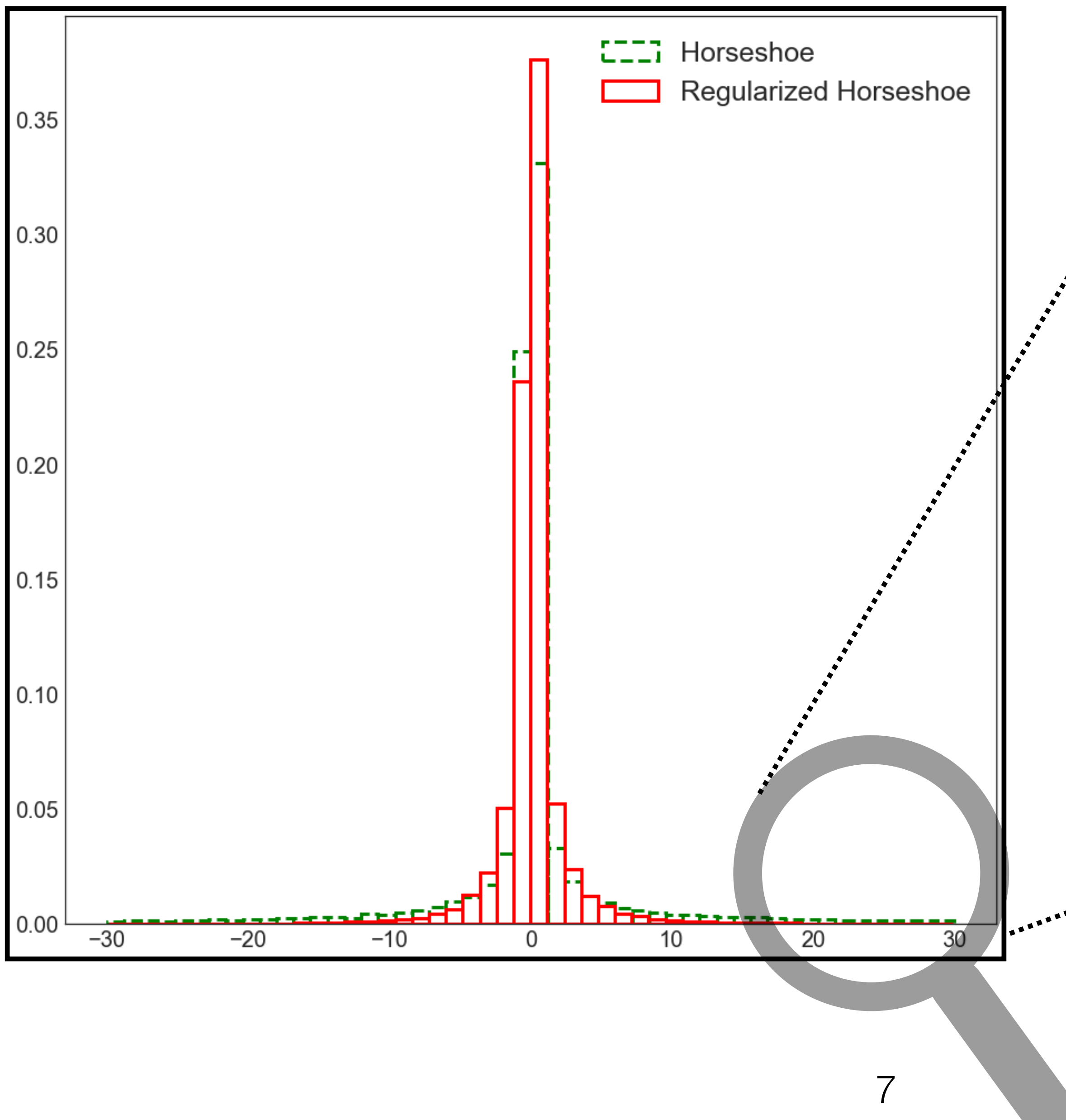
Ghosh & Doshi-Velez, 2017
Louizos et. al., 2017

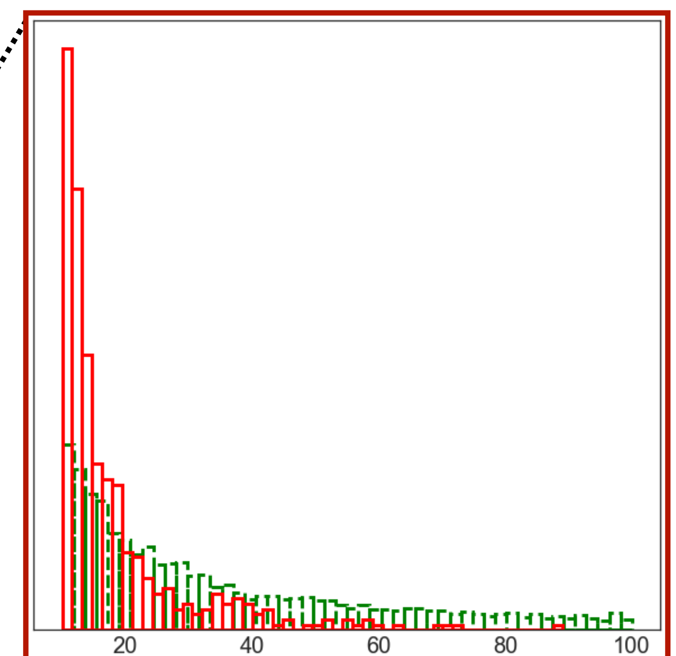
Regularized Horseshoe

$$p(w_{kk',l} \mid \tau_{kl}, v_l, \; c \;) \propto \mathcal{N}(w_{kk',l} \mid 0, \tau_{kl}^2 v_l^2) \; \mathcal{N}(w_{kk',l} \mid 0, c^2)$$

Equivalently,

$$w_{kk',l} \mid c, \tau_{kl}, v_l \sim \mathcal{N}(w_{kl} \mid 0, \tilde{\tau}_{kl}^2 v_l^2); \qquad \frac{1}{\tilde{\tau}_{kl}^2 v_l^2} = \frac{1}{c^2} + \frac{1}{\tau_{kl}^2 v_l^2}$$

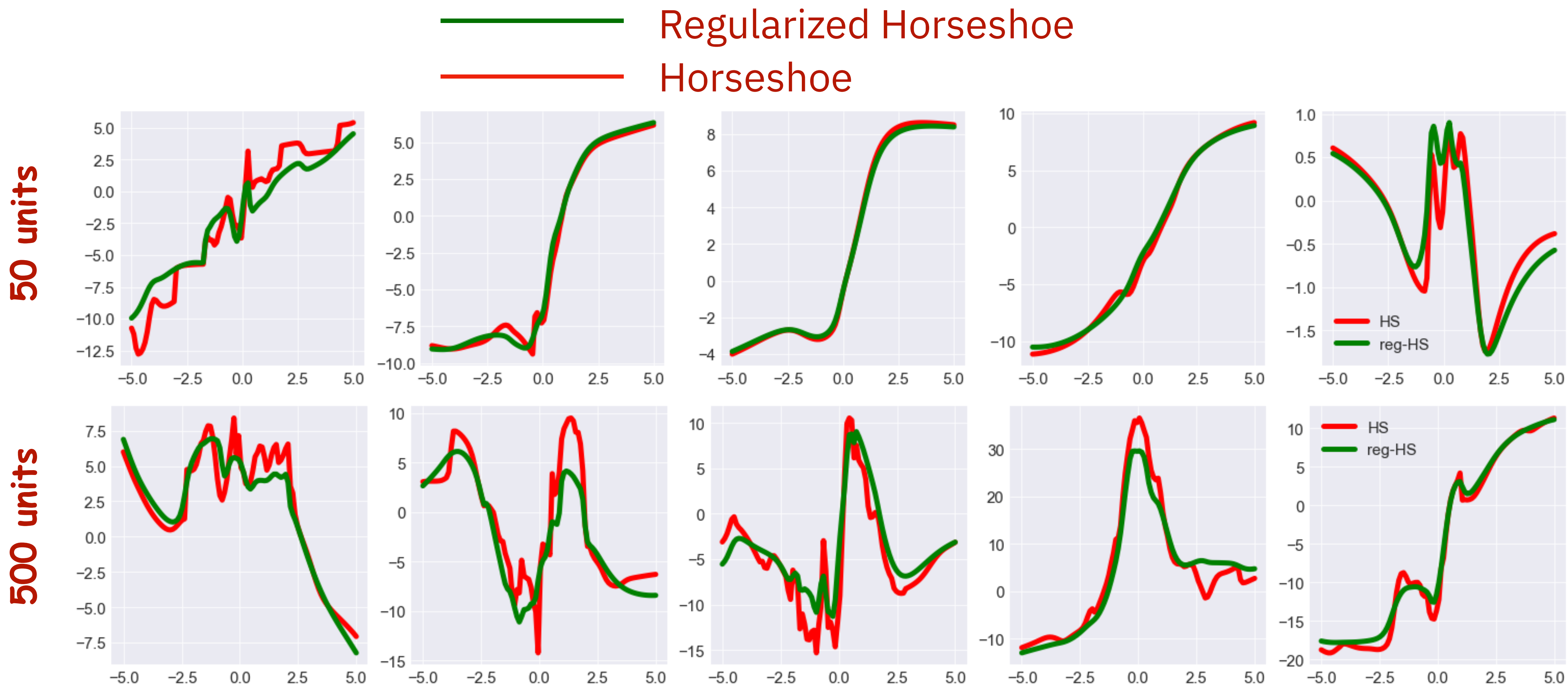




Piironen & Vehtari, 2017

Regularized Horseshoe BNNs

$$w_{kl} \mid \tau_{kl}, \upsilon_l, c \sim \mathcal{N}(0, (\tilde{\tau}_{kl}^2 \upsilon_l^2)\mathbb{I}), \quad \frac{1}{\tilde{\tau}_{kl}^2 \upsilon_l^2} = \frac{1}{c^2} + \frac{1}{\tau_{kl}^2 \upsilon_l^2}$$

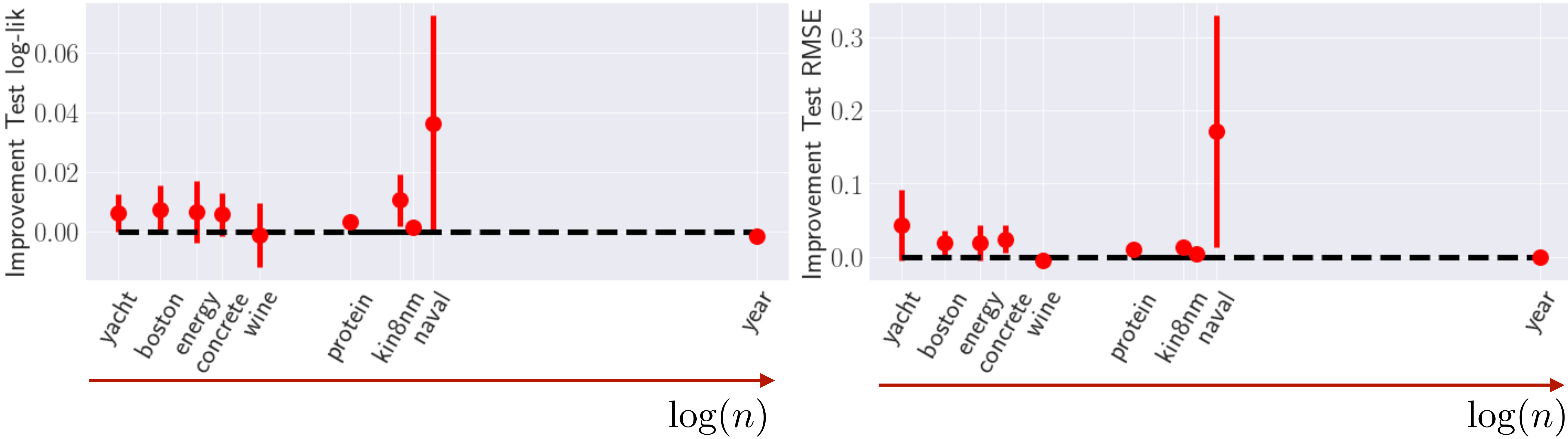


Random functions from single hidden layer (tanh) network with HS and reg-HS priors

Regularized Horseshoe BNNs

$$w_{kl} \mid \tau_{kl}, \upsilon_l, c \sim \mathcal{N}(0, (\tilde{\tau}_{kl}^2 \upsilon_l^2)\mathbb{I}), \quad \frac{1}{\tilde{\tau}_{kl}^2 \upsilon_l^2} = \frac{1}{c^2} + \frac{1}{\tau_{kl}^2 \upsilon_l^2}$$

UCI Regression Benchmarks *(Hernández-Lobato and Adams' 2015)*

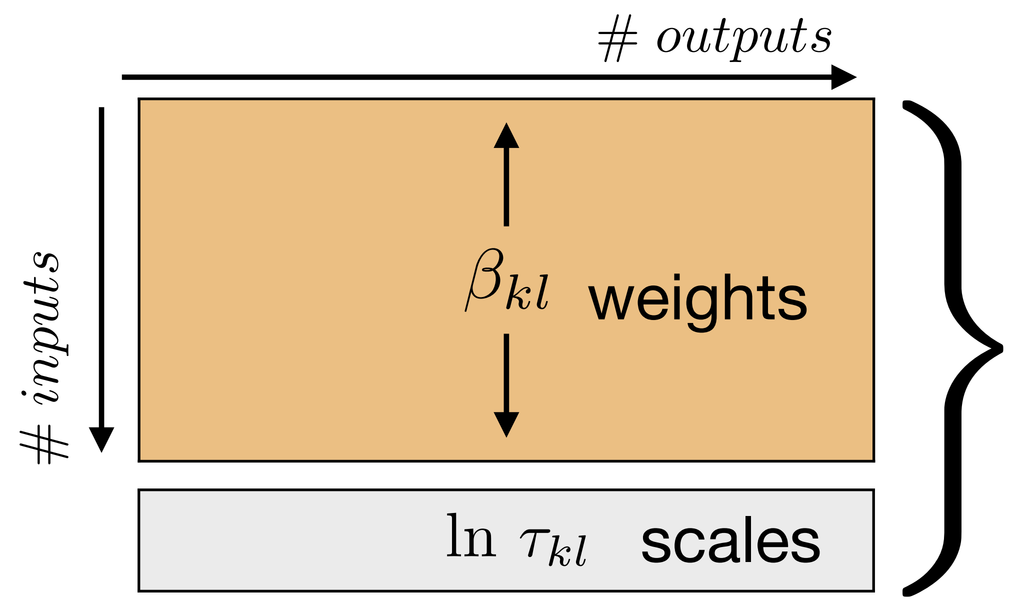


reg-HS BNNs improves predictive performance over HS BNNs for smaller datasets.

Relative improvement: (x - y)/ max(|x|, |y|)

Structured Variational Approximation

- Weights incident on a unit: $w_{kl} \mid \tau_{kl}, \upsilon_l, c \sim \mathcal{N}(0, (\tilde{\tau}_{kl}^2 \upsilon_l^2)\mathbb{I})$
- Non-centered Parameterization: $\beta_{kl} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ $\quad w_{kl} = \tau_{kl}\upsilon_l\beta_{kl}$
- Layer specific structured variational approximations:

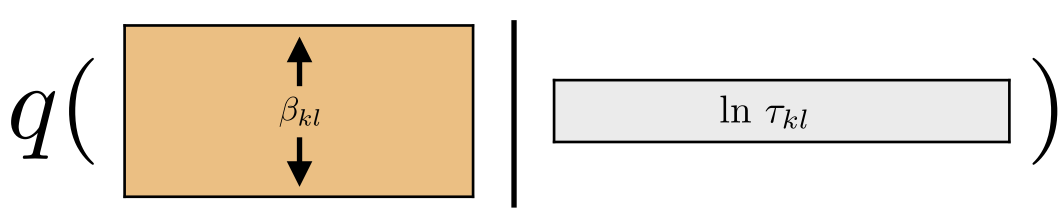


$$q(B) = \text{Matrix-Normal}(M, U, V)$$

$$U = hh' + \Psi$$

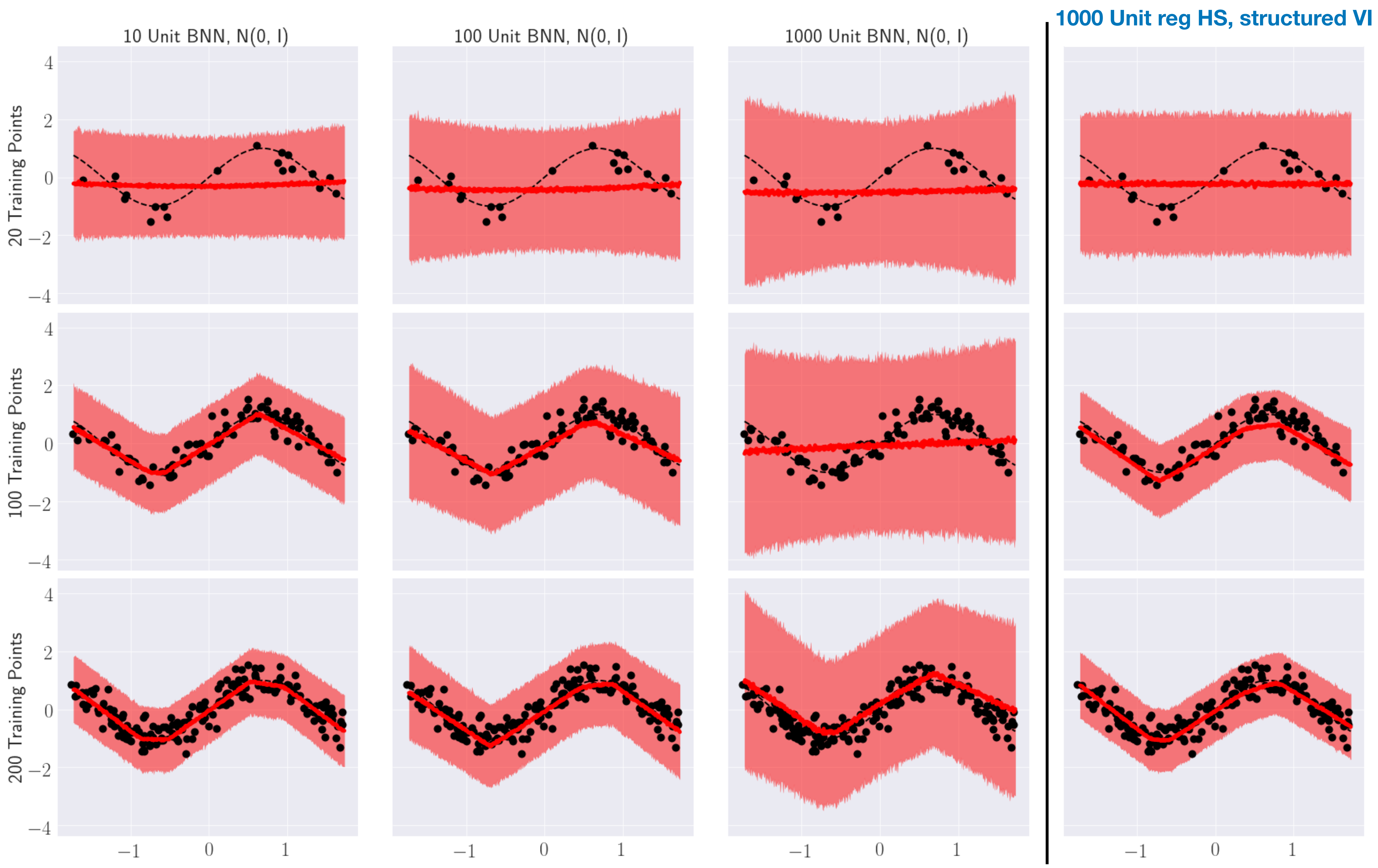
Low dimensional covariance maintains posterior structure between **weights** and **scales**.

- Local re-parameterization:



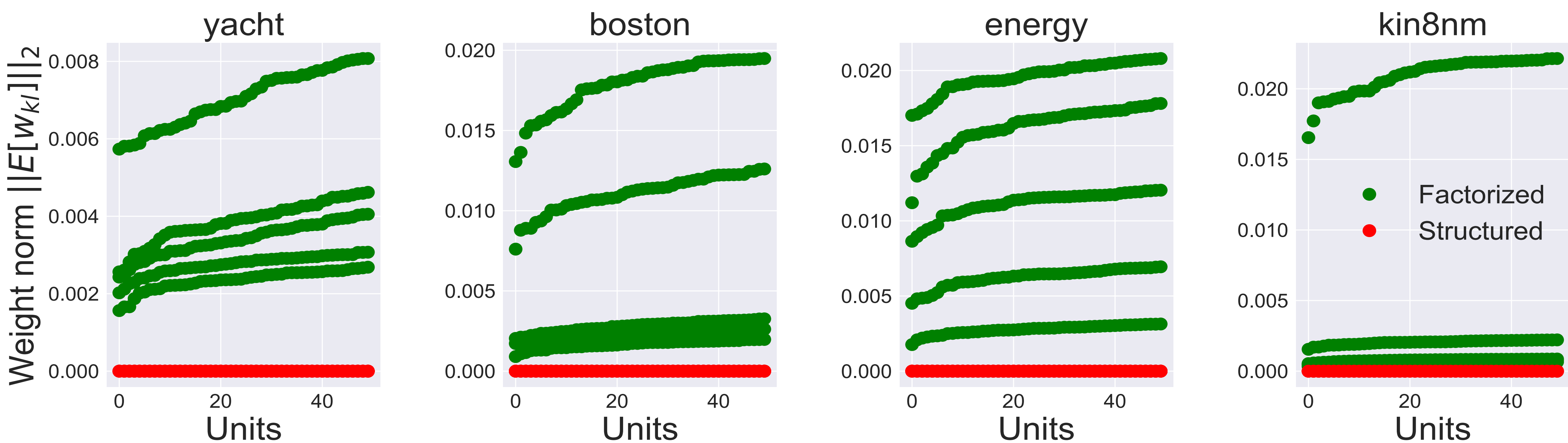
$$= \text{Matrix-Normal}(M_{\beta|\tau}, U_{\beta|\tau}, V_{\beta|\tau})$$

Synthetic Data: Better Fits

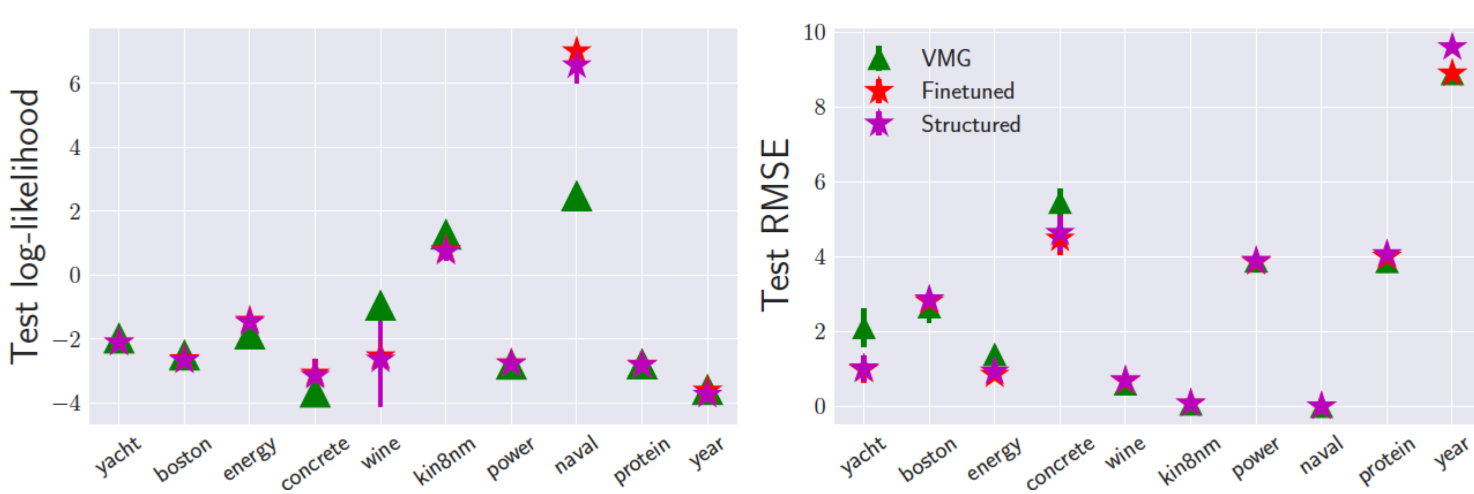


UCI Regression Tasks

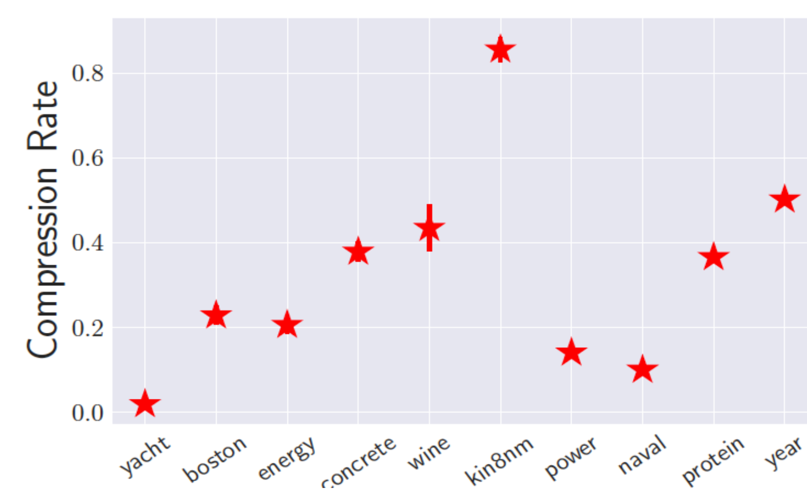
- Structured variational approximation -> stronger shrinkage, similar predictive performance



- Predictive Performance:



Comparisons with Variational matrix Gaussian (Louizos & Welling, ICML 2016)



$$q(\tau_{kl} v_l < \delta) > p_0$$

Pruning rule uses
the variational posterior

Summary

- (Regularized) Horseshoe Priors for BNNs can assist with model-selection
  - Recover small networks with similar performance to larger networks.
- Careful modeling of posterior structure between weights and scales is essential for reliable shrinkage.
- For more results, small data and reinforcement learning experiments, stop by the poster (#193)

Thanks!