Overpruning in Variational Bayesian Neural Networks

Brian L Trippe[1,2], Richard E. Turner[1]

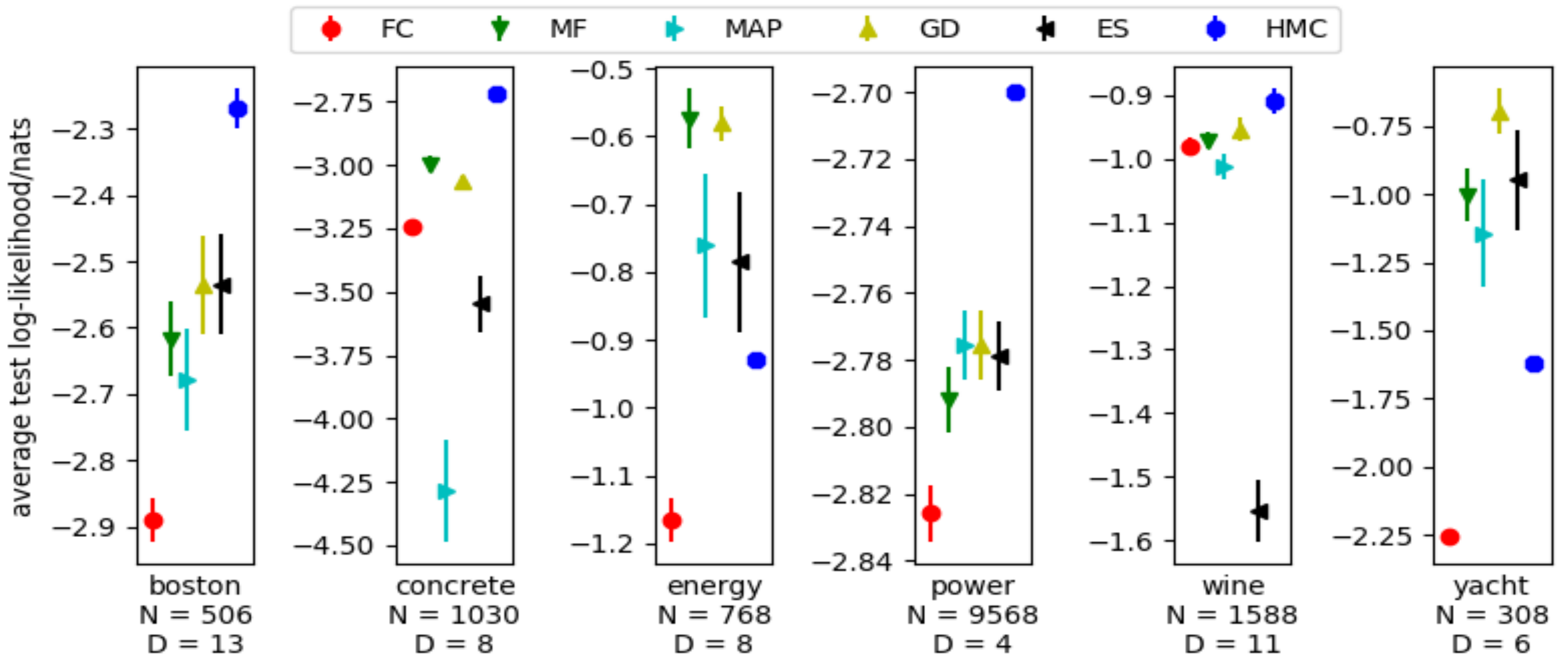
[1]University of Cambridge, [2]Massachusetts Institute of Technology

1. Overview

- Variational approximations to neural networks (NNs) do not resemble the posterior over parameters
- More expressive variational families often provide significantly worse performance than less expressive families
- Mean field approximations overprune (many hidden units turn off) due to looseness of the variational bound
- This results in a biased approximations in function space and leads to under-fitting

3.Variational Family Comparison

- We compare different variational families by test log-likelihood on benchmark regression tasks (higher is better)



GD: Gaussian Dropout (fixed variance)
MF: Mean Field Gaussian
FC: Full Covariance

ES: Early Stopping
MAP: Maximum a Posteriori
HMC: Hamiltonian Monte Carlo

- Often, more expressive approximations perform worse
- Hybrid Monte Carlo[1] does the best in most cases (but is expensive)
- Gaussian Dropout (fixed weight variances) performs well

5. Looseness of the Variational Bound Explains Overpruning

The variational free energy is the sum of the expected log likelihood of the training set and complexity penalties on the posterior over each weight. Pruning reduces the complexity penalty at the cost of explaining the data.

$$\mathcal{F}_{\text{VFE}} = -\mathbb{E}_{q(\theta)}\left[\log p(Y|X,\theta)\right] + \mathrm{D}_{\text{KL}}\big(q(\theta)||p(\theta|\alpha)\big)$$

$$= \underbrace{-\mathbb{E}_{q(\theta)}\left[\log p(Y|X,\theta)\right]}_{\text{Explain the data}} + \sum_{j=1}^{N} \underbrace{\mathrm{D}_{\text{KL}}\big(q(v_j)||p(v_j|\alpha)\big)}_{\substack{\text{KL- penalty} \\ \text{(H output weights)}}} + \sum_{i=1}^{D} \underbrace{\mathrm{D}_{\text{KL}}\big(q(w_{j,i})||p(w_{j,i}|\alpha)\big)}_{\substack{\text{KL- penalty} \\ \text{(HxD input weights)}}}$$

- Pruning allows input weights, $w_{i,j}$, to fit to the prior without increasing the variance of predictions.

$$p(w_{j,i}|v_j = 0, \mathcal{D}, \alpha) = \frac{p(w_{j,i}|v_j = 0, \alpha)p(\mathcal{D}|v_j = 0, \alpha)}{p(\mathcal{D}|v_j = 0, \alpha)} = p(w_{j,i}|\alpha)$$

- Pruning induces a conditional independence between predictions and lower layer weights, $w_{i,j}$. This reduces free energy by bringing $q(w|\mathcal{D},\alpha)$ close to $q(w|\mathcal{D},\alpha)$ [5].

References:

1. Radford M. Neal. Bayesian Learning for Neural Networks. PhD thesis, 1995.
2. Geoffrey E. Hinton and Drew Van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. ACM COLT, 1993.
3. Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *Icml*, 37:1613–1622, 2015.
4. Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning. *Icml*, 48:1–10, 2015.
5. Richard E Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. Inference and Estimation in Probabilistic TimeSeries Models, pages 109–130, 2011.

2. Variational Inference in NNs

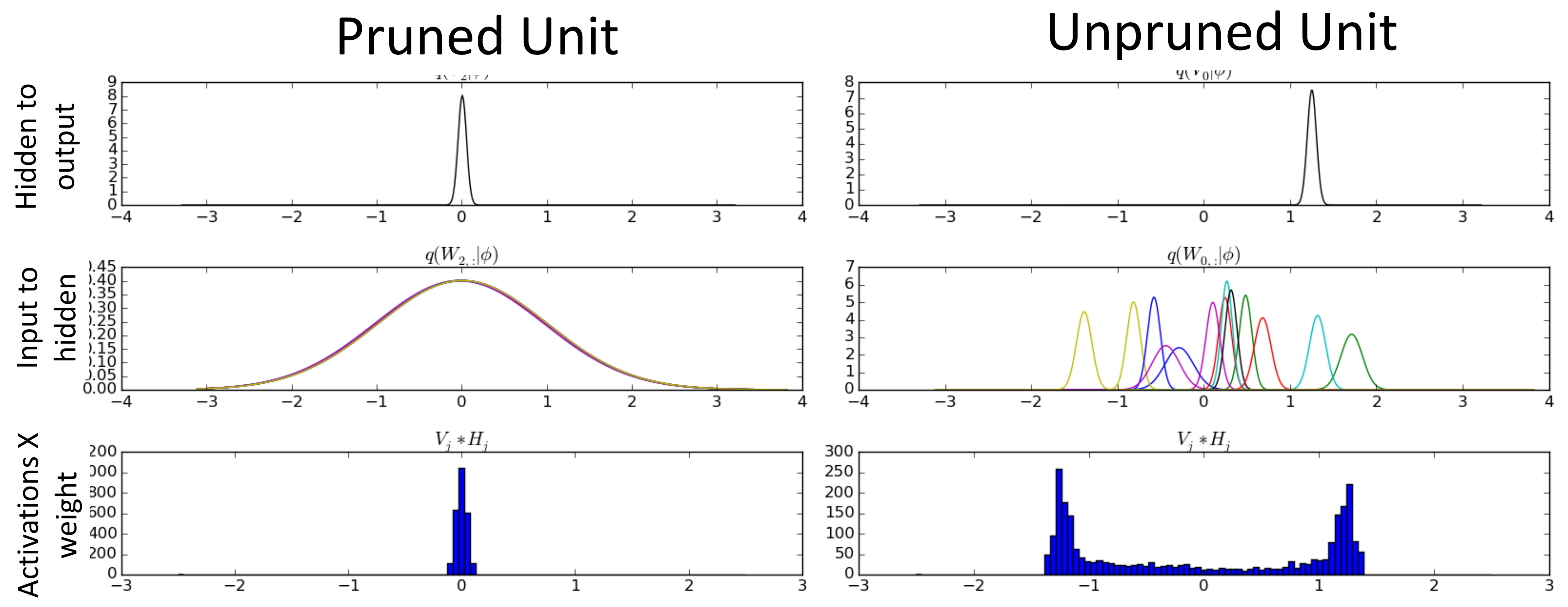
Infinitely wide Bayesian NNs are Gaussian processes, but inference is intractable[1]. We approximate infinite NNs with finite ones, and the exact posterior with a variational approximation].

$$\underset{q\in\mathcal{Q}}{\arg\min}\,\mathrm{KL}[q||p] = \underset{q\in\mathcal{Q}}{\arg\min}\,\mathcal{F}_{\text{VFE}}(q) = -\mathbb{E}_{q(\theta)}\left[\log p(Y,\theta|X,\alpha) - \log q(\theta)\right]$$

- We obtain gradient estimates and posterior predictions with MC sampling[3,4].

$$p(y_{\text{new}}|x_{\text{new}},\mathcal{D},\alpha) = \mathbb{E}_{p(\theta|\mathcal{D},\alpha)}[p(y_{\text{new}}|x_{\text{new}},\theta)] \approx \frac{1}{M}\sum_{i=1}^{M} p(y_{\text{new}}|x_{\text{new}},\theta_i)$$

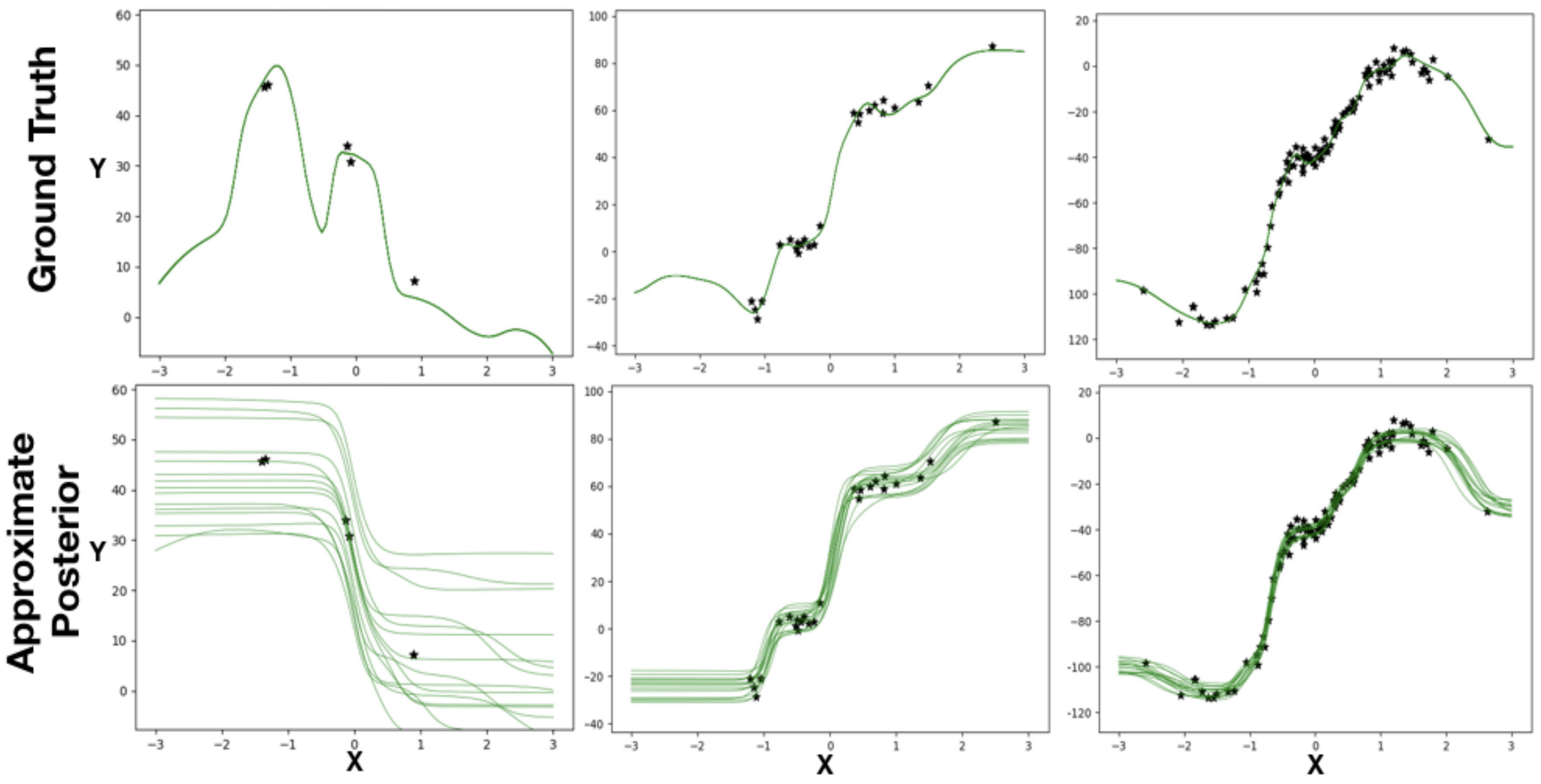
4. Pruning in Mean Field VI



- At convergence, **many units are pruned**: the output weight is at 0 with high confidence, the input weights are at the prior, and the hidden unit does not affect the function.
- In this case, only 11 out of 50 hidden units are active

6.Biased Posterior Over Functions

We assess pruning on toy datasets with functions and data sampled from the prior. We perform VI in the MF model, initializing to the ground truth.



N	5	25	100
Units Pruned	47/50	43/50	37/50

- **Overpruning results in pronounced bias and underfitting**
- Even correctly specified model with 'correct' initialization overprunes.
- More severe pruning occurs with small N and large observation noise
- Dropout/Gaussian Dropout[4] have fixed entropy; they do not prune and can outperform MF in practice

Acknowledgements

We would like to thank Thang Bui, Yingzhen Li, and Cuong Nguyen for insightful comments and discussion, the reviewers for constructive feedback, BLT thanks the Euretta J Kellett Fellowship for financial support.

For more information, contact btrippe@mit.edu