Dangers of Bayesian Model Averaging under Covariate Shift

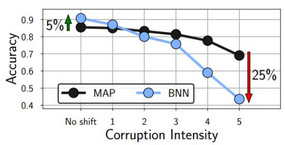
Pavel Izmailov Patrick Nicholson Sanae Lotfi Andrew Gordon Wilson

NYU

Paper Code

Overview

- We show that Bayesian model averaging (BMA) can be problematic under covariate shift in cases when linear dependencies in the inputs cause lack of posterior contraction.
- The same issue does not affect MAP and several approximate Bayesian deep learning methods.
- We propose a new prior that improves the robustness of BNNs.
- These issues could affect virtually any real-world application of Bayesian model averaging with neural networks.



Bayesian neural networks

Bayesian inference is especially compelling for deep neural networks!

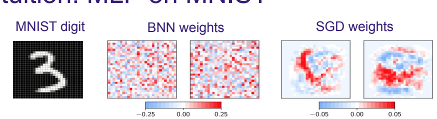
$$p(w|\text{Data}) \propto p(\text{Data}|w) \cdot p(w)$$

$$p_{BMA}(y|x) = \int p(y|w,x)p(w|\text{Data})dw \approx \frac{1}{n}\sum_i p(y|w_i,x)$$
$$w_i \sim p(w|\text{Data})$$

Covariate shift

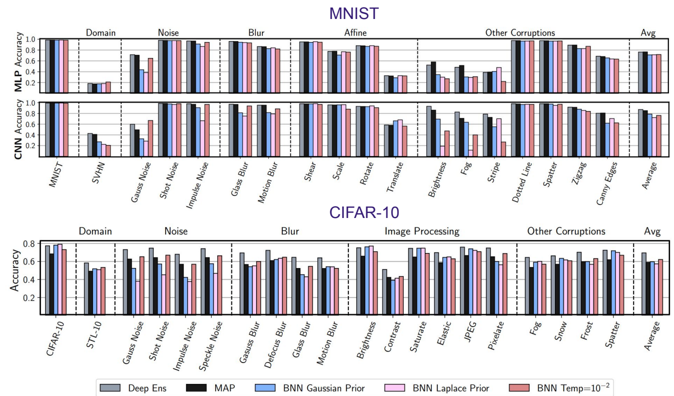
Target data distribution is different from the distribution used for training. $p_{\text{train}}(x,y) = p_{\text{train}}(x)p(y|x)$; $p_{\text{test}}(x,y) = p_{\text{test}}(x)p(y|x)$

Intuition: MLP on MNIST



- Weights in the first MLP layer corresponding to dead pixels have no effect on the likelihood.
- The posterior for these weights is the same as the prior.
- At test time due to noise dead pixels activate; the corresponding weights sampled from the prior now hurt predictions.
- MAP sets these weights to zero and ignores the dead pixels.

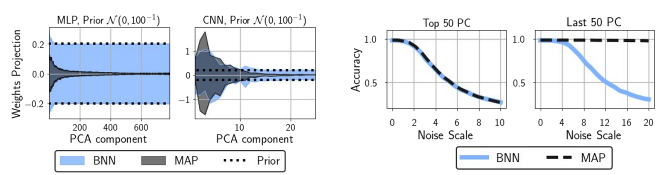
BNNs are not robust to covariate shift



BNNs underperform Deep Ensembles and MAP solutions over a wide range of shifts!

Theoretical explanation

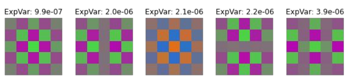
Theorem (Informal): *Suppose we use an i.i.d. Gaussian prior in a Bayesian MLP. Suppose there exists a constant linear combination in the input features. Then*
- *There will exist a direction in the parameter space such that the posterior along this direction coincides with the prior.*
- *The MAP solution will set this projection to zero.*
- *The BMA prediction will be susceptible to perturbations breaking the linear dependence, while the MAP solution will ignore them.*



Generalization to CNNs

Theorem (Informal): *Same result applies to convolutional layers, assuming there is a linear dependence in the dataset of all k x k patches, where k is the size of the convolutional filter.*

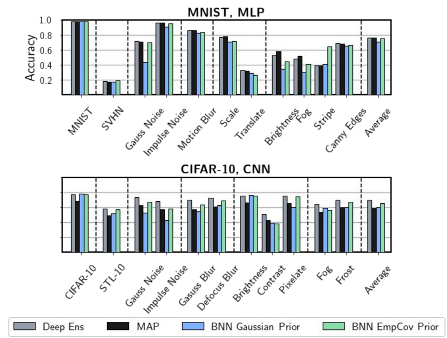
Low-variance directions on CIFAR-10 ⟶



Fix: *EmpCov* prior

Idea: Reduce prior variance along low-variance directions in data

EmpCov prior for the first MLP layer ⟶ $p(W_i^1) = \mathcal{N}\left(0, \frac{\alpha^2}{n-1}\sum_{k=1}^{n} x_k x_k^T\right)$



Which BDL methods are affected?

- This is a foundational issue with Bayesian model averaging.
- High-fidelity approximate inference, such as HMC, can be especially affected. VI and SG-MCMC can also be affected.
- MAP, Deep Ensembles, MC-Dropout, SWAG are unaffected.