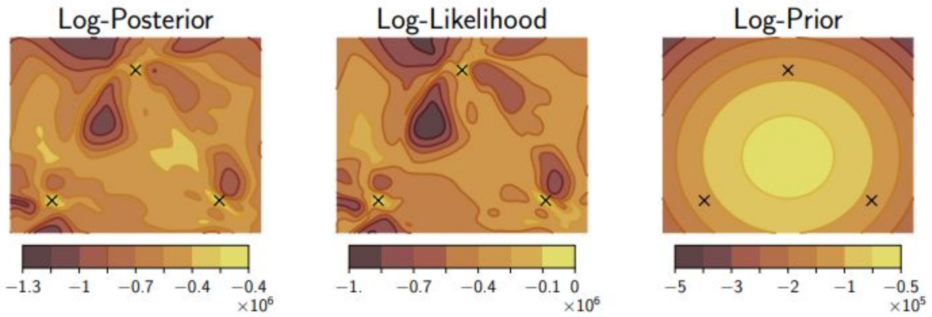
NYU

Google AI

What Are Bayesian Neural Network Posteriors Really Like?

Pavel Izmailov Sharad Vikram Matthew D. Hoffman Andrew Gordon Wilson

ICML | 2021



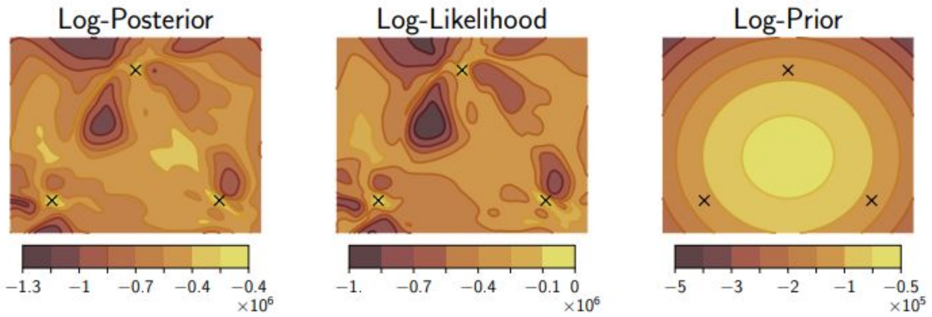
Overview

We perform approximate inference of the highest fidelity in Bayesian neural nets.

We answer many questions in Bayesian deep learning, often contradicting conventional wisdom:

- ❓ *Do BNNs perform well in practice?*

Overview

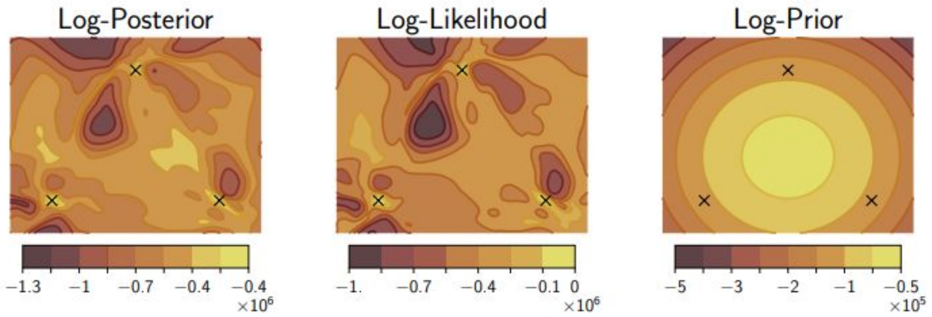


We perform approximate inference of the highest fidelity in Bayesian neural nets.

We answer many questions in Bayesian deep learning, often contradicting conventional wisdom:

- ❓ *Do BNNs perform well in practice?*
- ❓ *Do we need cold posteriors?*

Overview

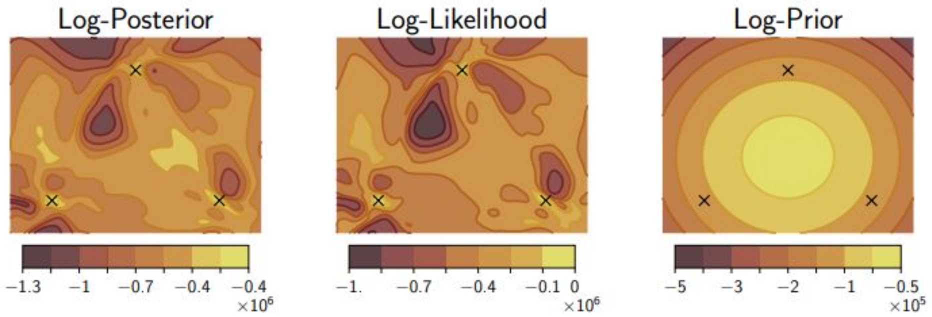


We perform approximate inference of the highest fidelity in Bayesian neural nets.

We answer many questions in Bayesian deep learning, often contradicting conventional wisdom:

- ❓ *Do BNNs perform well in practice?*
- ❓ *Do we need cold posteriors?*
- ❓ *Are BNNs robust to covariate shift?*

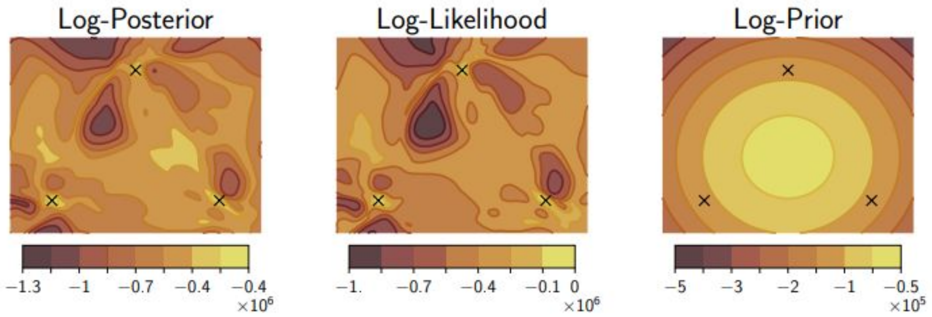
Overview



We perform approximate inference of the highest fidelity in Bayesian neural nets.

We answer many questions in Bayesian deep learning, often contradicting conventional wisdom:

- ❓ *Do BNNs perform well in practice?*
- ❓ *Do we need cold posteriors?*
- ❓ *Are BNNs robust to covariate shift?*
- ❓ *What is the effect of priors in BNNs?*



Overview

We perform approximate inference of the highest fidelity in Bayesian neural nets.

We answer many questions in Bayesian deep learning, often contradicting conventional wisdom:

- ❓ *Do BNNs perform well in practice?*
- ❓ *Do we need cold posteriors?*
- ❓ *Are BNNs robust to covariate shift?*
- ❓ *What is the effect of priors in BNNs?*
- ❓ *How good are different approximate inference methods?*

Bayesian neural networks

Bayesian Model Average:

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

$$w_i \sim p(w|\text{Data})$$

Bayesian neural networks

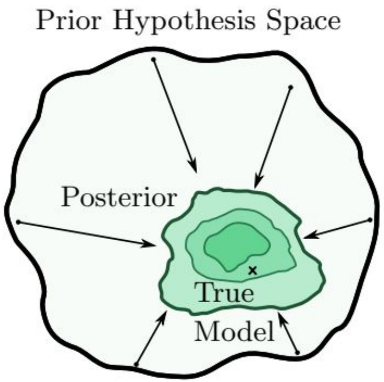
Bayesian Model Average:

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

$$w_i \sim p(w|\text{Data})$$

posterior likelihood prior

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



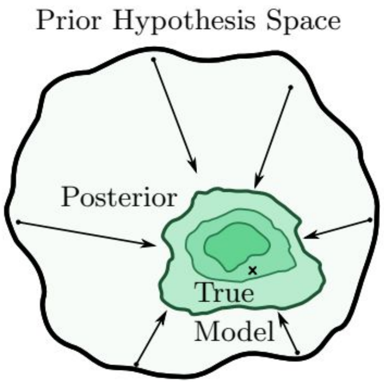
Bayesian neural networks

Bayesian Model Average:

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

$$w_i \sim p(w|\text{Data})$$

$$\underbrace{p(w|\text{Data})}_{\text{posterior}} \propto \underbrace{p(\text{Data}|w)}_{\text{likelihood}} \cdot \underbrace{p(w)}_{\text{prior}}$$



Bayesian inference is especially compelling for deep neural networks!

Bayesian neural networks

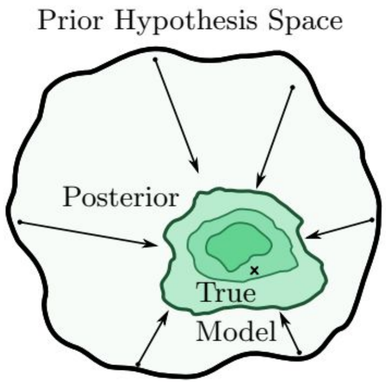
Bayesian Model Average:

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

$$w_i \sim \cancel{\boxed{p(w|\text{Data})}}$$

$$q(w|\text{Data})$$

$$\overset{\text{posterior}}{\boxed{p(w|\text{Data})}} \propto \overset{\text{likelihood}}{p(\text{Data}|w)} \cdot \overset{\text{prior}}{p(w)}$$



Bayesian inference is especially compelling for deep neural networks!

Bayesian inference is intractable for BNNs!
Have to do approximate inference

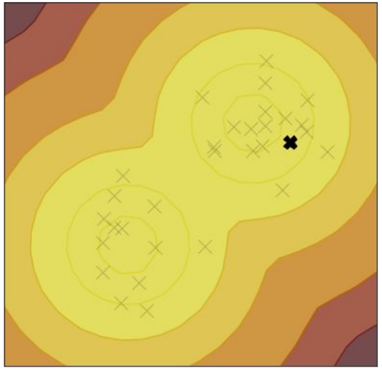
Hamiltonian Monte Carlo

Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from

Hamiltonian Monte Carlo

Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from

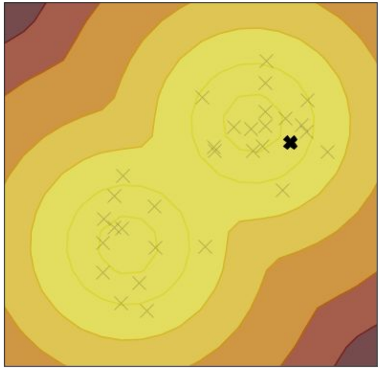
start at prev. sample



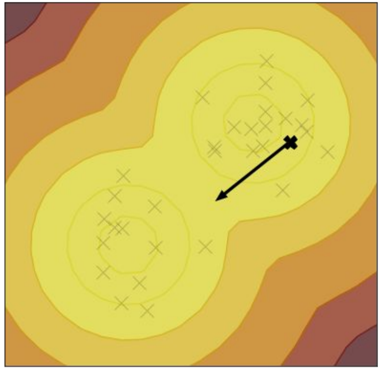
Hamiltonian Monte Carlo

Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from

start at prev. sample



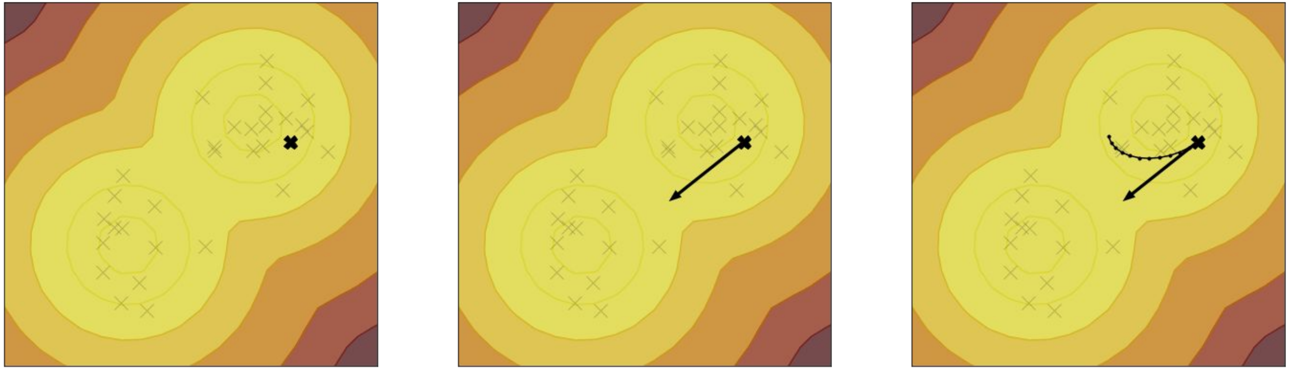
random momentum



Hamiltonian Monte Carlo

Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from

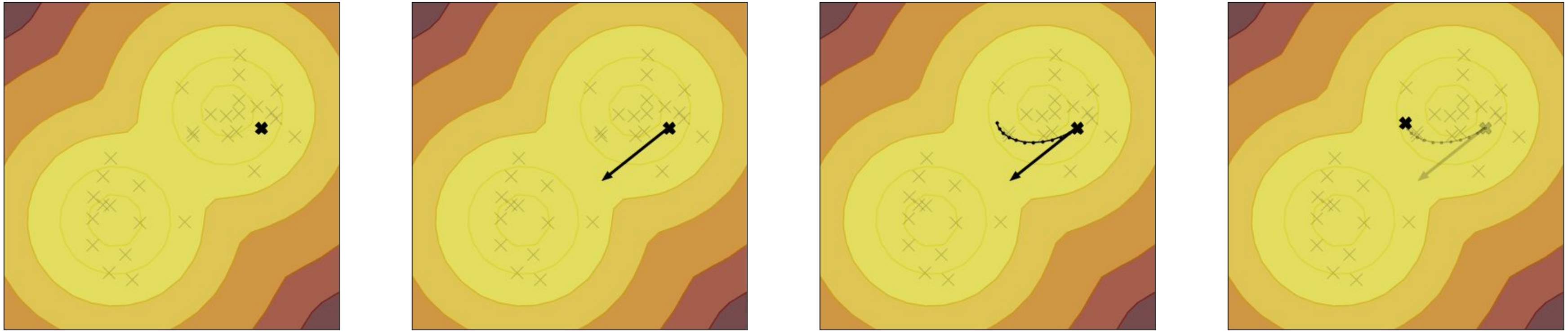
start at prev. sample | random momentum | simulate dynamics



Hamiltonian Monte Carlo

Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from

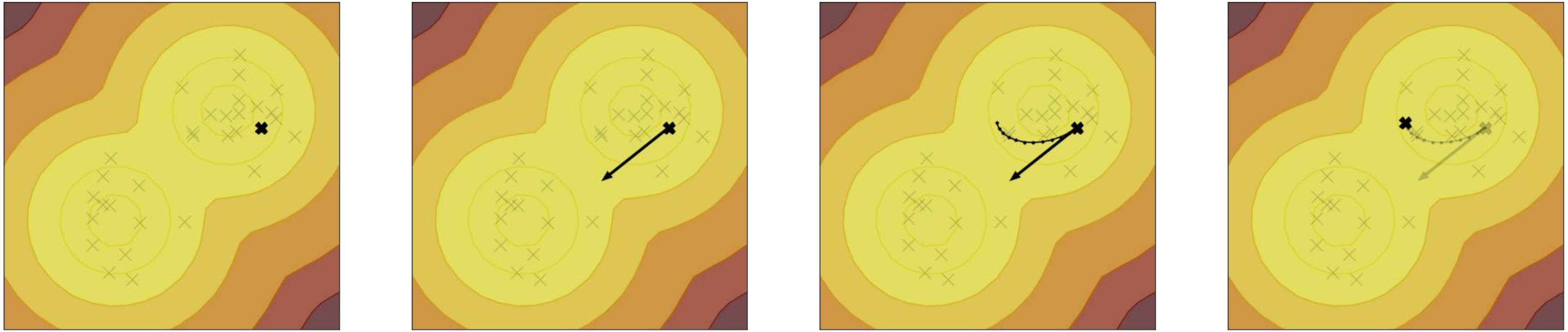
start at prev. sample | random momentum | simulate dynamics | accept / reject



Hamiltonian Monte Carlo

Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from

start at prev. sample | random momentum | simulate dynamics | accept / reject



+ Asymptotically exact
+ Well-studied and understood
+ Has been used in early BNNs

- Requires exact gradients
- Generally expensive

Computational complexity of HMC

Do the inference as accurately as possible, ignoring scalability and practicality

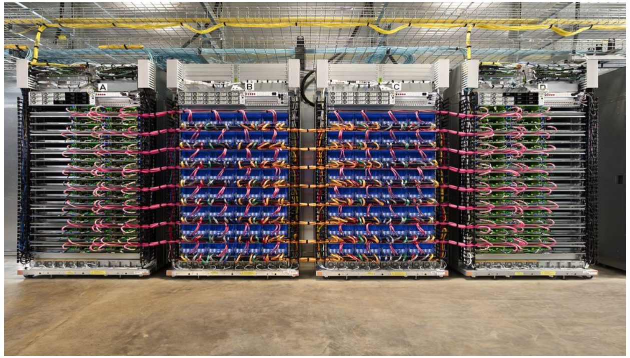
- Most recent papers on BNNs do no more than 1-5 thousand epochs
- For example, to approximate the posterior of a ResNet-20 on CIFAR-10 we spend *60 million epochs* of compute

Computational complexity of HMC

Do the inference as accurately as possible, ignoring scalability and practicality

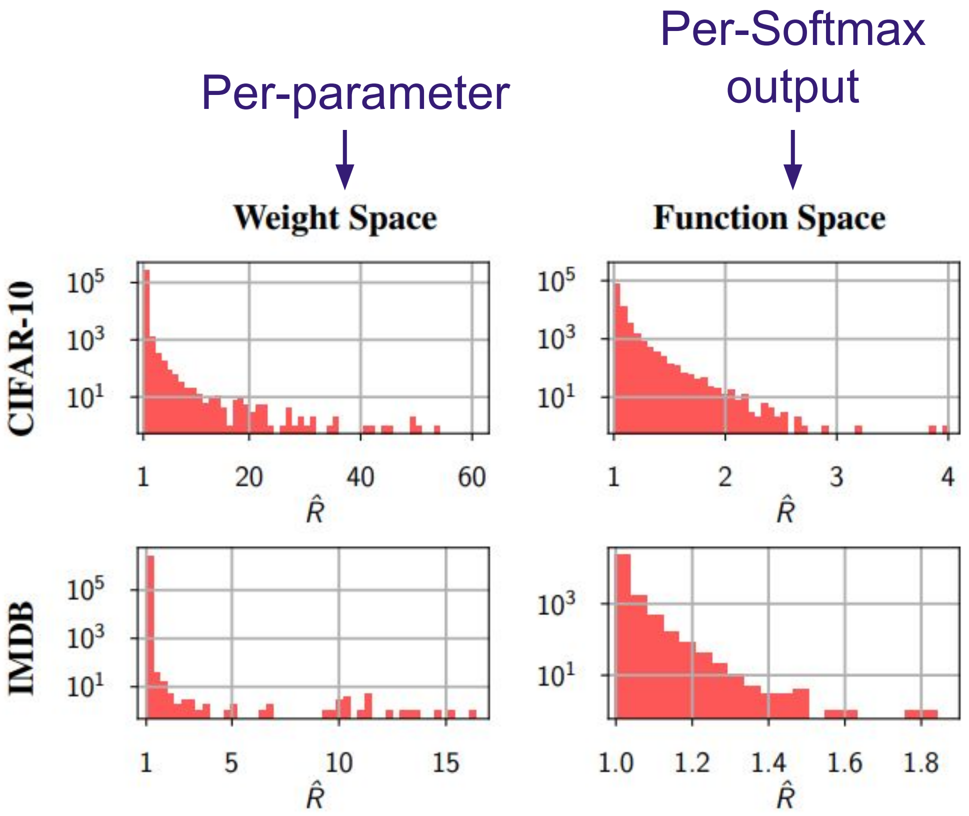
- Most recent papers on BNNs do no more than 1-5 thousand epochs
- For example, to approximate the posterior of a ResNet-20 on CIFAR-10 we spend *60 million epochs* of compute

To cope with extreme compute requirements we run HMC on 512 TPUs!



How well is HMC mixing?

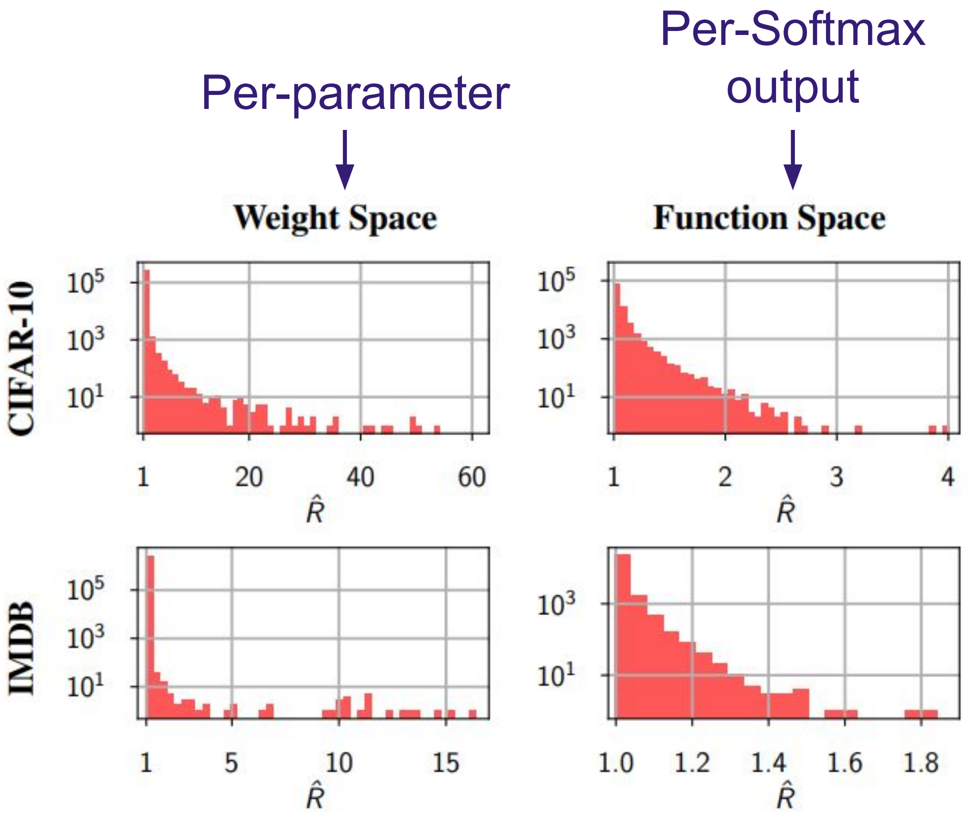
$$\hat{R} \approx \frac{\text{between-chain variance}}{\text{avg within-chain variance}}$$



How well is HMC mixing?

$$\hat{R} \approx \frac{\text{between-chain variance}}{\text{avg within-chain variance}}$$

Most R̂ are close to 1, especially in function space!

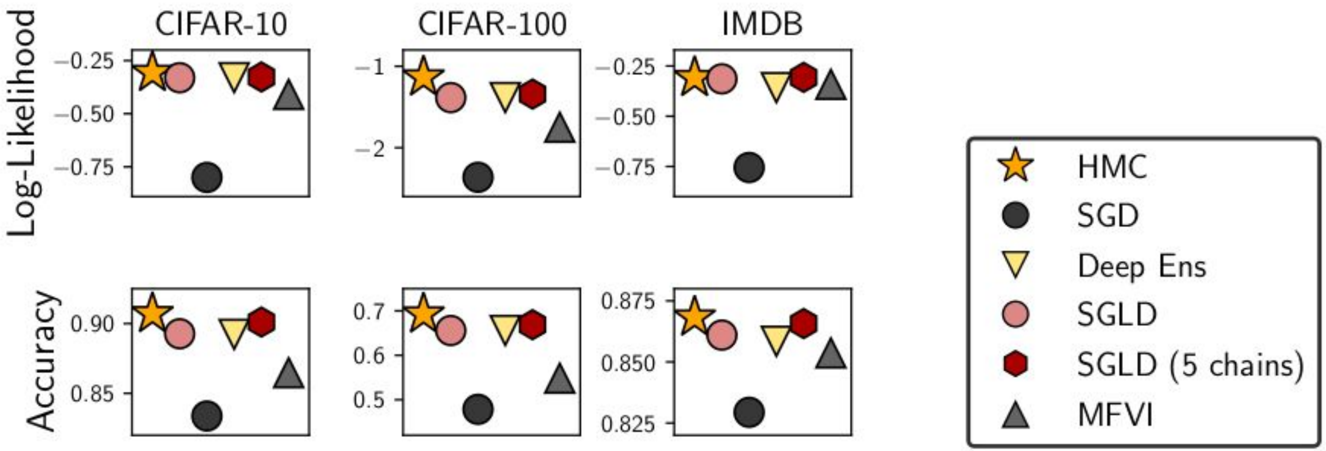


Answering Questions about Bayesian Neural Networks

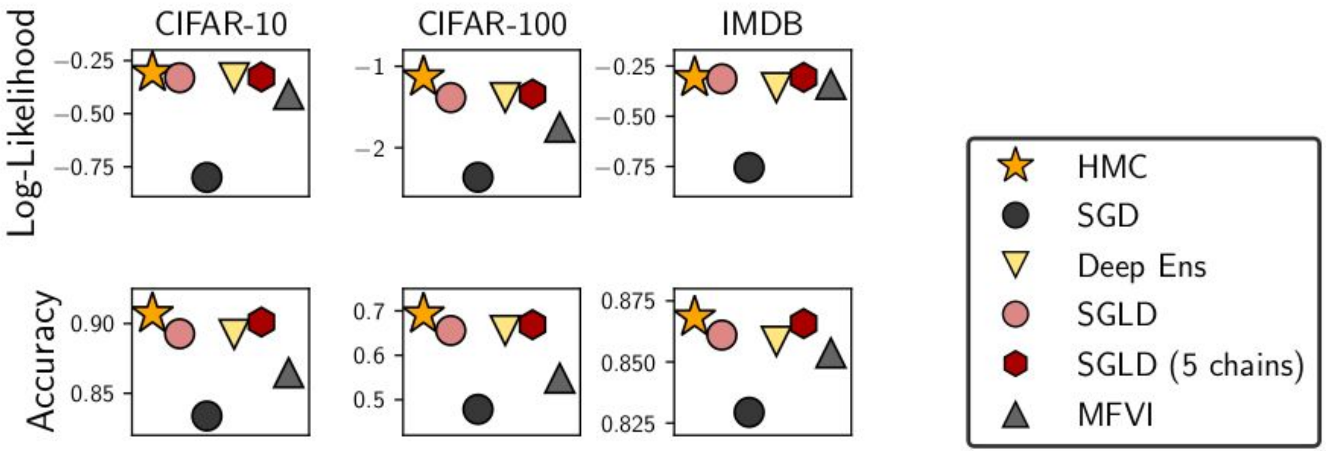
Answering Questions about Bayesian Neural Networks

🔭 🌌

Q1: Do BNNs perform well in practice?



Q1: Do BNNs perform well in practice?

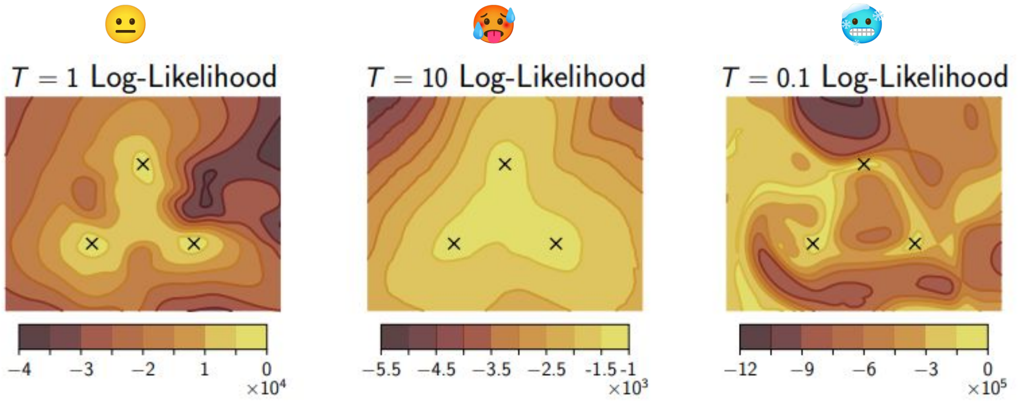


HMC BNNs outperform deep ensembles at temperature *T=1*!

Q2: Do we need cold posteriors?

$$p_T(w|\mathcal{D}) \propto \left(p(\mathcal{D}|w) \cdot p(w)\right)^{1/T}$$

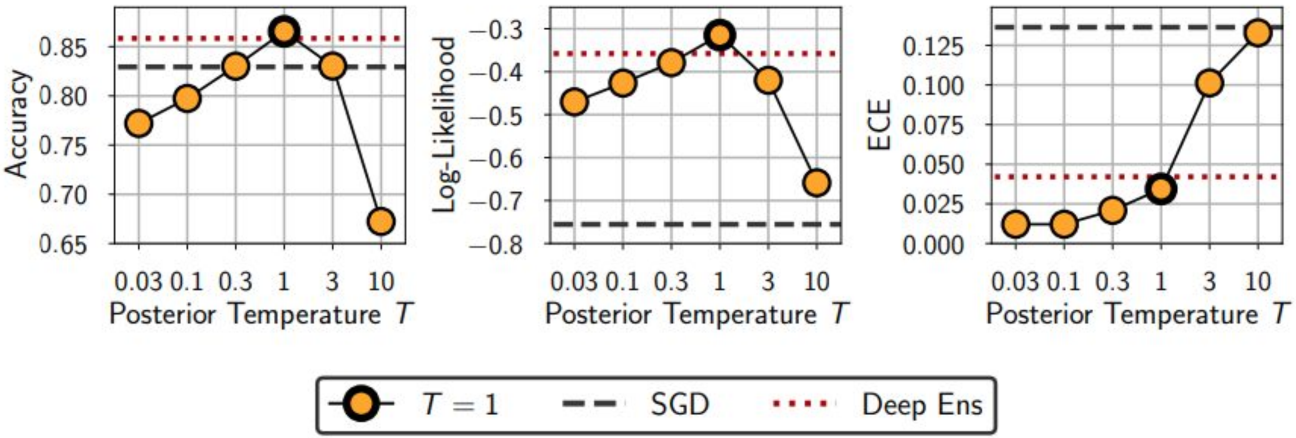
Cold posteriors effect by [Wenzel et al](): cold posteriors (temperatures T << 1) are needed to achieve good performance with BNNs



Cold posteriors → sharper distribution, concentrated on high-density points

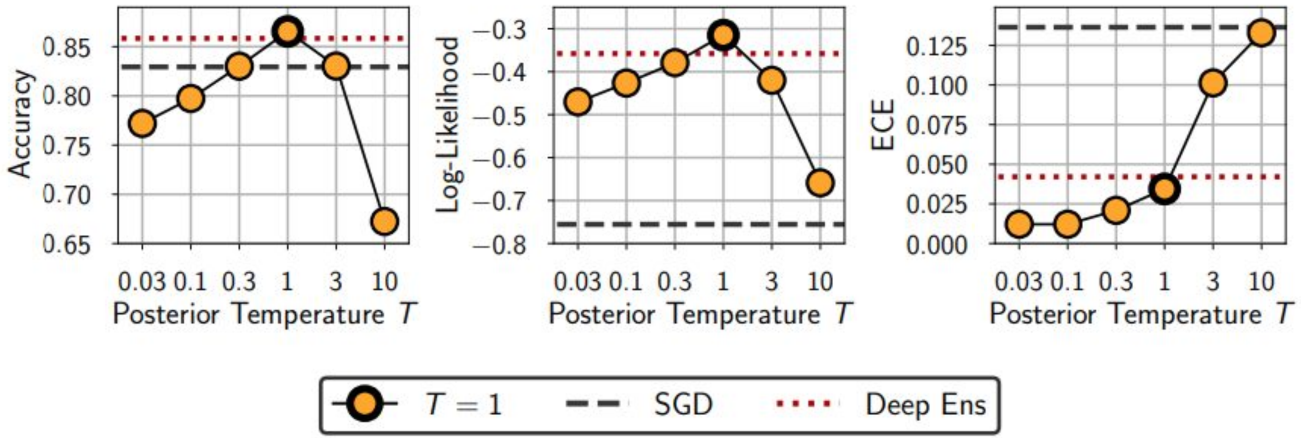
Q2: Do we need cold posteriors?

- We have already seen that BNNs can do well at T=1
- What is the effect of T then?



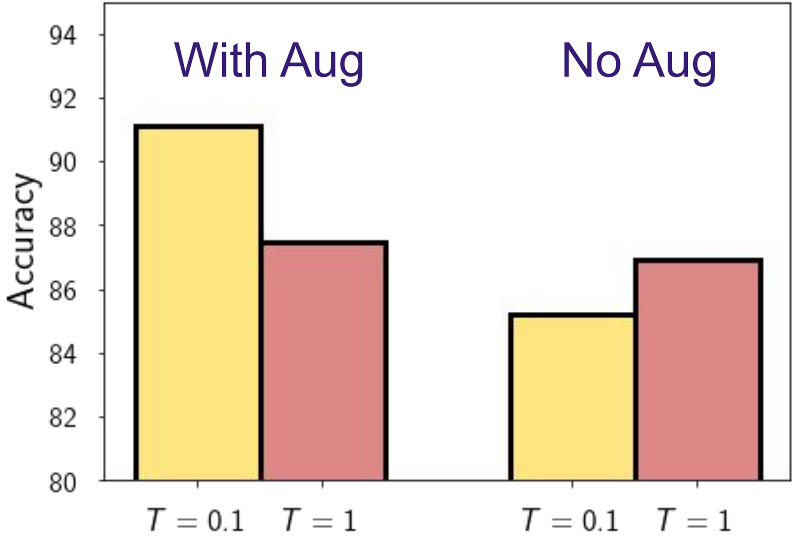
Q2: Do we need cold posteriors?

- We have already seen that BNNs can do well at T=1
- What is the effect of T then?



What’s the difference with Wenzel et al.?

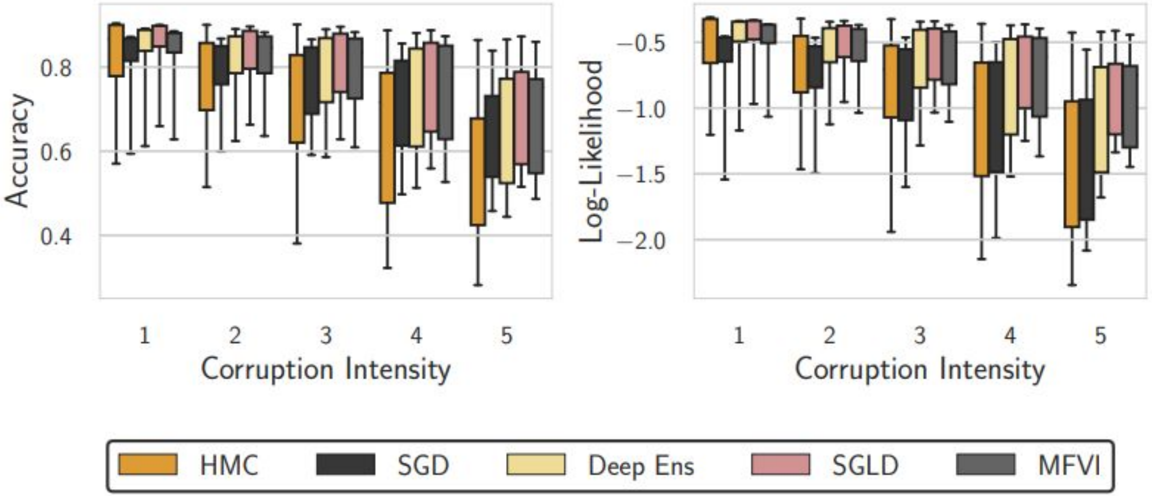
- Results using the original code of Wenzel et al. on CIFAR-10:



With no data augmentation, there is no cold posteriors effect.

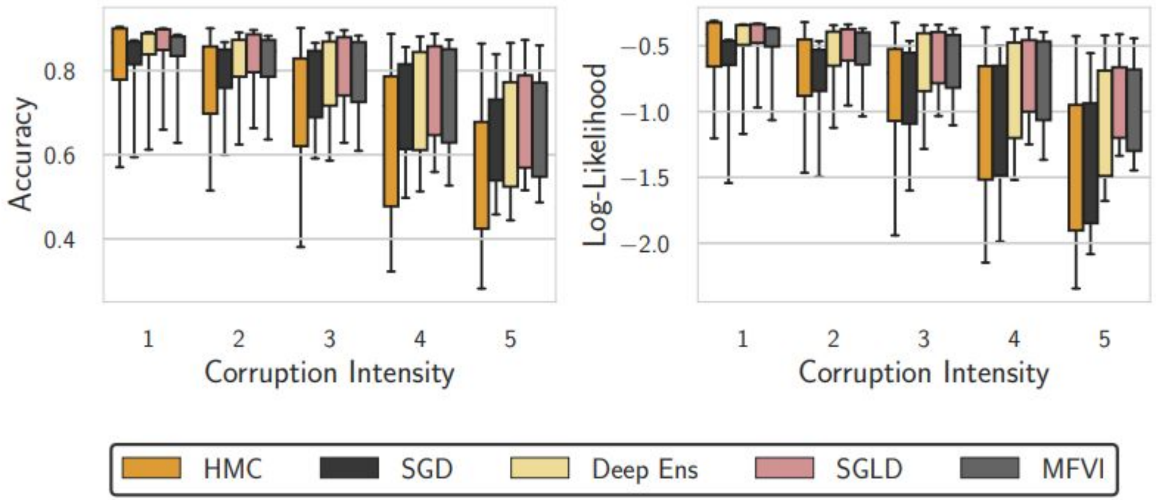
Q3: Are BNNs robust to covariate shift?

Train on CIFAR-10, test on CIFAR-10-C:



Q3: Are BNNs robust to covariate shift?

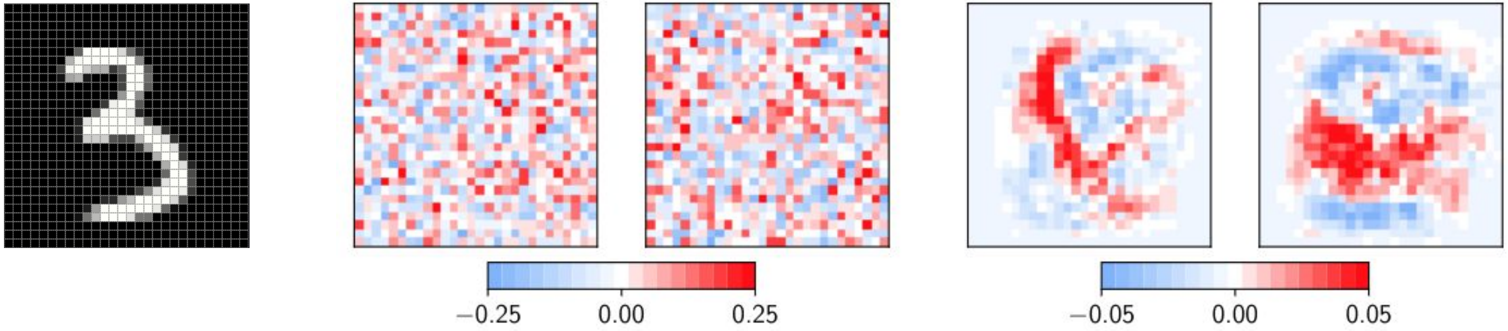
Train on CIFAR-10, test on CIFAR-10-C:



HMC BNNs are *terrible* on corrupted data!

Q3: Are BNNs robust to covariate shift?

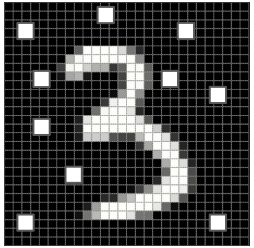
See "*Dangers of Bayesian model averaging under covariate shift*" by Izmailov, Nicholson, Lotfi, Wilson for a detailed explanation



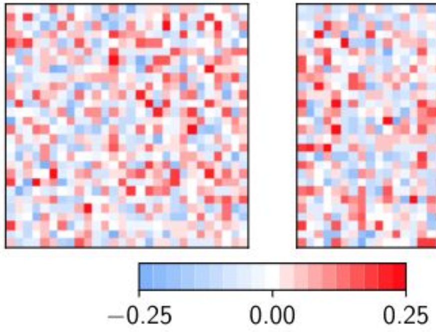
MNIST digit BNN weights SGD weights

Q3: Are BNNs robust to covariate shift?

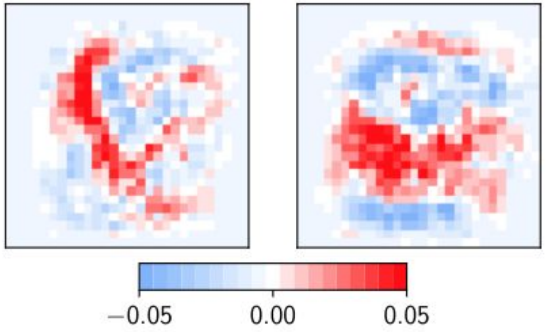
See “*Dangers of Bayesian model averaging under covariate shift*” by Izmailov, Nicholson, Lotfi, Wilson for a detailed explanation



Noisy MNIST digit



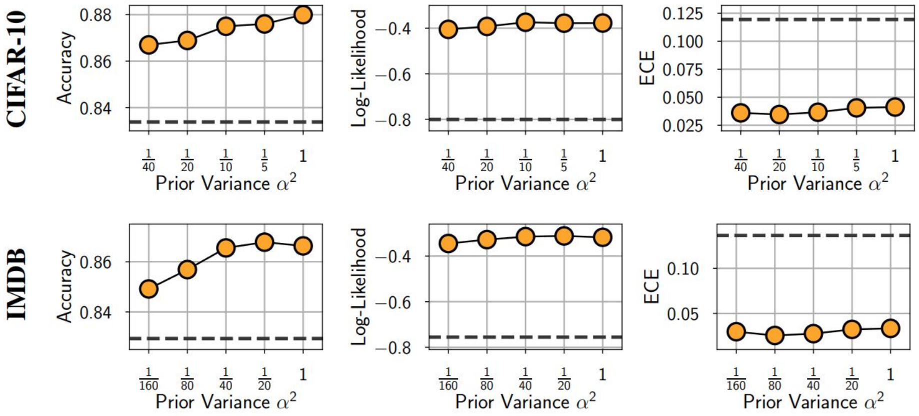
BNN weights



SGD weights

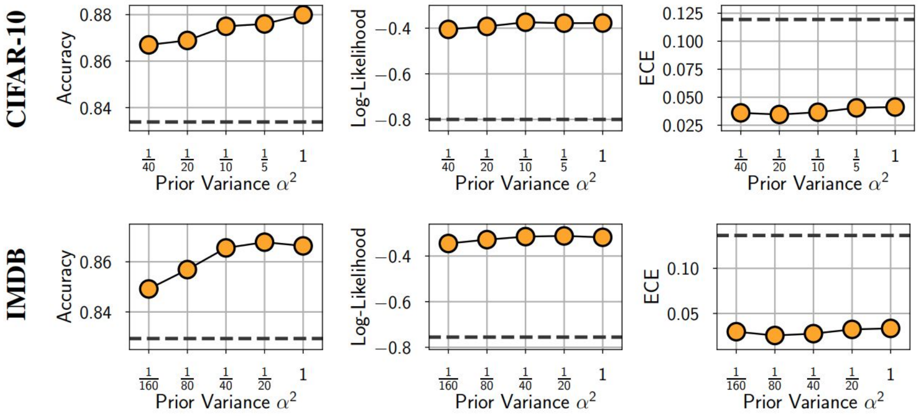
Q4: What is the effect of priors in BNNs?

Consider priors of the form $\mathcal{N}(0, \alpha^2 I)$.



Q4: What is the effect of priors in BNNs?

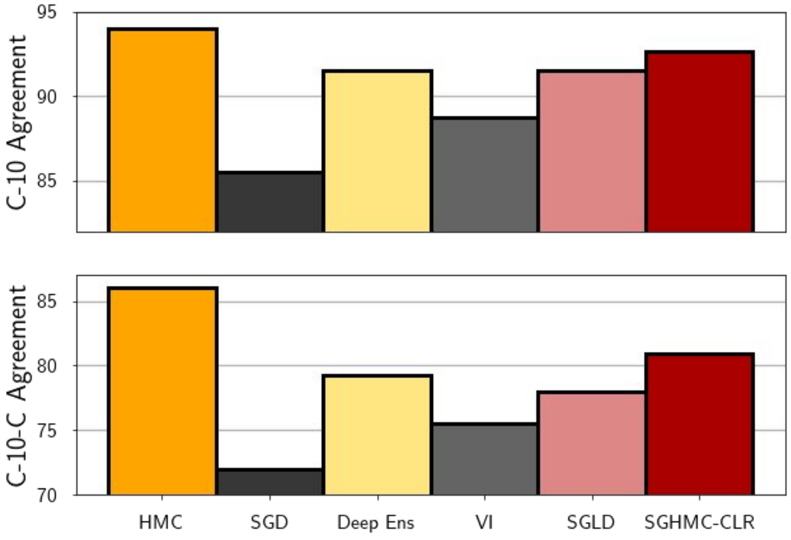
Consider priors of the form $\mathcal{N}(0, \alpha^2 I)$.



- High-variance Gaussian priors lead to strong performance
- The results are robust with respect to the prior scale

Q5: How good are approximate inference methods?

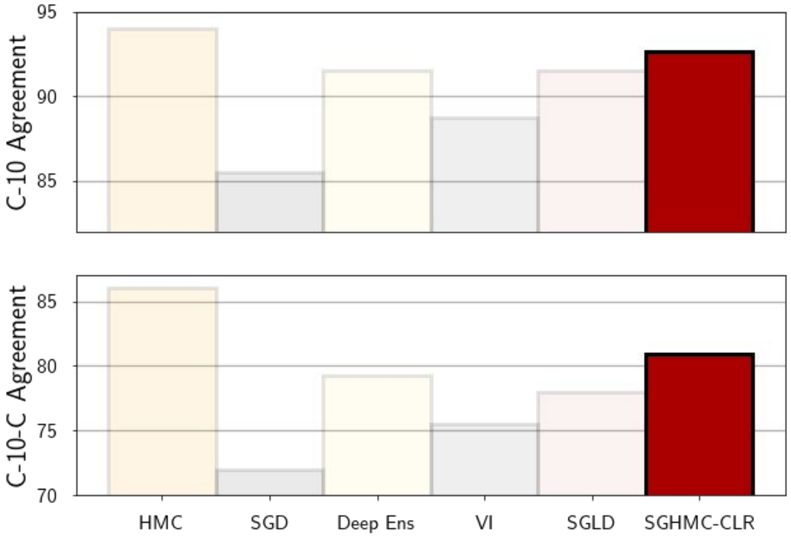
We compare the predictions of HMC to that of scalable BDL methods.



All scalable methods make predictions distinct from HMC

Q5: How good are approximate inference methods?

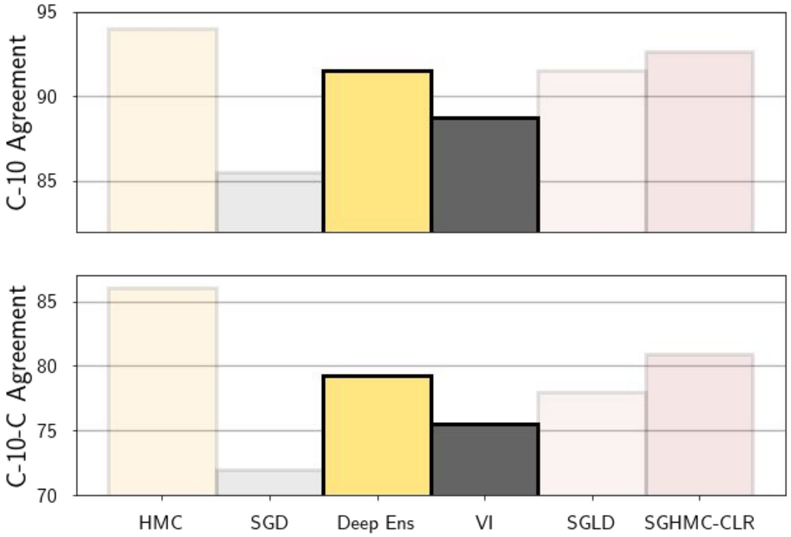
We compare the predictions of HMC to that of scalable BDL methods.



Advanced SGMCMC methods are closer to HMC than other methods

Q5: How good are approximate inference methods?

We compare the predictions of HMC to that of scalable BDL methods.



Deep ensembles are closer to HMC than VI!

Discussion

- BNNs outperform SGD and Deep Ensembles and do not require cold posteriors
- The cold posterior effect reported in prior work is largely an artifact of data augmentation
- BNNs are terrible when the test data is corrupted
- Deep ensembles are making more similar predictions to HMC BNNs compared to MFVI

We release our HMC samples!

We are organizing a NeurIPS 2021 competition on approximate inference in BDL!