Deep Ensembles as Approximate Bayesian Inference

https://cims.nyu.edu/~andrewgw/deepensembles/

Article by Andrew Gordon Wilson, Pavel Izmailov

With some content from Bayesian Deep Learning and a Probabilistic Perspective of Generalization (paper from same lab) - we'll briefly go through other parts of the paper at the end of the slides

What are deep ensembles?

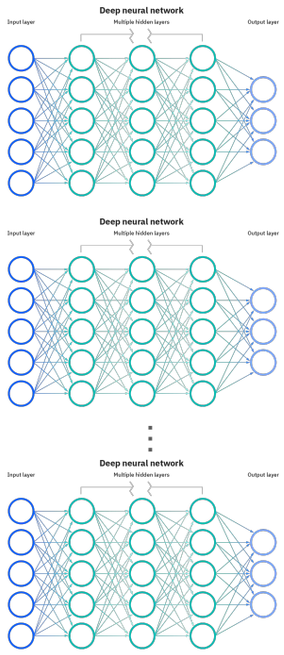
What are deep ensembles?

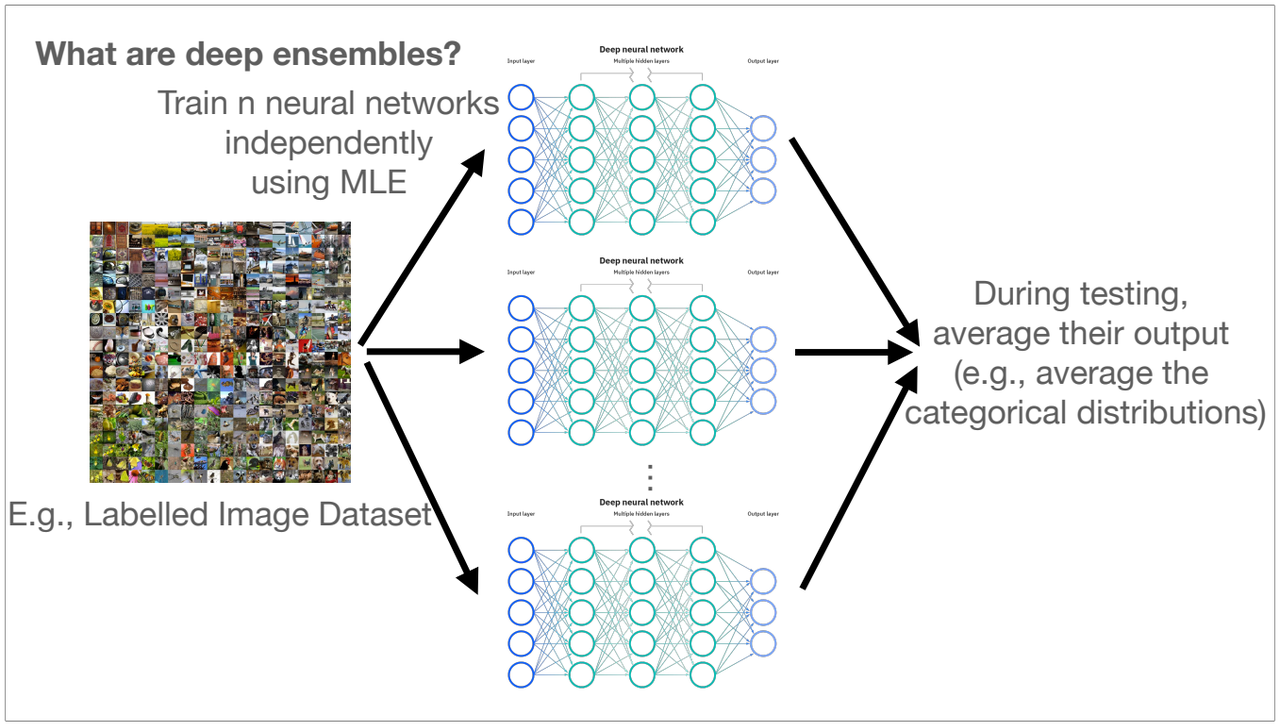
E.g., Labelled Image Dataset

What are deep ensembles?

Train n neural networks independently using MLE

E.g., Labelled Image Dataset

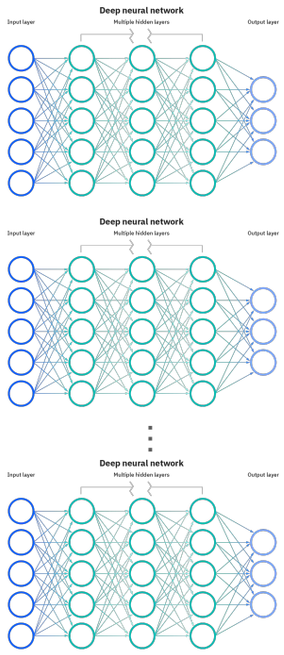



What are deep ensembles?
Train n neural networks independently using MLE
E.g., Labelled Image Dataset
Deep neural network
Input layer
Multiple hidden layers
Output layer
Deep neural network
Input layer
Multiple hidden layers
Output layer
Deep neural network
Input layer
Multiple hidden layers
Output layer
During testing, average their output (e.g., average the categorical distributions)

What are deep ensembles?

Train n neural networks independently using MLE

E.g., Labelled Image Dataset



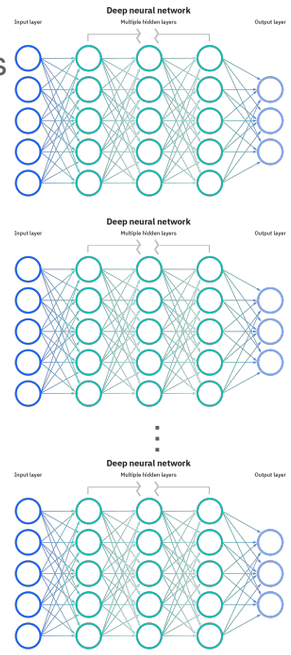
During testing, average their output (e.g., average the categorical distributions)

Improves upon single network in terms of test performance!

What are deep ensembles?

Train n neural networks independently using MLE

E.g., Labelled Image Dataset



But is this Bayesian? (And the article and paper try to argue for "yes")

During testing, average their output (e.g., average the categorical distributions)

Improves upon single network in terms of test performance!

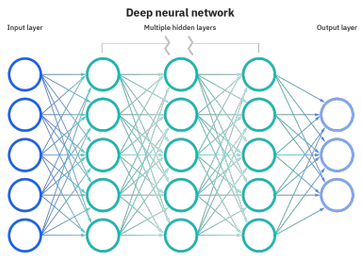
Big picture recap

What we want to evaluate the BMA:

$$p(y \mid x, D) = \int p(y \mid x, \theta) p(\theta \mid D) d\theta$$

Model's output Posterior Density

But neural network can have millions of parameters!



Must **resort** to approx. Bayesian inference

Popular approaches:
- MCMC
- Deterministic methods

(We will likely read about them in later weeks!)

Popular approaches: directly ported from bayesian inference for simpler models

MCMC: draws representative sample from the posterior over parameters
Deterministic methods like variational inference: captures a single mode of a possibly multimodal posterior

A third perspective

A third perspective

$$p(\theta \mid D)$$

A third perspective

A posterior over a million dimensional space

$p(\theta \mid D)$

10-100 samples from this space

A third perspective

A posterior over a million dimensional space

$p(\theta \mid D)$

10-100 samples from this space

What is the best way to allocate these samples?

What is not the best way to allocate these samples?

A third perspective

A posterior over a million dimensional space

$p(\theta \mid D)$

10-100 samples from this space

What is the best way to allocate these samples?

What is not the best way to allocate these samples?

Redundancy
Sampled solutions are very similar
(cannot provide good estimate of
epistemic uncertainty
and tend to be overconfident)

Problem with variational methods

A third perspective

A posterior over a million dimensional space

$p(\theta \mid D)$

10-100 samples from this space

What is the best way to allocate these samples?

What is <u>not</u> the best way to allocate these samples?

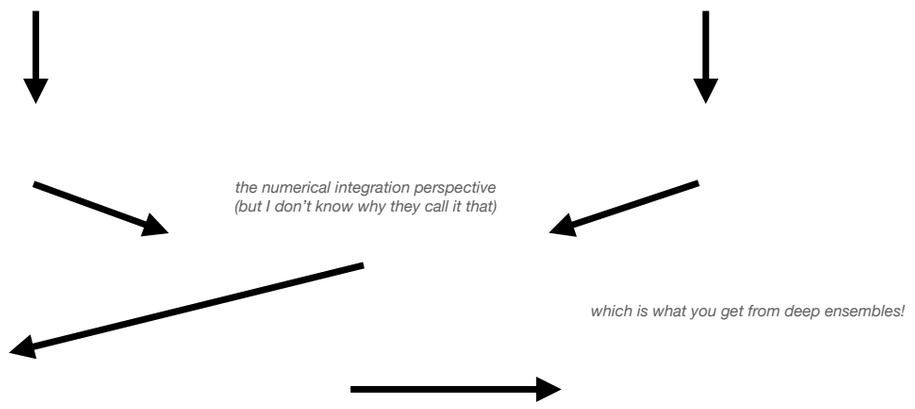
Redundancy
Sampled solutions are very similar (cannot provide good estimate of epistemic uncertainty and tend to be overconfident)

Problem with variational methods

Low density
Solutions with low posterior density contribute little to the true BMA

Problem with MCMC

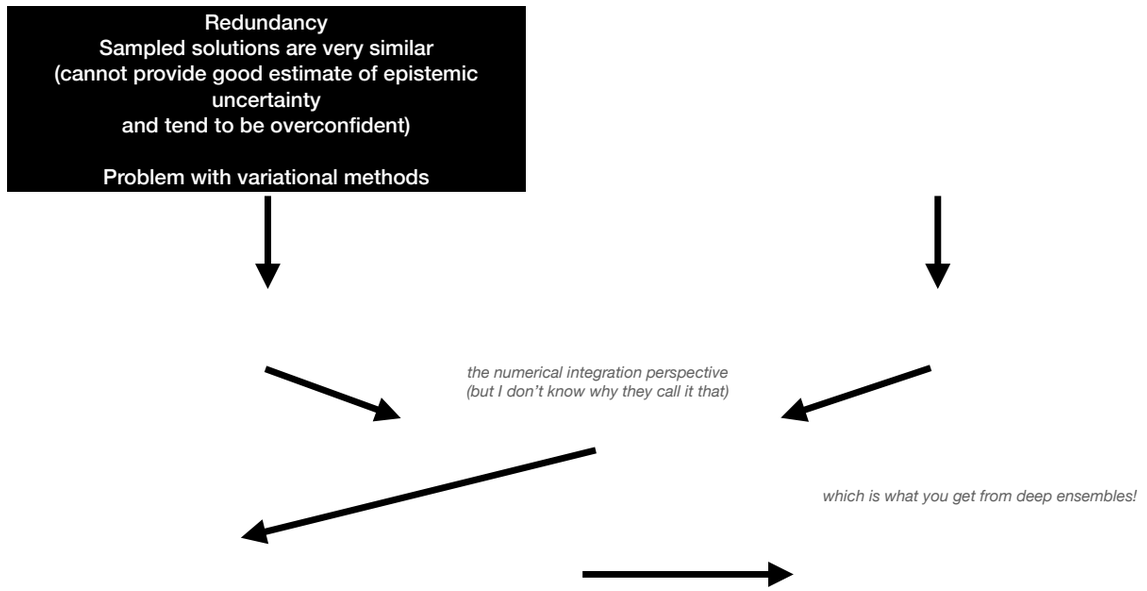
Potential resolution & connection to deep ensembles



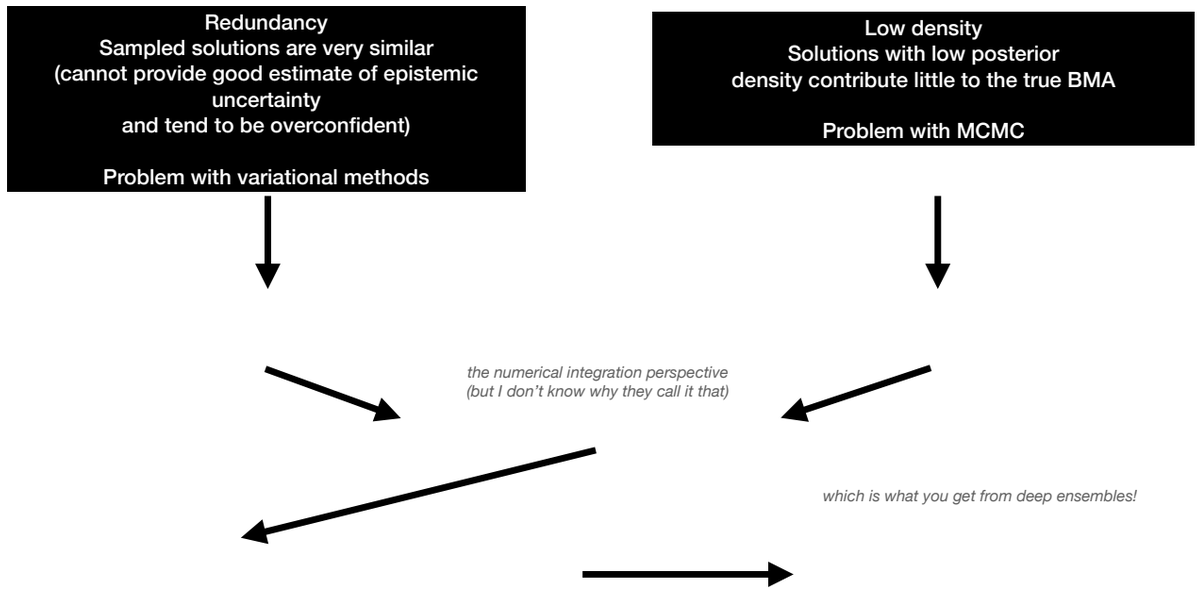
Some people have tried to argue that deep ensembling is not Bayesian.

But the authors have made several good arguments against that. Basically, all bayesian inference for deep learning must be approximate, meaning that it is one or the other not exactly doing the integral, but is indeed inspired by such approach. And methods fall on a spectrum, people shouldn't divide work arbitrarily into Bayesian vs not Bayesian.

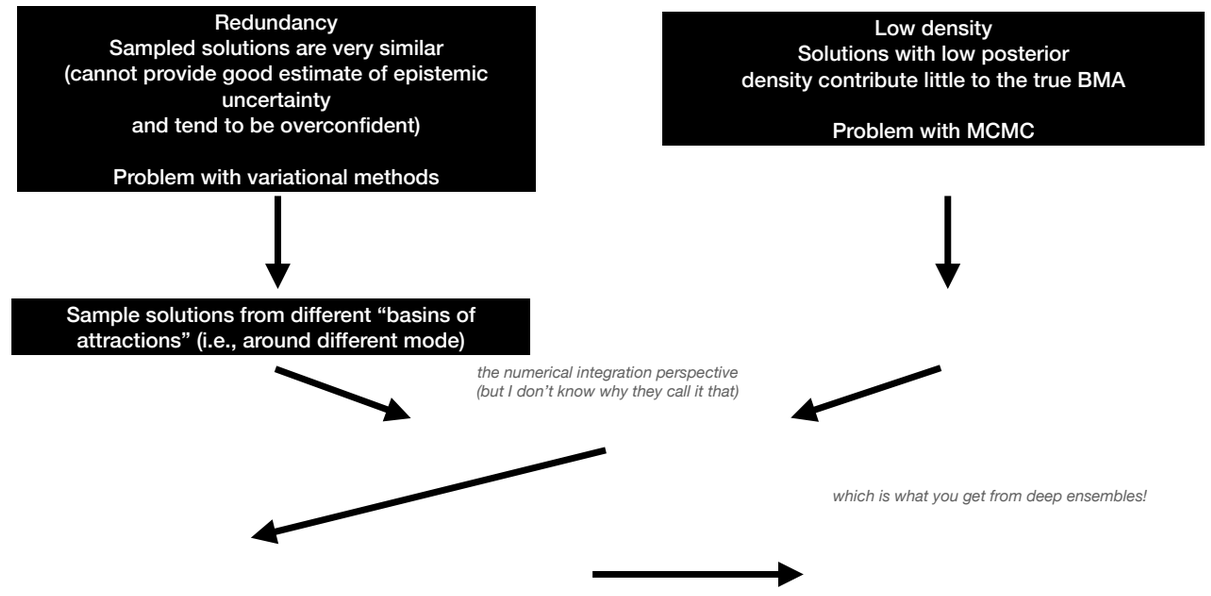
Potential resolution & connection to deep ensembles



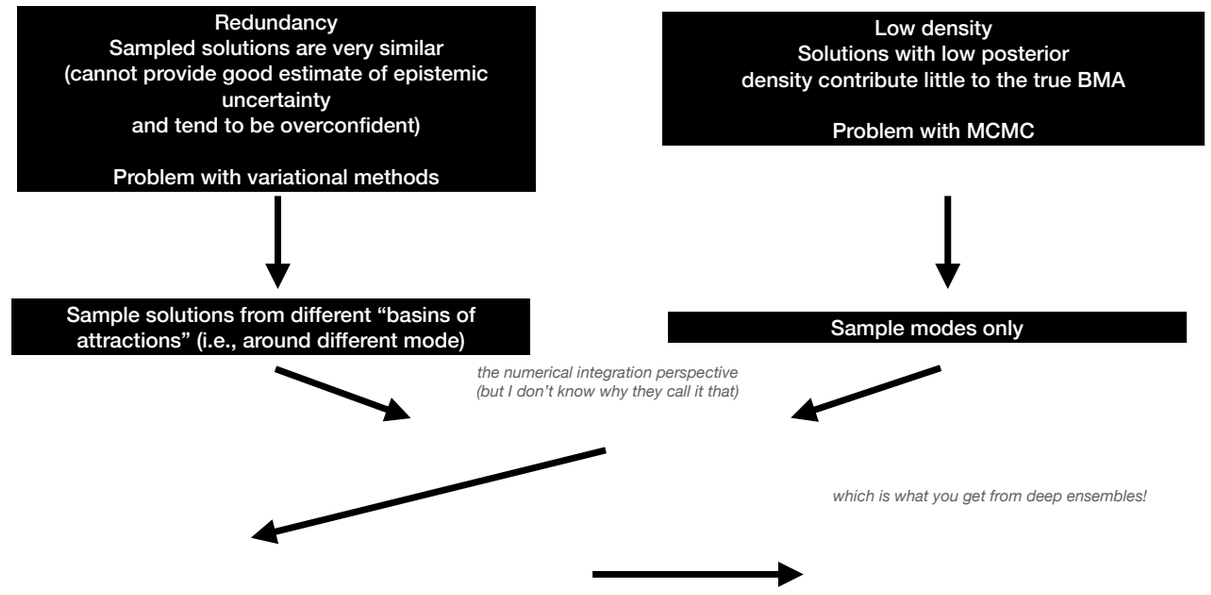
Potential resolution & connection to deep ensembles



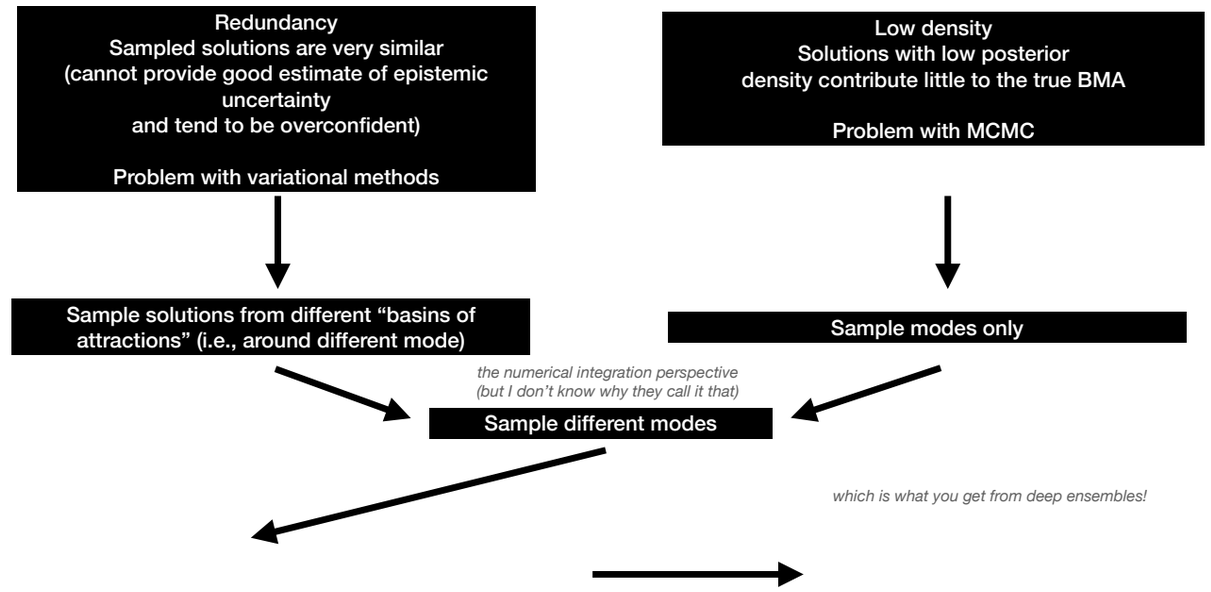
Potential resolution & connection to deep ensembles



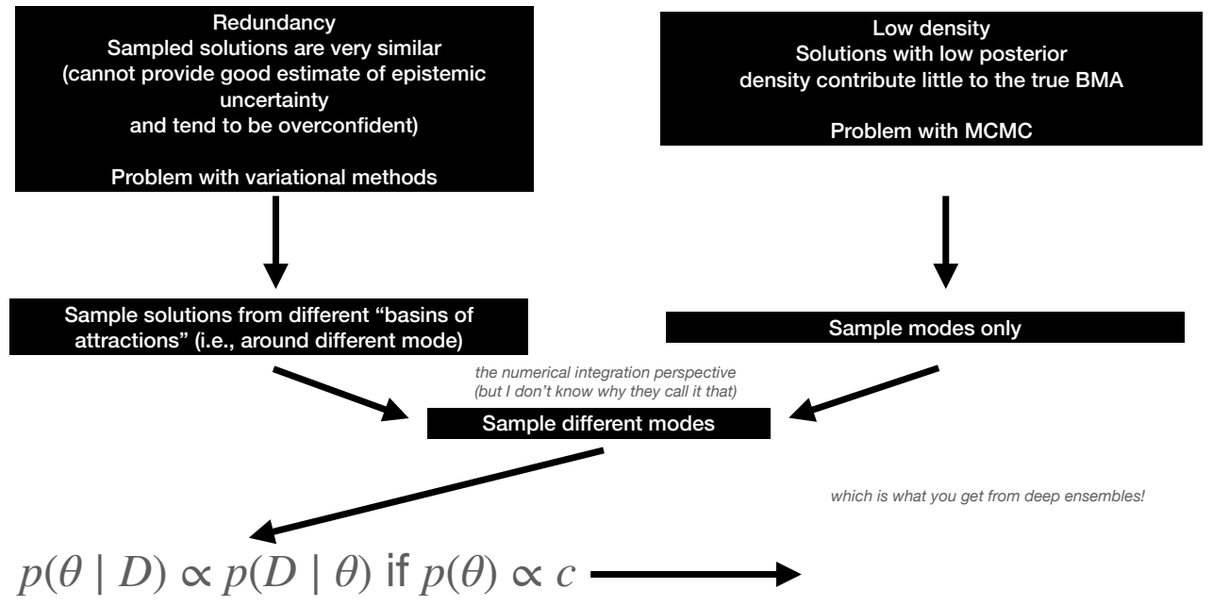
Potential resolution & connection to deep ensembles



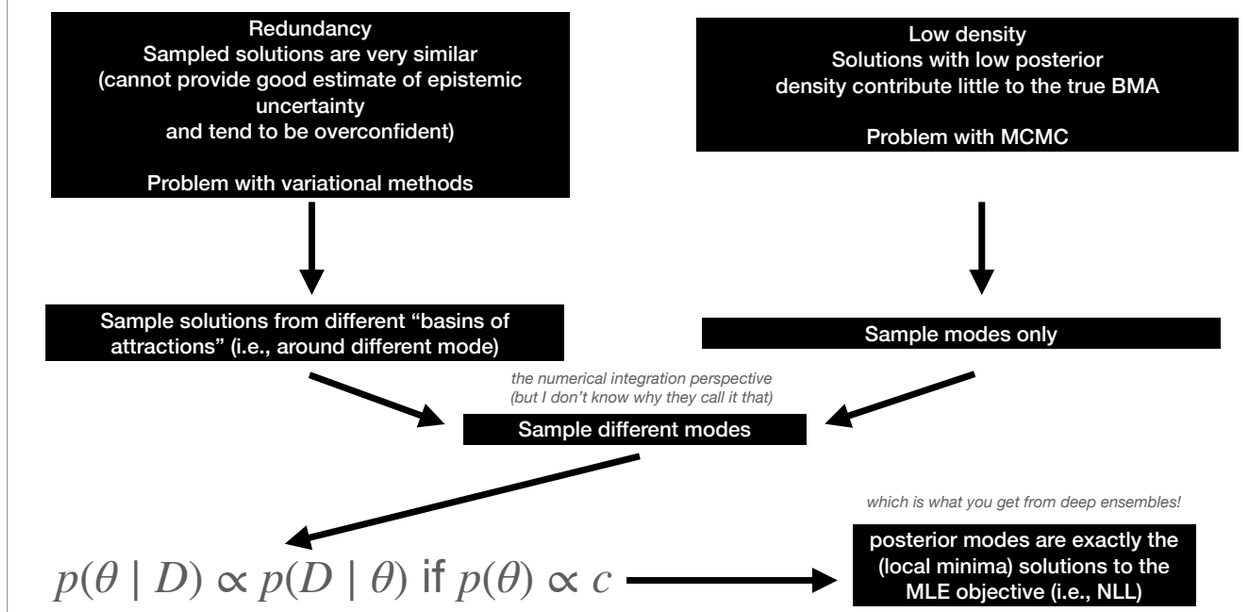
Potential resolution & connection to deep ensembles



Potential resolution & connection to deep ensembles



Potential resolution & connection to deep ensembles



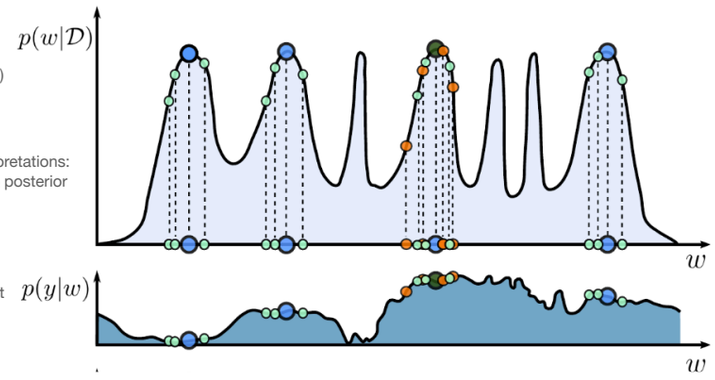
A conceptualization of what we just talked about and how this inspires a new approach

In reality may not be equally tall (but experiments suggest that they should be similar because variance in accuracy is quite small)

recall

so this y-axis have 2 interpretations:
- MLE training objective & posterior

y is some unseen data point



Deep Ensembles VI

I've blocked some portions of the plot to focus on some more relevant stuff, but check out the paper for more details!

A conceptualization of what we just talked about and how this inspires a new approach

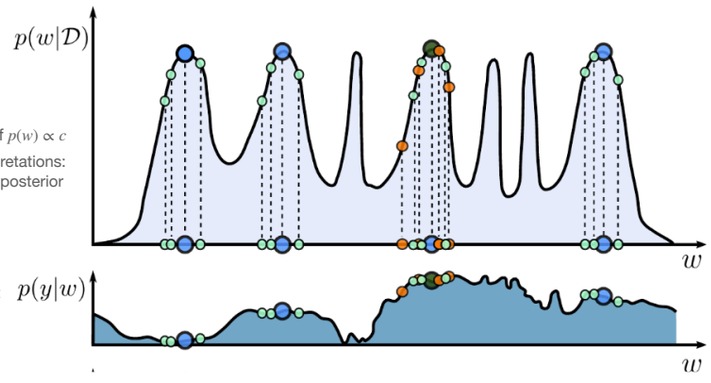
In reality may not be equally tall (but experiments suggest that they should be similar because variance in accuracy is quite small)

recall

$p(w \mid D) \propto p(D \mid w)$ if $p(w) \propto c$

so this y-axis have 2 interpretations:
- MLE training objective & posterior

y is some unseen data point



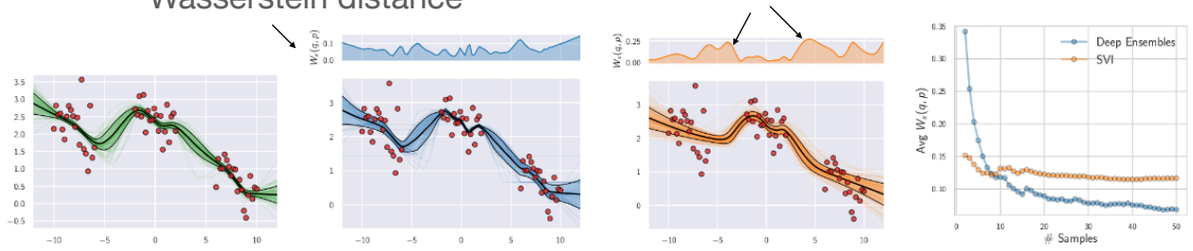
I've blocked some portions of the plot to focus on some more relevant stuff, but check out the paper for more details!

An experiment

Idea to get across: deep ensembles approximate the posterior predictive better than methods that focus on high-fidelity approximation of a single mode

Conditional Wasserstein distance

Over confidence for unseen regions



(a) Exhaustive HMC

combining 200 HMC chains

(b) Deep Ensembles

50 independently trained network

Look pretty similar to (a)!

(c) Variational Inference

factored VI

(d) Distance to BMA

Deep ensembles seems worse for smaller number of samples (?), but asymptotic performance is better

Another experiment

Deep ensembles & SWA only average over the modes

"We show that simple averaging of multiple points along the trajectory of SGD, with a cyclical or constant learning rate, leads to better generalization than conventional training."

MultiSWAG is more like the green dots (tend to improve test NLL compared to MultiSWA)

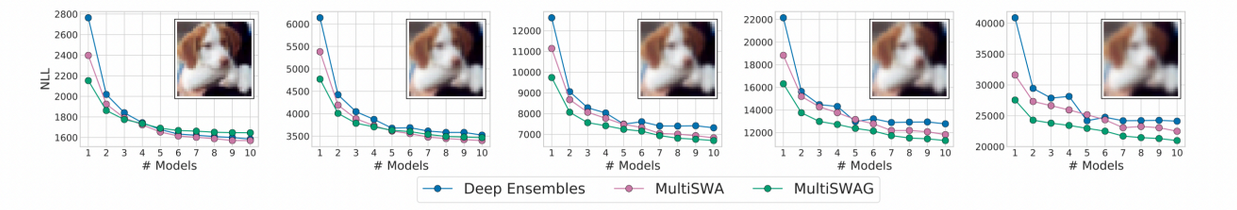
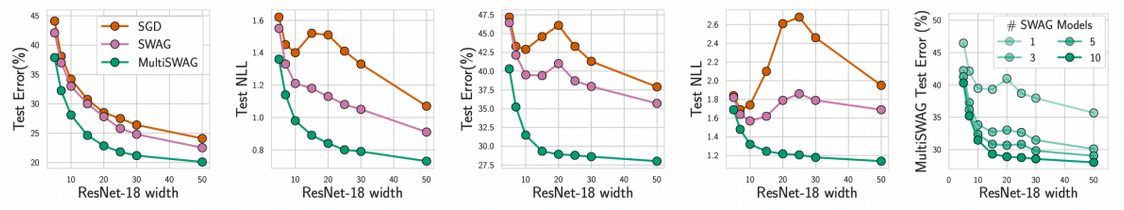


Figure 5. Negative log likelihood for Deep Ensembles, MultiSWAG and MultiSWA using a PreResNet-20 on CIFAR-10 with varying intensity of the *Gaussian blur* corruption. The image in each plot shows the intensity of corruption. For all levels of intensity, MultiSWAG and MultiSWA outperform Deep Ensembles for a small number of independent models. For high levels of corruption MultiSWAG significantly outperforms other methods even for many independent models. We present results for other corruptions in the Appendix.

Yet another experiment: Mitigating the mysterious double descent

*"Double descent (e.g., Belkin et al., 2019) describes **generalization error that decreases, increases, and then again decreases**, with increases in model flexibility. The first decrease and then increase is referred to as the **classical regime**: models with increasing flexibility are increasingly able to capture structure and perform better, until they begin to overfit. The next regime is referred to as the **modern interpolating regime**. The existence of the interpolation regime has been presented as mysterious generalization behaviour in deep learning."*



(a) True Labels (Err) (b) True Labels (NLL) (c) Corrupted (Err) (d) Corrupted (NLL) (e) Corrupted (# Models)

Figure 8. **Bayesian model averaging alleviates double descent. (a):** Test error and **(b):** NLL loss for ResNet-18 with varying width on CIFAR-100 for SGD, SWAG and MultiSWAG. **(c):** Test error and **(d):** NLL loss when 20% of the labels are randomly reshuffled. SWAG reduces double descent, and MultiSWAG, which marginalizes over multiple modes, entirely alleviates double descent both on the original labels and under label noise, both in accuracy and NLL. **(e):** Test errors for MultiSWAG with varying number of independent SWAG models; error monotonically decreases with increased number of independent models, alleviating double descent. We also note that MultiSWAG provides significant improvements in accuracy and NLL over SGD and SWAG models. See Appendix Figure 13 for additional results.

However, our perspective of generalization suggests that performance should monotonically improve as we increase model flexibility when we use Bayesian model averaging with a reasonable prior.

SWAG alleviates double descent, MultiSWAG with 10 models eliminates double descent.

Also want to highlight the performance gain.

Bayesian deep learning cannot overfit!

The single most important takeaway

For bayesian deep learning, we need to do approximate inference.

For sampling, we do not need a representative sample from the posterior over parameters.