Myths and Reality in Bayesian Deep Learning

Andrew Gordon Wilson

https://cims.nyu.edu/~andrewgw
New York University

International Society for Bayesian Analysis (ISBA) World Meeting
Session on Bayesian Deep Learning
Montreal, Canada
June 29, 2022

Collaborators:
Pavel Izmailov, Sanae Lotfi, Sanyam Kapoor, Wesley Maddox, Matt Hoffman, Sharad Vikram

Myths and Reality

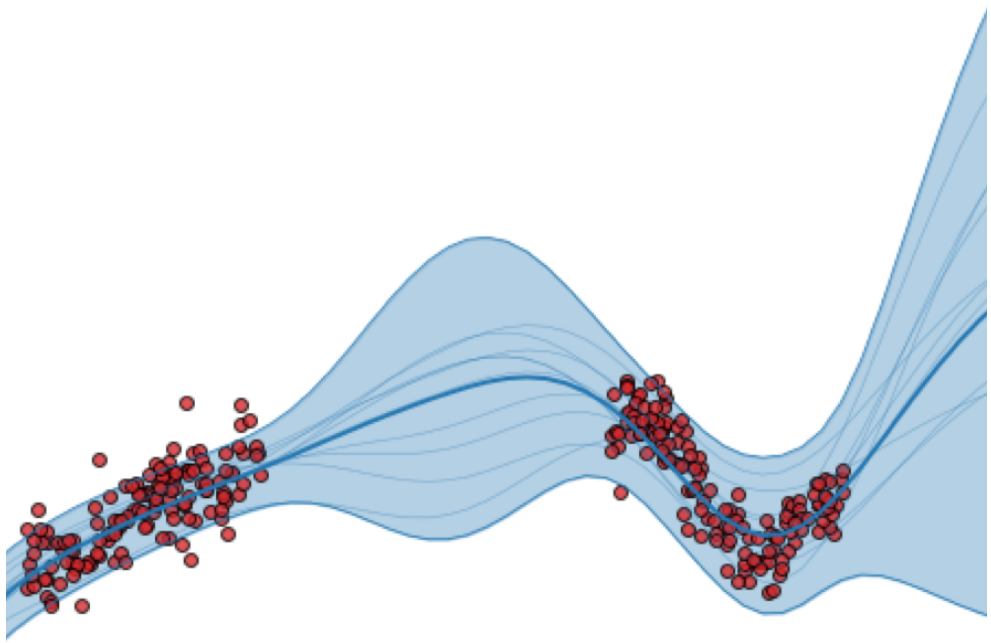
- In recent years, approximate inference for Bayesian deep learning has become *highly practical*, often improving accuracy and performance over conventional training, *with essentially no runtime overhead*.
- However, **many false narratives** persist about Bayesian deep learning.
- In this talk we will re-examine conventional wisdom:
  - *Does Bayesian deep learning provide good results?*
  - *Is Bayesian deep learning scalable?*
  - *Is Bayesian deep learning successfully applied in the real world?*
  - *Are standard (e.g. Gaussian) priors bad?*
  - *Are "cold posteriors" problematic?*
  - *Are "deep ensembles" a non-Bayesian competitor?*
  - *Are there practical challenges?*

Key references:

[1] *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*. Wilson & Izmailov, NeurIPS 2020.
[2] *What are Bayesian Neural Network Posteriors Really Like?* Izmailov et. al, ICML 2021.
[3] *Dangers of Bayesian Model Averaging under Covariate Shift*. Izmailov et. al, NeurIPS 2021.
[4] *On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification*. Kapoor et. al, arXiv 2022.

What is Bayesian learning?

- The key distinguishing property of a Bayesian approach is **marginalization** instead of optimization.
- Rather than use a single setting of parameters **w**, use all settings weighted by their posterior probabilities in a *Bayesian model average*.



Marginalization: why we should do Bayesian deep learning

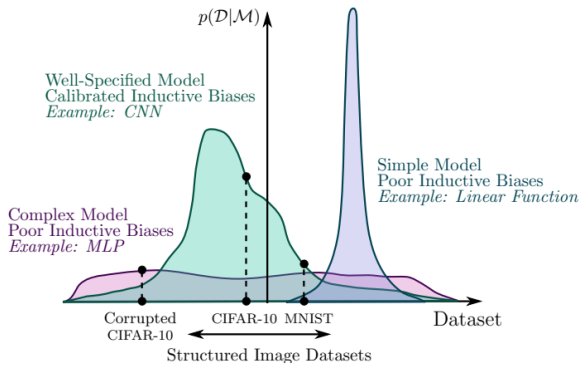
Sum rule: $p(x) = \sum_x p(x, y)$. **Product rule:** $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$.

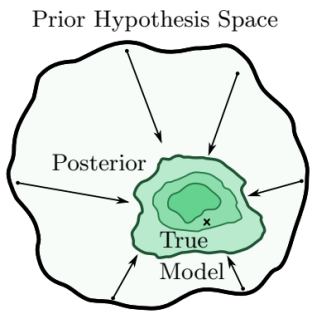
$$p(y|x_*, \mathbf{y}, X) = \int p(y|x_*, \mathbf{w})p(\mathbf{w}|\mathbf{y}, X)d\mathbf{w}\,. \tag{1}$$

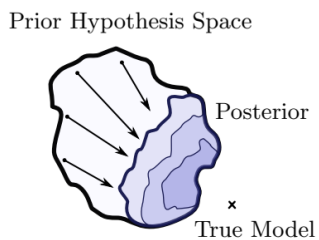
- Think of each setting of $\mathbf{w}$ as a different model. Eq. (1) is a *Bayesian model average*, an average of infinitely many models weighted by their posterior probabilities.
- Automatically calibrated complexity even with highly flexible models.
- Can view classical training as using an approximate posterior $q(\mathbf{w}|\mathbf{y}, X) = \delta(w = w_{\text{MAP}})$.
- Typically more interested in the induced distribution over **functions** than in parameters $\mathbf{w}$. Can be hard to have intuitions for priors on $p(\mathbf{w})$.

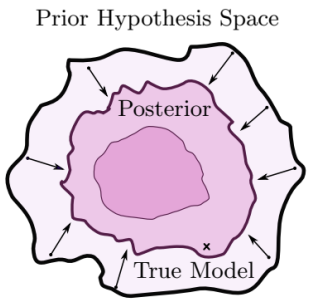
Model construction from a Bayesian perspective

- The ability for a system to learn is determined by its *support* (which solutions are a priori possible) and *inductive biases* (which solutions are a priori likely).
- From this perspective, it is not surprising that "over-parametrized" models perform well. We should not shy away from flexibility, and we should not conflate flexibility and model complexity.







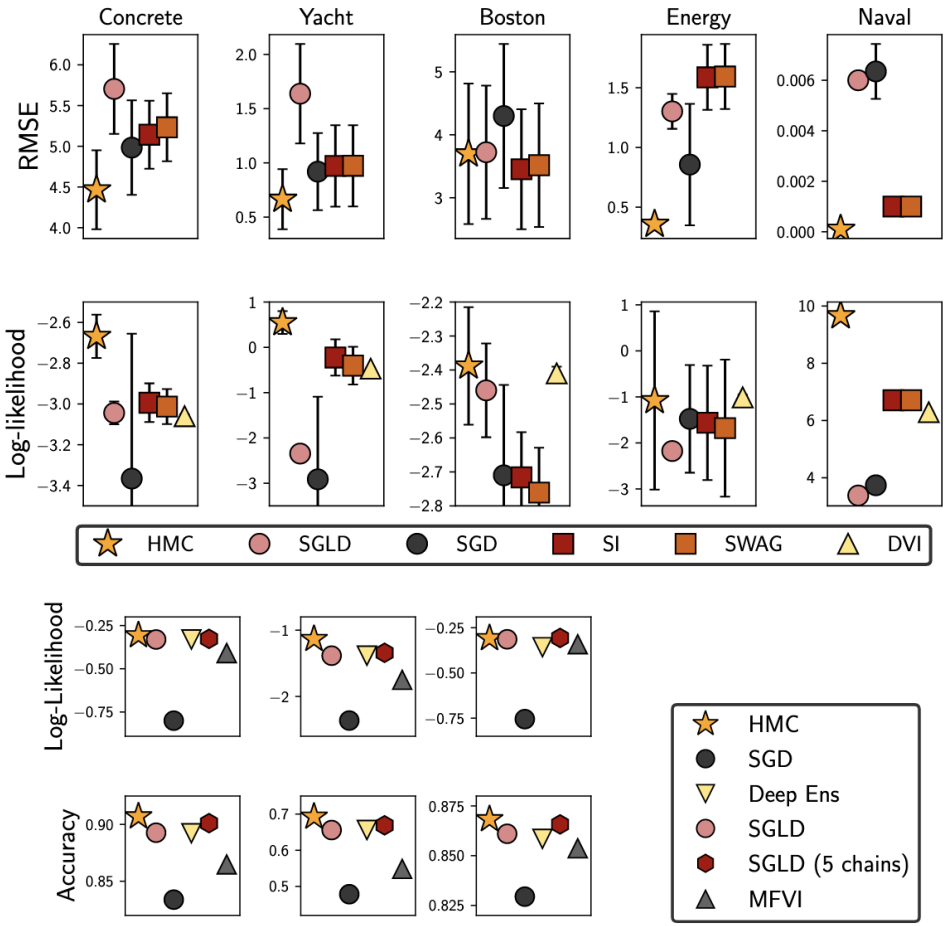


Bayesian Deep Learning and a Probabilistic Perspective of Generalization
Wilson and Izmailov, NeurIPS 2020

Is Bayesian Deep Learning Practical?

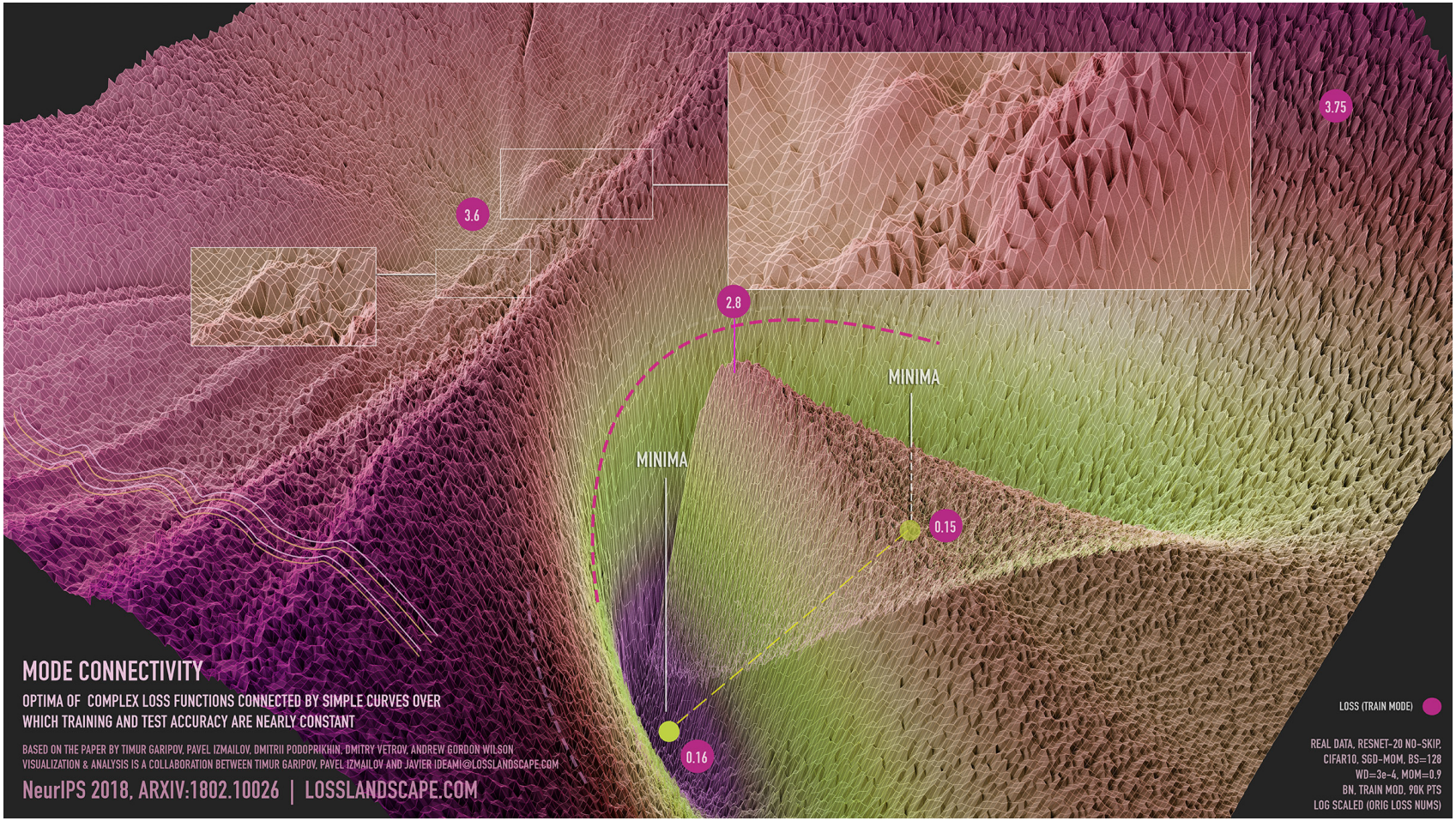
High Quality Inference is a Worthy Goal

- What happens in the best case?
- **High fidelity Hamiltonian Monte Carlo (HMC) inference** (distributed over hundreds of TPUs) provides notably **better generalization** than alternatives.



What are Bayesian Neural Network Posteriors Really Like?. Izmailov et. al, ICML 2021.

Scalable Inference by Exploiting Loss Surface Geometry

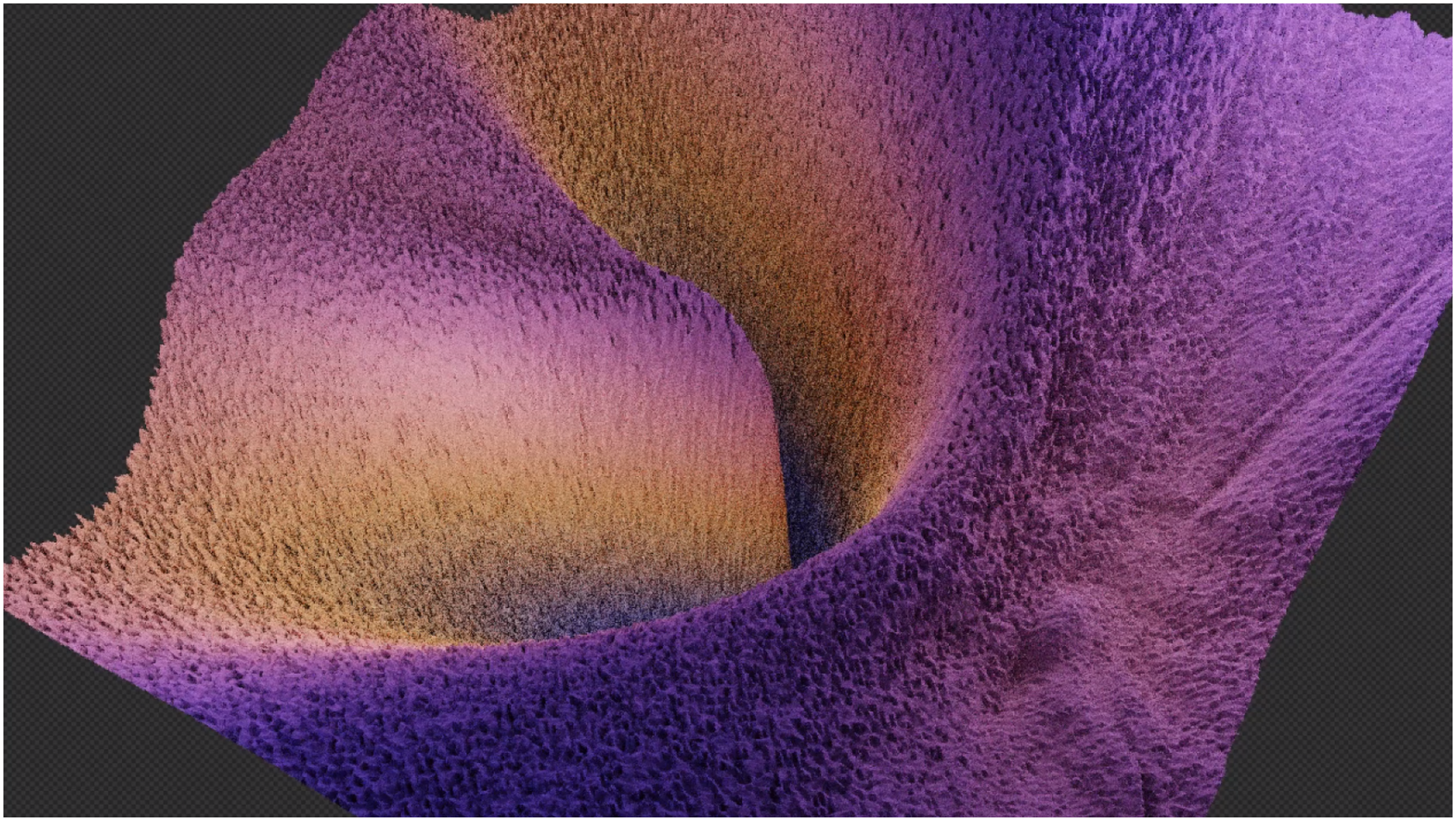


Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

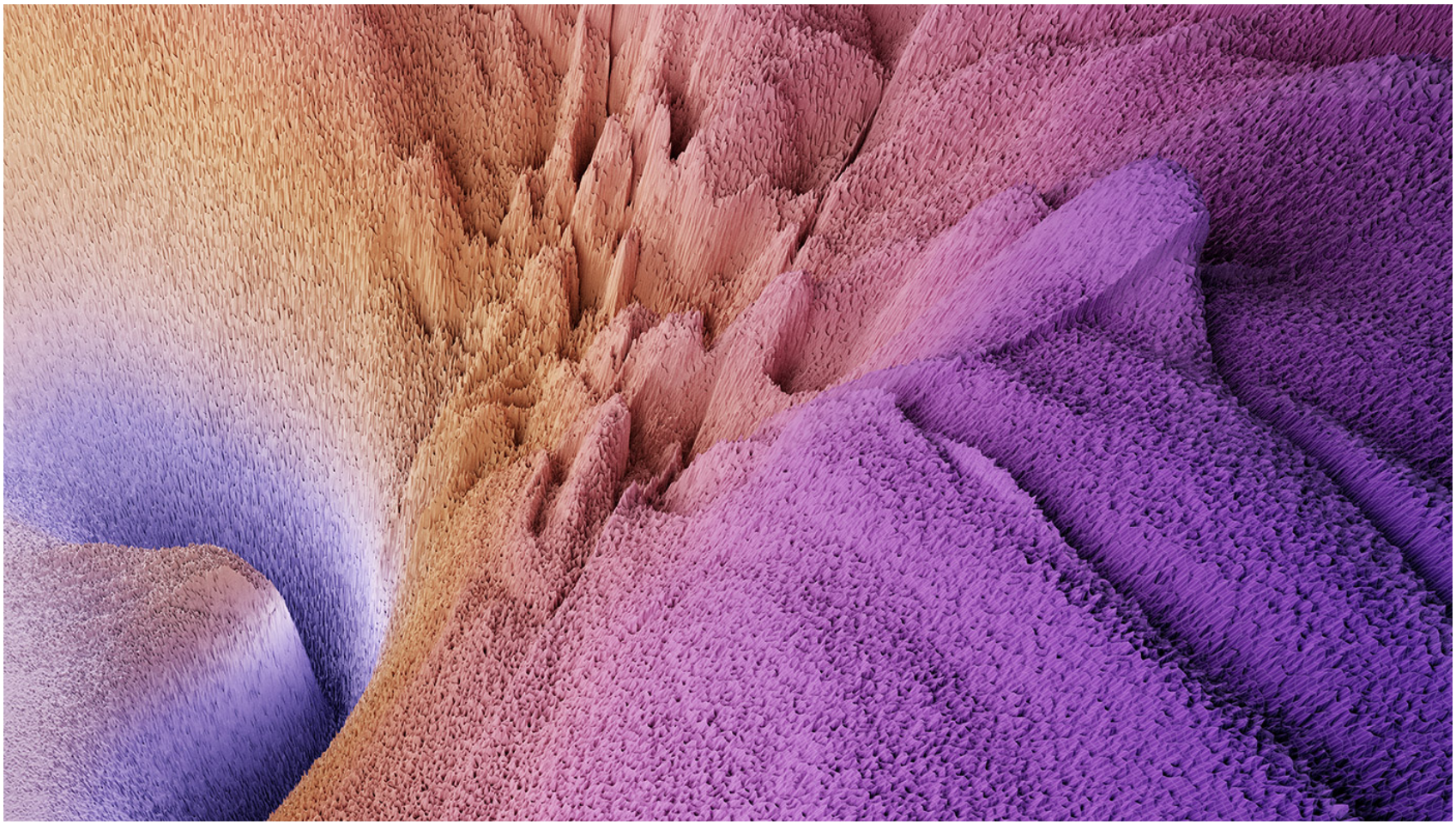
T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, A.G. Wilson

NeurIPS 2018

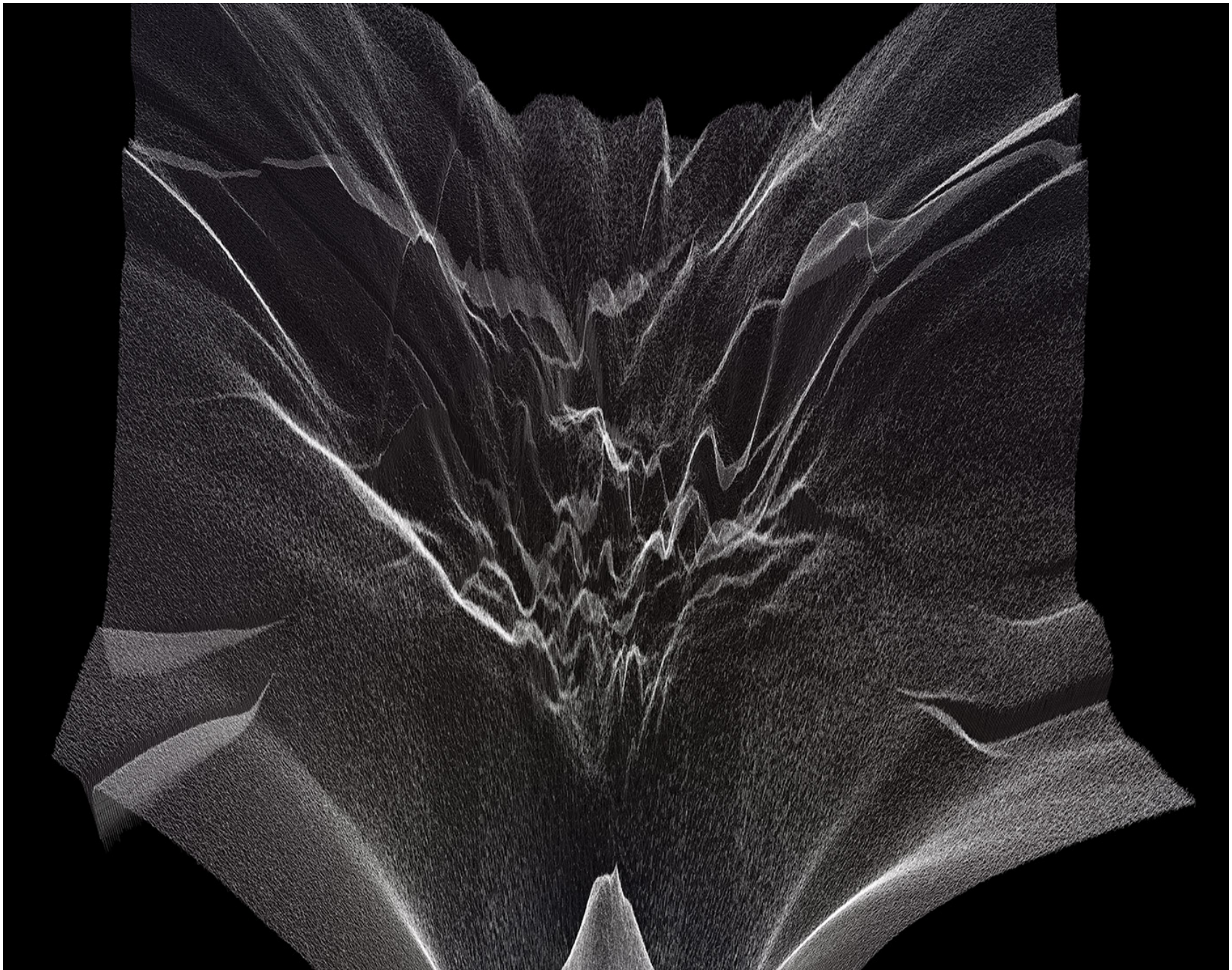
Loss Valleys



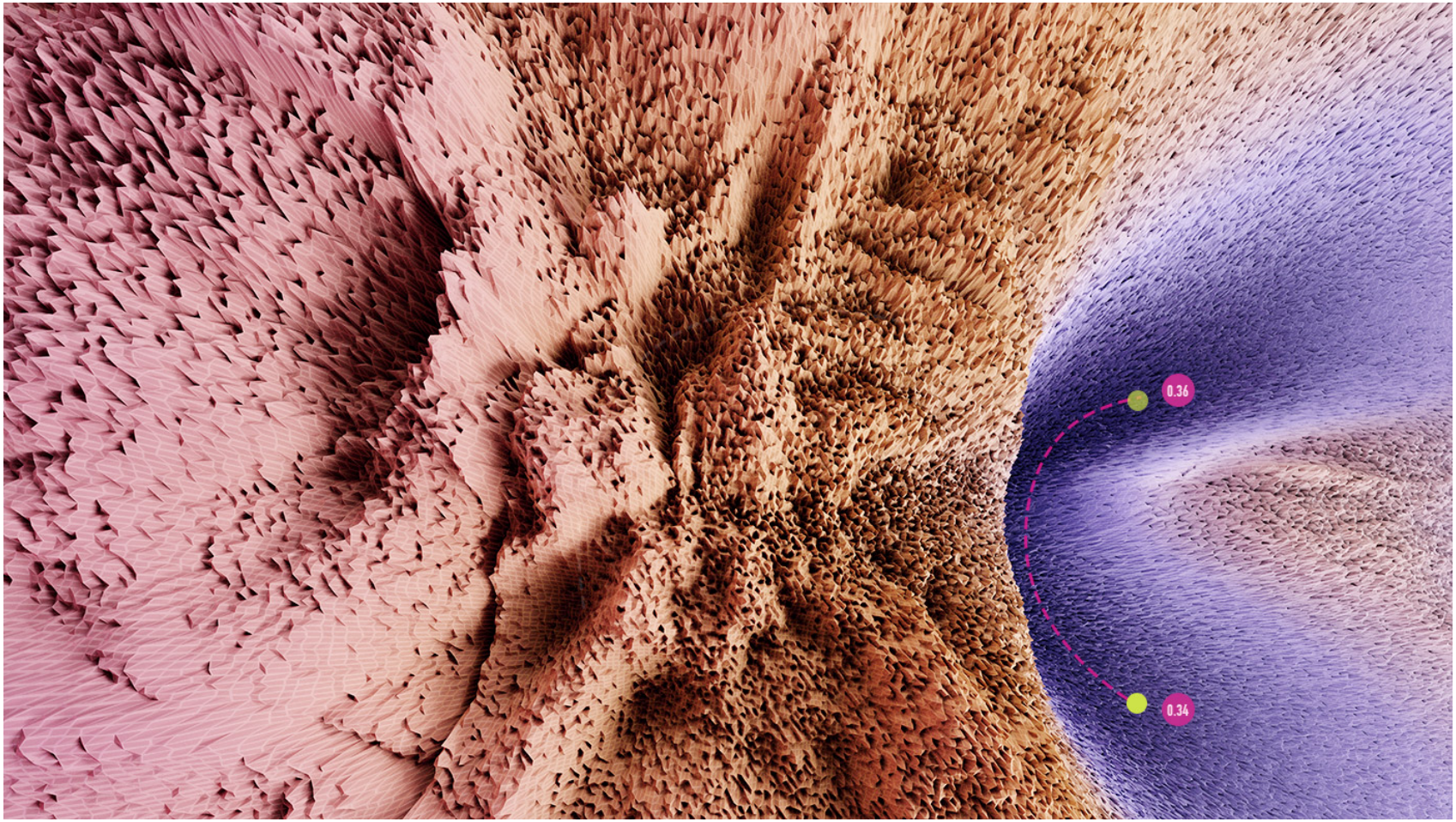
Loss Valleys



Loss Valleys

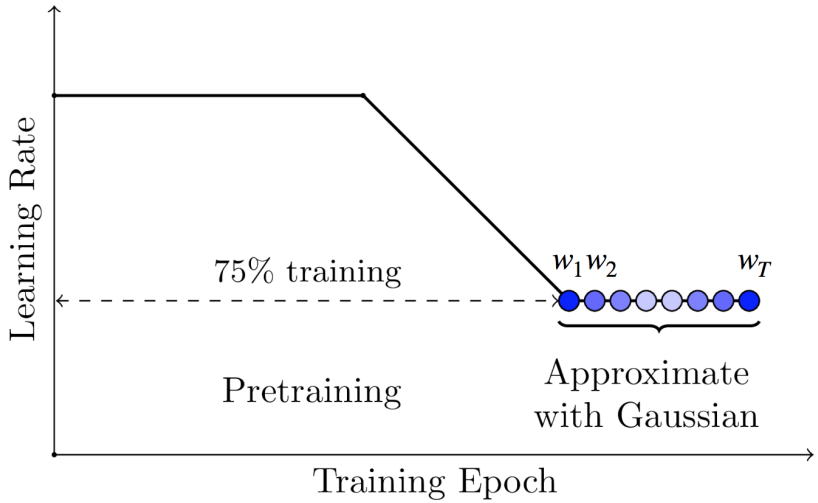


Loss Valleys



Uncertainty Representation with SWAG

1. Leverage theory that shows SGD with a constant learning rate is approximately sampling from a Gaussian distribution.
2. Compute first *two* moments of SGD trajectory (SWA computes just the first).
3. Use these moments to construct a Gaussian approximation in weight space.
4. Sample from this Gaussian distribution, pass samples through predictive distribution, and form a Bayesian model average.



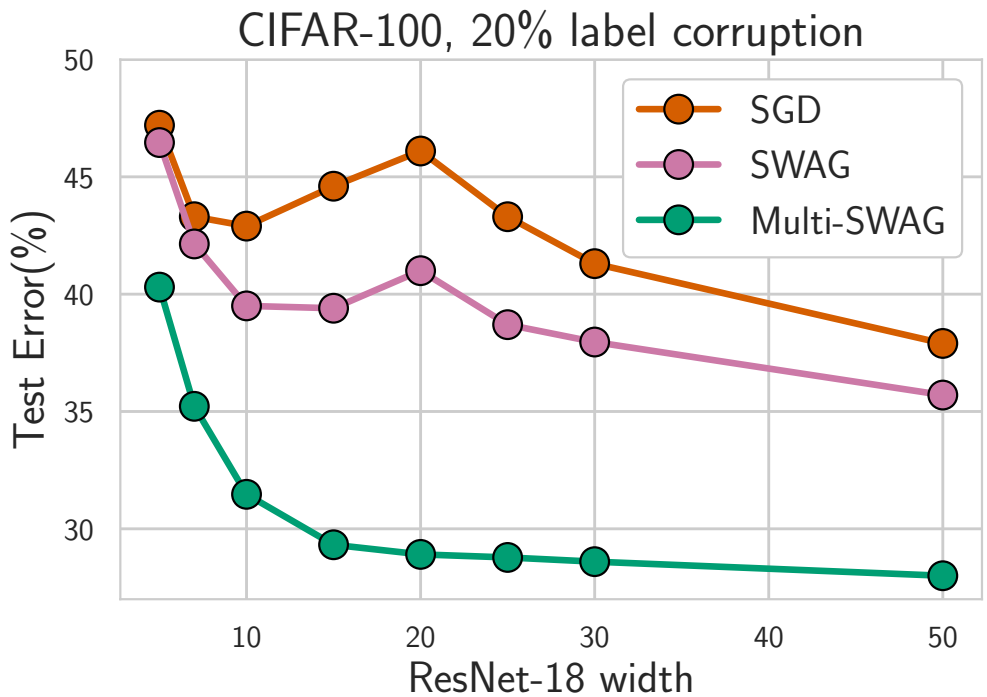
$$p(y_*|\mathcal{D}) \approx \frac{1}{J}\sum_{j=1}^{J} p(y_*|w_j)\,, \quad w_j \sim q(w|\mathcal{D})\,, \quad q(w|\mathcal{D}) = \mathcal{N}(\bar{w}, K)$$

$$\bar{w} = \frac{1}{T}\sum_t w_t\,, \quad K = \frac{1}{2}\left(\frac{1}{T-1}\sum_t (w_t - \bar{w})(w_t - \bar{w})^{\mathrm{T}} + \frac{1}{T-1}\sum_t diag(w_i - \bar{w})^2\right)$$

SWA now natively supported in PyTorch 1.6!

SWAG: *A Simple Baseline for Bayesian Uncertainty in Deep Learning*. Maddox et. al, NeurIPS 2019.
SWA: *Averaging Weights Leads to Wider Optima and Better Generalization*. Izmailov et. al, UAI 2018.

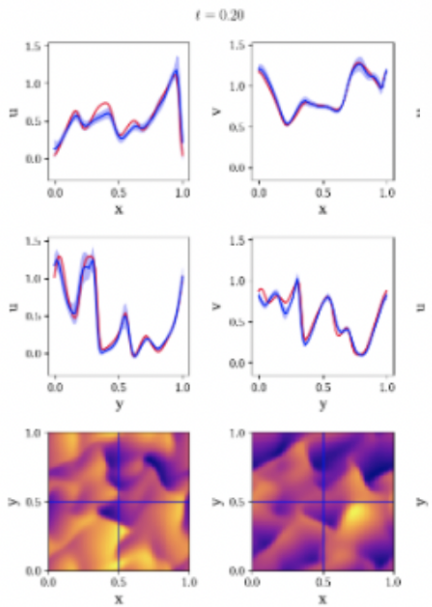
MultiSWAG



Bayesian Deep Learning and a Probabilistic Perspective of Generalization. Wilson & Izmailov, NeurIPS 2020.

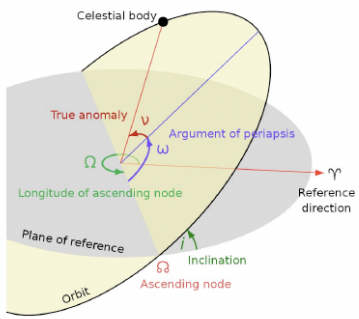
Applications of Bayesian Deep Learning

There are *many* compelling applications. Some applications of SWAG.

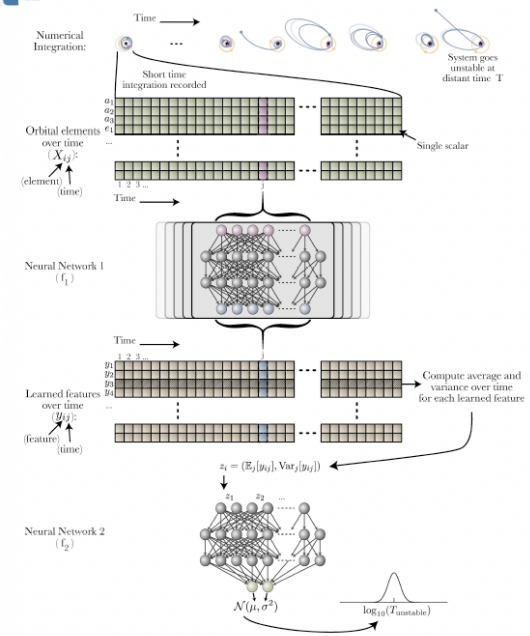


Modelling PDE Dynamics,
Geneva & Zabaras, '19

Federated learning, Chen & Chao, '21
Image:https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/



Planetary dynamics,
Cranmer et al, '21



Slide credit: Wesley Maddox

Are Standard (Gaussian) Priors Bad?

Neural Network Priors

A parameter prior $p(w) = \mathcal{N}(0, \alpha^2)$ with a neural network architecture $f(x, w)$ induces a structured distribution over *functions* $p(f(x))$.

Deep Image Prior

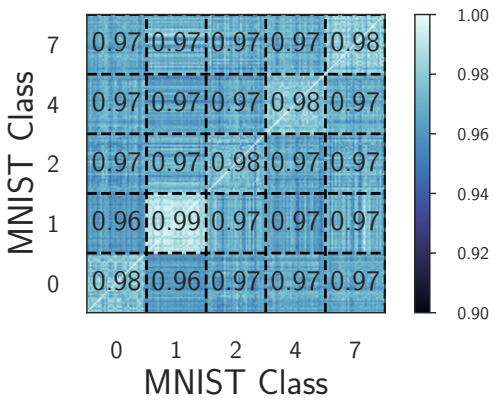
- *Randomly initialized* CNNs *without training* provide excellent performance for image denoising, super-resolution, and inpainting: a sample function from $p(f(x))$ captures low-level image statistics, before any training.

Random Network Features

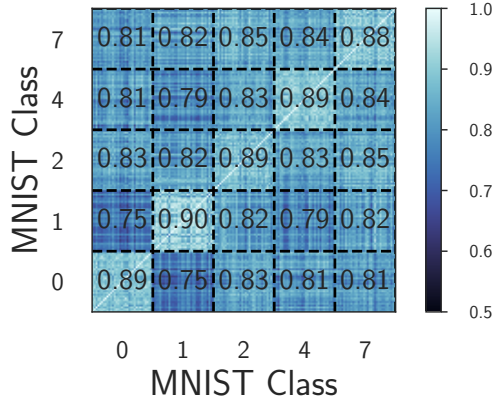
- Pre-processing CIFAR-10 with a *randomly initialized untrained* CNN dramatically improves the test performance of a Gaussian kernel on pixels from 54% accuracy to 71%, with an additional 2% from ℓ_2 regularization.

Deep Image Prior. Ulyanov, D., Vedaldi, A., Lempitsky, V. CVPR 2018.
Understanding Deep Learning Requires Rethinking Generalzation. Zhang et. al, ICLR 2016.
Bayesian Deep Learning and a Probabilistic Perspective of Generalization. Wilson & Izmailov, NeurIPS 2020.

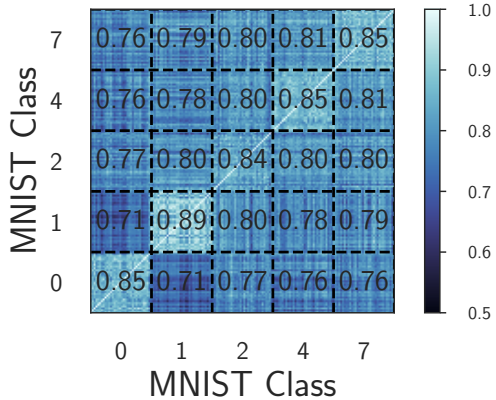
Prior Class Correlations



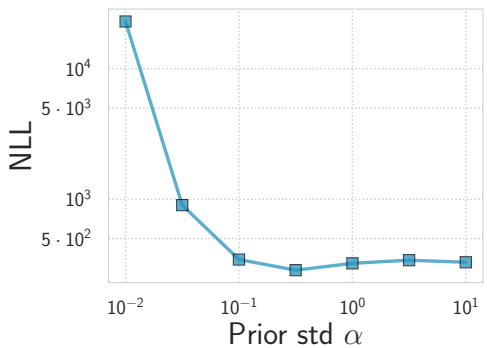
(a) $\alpha = 0.02$



(b) $\alpha = 0.1$



(c) $\alpha = 1$



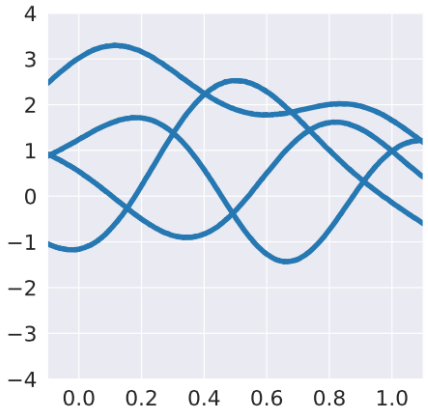
(d)

Many good results with Gaussian priors

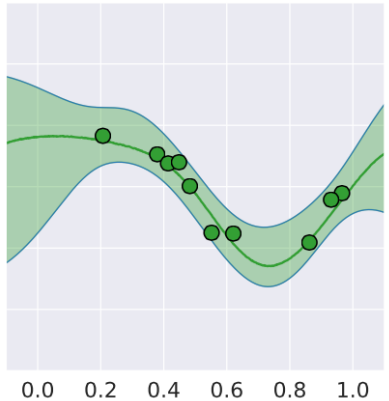
We've already seen many good results:

- HMC experiments
- SWAG experiments
- Double descent result

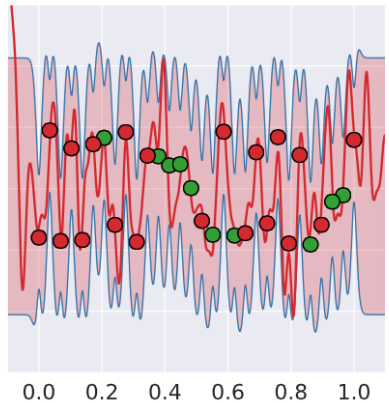
Rethinking Generalization



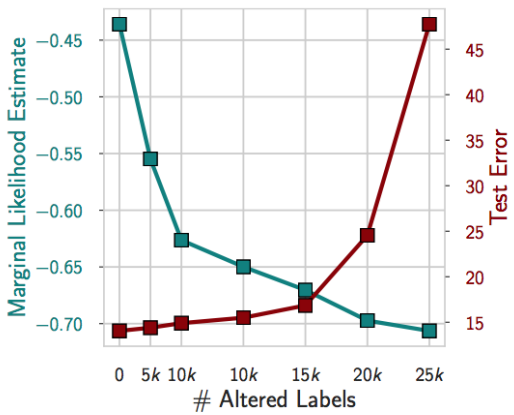
(a) Prior Draws



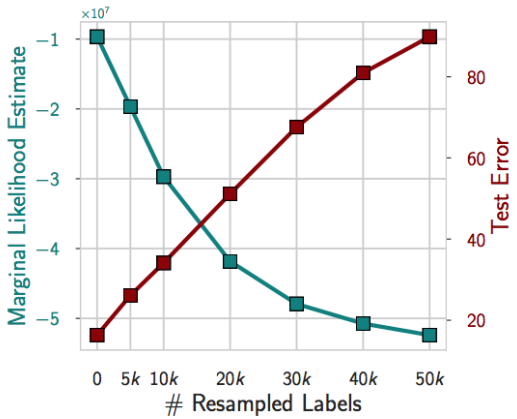
(b) True Labels



(c) Corrupted Labels



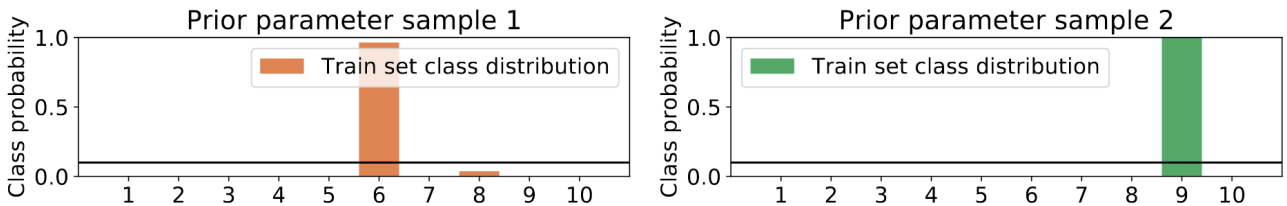
(d) Gaussian Process



(e) PreResNet-20

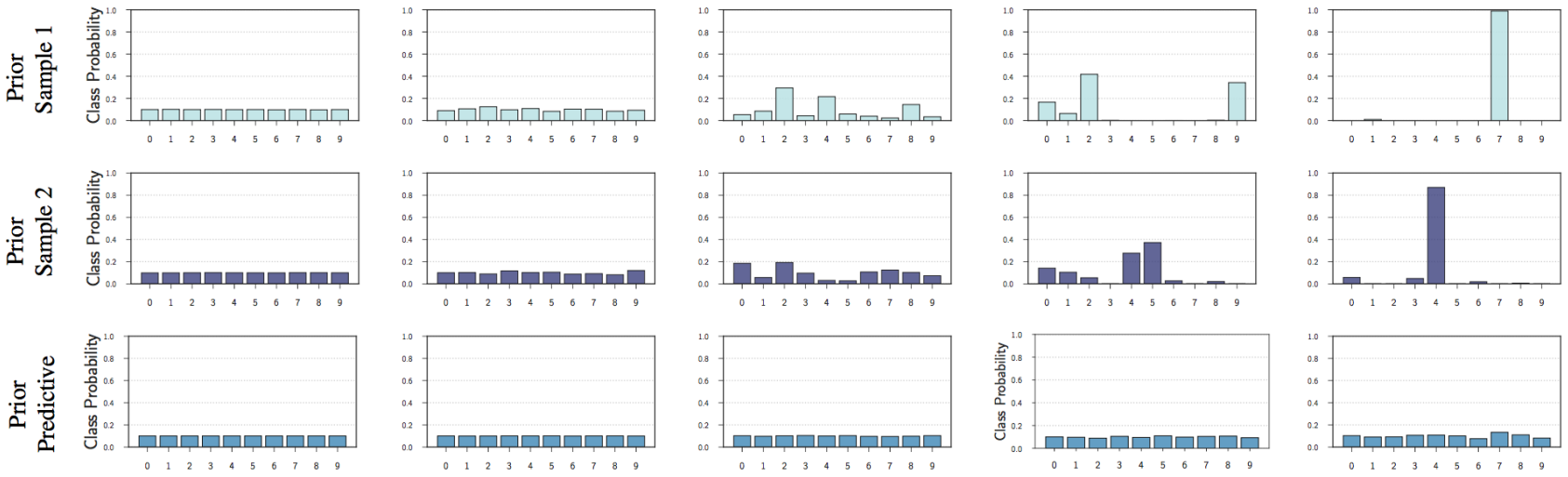
Prior Misspecification?

Wenzel et. al suggest $p(w) = \mathcal{N}(0, I)$ prior misspecification, showing that sample functions $p(f(x))$ seem to assign one label to most classes on CIFAR-10.

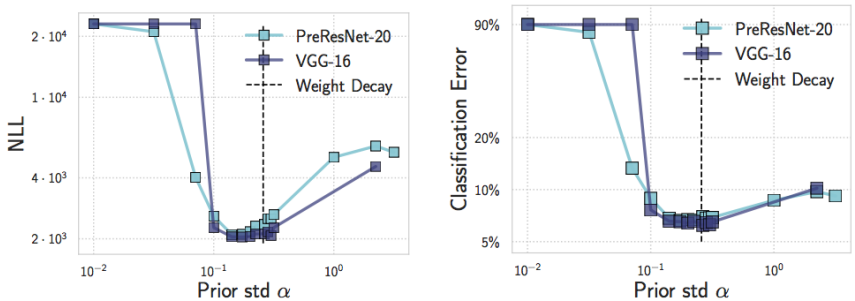


How good is the Bayes posterior in deep neural networks really? Wenzel et. al, ICML 2020.

Changing the prior variance scale α



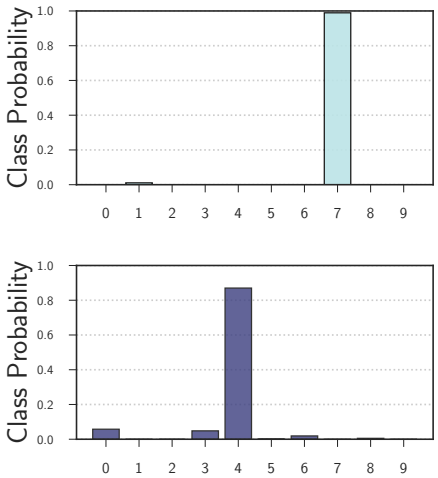
(a) $\alpha = 0.01$ (b) $\alpha = 0.1$ (c) $\alpha = 0.3$ (d) $\alpha = 1.$ (e) $\alpha = \sqrt{10}$



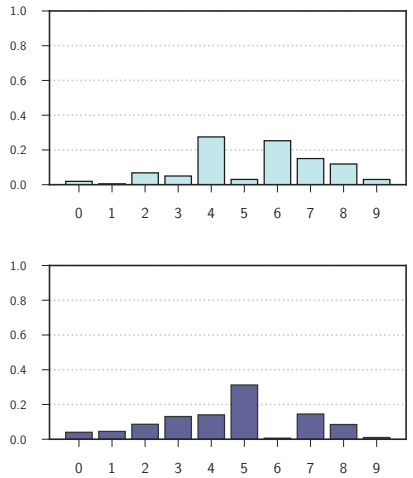
(f) (g)

Bayesian Deep Learning and a Probabilistic Perspective of Generalization. Wilson & Izmailov, 2020.

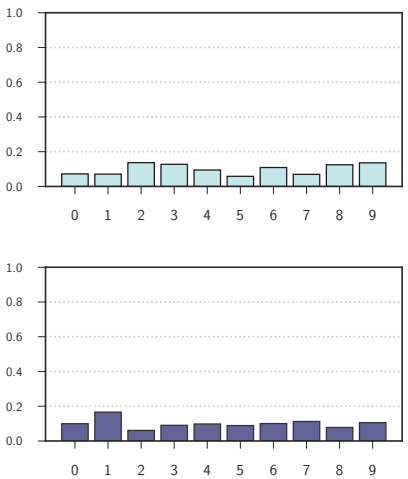
The effect of data on the posterior



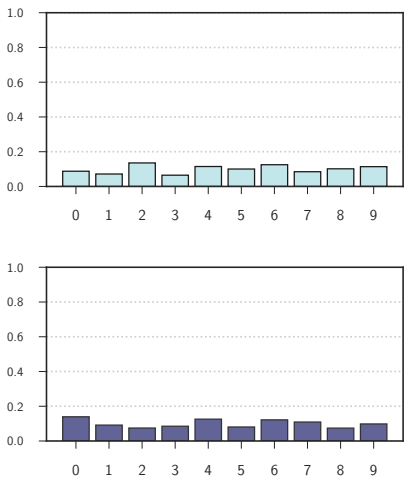
(e) Prior ($\alpha = \sqrt{10}$)



(f) 10 datapoints



(g) 100 datapoints



(h) 1000 datapoints

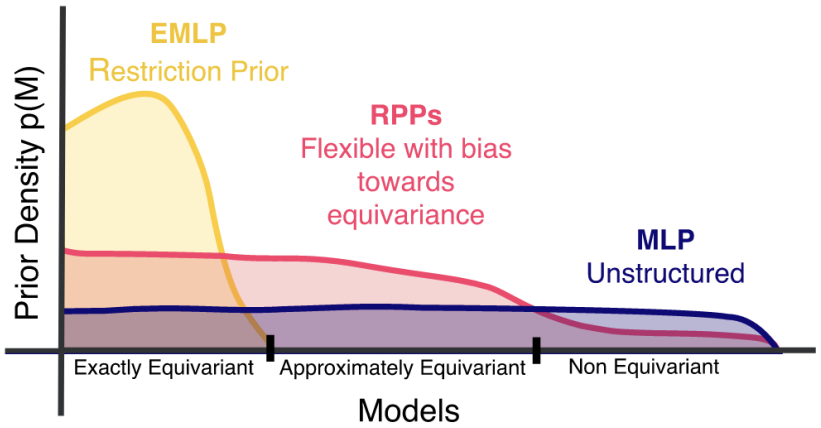
Effects of Using Different Priors with HMC Inference

PRIOR	GAUSSIAN	MOG	LOGISTIC
ACCURACY	0.866	0.863	**0.869**
ECE	0.029	0.025	**0.024**
LOG LIKELIHOOD	-0.311	-0.317	**-0.304**

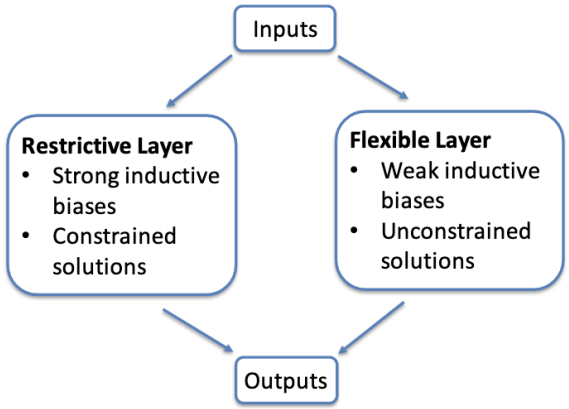
- CNN-LSTM on IMDB.
- Heavy tailed logistic prior slightly better, but not much difference.
- **Architecture is much more important for determining the prior over *functions*, than the precise details of the distribution over weights.**

Residual Pathway Priors

- Build priors in *function space*. We can encode approximate symmetries with *residual pathway priors*.



(a) Priors over Equivariant Solutions



(b) Structure of RPP Models

$$p(w) = \mathcal{N}(0, (\sigma_a^2 + \sigma_b^2)QQ^T + \sigma_b PP^T)$$

- Q is an orthogonal matrix representing the equivariance subspace
- P is the orthogonal complement of Q

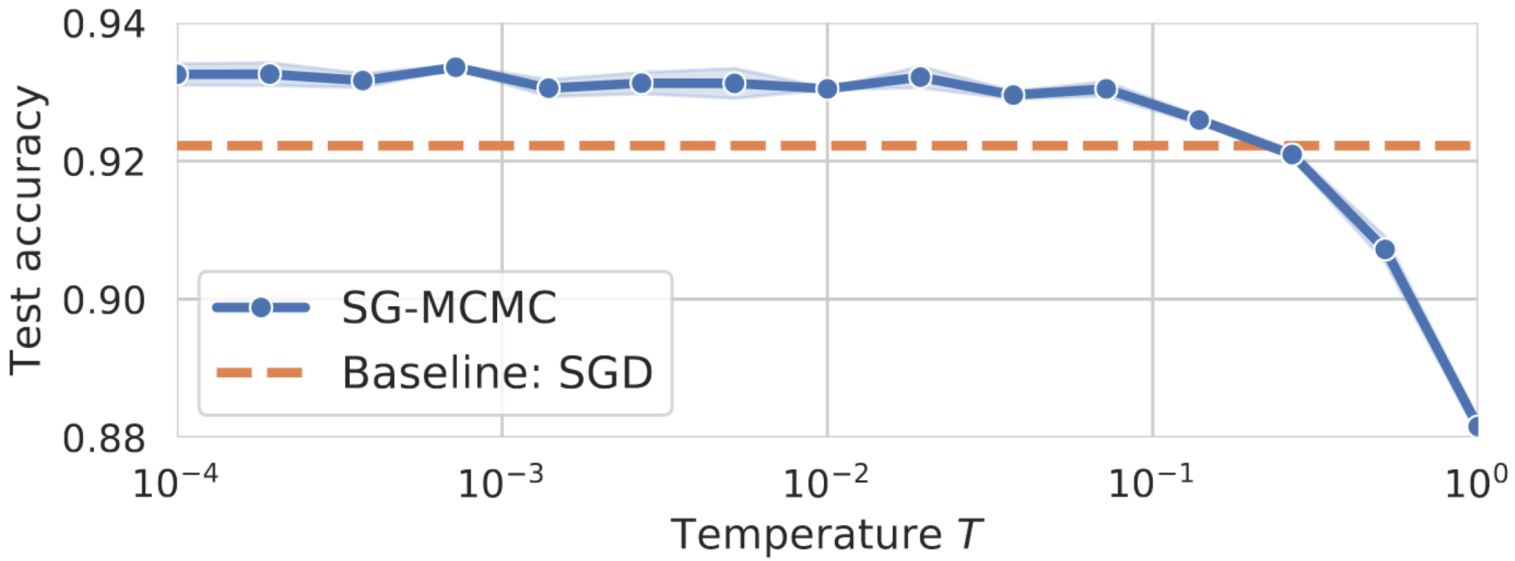
Residual Pathway Priors for Soft Equivariance Constraints. Finzi et. al, NeurIPS 2021.

Cold Posteriors

Are Cold Posteriors Problematic?

What are Cold Posteriors?

Wenzel et. al (2020) highlight the result that for $p(w) = N(0, I)$ *cold posteriors*, raised to a power $1/T$ with $T < 1$ often provide improved performance.

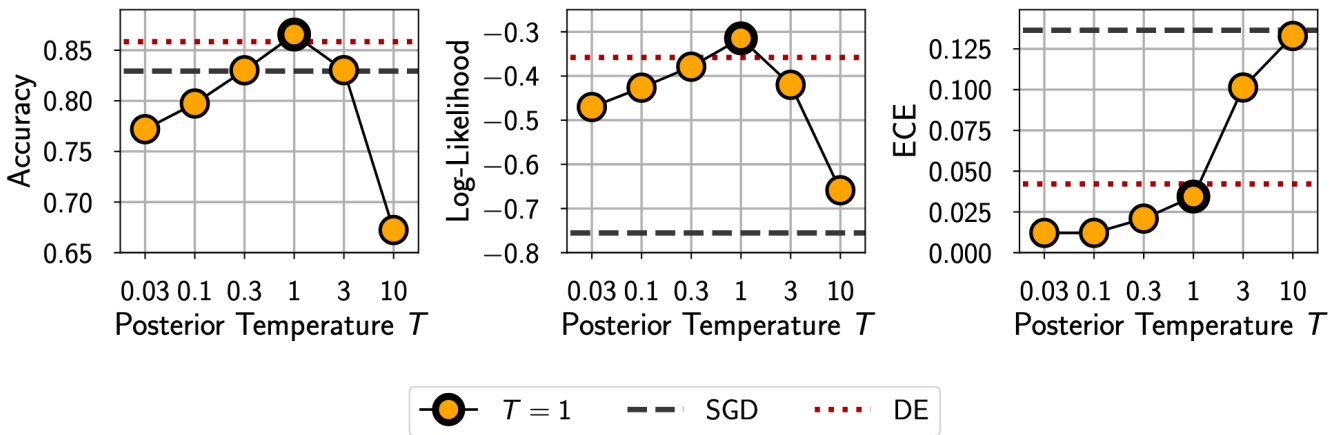


How good is the Bayes posterior in deep neural networks really? Wenzel et. al, ICML 2020.

Should we really be surprised?

- It would be surprising if $T = 1$ was the best setting of this hyperparameter.
- Our models are certainly misspecified, and we should acknowledge that misspecification in our estimation procedure by learning T. Learning T is not too different from learning other properties of the likelihood, such as noise.
- A tempered posterior is a more honest reflection of our prior beliefs than the untempered posterior. Bayesian inference is about honestly reflecting our beliefs in the modelling process.

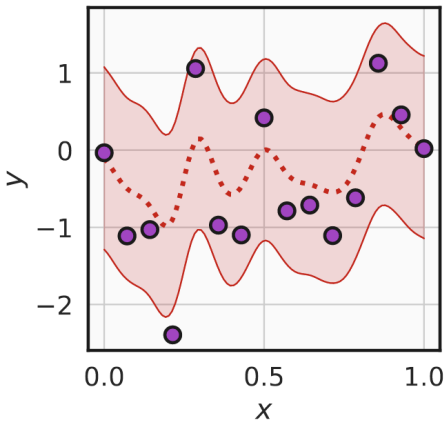
Role of Data Augmentation



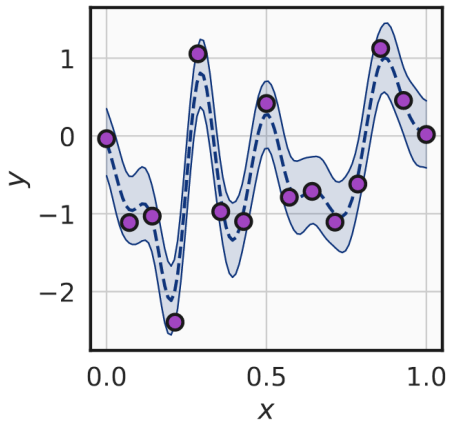
- All the cold posteriors in *Wenzel et. al* disappear when we remove data augmentation.

What are Bayesian Neural Network Posteriors Really Like? Izmailov et. al, ICML 2021.

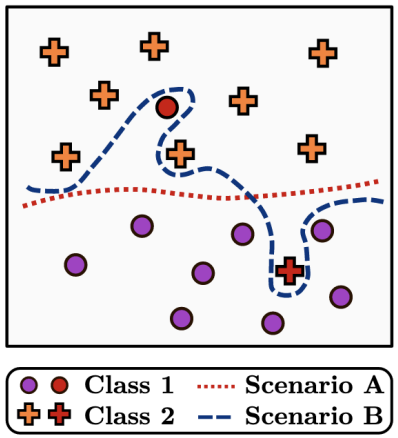
Resolution



(a) GP regression, $\sigma^2 = 1$

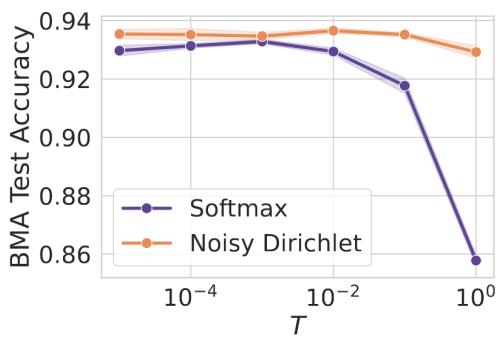


(b) GP regression, $\sigma^2 = 10^{-2}$



(c) Classification

- We profoundly underestimate **aleatoric uncertainty** with standard likelihoods.
- With SGLD, data augmentation has the effect of raising the likelihood to the power of $1/K$ for K augmentations.
- Cold posteriors better reflect our *honest beliefs* about aleatoric uncertainty. But it isn't the only way!



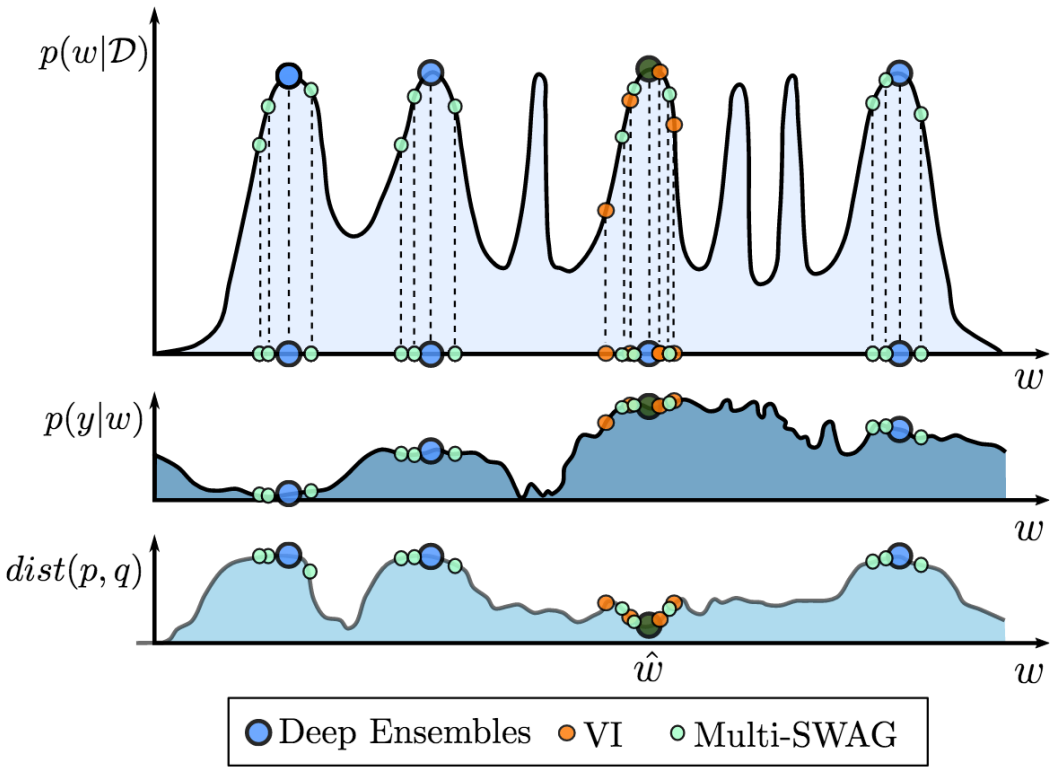
On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification. Kapoor et. al, arXiv 2022.

What is the competition?

Are "deep ensembles" a non-Bayesian competitor?

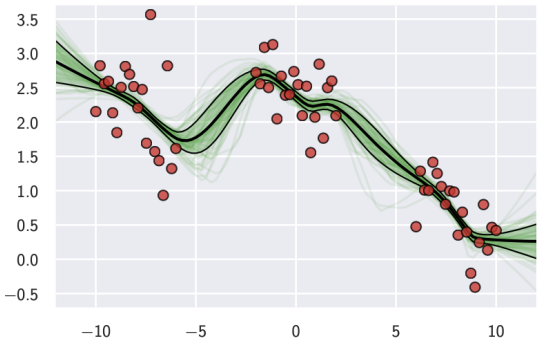
Better Marginalization

$$p(y|x_*, \mathcal{D}) = \int p(y|x_*, w)p(w|\mathcal{D})dw\,. \tag{2}$$

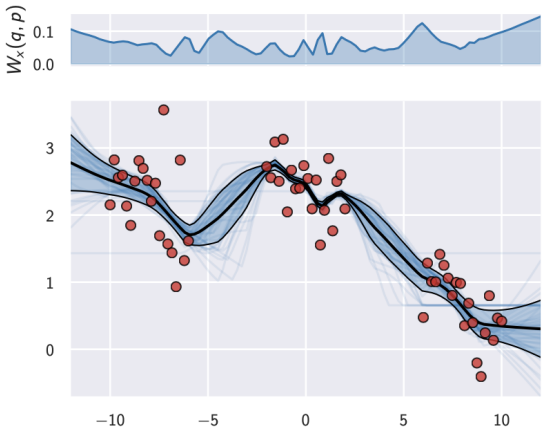


- Be wary about viewing BMA through the prism of simple MC.
- We want to best estimate the BMA integral given computational constraints.
- View BMA estimation as active learning, rather than posterior sampling.

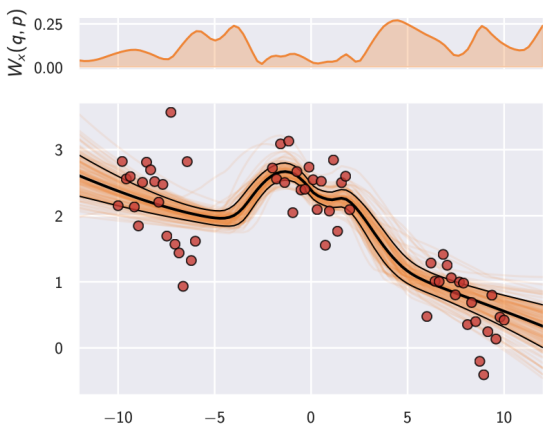
Better Marginalization: Deep Ensembles



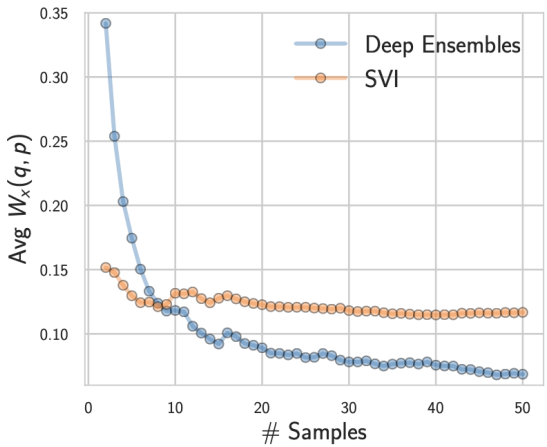
(a) HMC



(b) Deep Ensembles

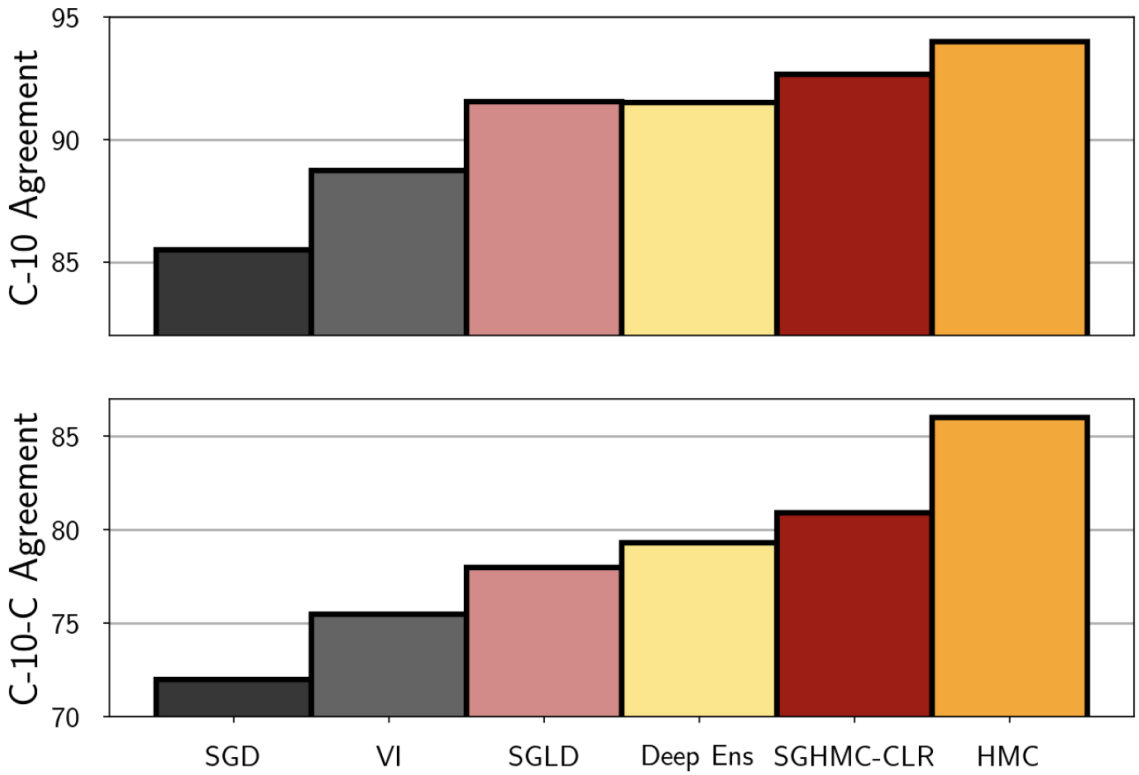


(c) Variational Inference



(d) Distance to BMA

Fidelity of Approximate Marginalization Procedures

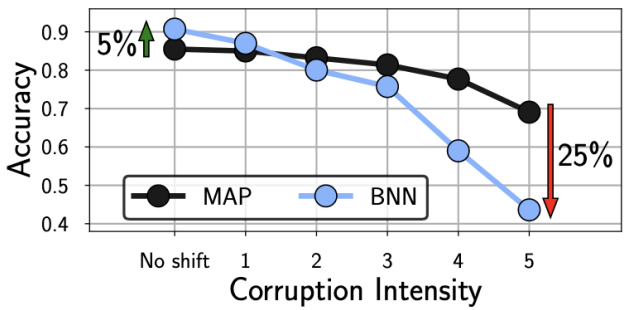


Additionally, in the approximate inference for neural networks competition at NeurIPS 2021, the winning entries were all based on deep ensembles

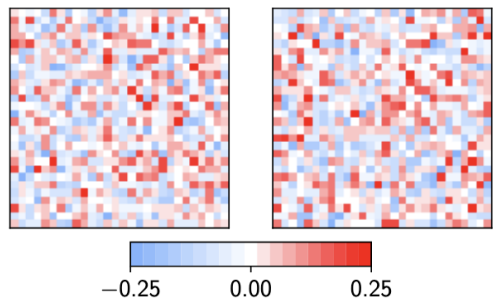
But are there challenges in BDL?

Are there practical challenges to adopting BDL?

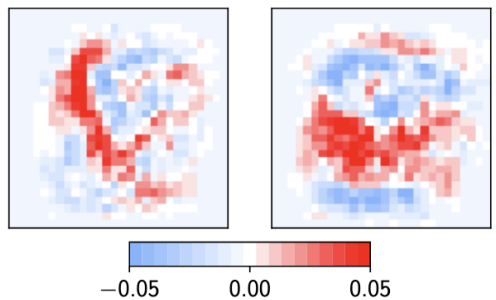
Bayesian Model Averaging Under Covariate Shift



(a) ResNet-20, CIFAR-10-C

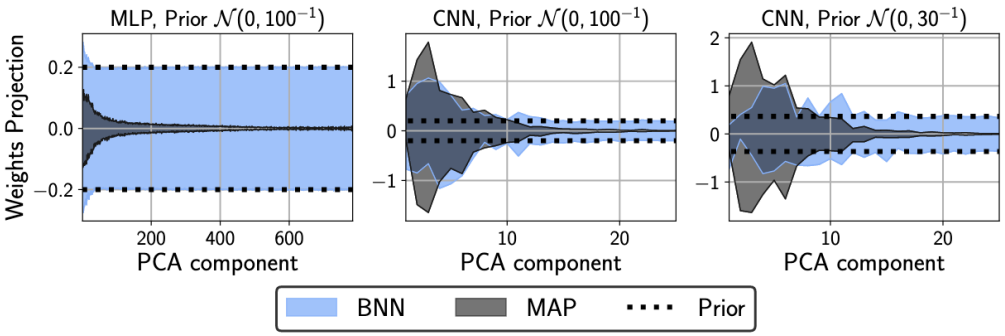


(b) BNN weights

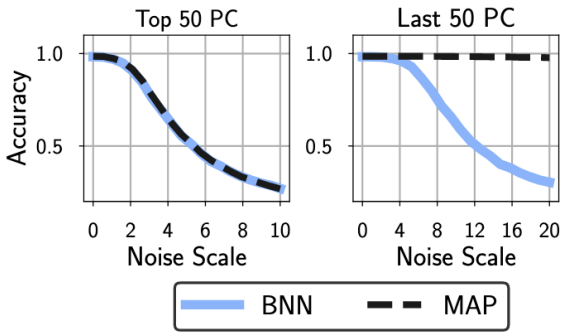


(c) MAP weights

Principal Components



(a) Weight projections

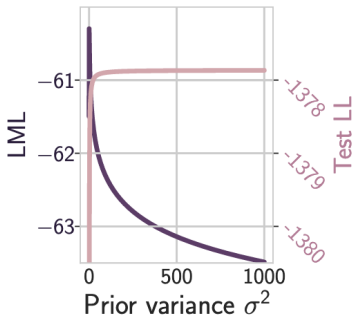


(b) Robustness along PCs

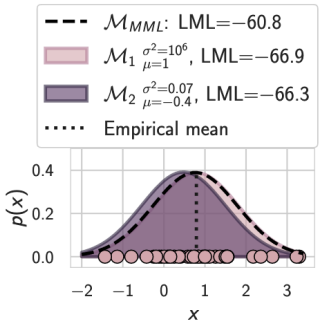
- Idea: create a prior for first-layer weights that is aligned with the principal components.
- $p(w) = \mathcal{N}(0, \alpha K + \epsilon I)$, $K = \frac{1}{n-1} \sum_{i=1}^{n} x_i x_i^T$

Dangers of Bayesian Model Averaging under Covariate Shift. Izmailov et. al, NeurIPS 2021.

Bayesian Model Selection



(a) Fixed $\mu = 0$



(b) LML prefers $\mathcal{M}_2$

- The marginal likelihood (evidence) answers the question "what is the probability that my prior model generated the training data?"
- This question is fundamentally different than "will my trained model provide good generalization?"
- E.g., $p(x) = N(\mu, 1), p(\mu) = \mathcal{N}(0, \sigma^2)$. The marginal likelihood quickly decreases with increases in σ past a certain point, but the posterior predictive is unaffected, because the prior becomes weak as we increase σ. **As a consequence, the marginal likelihood can have a strong preference between models with virtually identical predictive distributions**.
- In Bayesian deep learning, we also have to approximate the marginal likelihood, which has its own issues.

Many examples, and some partial remedies, in ***Bayesian Model Selection, the Marginal Likelihood, and Generalization, Lotfi et. al, ICML 2022***.

Discussion

- There is a lot of false conventional wisdom around Bayesian deep learning.
- BDL is making great practical advances, and many of the perceived limitations (tractability, priors, cold posteriors, ...) are not in fact limitations.
- However, there are real challenges to practical adoption, such as robustness to distribution shift. The challenges are just not what we normally discuss.