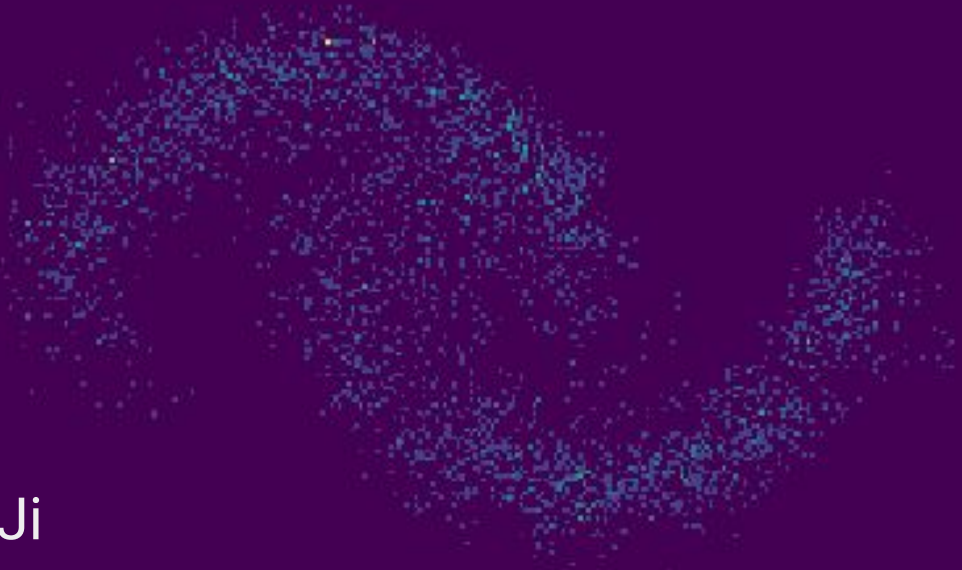
Samples under density uncertainty; E. Jiang 2022



# Bayesian Uncertainty for Likelihood-Defining Neural Networks using Importance Sampling

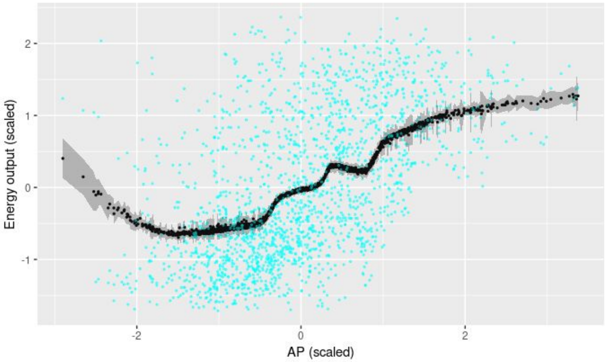
**Presenters:** Ryan Wang, Yichen Ji

**Acknowledgements:** Haining Tan, Eric Jiang, Dr. Scott Schwartz

# UQ4DL: Uncertainty Quantification for Deep Learning

- DL increasingly considers Epistemic (**Model**) + Aleatoric (**Data Randomness**) Uncertainty

Epistemic Uncertainty

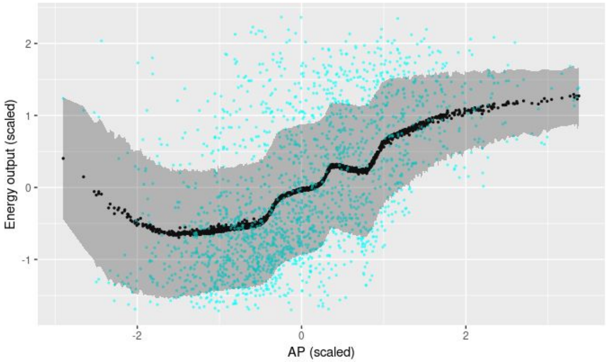



$$p(\theta|D) \propto p(\theta)p(D|\theta)$$

Posterior MCMC sampling / VI approximation =

Computational Challenges in DNNs

Aleatoric Uncertainty

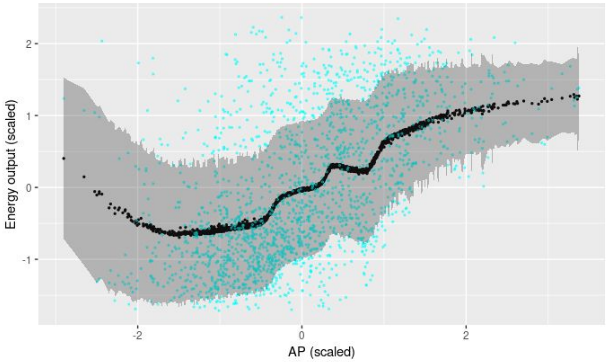



$$p(y|\theta)$$

Data Generating Mechanism =

Likelihood

Predictive Uncertainty




$$p(y|x,D) = \int_\theta p(y|x,\theta')p(\theta'|D)d\theta'$$

Epistemic + Aleatoric Uncertainty =

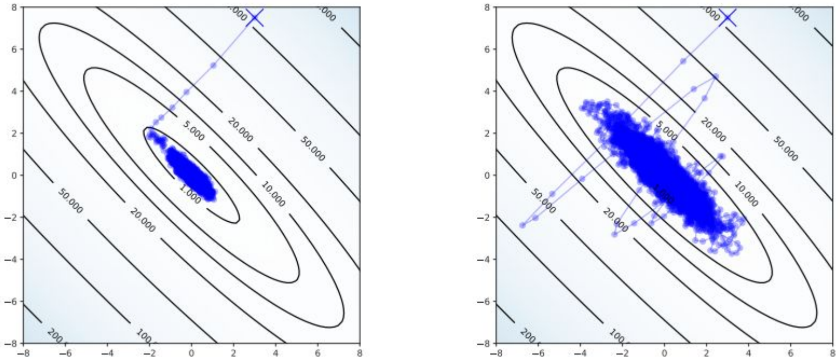
Predictive Uncertainty

# SWA-Gaussian (SWAG): a Baseline Model

- Key geometric observation

  The posterior distribution over NN parameters is close to Gaussian on SGD trajectory subspaces *(see [Izmailov, et al. 2020] for construction details)*

SGD trajectory on quardratic problem without (left) + with (right) momentum



- SGD trajectories indicate orientation in NN weight posterior approximation
  - SWAG exploits this information to compute low-rank plus diagonal covariance matrix

$$p(w|\mathcal{D}) \approx \mathcal{N}\left(\theta_{\mathrm{SWA}}\,, \frac{1}{2} \cdot (\Sigma_{\mathrm{diag}} + \Sigma_{\text{low-rank}})\right)$$

SWAG posterior curvature approximation

# Likelihood As Importance Weights in Bayes-IS

1 **Posterior Importance Sampling**

$$\mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}[g(\theta)] = \int g(\theta)p(\theta|\mathcal{D})d\theta$$
$$= \int g(\theta)p(\theta)\frac{p(\theta|\mathcal{D})}{p(\theta)}d\theta \implies IW(\theta) = \frac{p(\theta|\mathcal{D})}{p(\theta)} \propto f(\mathcal{D}|\theta)$$

2 **with prior proposals**

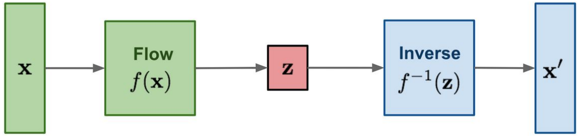
3 **Bayesian Sequential Updating**

$$p(\theta|x_1, x_2) \propto p(\theta)f(x_1|\theta)f(x_2|\theta) \propto q(\theta)f(x_2|\theta)$$

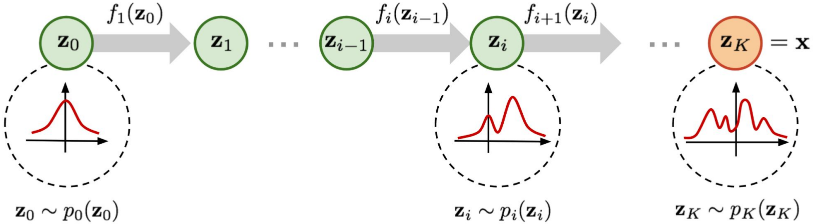
4 *an approximate

**SWAG as intermediate prior** $q(\theta)$

5 $IW(\theta)$ **for IS is available for** <u>any likelihood-defining NN</u> like a normalizing flow (NF)



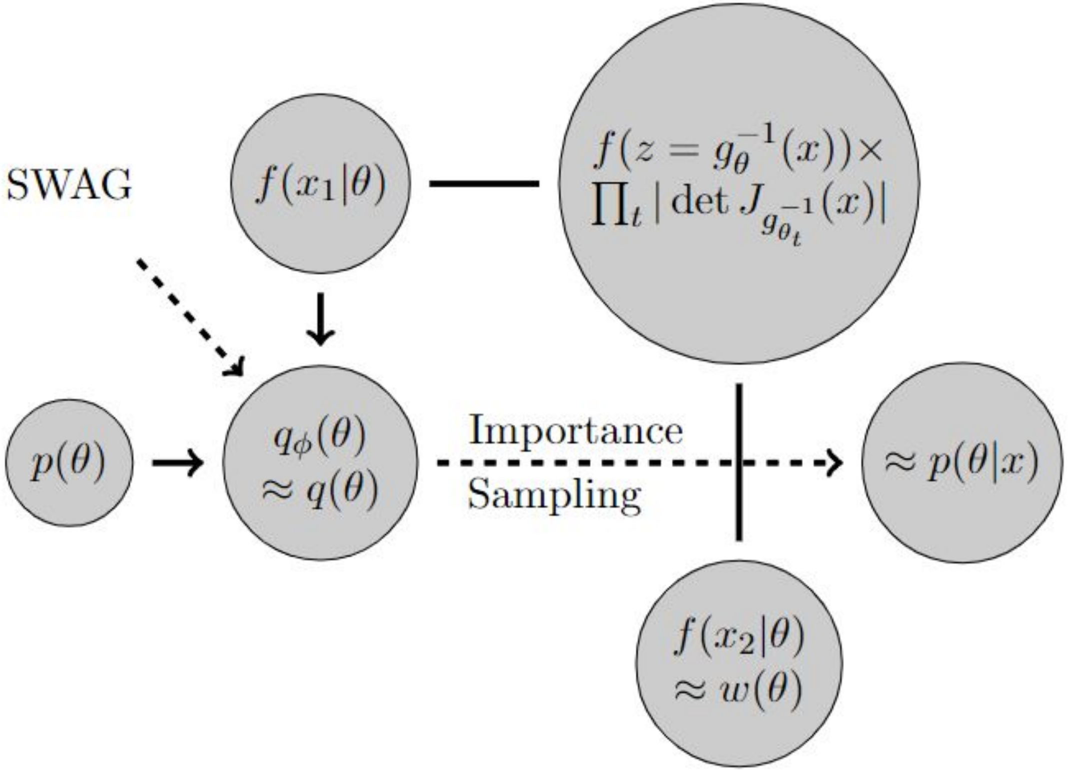

https://lilianweng.github.io/posts/2018-10-13-flow-models/ (L. Weng, 2018)



# Related Work

- Importance sampling during training to improve performance
  - Importance Weighted Autoencoders (Y. Burda  et al., 2016); IS improved training objective
  - Deep Learning with IS (A. Kartharaopoulos, 2019); variance reduction and convergence speed
- Importance sampling posteriors for likelihood-free simulations
  - Sequential Neural Likelihood (G. Papamakarios et al., 2019); NF substitution for non-tractable simulations to IS sample posterior of simulation (but not NF) parameter MCMC samples
- Non importance sampling posterior inference on generative models
  - Likelihood-Free GAN Inference via ABC scoring rules (L. Pacchiardi et al. 2022); inference for GANs without NF density estimation
- Importance sampling based on tractable NF sampling and likelihood evaluation
  - Neural Importance Sampling (T. Muller et al., 2019); NICE (an NF architecture) as proposals
- Importance sampling to increase normalizing flow flexibility
  - Stochastic Normalizing Flows (H. Wu et al. 2020); IS to reweight stochastic perturbations

# Our Methodology/Work



# What's Special About Our Work?

- SWAG Gaussian approximations may be too rigid
- SWAG followed by IS as **Bayesian sequential learning** replaces approximation with re-weighted samples that flexibly fine-tune **final** posterior inference

**Algorithm 1** Bayes-IS With SWAG

```
1: θ_SWA, D̂, Σ_diag ← Train-SWAG(nnet, D_SWA)
2: p(θ) ← SWAG(θ) = N(θ_SWA, ½Σ_diag + D̂D̂ᵀ/(2(K−1)))  (W. Maddox, 2019)
3: IW ← []
4: G_θ ← []
5: for i ← 1, 2, ..., S do
6:     θ̃_i ~ p(θ)
7:     f(D_LIK|θ̃_i) ← nnet_θ̃_i(D_LIK)
8:     IW[i] ← f(D_LIK|θ̃_i)
9:     G_θ[i] ← g(θ̃_i)
10: end for
11: return IW, G_θ
```

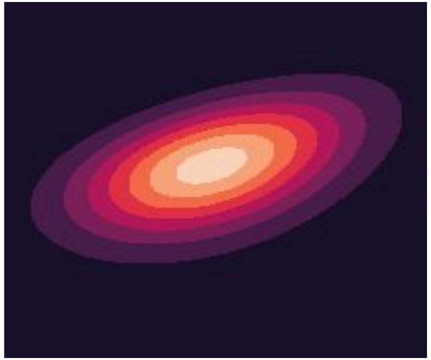
$p(\theta) \leftarrow \mathrm{SWAG}(\theta) = \mathcal{N}\left(\theta_{\mathrm{SWA}}, \frac{1}{2}\Sigma_{\mathrm{diag}} + \frac{\hat{D}\hat{D}^{\top}}{2(K-1)}\right)$

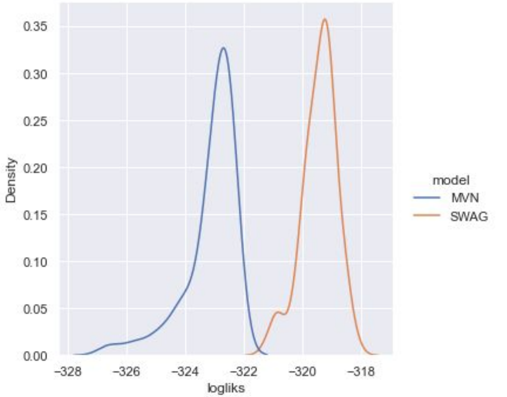
**Algorithm 2** Bayes-IS Averaging

```
1: NIW ← Normalize(IW)
2: E_p(θ|D)[g(θ)] = 0
3: for i ← 1, 2, ..., S do
4:     E_p(θ|D)[g(θ)] ← E_p(θ|D)[g(θ)] + E_p(θ|D)[g(θ)] × NIW[i]
5: end for
6: return E_p(θ|D)[g(θ)]
```

- Bayes-IS sampled likelihood-defining NNs reweighted by their likelihood better reflect the **true posterior** and hence **uncertainty characterization**
    - which improves Monte Carlo integration for predictive tasks via Bayes-IS averaging

# First, How Effective is the SWAG Posterior?



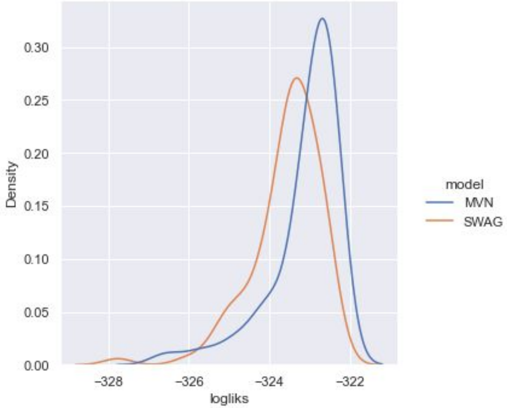



**Experiment Setup:**

1. Sample from MVN data
2. SWAG fit NF parameters
3. The data likelihood is a proxy for how well SWAG recovers true MVN

**Results:**

- Parameters sampled from SWAG define NFs that fit the true distribution exceptionally well; however, this is very sensitive to hyperparameter choices **like NN size** → →




Credits: Haining Tan

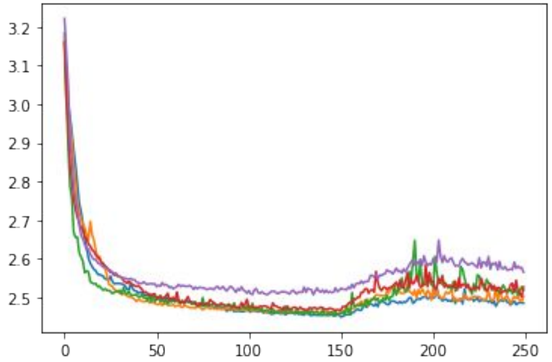
# What is SWAG Sensitive to?

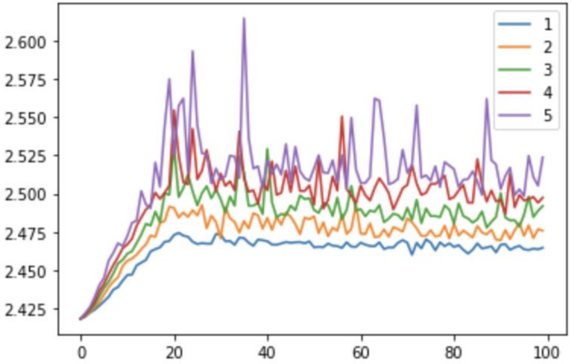
**Experiment Setup:**

- Train SWAG NF on various optimizers, learning rates, and learning rate schedulers

**Results:**

- SWAG is sensitive to these choices; but, learning rate strongly controls loss surface exploration
- SWAG covariance is only consistent when used to explore the same trained model initialization (due to non identifiability from model symmetry)




Credits: Eric Jiang

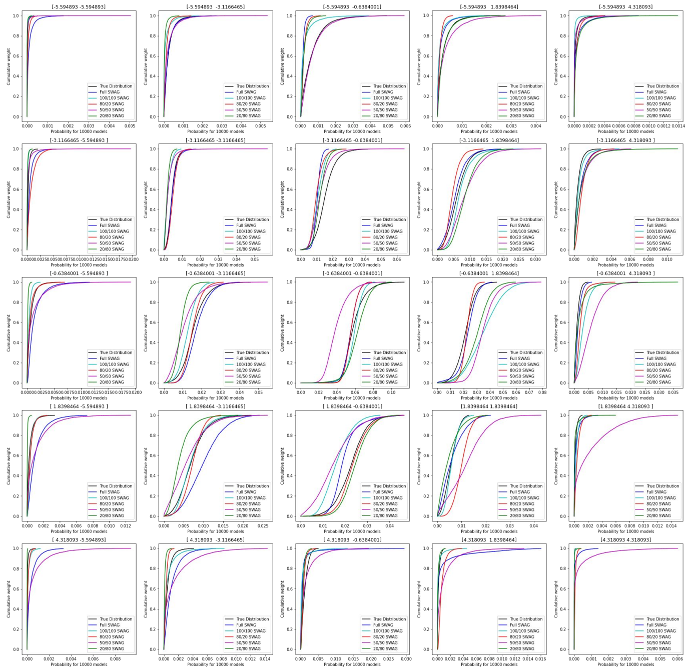
**SWA-Gaussian is a pretty good approximation of the true posterior!**

***Can we do better?***

# Can IS Fix Rigid MVN Assumptions?

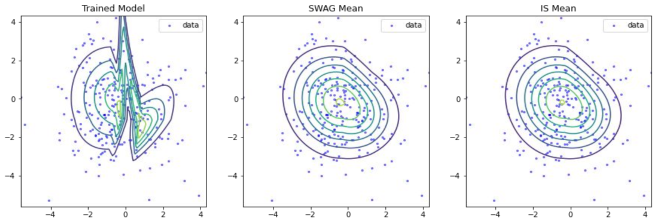
**Experiment Setup:** Finding a "gold standard"(?)

1. Sample exponential (non-MVN) NF parameters
2. Sample data from NF defined by "true" weights
3. Fit (overly rigid?) SWAG NF to sampled data and explore various sequential learning splits
4. Evaluate differences in likelihood distributions




**Results:**

- Bayes-IS most closely matches "gold standard" uncertainty characterization (but just slightly*)
    - 80SWAG/20IS consistently matched "gold standard"



# What Happens When We Don't Have an Exact Likelihood?

- Using loss in importance weighting has been explored for training
  - (**Robust, Approximate Importance Sampling** T. Johnson, 2018)
  - We use loss as a proxy for likelihood in our weighting scheme
- Regression may optimize MSE (for negative Gaussian log-likelihood)
  - Bayes-IS with MSE loss did not greatly change uncertainty characterization
  - (but perhaps SWAG is already modelling true posterior well)
- Classification may optimize Cross-Entropy (for negative ber/cat log-lik.)
  - Bayes-IS weighted averaging actually improves on MNIST classification for a basic ConvNet; however, how to characterize and evaluate model uncertainty here remains an open question

# Current Conclusions

- SWAG **approximates the true posterior** over parameters very well
- SWAG is **sensitive** to various hyperparameters, especially learning rate
- Bayes-IS **may improve on the uncertainty characterization** in 2D(+?) density estimation (small, but consistent improvement)
- Bayes-IS **weighted averaging may help improve predictions** when using approximate importance sampling (loss as importance weights)

# Future Work

- Exploring uncertainty in more (statistical) examples:
  - Regression/Conditional density estimation
  - Hierarchical models
  - Outlier detection
  - Image classification (CIFAR, MNIST)

- Characterizing uncertainty in more unique tasks (like classification) and how they compare to standard Bayesian methods:
  - Variational inference methods like Bayes-by-Backprop
  - Markov Chain Monte Carlo methods

- Developing framework which guides tuning parameter selection

# Thank you for listening!