A Simple Baseline for Bayesian Uncertainty in Deep Learning

Wesley Maddox Pavel Izmailov Timur Garipov
Dmitry Vetrov Andrew Gordon Wilson

NEW YORK UNIVERSITY SAMSUNG

Bayesian Deep Learning

- Combines models for better predictions
- Incorporates prior knowledge and domain expertise
- Provides reliable uncertainty representation

$$p(w \mid Data) = \frac{p(Data \mid w)p(w)}{\int p(Data \mid w')p(w')dw'}$$

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

Challenge: Hard to capture the geometry of the intractable posterior in DNNs with approximations

SGD Trajectory

SGD can capture the geometry of the loss function.
Example:

- Quadratic loss
- Isotropic Gaussian noise in the gradients
- Small step size

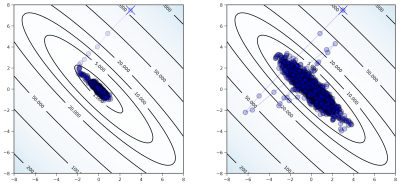


Figure: Quadratic loss contour plot and iterates of SGD (**left**) without momentum and (**right**) with momentum.

- SGD can capture the shape of the posterior
- Momentum only affects scale
- Idea: use SGD trajectory for Bayesian deep learning

SWA-Gaussian (SWAG)

Training:

- Pre-train a model with e.g. SGD
- Keep running SGD with a high constant learning rate
- Approximate iterates with a Gaussian

Mean: SWA-solution

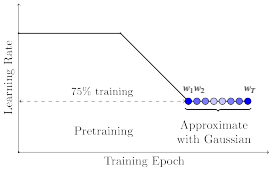
$$w_{SWA} = \bar{w} = \frac{1}{T}\sum_i w_i$$

Covariance: low-rank + diagonal

$$\Sigma_{SWAG} = \frac{1}{2}(\Sigma_{diag} + \Sigma_{low-rank})$$

$$\Sigma_{diag} = \frac{1}{T-1}\sum_i \mathrm{diag}(w_i - \bar{w})^2$$

$$\Sigma_{low-rank} = \frac{1}{T-1}\sum_i (w_i - \bar{w})(w_i - \bar{w})^T$$

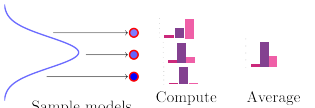


Evaluation:

- Sample models from the Gaussian approximation
- Compute predictions for each model
- Average predictions

$$w_i \sim N(w \mid w_{SWA}, \Sigma_{SWAG})$$

$$p_{BMA}(y \mid x) = \sum_i p(y \mid w_i, x)$$



SWA-Gaussian (**SWAG**):

- A simple and scalable method for Bayesian deep learning
- Fits SGD iterates with a low-rank + diagonal Gaussian distribution
- Captures the geometry of the posterior in the subspace of SGD
- Improves predictions and uncertainty on ImageNet scale

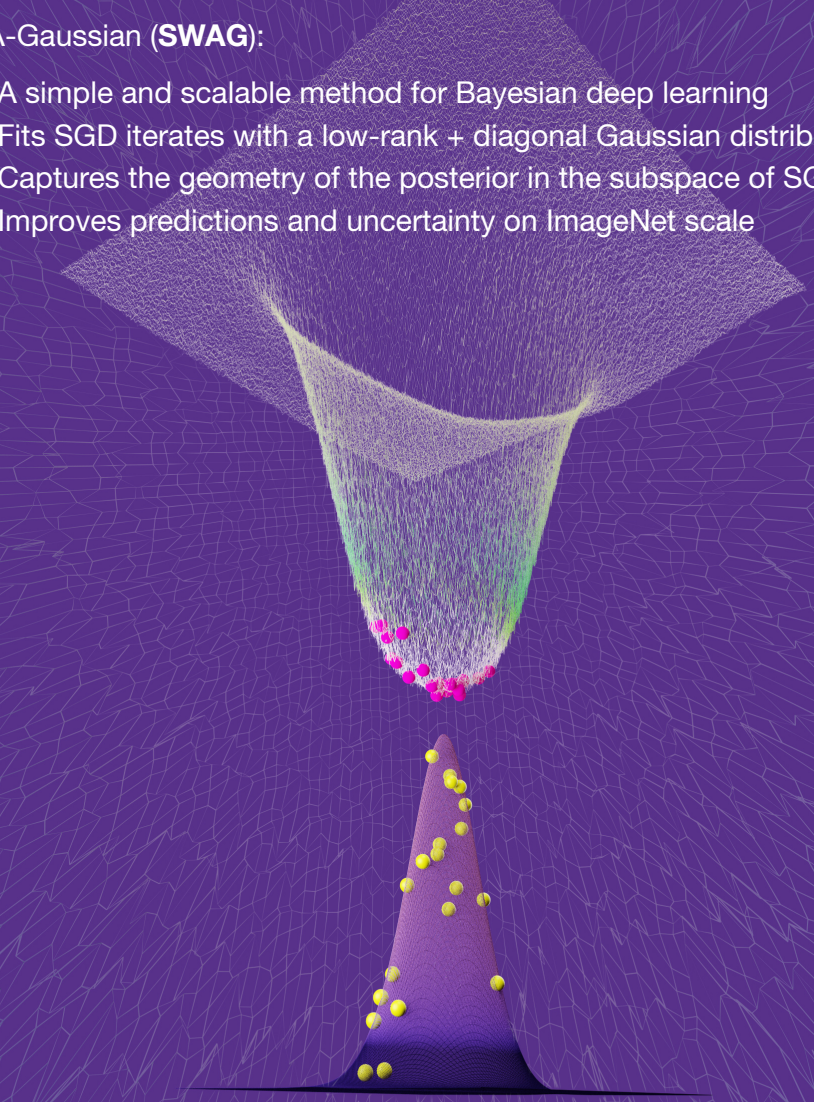


Figure: Posterior log-density surface of a PreResNet-20 trained on CIFAR-10 (**top**) and SWAG approximation (**bottom**) in the subspace spanned by top-2 principal components of SGD trajectory. Illustration produced in collaboration with Javier Ideami (*losslandscape.com*)

Experiments

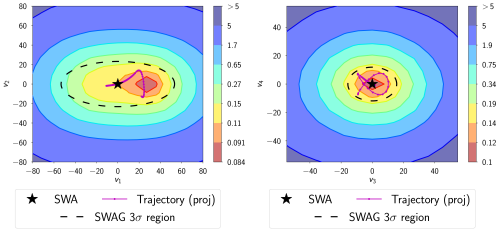


Figure: Posterior log-density surface and SWAG 3-sigma region for a PreResNet-164 trained on CIFAR-100. **Left:** subspace spanned by PCA components #1,2 and **Right:** components #3,4. SWAG captures the geometry of the posterior.

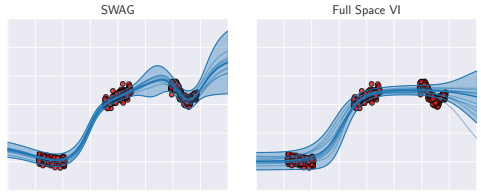


Figure: SWAG and Variational Inference (FFG) predictive distributions for a synthetic 1D regression problem. VI fails to represent epistemic uncertainty.

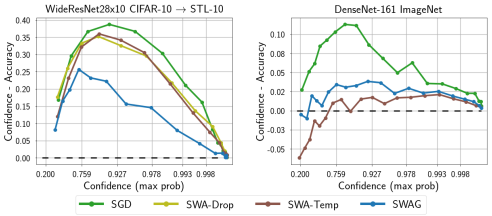


Figure: Reliability diagrams for (**left**) transfer learning setting from CIFAR-10 to STL-10 and (**right**) ImageNet. SWAG produces better calibrated uncertainties, especially under distribution shift.

Method	PTB val	PTB test	WikiText-2 val	WikiText-2 test
NT-ASGD	61.2	58.8	68.7	65.6
SWA	59.1	56.7	68.1	65.0
SWAG	**58.6**	**56.26**	**67.2**	**64.1**

Table: Validation and test perplexities for 3-layer LSTM.

References:

- *Izmailov et al, 2018. Stochastic Weight Averaging, UAI.*
- *Izmailov et al, 2019. Subspace Inference for Approximate Bayesian Deep Learning, UAI.*
- *Mandt et al, 2017. SGD as Approximate Bayesian Inference, JMLR.*

Code Paper Video