Fast Uncertainty Estimates and Bayesian Model Averaging of DNNs

1

Wesley Maddox[1] Timur Garipov[2,3] Pavel Izmailov[1] Dmitry Vetrov[4,5] Andrew Gordon Wilson[1]

[1]Cornell University, [2]Samsung AI Center in Moscow, [3]Lomonosov Moscow State University,
[4]Higher School of Economics, [5]Samsung-HSE Laboratory,

Motivation and Contribution

We want to capture information about the uncertainty of deep neural network (DNN) predictions.

We extend Stochastic Weight Averaging (SWA) [1] by forming a Gaussian distribution around the SWA mean.

SWA-Gaussian (SWAG) produces reliable uncertainty estimates, while maintaining accuracy in Bayesian model averaging.

Methods

Stochastic weight averaging [1] uses the average of the weights of SGD to compute predictions for DNNs.: $\theta_{\text{SWA}} = \sum_{i=1}^{K} \theta_i$.

For convex (and other nice problems), Polyak-Ruppert averaging is asymptotically normal: $\bar{\theta} \sim N(\theta_{\text{true}}, H^{-1}SH^{-1})$, motivating the use of a Gaussian distribution [2].

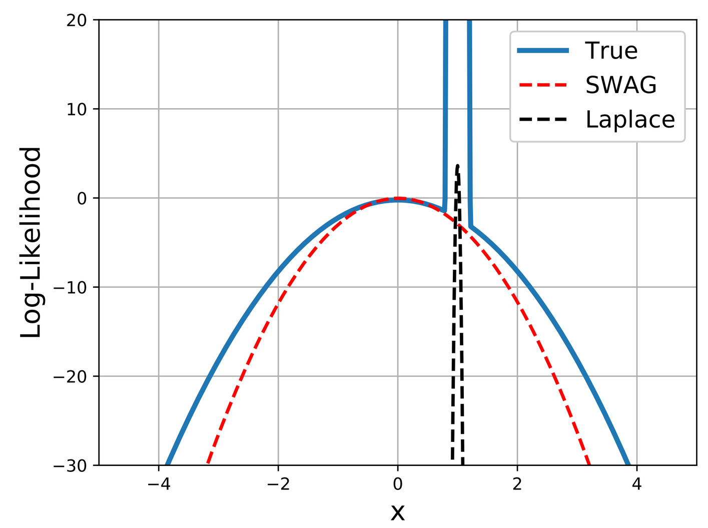
SGD iterates with a constant learning rate are also thought to behave in an approximately Gaussian manner [4].

- **SWAG**: $\theta \sim N(\theta_{\text{SWA}}, XX^T), \quad X_{i.} = (\theta_i - \theta_{\text{SWA}_i})$.
- **SWAG-Diagonal**: $\theta \sim N(\theta_{\text{SWA}}, \sum_{i=1}^{K} \theta_i^2 - \theta_{\text{SWA}}^2)$.

Use as an approximate posterior distribution over θ

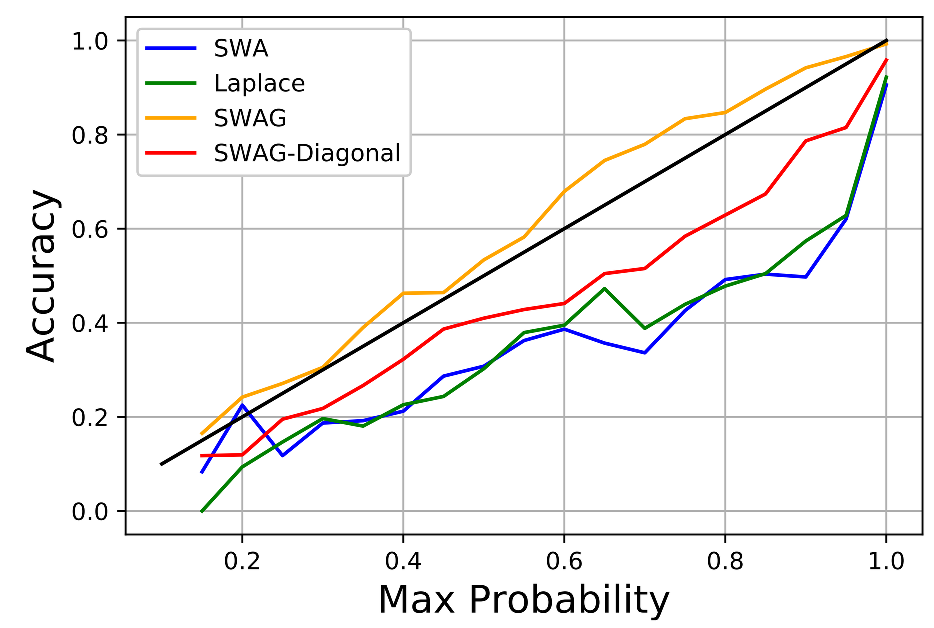
Other Gaussian posterior approximations:

- *Laplace*: $N(\theta_{MAP}, \sigma H^{-1})$ (H^{-1} is very expensive...)
- *Variational Bayes*: $N(\mu, \Sigma)$ (which μ, Σ?)



True likelihood is weighted mixture of Gaussians. SWAG is width-finding, not mode-finding.

Model Calibration



Calibration plots on CIFAR 100 with VGG16 architecture. **update**

Method	ECE
Laplace	0.7604
SWA	0.7650
SWAG-Diagonal	0.7093
SWAG	**0.6001**

Expected calibration error on CIFAR 100. Lower is better.

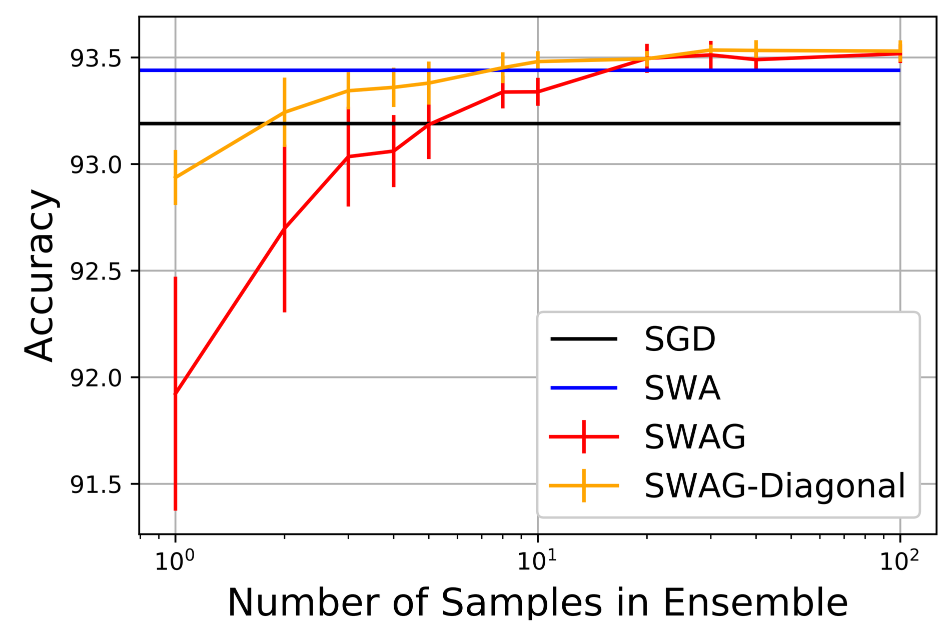
Bayesian Model Averaging

Training only requires memory overhead (for storage).
Test time is just K forwards passes (+ cheap sampling).

Predictions are made using Bayesian model averaging:

$$p(y^*|y) = \mathbb{E}_{p(\theta|y)}(p(y^*|\theta)) \approx \frac{1}{K}\sum_{i=1}^{K} p(y^*|\theta_i), \;\; \theta_i \sim q_{\text{SWAG}}(\theta|Y).$$

Dataset (epochs)	SGD, point	SWA	SGD, empirical	SWAG, 30 samples
CIFAR-10 (300)	93.19 ± 0.22	93.44 ± 0.09	$\mathbf{93.64} \pm 0.14$	$\mathbf{93.57} \pm 0.15$
CIFAR-10.1	84.93 ± 0.32	86.14 ± 0.59	85.78 ± 0.20	$\mathbf{86.24} \pm 0.67$
CIFAR-100 (300)	73.29 ± 0.38	74.04 ± 0.25	$\mathbf{74.74} \pm 0.26$	$\mathbf{74.57} \pm 0.39$

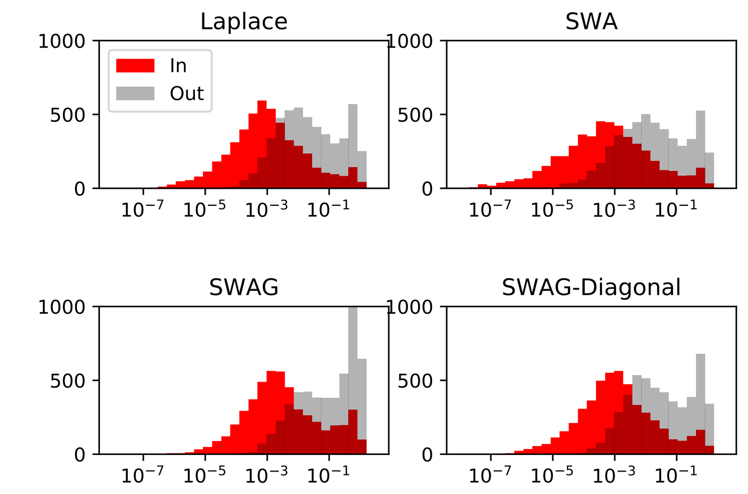


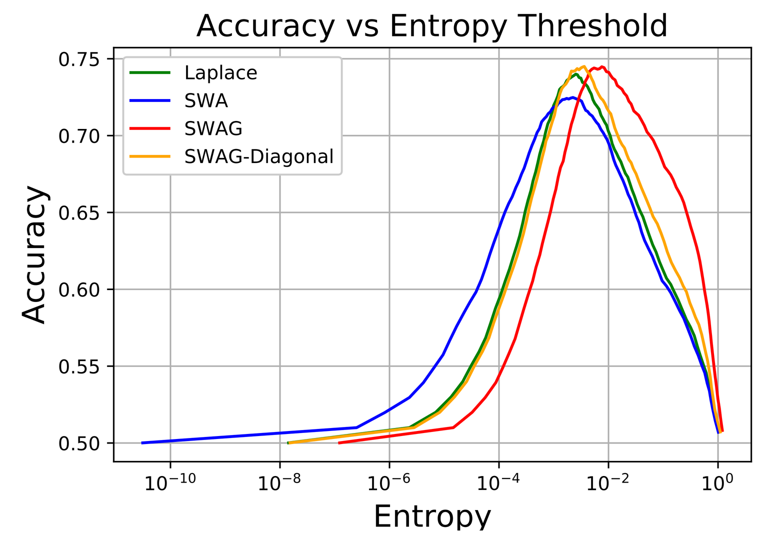
Number of samples from approx. posterior vs accuracy. VGG16 on CIFAR10.

The distribution of SGD iterates is empirically well approximated by SWAG.

Out of Distribution Uncertainty

We trained VGG16 on 5 classes from CIFAR10, and then tested on all 10 classes. Entropy, $\sum_{i=1}^{5} p(y=i)\log p(y=i)$, should be higher if the model is unsure.





Distribution of entropies on CIFAR 5 task. Accuracy for a given entropy threshold.

Conclusions

- Principled method for approximate Bayesian inference that scales well for DNNs.
- SGD posterior appears Gaussian, as theory predicts [2,4]; can also interpret SWA as a posterior mean.

Future Work

- Expand theoretical motivation, like in [3].
- Comparisons with other methods for approximate Bayesian inference
  - Laplace, Variational Bayes, MC Dropout, etc…
- Other problems: penalized regression [3], adversarial robustness, ImageNet, image segmentation

Code

`https://github.com/wjmaddox/swa_uncertainties`

References

[1] Izmailov et al, Averaging Weights Leads to Wider Optima and Better Generalization. UAI, 2018.

[2] S. Asmussen and P.W. Glynn. Stochastic simulation: algorithms and analysis. Springer, New York. 2007.

[3] Chen et al, Statistical Inference for Model Parameters in Stochastic Gradient Descent. arXiv: 1610.08637, Oct. 2016.

[4] Mandt et al, Stochastic Gradient Descent Performs Approximate Bayesian Inference. JMLR, 2017.

Acknowledgements

Ruqi Zhang, Jacob Gardner. WM is supported by an NSF Graduate Research Fellowship.