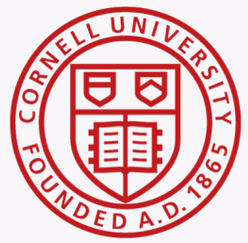
FAST UNCERTAINTY ESTIMATES AND BAYESIAN MODEL AVERAGING OF DNNS

WESLEY MADDOX

JOINT WORK WITH TIMUR GARIPOV, PAVEL IZMAILOV, DMITRY VETROV, ANDREW GORDON WILSON



SUMMARY

- Stochastic Weight Averaging **(Izmailov et al, UAI, 2018)** computes first moment of weights given from SGD iterates with a modified learning rate schedule.

- We propose to keep the variance as well to form a **Gaussian approximation in weight space.**

- Sample from Gaussian to compute **Bayesian model averages and estimate uncertainty.**

- **Theoretically motivated** from results on SGD & relation of iterates to Gaussian distribution (Ruppert, 1992 and Mandt et al, 2017).

APPROXIMATE BAYESIAN INFERENCE

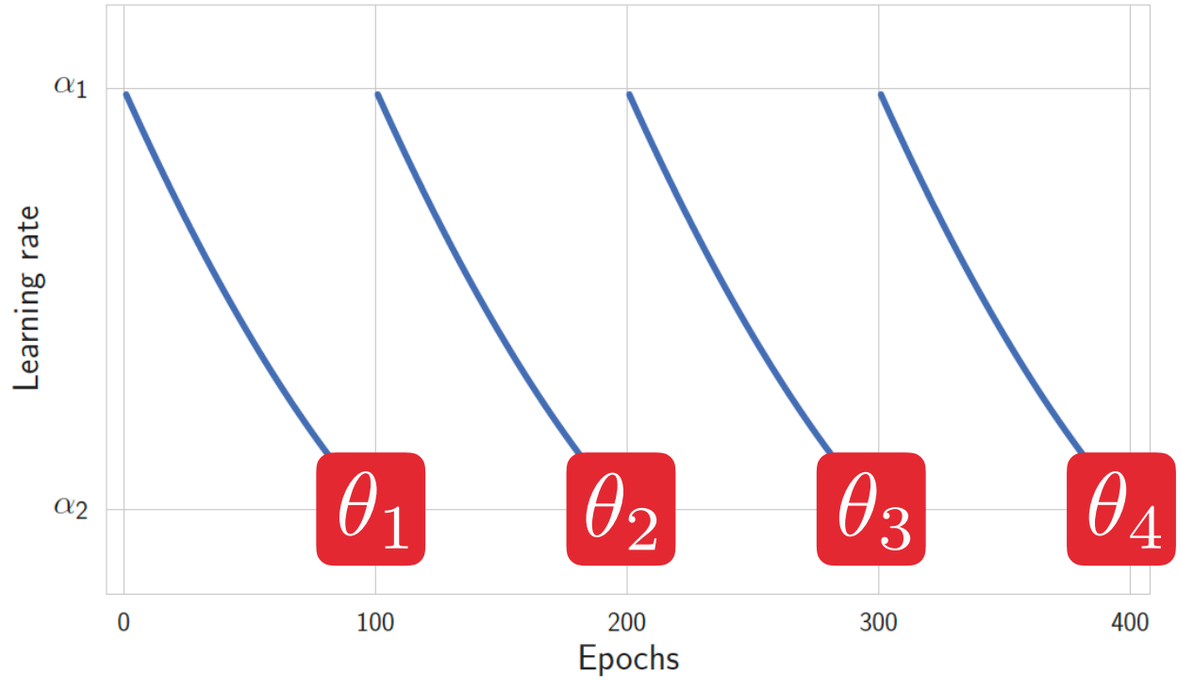
- Why?
  - Compute intractable integrals

$$p(y^*|y) = E_{p(\theta|y)}(p(y|\theta))$$

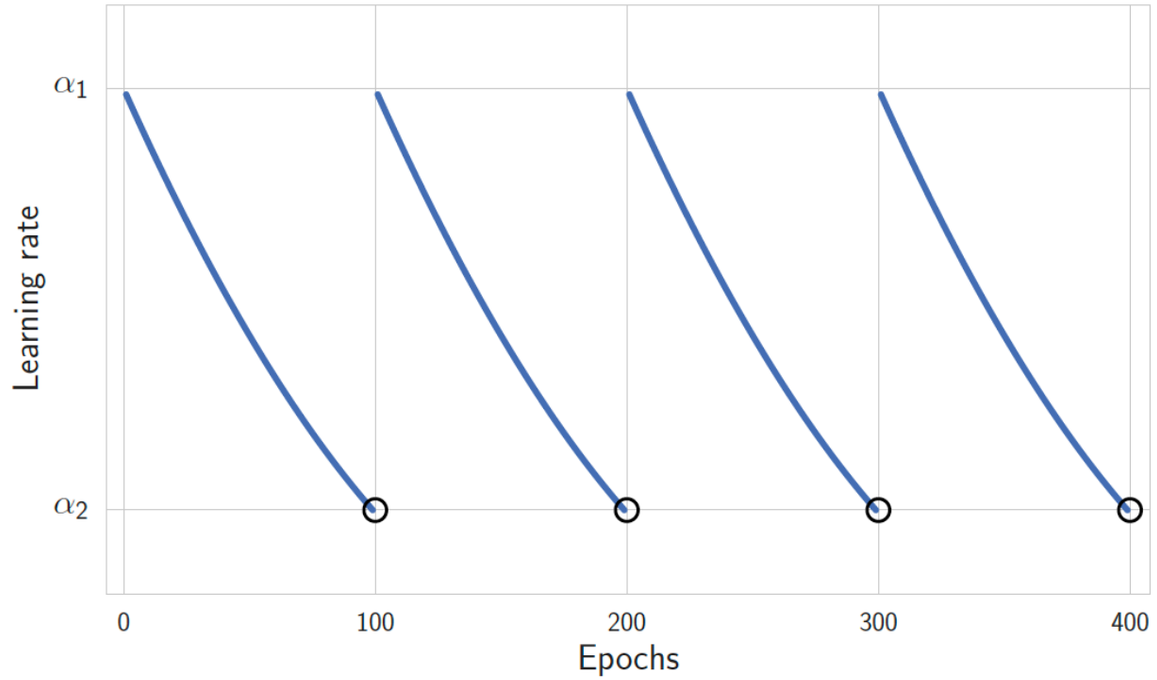
  - Uncertainty quantification
- How?
  - Laplace: $p(\theta|y) \approx N(\theta_{MAP}, (H(\theta_{MAP}) + \lambda I)^{-1})$
  - Variational Bayes: $p(\theta|y) \approx N(\mu, S)$
  - Markov Chain Monte Carlo

STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)

Cycle the learning rate of SGD.



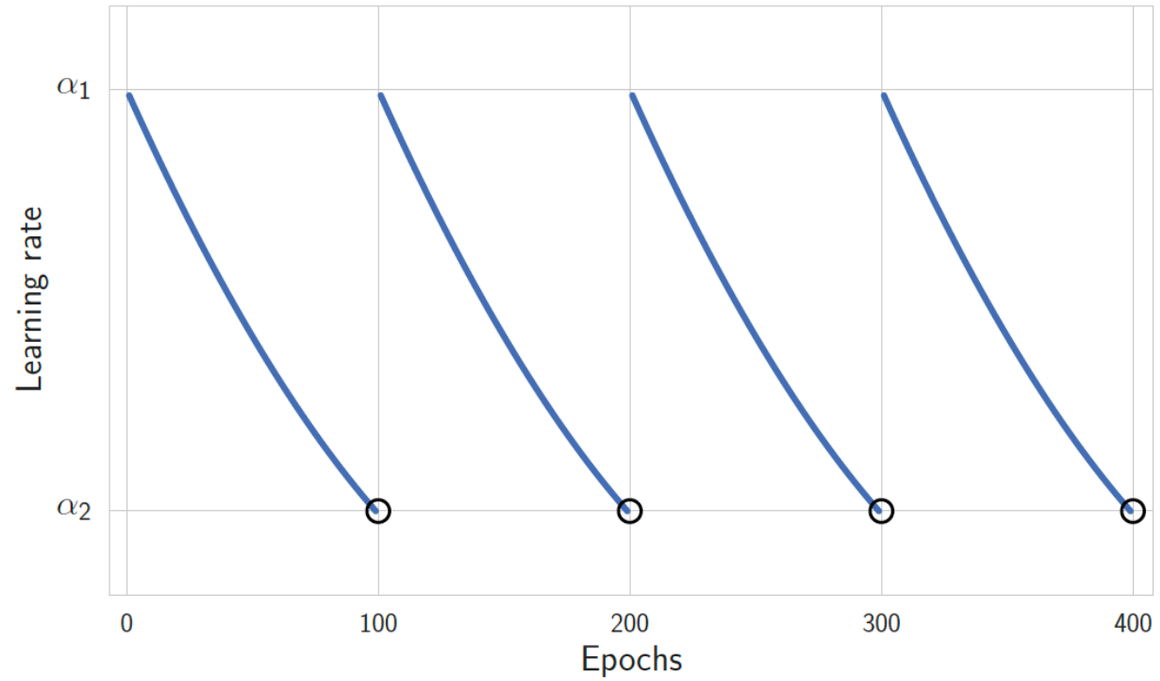
STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)



Average models at end of cycles.

θ_i

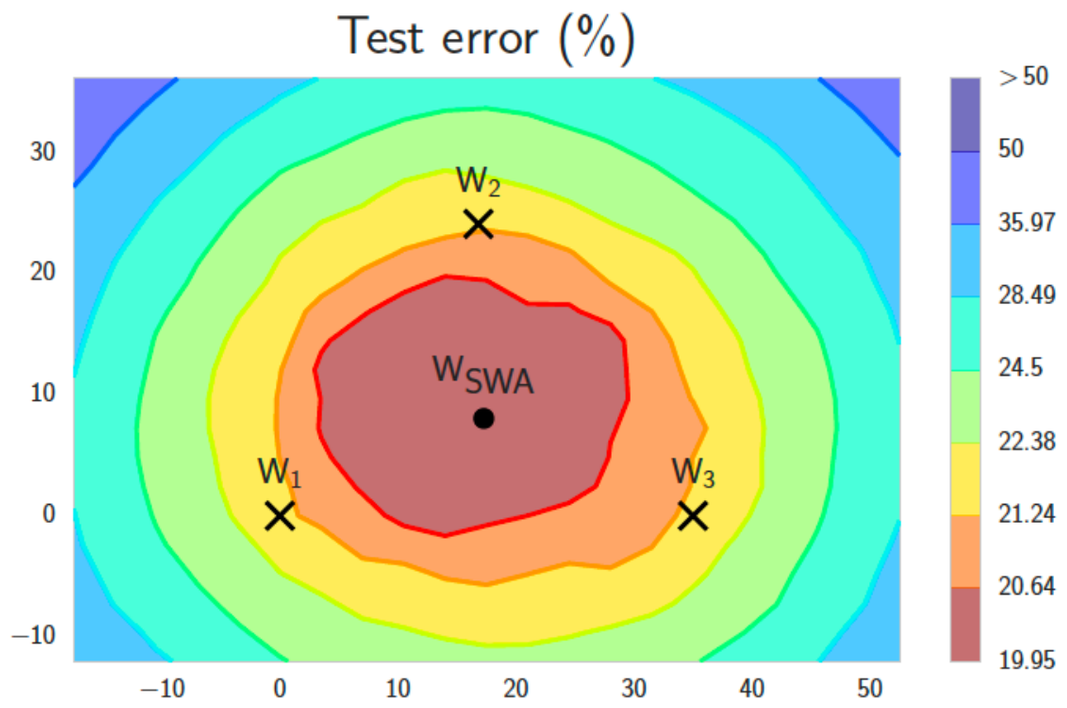
STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)



Average models at end of cycles.

$$\theta_{swa}$$

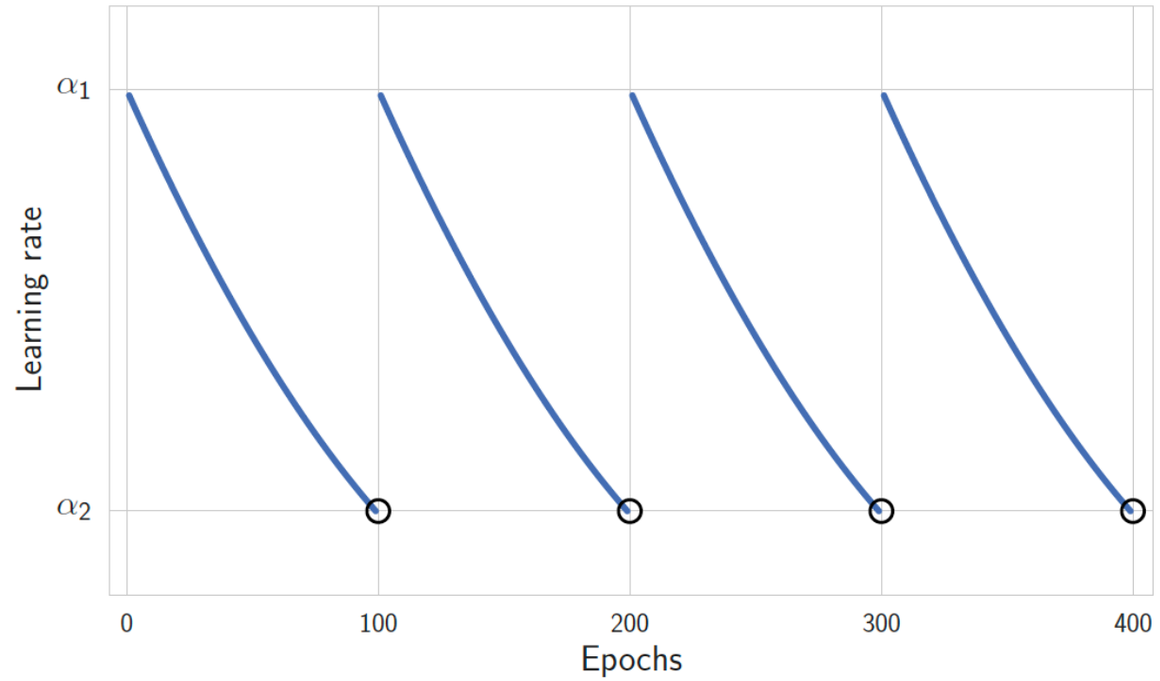
STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)



Average models at end of cycles.

$$\theta_{swa}$$

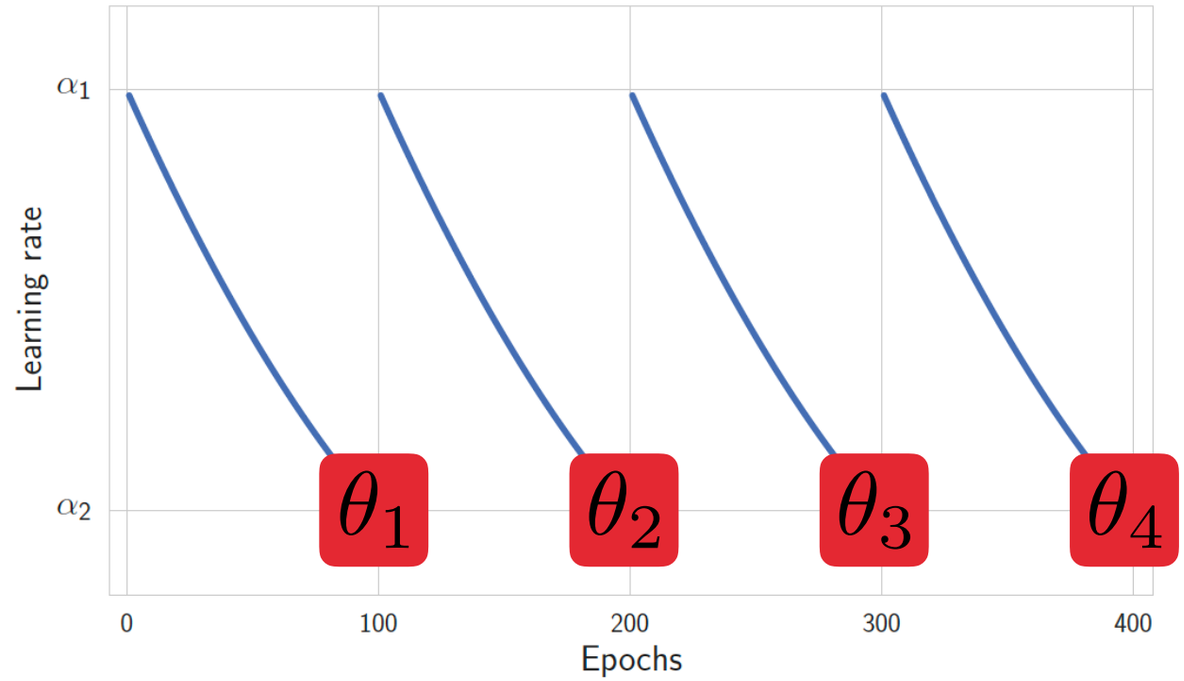
STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)



Average models at end of cycles.

$$\theta_{swa}$$

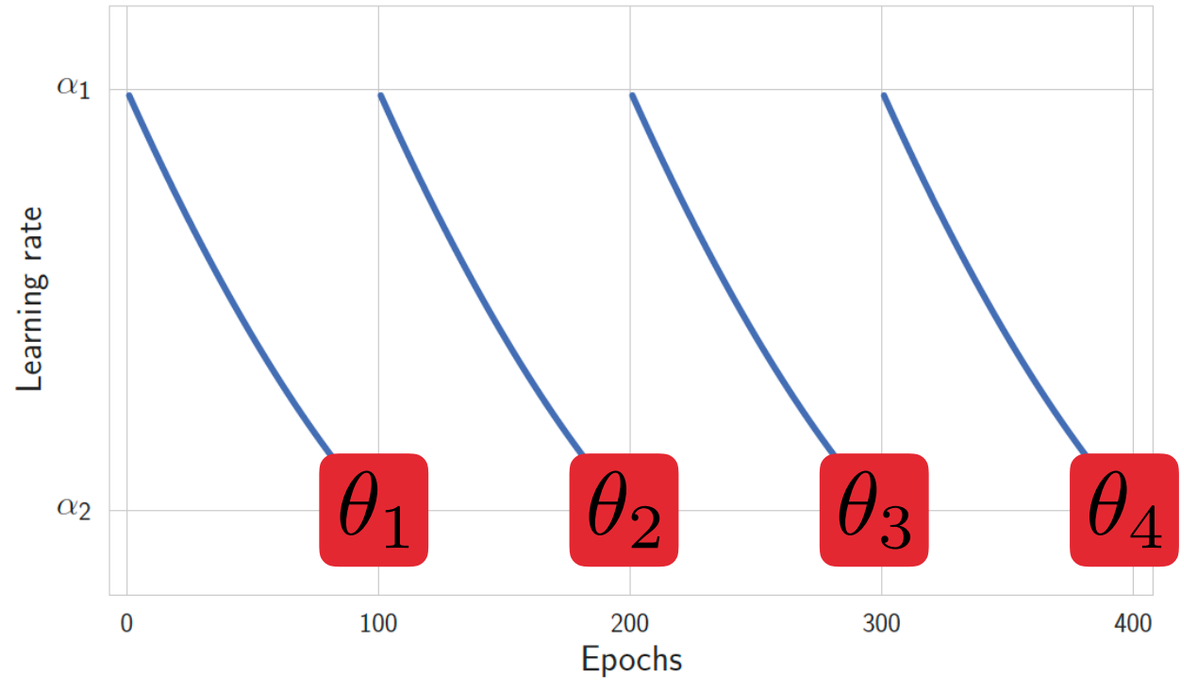
STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)



θ_{swa}

$X = [\theta_i - \theta_{swa}]_{i=1..s}$

STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)

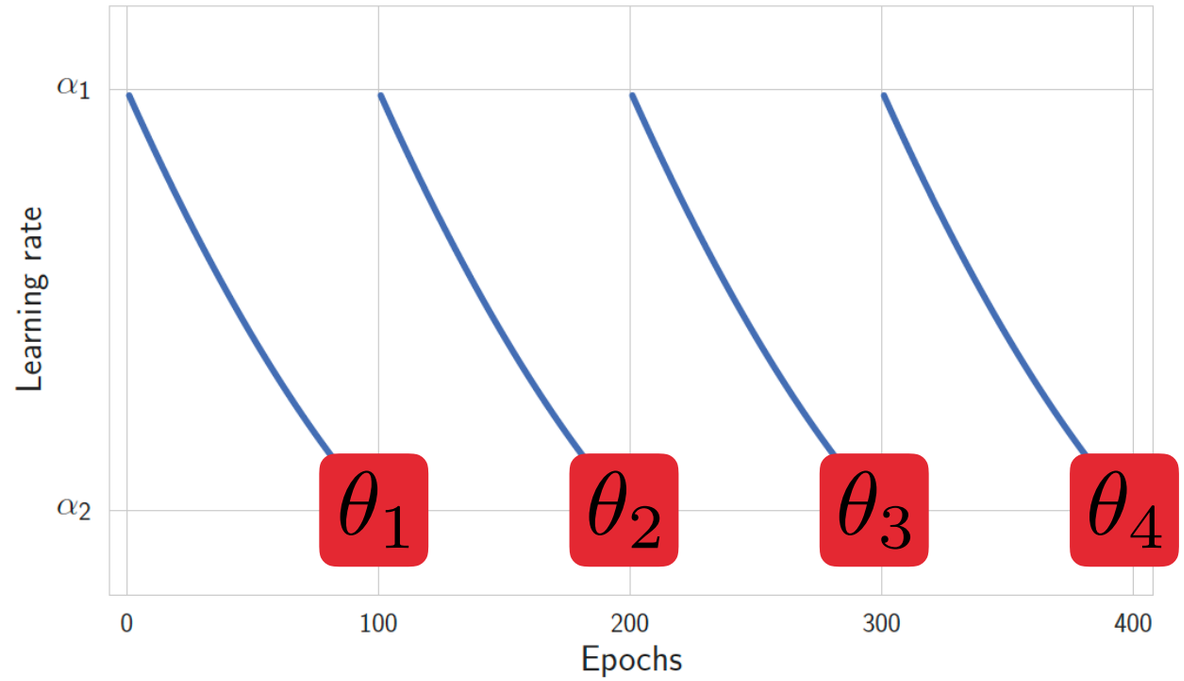


θ_{swa}

$X = [\theta_i - \theta_{swa}]_{i=1..s}$

$\theta_{swag} \sim \mathcal{N}(\theta_{swa}, XX^\top)$

STOCHASTIC WEIGHT AVERAGING (IZMAILOV ET AL 2018)



θ_{swa}

Diagonal version:

$$\theta_{\mathrm{swa,diag}} \sim N\left(\theta_{\mathrm{swa}}, \sum_{i=1} \theta_i^2 - \theta_{\mathrm{swa}}^2\right)$$

See also Liu et al, 2018, UDL Workshop

BAYESIAN MODEL AVERAGING WITH SWAG

- Monte Carlo estimates of predictions: $p(y^*|y) \approx \frac{1}{K}\sum_{i=1}^{K} p(y^*|\theta_i), \theta_i \sim q_{SWAG}(\theta|y)$
- Results in approximate Bayesian inference.

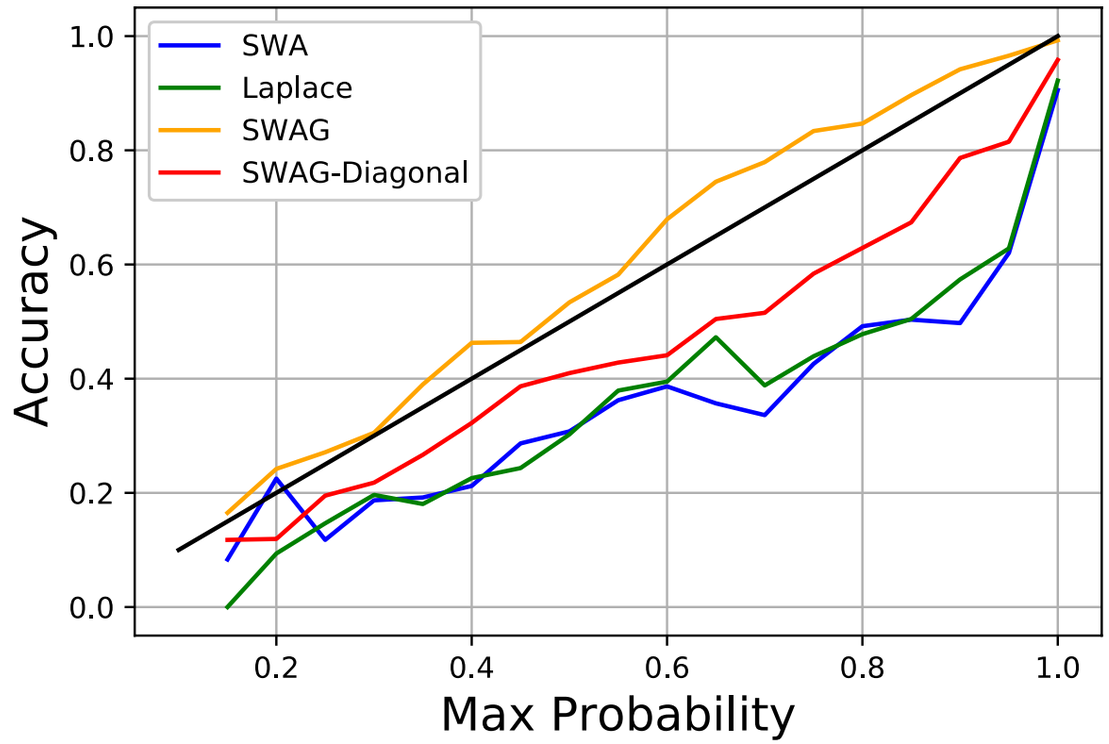
Dataset (epochs)	SGD, point	SWA	SGD, empirical	SWAG, 30 samples
CIFAR-10 (300)	93.19 ± 0.22	93.44 ± 0.09	$\mathbf{93.64} \pm 0.14$	$\mathbf{93.57} \pm 0.15$
CIFAR-10.1	84.93 ± 0.32	86.14 ± 0.59	85.78 ± 0.20	$\mathbf{86.24} \pm 0.67$
CIFAR-100 (300)	73.29 ± 0.38	74.04 ± 0.25	$\mathbf{74.74} \pm 0.26$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

MODEL CALIBRATION

- Can Bayesian methods help fix the model calibration problem for DNNs?

- Expected calibration error (Naeini et al 2015)

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

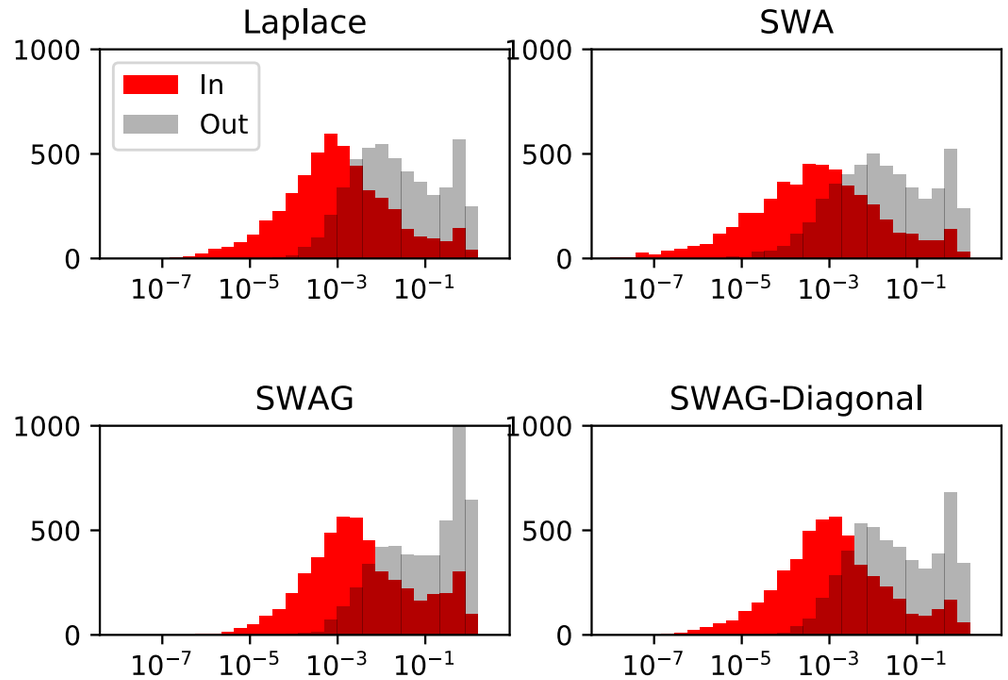


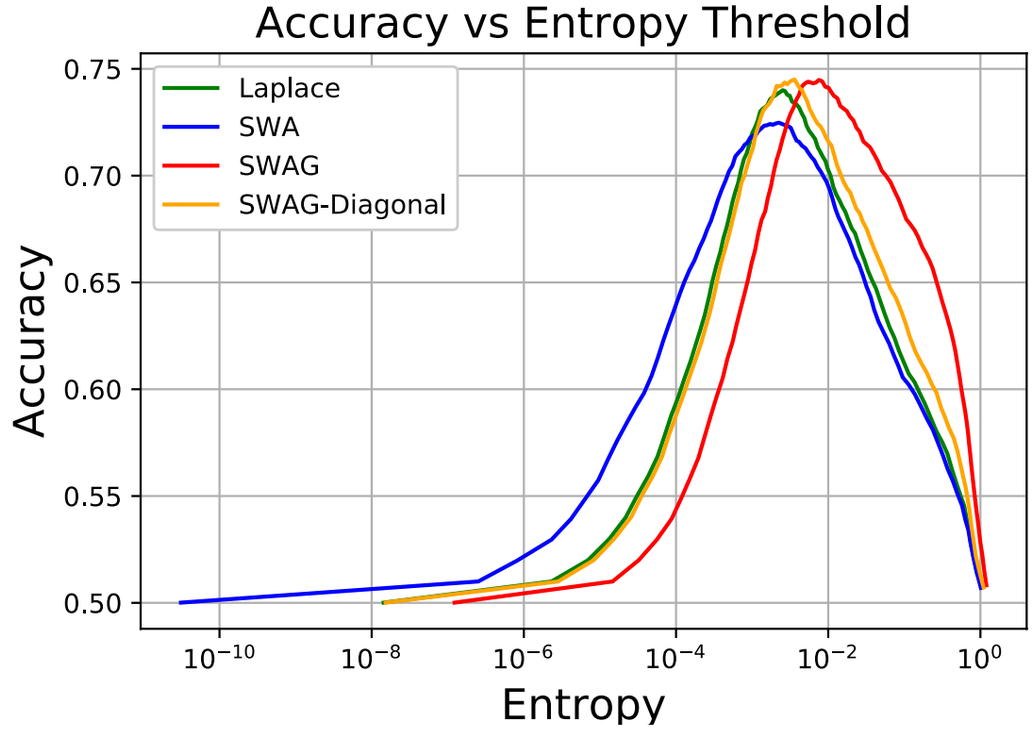
VGG16 on CIFAR100.

OUT OF DISTRIBUTION UNCERTAINTY

- Train DNN on 5 classes of CIFAR 10 (in-class), test on in-class test set and other examples (out-of-class).
- How well can SWAG tell them apart?

Method	AUC
Laplace	80.41
SWA	79.26
SWAG-Diagonal	**80.84**
SWAG	**80.86**





CONCLUSIONS & FUTURE WORK

- **Code at: https://github.com/wjmaddox/swa_uncertainties**
- Theory: connection to Polyak-Ruppert Averaging (see Chen et al 2016).
- Comparisons with other methods for approximate Bayesian inference.
  - Laplace, Variational Bayes, etc…
- Adversarial attacks and defenses.
- Check out our poster…

REFERENCES

- X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical Inference for Model Parameters in Stochastic Gradient Descent. arXiv: 1610.08637, Oct. 2016.
- P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. In UAI , 2018.
- J. Liu, S. Tripathi, U. Kurup, M. Shah, Make (Nearly) Every Neural Network Better: Generating Neural Network Ensembles by Weight Parameter Resampling. In UAI Workshop on Uncertainty in Deep Learning, 2018.
- M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In AAAI , pages 2901–2907, 2015
- B. T. Polyak and A. B. Juditsky. Acceleration on Stochastic Approximation by Averaging. SIAM Journal on Control and Optimization , 30(4):838–855, July 1992.
- D. Ruppert. Efficient Estimators from a Slowly Convergent Robbins-Munro Process. Technical Report 781, Cornell University, School of Operations Research and Industrial Engineering, 1988.

ASYMPTOTIC MOTIVATION OF SWAG

- Polyak-Ruppert Averaging (Ruppert 1988; Polyak & Juditsky, 1992)
  - Average the iterates of SGD
  - Asymptotic distribution (around stationary point):

$$\frac{1}{T}\sum_{i=1}^{T}\theta_i \approx N(\theta, H(\theta)^{-1}SH(\theta)^{-1})$$

- Laplace approximation uses Gaussian around MAP with covariance H(\theta)

SWA WITH GAUSSIANS (SWAG): FORMULAS

- (diagonal) **Laplace**: compute Hessian diagonals
- **SWAG**: diagonal approx.

$$N(\theta_{SWA}, \bar{\theta^2} - \theta_{SWA})$$

- **SWAG-LR**: SWA + covariance

$$N(\theta_{SWA}, XX'), X_j = (\theta_j - \theta_{SWA_j})$$

- **SWAG-Hessian**: compute Hessian diagonal + covariance
- $$N\left(\theta_{SWA}, H_{ii}^{-1} XX' H_{ii}^{-1}\right)$$

EXPECTED CALIBRATION ERROR

$$E_{\hat{P}}\Big[\Big|P(\hat{Y}=Y|\hat{P}=p)-p\Big|\Big]$$

$$ECE=\sum_{m=1}^{M}\frac{|B_m|}{n}\Big|acc(B_m)-conf(B_m)\Big|$$

From Naeini et al 2015, also Guo et al ICML 2017 “On Calibration of Modern Neural Networks”