We would like to thank all the reviewers for thoughtful feedback. As the reviewers pointed out, SWAG is a very practical Bayesian deep learning method readily applicable to ImageNet-scale problems. SWAG achieves strong results on image classification, tabular regression and language modeling, out-performing strong and elaborate Bayesian deep learning methods. We also explicitly demonstrate that SWAG can capture the shape of the posterior (along certain directions) in Section 4, which justifies using SWAG as an approximation to the posterior distribution. We believe that our paper (i) sets a strong baseline for Bayesian deep learning and (ii) motivates researchers in the field to conduct realistic evaluations on large-scale datasets and models, and (iii) use loss surface visualizations to show that the approximate posterior distribution captures the shape of the true posterior.

Inspired by reviewer suggestions, we ran two additional experiments. First, we evaluated **ensembles of SGD iterates** that were used to construct the SWAG approximation for all of our CIFAR models. We report NLLs in the table:

	CIFAR-100		CIFAR-10	
Architecture	SWAG	SGD-Ens	SWAG	SGD-Ens
VGG-16	0.9480 ± 0.0038	$\mathbf{0.8979} \pm 0.0065$	0.2016 ± 0.0031	$\mathbf{0.1883} \pm 0.002$
PreResNet-164	$\mathbf{0.6595} \pm 0.0019$	0.7839 ± 0.0046	$\mathbf{0.1232} \pm 0.0022$	0.1312 ± 0.0023
WideResNet28x10	$\mathbf{0.6078} \pm 0.0006$	0.7655 ± 0.0026	$\mathbf{0.1122} \pm 0.0009$	0.1855 ± 0.0014

SWAG loses on VGG-16, but wins by a large margin on the larger PreResNet-164 and WideResNet28x10; the results for accuracy and ECE are analogous. We will include these results as well as results on ImageNet and transfer learning in the camera-ready version. Second, we evaluated **ensembles of independently trained SGD solutions** on PreResNet-164 on CIFAR-100. We found that an ensemble of 3 SGD solutions has high accuracy (82.1%), but only achieves NLL 0.6922, which is *worse than a single SWAG solution*. An ensemble of 5 SGD solutions achieves NLL 0.6478, which is *competitive with a single SWAG solution, that requires* $5\times$ *less computation to train*. Moreover, we can similarly ensemble independently trained SWAG models; an ensemble of 3 SWAG models achieves NLL of 0.6178.

R1: We thank the reviewer for the thoughtful and positive review. In addition to the new results we discuss above, we note that in appendix Figure 3a we show that in terms of accuracy SWAG outperforms an ensemble of SGD iterates. We would also like to note that in many problems, such as incremental learning (see e.g. [1]), it is desirable to represent uncertainty over weights as a closed form distribution, rather than just storing samples. Further, we can produce an arbitrary number of samples from a fixed SWAG approximation, and in appendix Figure 3b, we show that NLL of the ensemble continues to improve as we add more samples. With just using ensembles of SGD iterates, we cannot cheaply increase the ensemble.

R2: Thank you for the thoughtful and positive review. See the above comparison with SGD-ensembles. [2] demonstrated that high-frequency ensembles of SGD iterates typically outperform snapshot ensembles, so we focus on the former.

R3: While we value the feedback, and are happy you appreciate the quality of the work, we do not agree that the paper should be rejected unless SWAG is not called "an approximation to Bayesian learning". The proposed method is unequivocally an approximate Bayesian inference approach, exactly analogous to the Laplace approximation or variational methods. Similar to many such canonical approximate Bayesian inference procedures, we use a Gaussian approximation to the posterior, but centred on the SWA solution, with curvature defined by the SGD trajectory; for comparison, the Laplace approximation uses a Gaussian centred on the SGD solution with curvature defined by the Hessian of the posterior log-density at that point. Whether or not the posterior is truly Gaussian (as modeled by Laplace or SWAG), or whether the Gaussian should be centred at an SGD solution (as in Laplace), or what its curvature should be, or whether the stationary distribution of SGD is Gaussian, are reasonable questions for Laplace, variational approaches, SWAG, and many other methods, but orthogonal to whether these methods provide approximate Bayesian inference. It is fair to question the assumptions – indeed we do so ourselves in the paper, and provide exhaustive empirical support in Section 4 – but calling SWAG an approximate Bayesian method is factually correct and thus not a fair reason for rejection. Moreover, the assumptions of our procedure are much milder than many standard approximate Bayesian inference procedures, such as the widespread mean-field variational approximations which assume fully factorized posteriors. While, as you mention, it is possible to construct special cases where the stationary distribution does not capture the shape of the posterior (Section 6.2 of [3]), in general these distributions are tightly constrained as in equation (13) of [3]. In Section 4 of the paper (in particular Figures 1 and Appendix Figure 2) we go beyond many works employing Gaussian posterior approximations to explicitly demonstrate that the posterior for our applications is approximately Gaussian in the PCA subspace of the SGD trajectory and SWAG is able to capture its shape.

We evaluated ensembles of independent SGD solutions as you suggested; please see the discussion above.

References:

[1] Kirkpatrick, James, et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017.
[2] Garipov, Timur, et al. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. NeurIPS, 2018.
[3] Mandt, Stephan, et al. Stochastic Gradient Descent as Approximate Bayesian Inference. JMLR, 2017.