We thank the reviewers for their helpful feedback. We were glad to see that all reviewers understood and remarked on both our theoretical contributions showing an equivalence between a notion of training speed and the Bayesian model evidence in linear models, and our empirical results confirming 1) the theory in the linear setting and 2) showing similar mechanisms at play in neural networks. In particular, **R1**: The connection between training speed and the model evidence is interesting, and [...] the main idea is original; and **R2**: I believe the authors' observations and conclusions are worthy of attention. We now address some concerns.

R1: 'The experiments are conducted only in toy dataset.' We have replicated the DNN experiments (S4.2) on the CIFAR-10 dataset and observe similar results as for FashionMNIST; as can be seen in Figure 1 below.

R1: 'Equation (2) assumes the update follows a Bayesian updating procedure. However [...]' We leverage the work of Matthews et al. [citation 15 in the paper] demonstrating that in the linear setting, gradient descent is approximately equivalent to Bayesian posterior updating. Our experiments on linear models all follow this procedure, consistent with our theory. Extending our results to multi-epoch SGD and the nonlinear setting is an exciting avenue for future work.

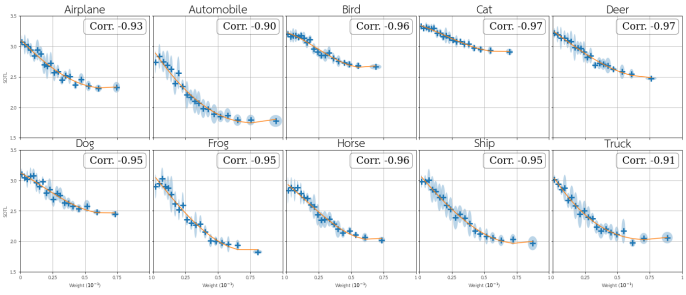

Figure 1: Replication of Figure 5 on CIFAR-10.

R2: 'The second estimator is an unbiased estimator of the (log) marginal likelihood (ML)'. We thank the reviewer for bringing this potential source of confusion to our attention. We affirm that *both* estimators compute unbiased estimates of members of a family of lower bounds on the marginal likelihood (defined with respect to the number of samples J) and so will produce biased estimates of the marginal likelihood. We can derive this result using Jensen's inequality as follows: $\mathbb{E}_{\theta_j^i \sim P(\theta|\mathcal{D}_{<i})}[\sum_{i=1}^n \log \frac{1}{J} \sum_{j=1}^J P(\mathcal{D}_i|\theta_j^i] \leq \sum_{i=1}^n \log \mathbb{E}_{\theta_j \sim P(\theta|\mathcal{D}_{<i})} \frac{1}{J}[\sum_{j=1}^J P(\mathcal{D}_i|\theta_j^i)] = \sum_{i=1}^n \log P(\mathcal{D}_i|\mathcal{D}_{<i})$. I deduced [...] this should have been made clear by the authors. We will clarify the number of samples used to estimate $\log P(\mathcal{D}_i|\mathcal{D}_{<i})$ in our revisions; thanks for highlighting this. [...] this leaves the question about their experiments in Figure 2 which they presented without any reference to lack of theoretical results. Because both discussed estimators are for lower bounds to the log ML (and we know that Jensen's inequality becomes tighter with reduced variance), the results of Figure 2 are in fact consistent with our theory, which is developed in section 3.2.1.

R2: 'interpretation of [Jiang et al.].' We agree that the wording of Jiang et al. is ambiguous; we have verified the numerical results from the paper and are reasonably confident that our interpretation is correct.

'I was not able to ascertain how the result of Theorem 2 is used in the text, I'd be happy if the authors could clarify.' Theorem 2 is the noiseless analogue of Theorem 1, and we include it to demonstrate that it is still possible to obtain an estimate of the model evidence based on training statistics for linear models in the noiseless setting. The reviewer is correct that we do not follow up on this result empirically; this is to save space for experiments that more clearly link training speed, generalization, and the marginal likelihood.

R1: 'There should be more baselines in the experiments sections [...] LASSO [...]' We agree that a comparison between Bayesian model selection and other model selection baselines is an important line of work; unfortunately, a thorough empirical analysis of the generalization performance of Bayesian model selection compared to non-Bayesian baselines was outside the scope of this work. Because we show that our method obtains a similar ranking over models as Bayesian evidence maximization, we can extrapolate that it will exhibit similar strengths and weaknesses as exact marginal likelihood maximization when compared against other baselines such as LASSO.

R2: 'I found that the transition to the neural networks remains a bit confusing.' We agree that the analysis of neural networks is quite distinct from our results in the linear setting, and we believe that bridging from linear models to neural networks via a discussion of infinitely wide neural networks (the neural tangent kernel regime) will clarify this transition. Our sample-then-optimize lower bound estimator for linear models can be applied in a straightforward way to gradient descent on infinitely wide networks trained using the procedure outlined in [1] (posted after our submission to NeurIPS). We have updated the paper to discuss this corollary of our results as a way to better motivate and provide context for the neural networks results.

'how much the results support the marginal likelihood-based model selection hypothesis, or whether they should more safely be observed only as a relationship between SOTL and generalization.' We thank the reviewer for highlighting this. Our empirical results for DNNs are presented as evidence that the mechanism that we observe in linear models seems to be at work in deeper models as well, as a way to motivate further exploration of the ideas presented here. We will clarify this aspect in our paper.

R3: 'Figures 1,2, and 3 needs clearer explanation about the experiment setup.' We thank the reviewer for bringing up this source of confusion and will clarify this in our edits: due to the rebuttal's page limit, we defer to the discussion of experimental setup in the appendix if more details are needed.

[1]Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel.