Our work "establishes interpretations of SGD and Adam-family optimizers from a Bayesian filtering perspective" (R3). It is "the first to demonstrate how viewing optimization as Bayesian inference requires modeling temporal dynamics" (R4) and results in an algorithm that is "easy to implement and is computationally efficient" (R4) and "benefits from good optimization and generalization properties" (R2). Finally, the reviewers recognised the potential impact of our method "A unifying formulation with connections to Bayesian inference may help to improve that situation of the field, and help build significantly better methods in the future" (R3)

Shared points. Gaussianity (R1, R2). We included considerable empirical analysis of the Gaussianity in appendices A+B, and Figs A1-7. All show excellent empirical agreement with Gaussianity in this setting. Note, this is partly because we care about minibatch gradients, which are averages of 128 independent training-example-gradients, so central-limit arguments push the distribution towards Gaussianity (but without taking the limit, all we can do is establish empirical agreement). **Empirical results (R1, R3).** While our method improves over all baseline adaptive methods, and often also SGD, we agree that these improvements are not spectacular. Our major contribution is to give a Bayesian approach that "interpolates between Adam (a state of the art optimizer) and vanilla gradient descent, and also recovers Adam W" (R4), and therefore explains the excellent performance of these SOTA methods. We would hope that our rigorous approach would give some performance improvements, which it does, but the underlying similarities to these SOTA methods (which become exact in various asymptotic limits) imply that we cannot expect huge performance differences. **Approximations (R2, R4).** Inevitably, there are approximations and heuristics (including $\eta/(2\sigma^2)$) required to obtain an efficient and effective method in these extremely challenging high-dimensional settings. However, these are much less concerning than those in past work (Khan et al 2018) that required "unintuitive modifications to derive Adam's root mean square normalizer"(R3). We believe that these issues can be resolved by using a more complex dynamical prior over weights. However, this approach introduces considerable additional complexity, which is simply too much for the first paper "to demonstrate how viewing optimization as Bayesian inference requires modeling temporal dynamics"(R4). **Conclusion.** I would urge the reviewers to consider the value of the approach broadly, and in particular that "A unifying formulation with connections to Bayesian inference may help to improve that situation of the field, and help build significantly better methods in the future." (R3). For instance, one particularly exciting extension is to infer a posterior over a full weight matrix, rather than each element separately. This results in a K-fac variant of Adam, where we precondition updates by the *square root* of the inverse Fisher Information, rather than the inverse Fisher Information as in standard natural gradient approaches. This approach can be expected to offer big benefits in terms of stability and generalisation error, just as Adam, with a RMS gradient normaliser, offers big benefits over using a squared-gradient normaliser. However it is difficult to ask my students to work on these exciting and challenging extensions when this foundational work remains unpublished.

R1. 3 See "Gaussianity" above. See A8-12 for much more in-depth analysis of the results, including training losses and accuracies, also see "Empirical results" above. We have updated the paper to include a discussion of computational and time complexity, both are $\mathcal{O}(N)$, where N is the number of parameters. Practically, performance is very similar to standard methods such as Adam. **5** We have updated the manuscript to introduce Algos 1 and 2.

R2. 3.1 In the ideal case you shouldn't use a factorised model, and 77-81 aren't trying to motivate a factorised model. But the high-dimensionality of typical CNNs *forces* approximate, factorised models. 77-81 are only arguing that if we must use a factorised model, we should use one with dynamics. Also, see "Conclusions" above for non-factorised future work. **3.2** See "Gaussianity" above. **3.3** Eq 12 should not yet reflect gradient-based optimization, as it only describes the prior distribution under which we perform inference. The multiplicative decay is necessary because if we just had noise, it would imply that a-priori, the weights slowly grow to infinity. **3.4** The Hessian substitution is standard in the literature (e.g. Khan et al. 2018), but we agree that its improvement is an important avenue for future research. **3.5** See "Approximations" above. **Minor** 1. Agreed, but a few people get very confused on this point. 2. Fixed. 3. Fixed.

R3. 3.1 We have written a new section introducing filtering methods. **3.2** We agree, these plots aren't the main point, but it remains is valuable to show that our method indeed achieves somewhat improved performance (see "Empirical results" above). **3.3** Many of our existing plots support the main argument of the paper: we have detailed plots showing the Gaussianity assumptions of the method hold (Fig A1-7) and showing that our steady-state limits hold in practice (Fig 3). **3.4** We did not need to run for longer, but it is useful because it gives strictly more information about the performance of the method. The results do not change markedly if we run for 200 epochs. **4** Thanks! We have cited [1] and been more careful about the "FI" terminology. **5+8** Thanks! A number of great points, all fixed.

R4. **3.1** See "Approximations" above. **3.2** 0.1 is the default initial learning rate for SGD in these models/datasets, and is the best for SGD in the hyperparameter search. **4.** Excellent point: the momentum is not trivial. We intend to address this in future work, either by introducing a more complex generative model (see "Approximations" above), but doing this rigorously is too complex for this first paper. **5.1** We have cleared up these readability issues especially regarding $\mathbf{Q}$ and $\mathbf{H}$. **5.2** We have added some additional plots showing how μ (especially changes in μ) interact with σ. **5.3** We have decluttered Fig 4 by plotting performance every 5 epochs.