We would like to thank the reviewers for their time and thoughtful comments.

We would first like to address some clarifications:

- Using the KF does *not* make this a model-based RL method. The KF is not modeling the dynamics of the environment, rather it is estimating (with uncertainty), the local loss landspace.
- To clarify some questions on the assumptions: the assumptions of the filter depend on the model and the reward landscape, since these are the two items that influence the gradient distribution.
- Present assumptions in terms of RL.
- What do the matrices of KF represent for RL.
- TODO: Relationship to Adam: we view our method as complementary to an optimizer, like Adam. The optimizer's job is to take the gradient at a given iteration (perhaps a history of gradients as well) and perform a parameter update that improves some objective. The KF in our experiments is used to estimate the gradient at each of these iterations, not perform the update as well. Only the mean from the KF is passed to the optimizer; the variances of the KF are used only to estimate the errors.

0.1 Complexity concerns

Several reviewers noted potential issues in scaling to larger models. Yes, diagonal or block-diagonal approximation is an option for larger models and in these cases matrix multiplication and inversion become computationally simpler. But more generally, in control tasks there is much research suggesting that (generalized) linear models are sufficiently expressive. We chose to focus on linear models because (1) is it straightforward to ensure the assumptions hold and (2) they are sufficiently expressive for the vast majority of real-world control tasks (THIS IS TOO LARGE OF A CLAIM).

0.2 Additional experimental evaluation

The reviewers also noted that our experimental results section seems a bit thin and contained only a toy problem. We focused on a simple problem, as opposed to directly running the algorithm on more complicated tasks such as the MuJoCo benchmarks, because the simple LQR setup permits straight-forward evaluation and the optimal controller is known. We also focused on a generic policy gradient method (without 2nd order derivatives or a critic) because we wanted to isolate, as much as possible, the effects of the KF estimator.

Based on reviewer comments we ran additional experiments that compare the effects of adding a critic and also perform evaluation on more complicated benchmark tasks.

* Experiments: * Compare to additional variance reduction techniques: * This is a valid rebuke, though we believe our method should operate in an orthogonal fashion to other approaches and did not incorporate variance reduction techniques with or without KF. * Optimal baseline * Linear value function baseline. * Compare to SGD baseline. * Additional tasks: * Cartpole - linear. * Walker mujoco (RBF)? * Testing on non toy problems with a model that breaks assumptions: to test our initial work we wanted to focus on a model and task that would clearly express the inner workings of this method. A larger model, and more complicated tasks make it much more difficult to analyze anything beyond empirical performance.