**To Reviewer 1:**

R1: We do agree that our approach is complex and involves "multiple approximation steps". However, we view it as a contribution rather than "weakness", because it enables multi-fidelity optimization with a highly expressive DNN model. The experiments have shown not only better optimization performance, but also much higher computational efficiency (Fig. 2).

C2: On *Park1*, the NN "finds the global optimum after one query point"— How is this significant... "

R2: Great question. First, the quality of the query point very much depends on the accuracy of the surrogate model. The first query of our NN (trained with initial points) has already been very close to the optimum point (Fig. 1b), showing its superior quality in approximating the objective (contrast to MF-MES using the same acquisition function). Second, adding one more point, our NN coincides with the objective in the optimum value (Fig. 1f), confirming its capability of approximating/estimating the objective.

C3: Details about hyper-parameter selection; no liberty to choose a heldout dataset in practice.

R3: We did not use heldout datasets. We randomly split the initial data points multiple times, each time with half for training and the remaining half for test. We optimized the hyper-parameters to minimize the average test error. We will supplement these details.

**To Reviewer 2:** Thanks for suggesting us to test high-dimensional optimization problems. We will do it. Extending our approach to other acquisition functions is not straightforward and we have included it in our future research plan. We will release our code.

C4: Compare with multi-task BO in (Perrone et al. 2018); contrast to the multi-task modeling choice

R4: Thanks for providing this great reference. We will cite and compare with it. Note that we have compared with SOTA MF-BO based on multi-task GPs (MF-MES). The chain structure of our model can flexibly capture the strong, complex correlations (*e.g.,* nonlinear, nonstationary) between successive fidelities, so as to enhance the function approximation in each (higher) fidelity. The multi-task model in (Perrone et al. 2018) views each fidelity (task) as symmetric and does not reflect the monotonicity of function accuracy/importance along with the fidelities. More important, it does not model the correlation between fidelities — given the shared bases, different fidelities are assumed to be independent. Note that MF-MES assumes a linear fidelity correlation and has been worse than our method, implying that accounting for the complex relationships of the fidelities is necessary and beneficial.

C5: Why use different kernels for competing methods? What is the difference between SE, ARD and RBF?

R5: Great question. For all the competing methods, we used the original implementations and settings that give the best results reported in their papers. The used kernels, however, have minor differences. SE is RBF multiplied by a coefficient (called "amplitude"); ARD introduces a length scale parameter for each input dimension while RBF uses one for all.

C6: Why do not the methods start with the same regret? aren't all methods using the same initial training points...

R6: The initial training sets are the same for all the methods. We started to report the regrets after the first query. These methods learn different models to approximate the objective, and use different procedures to calculate the acquisition function and select the query points. As a result, their regrets are unlikely to be identical, even at the beginning.

**To Reviewer 3:** Regarding the initial regrets, please see R6.

C7: Lack of ablation study

R7: We followed the paper of SOTA MF-BO (MF-MES) to perform comprehensive experiments in both synthetic and real datasets. We followed the DNN-BO by Snoek et. al. 2015 to use another BO to select the hyper-parameters, and to evaluate the computational efficiency. Hence, we argue that our experiments are enough. However, we agree that more studies will be helpful and will add more.

C8: The work is not novel because the multi-fidelity model (Meng & Karniadakis, 2019) was proposed and the acquisition function and calculation are borrowed from (Wang et. al. 2017).

R8: First, we agree that our MF model structure resembles (Meng & Karniadakis, 2019) in some degree (though the latter still has quite a few critically distinct components, *e.g.,* differential NN and linear correlations). We would love to cite and discuss about it. However, the two models are totally different in the goal (integrating PDEs and multi-fidelity data for prediction *vs.* optimizing blackbox functions with multi-fidelity queries), training objective (summation of several heuristic square losses *vs.* a principled variational evidence bound), and inference (pure point estimations *vs.* Bayesian inference with uncertainty quantification/propagation). It is **baseless** and **unfair** to claim that our work has NO novelty just because of some lightly related and essentially different work.

Second and more important, the critical contribution of our work is that we use quadrature and moment-matching to develop a highly efficient and tractable approach to calculate and optimize the widely used MES acquisition function for DNNs (Sec. 4). This task is very challenging and our solution is never covered by the prior work. The reviewer missed our key contribution and claimed our calculation is just borrowed from (Wang, et. al. 2017) (applicable to a single-fidelity GP only), which is **wrong** and **irrational**. Following the reviewer's reasoning, the many excellent BO works (see Sec.5 Related Work) that use MES principle should all be judged as "not novel" and "incremental".

**To Reviewer 4:**

C9: The model is not novel due to the deep GP mode in (Cutajar et. al., 2019); "the paper focuses more on a multi-fidelity modeling rather than multi-fidelity optimization framework"; "the comparison seems to be unfair since other MFBO use only GP models"

R9: We are astonished by these **baseless** and **erroneous** conclusions! First, despite some high-level similarity, our model and (Cutajar et. al. 2019) belong to distinct families: DNN and GP. The goals and inferences are totally different. Following your reasoning, deep GP should be viewed as not novel because it plagiarizes the idea of "deep" architectures in DNNs! Second, the dominant space of our paper is used to introduce our efficient and tractable approach to compute the acquisition function for multi-fidelity optimization (page 4-6). This is our major contribution (see R1&R9). The multi-fidelity modeling only takes half a page (page 3). How do you conclude our work " focuses more on a multi-fidelity modeling rather than multi-fidelity optimization framework"? Third, we compared with GP based methods, because the SOTA MFBO are all based on GPs! This is for a fair comparison. We will compare with the multi-task model suggested by Reviewer 2. However, preserving results of the GP based MFBO is necessary to ensure the fairness and comprehensiveness.