We thank the reviewers for their thoughtful comments! We are encouraged that they find the empirical comparison of our method, L-GSO, detailed and thorough, based on a mix of simple and complex well-motivated domain problems (R1, R2, R3), noted broad applicability in the scientific domain and appreciated the physics experiment (R2), clear hyperparameter choice (R1) and simplicity of the implementation, yet performing well on a range of problems (R2, R3). We agree that we might have been under-selling scalability of our method (R3) w.r.t. parameters of the surrogate model, though this was not our goal. We are glad the text is clearly written (R2, R4) with a broad literature review (R1, R2).

(R1, R2) Scalability w.r.t. simulator parameters ψ dimensionality: From empirical observations, L-GSO generally scales linearly with the dimensionality of ψ. However, in cases where $\psi \in \mathbb{R}^D$ lies on a manifold of dimension d lower than the extrinsic dimension of the space D, L-GSO requires less than D samples for producing a reasonable gradient. We note this on lines [146-149].

(R2) Scalability w.r.t. surrogate parameters θ dimensionality: As R3 noted, L-GSO scales easily with the dimensionality of θ due to using SGD on model parameters. Also, our method is quite robust to the choice of the architecture, as for all experiments simple and more complex, we have used an identical GAN architecture with 60k parameters.

(R1) Adding baselines for Submanifold Hump and NN Weight Optimization problems: All baselines have been applied to these problems, but for visual clarity, we did not show comparisons in Fig. 5, which focuses on comparing the speed of convergence w.r.t. to the num. of ψ samples. Other methods did not perform as well, but can be added to the figure. Similarly, for visual clarity, error bands were not included for all baselines in Fig. 2. We will add plots for all baselines with respective error bands with an increased number of re-runs into appendix as requested.

(R1, R3) Adding baselines and longer L-GSO training on physical problem: For the physics experiment, simulation calls are extremely costly; 120 steps took $\sim$300h of which $\sim 70\%$ was simulation time. This physical example was a proof-of-concept of applicability of our method for real physics problems. A comparison to BO is listed on line [339], but rerunning other baselines is not feasible on short timescales.

(R1, R2) Total runtime of the algorithm: For NN Weights Optimization problem one epoch takes $\sim$2m, totalling $\sim$12h for training (on 23k samples) of which nearly all time was GAN training. BO runtime was 113h, using only 4k samples before termination. Other baselines (LTS/Void/GES) were faster(<1h), b/c the toy problems were not simulator time dominated. For problems dominated by simulation time, the longer training time for L-GSO relative to LTS/Void/GES is less consequential than the number of needed simulator calls (as in the physics example).

(R1, R3) Adaptive/active learning for sampling process/other trust regions types: We agree, adaptive/active learning is a key next step. We thought it important to first establish the method, even w/o adaptive sampling schemes. We also thank R3 for pointing out the sampling region formula typo (trust region is square).

(R1, R2) Objective function R must be differentiable w.r.t. y: It is correct, however on line [61-63] we note that this assumption can be relaxed by incorporating the objective into the simulator, i.e. re-assigning $y' = R(y)$ and $R' = y'$.

(R1) Is the surrogate model initialized from scratch for every gradient step? Yes, this is discussed on line [127].

(R1, R3) Combining Bayesian Optimization with our approach: We agree, gradient information from our surrogate could be used in down-stream optimization like BO, but we have not investigated this yet. This could also improve the exploration vs. exploitation, as SGD focuses on exploitation.

(R2) Are there any theoretical convergence guarantees?: We do not have explicit theoretical guarantees. However, we empirically observe unbiased gradients in Fig. 4. We can appeal to the theory of SGD convergence which state that, subject to some conditions, it is guaranteed to converge (arXiv:1805.08114). We will discuss this in the revised paper.

(R2) How does the generative model help compute gradients?: The extra assumptions come from defining a model class (the gen model) that restricts the set of potential solutions and allows interpolation. The gen model is building a continuous approximation of the simulator distribution from the samples and thus implicitly regularizing the solution, which is not done with numerical methods. So the gen model has more information in terms of the implicit prior defined by choice of model and optimization scheme. We believe this relates to recent work such as [arXiv:1809.04542,arXiv:1811.03259] on inductive bias and generalization in deep generative models in GANs. On lines [207-215], we also briefly discuss that our findings on the effectiveness of our method in case of low intrinsic dimensionality are consistent with papers on adaptivity of deep learning models to intrinsic dimensionality. On matching / using gradients, this could indeed be helpful to incorporate them when available.

(R4) Novelty w.r.t. MetricGAN: We respectfully disagree with the reviewer, while MetricGAN (MG) is related (and can be noted in related work), it is not directly applicable to black box optimization (BBO). MG aims to find optimal parameters of the generator network while approximating a fixed black box metric (function) with a regression-based surrogate, while we aim to optimize the black box function itself, which are quite different tasks needing different algorithm considerations. In no place does MG optimize over simulator parameters to generate better observations, it fits to fixed observations. MG could potentially be adjusted for the BBO setting, but this is clearly outside of the scope of the existing MG work, and we can not be expected to undertake R&D to make MG applicable to this setting. We also disagree with the notion that empirical studies, which compare to a variety of baseline models and show favourable performance, is weak based solely on not comparing to MG which is not directly applicable without further research. Please consider reviewing the work more holistically and in the broader setting of established BBO methods, rather than only through the lens of a single additional method which is not directly applicable.