We thank the reviewers for their detailed and comprehensive review. We are glad that the reviewers have found our algorithm "elegant"(R1) and "simple"(R3), with "'solid empirical evaluations"(R1) and "highly encouraging results"(R3). However, the reviewers have raised concerns particularly regarding comparisons to existing literature (R2, R4). We address the most salient points of feedback below, and will incorporate all feedback in our paper.

General Feedback

Significance of results: In this work, we propose a simple automatic curriculum technique (VDS) in the goal-conditioned RL setting that samples goals according to the epistemic uncertainty of learned value functions. The simplicity of this method allows us to use it on a range of domains from simple ones like maze navigation to challenging ones such as multi-fingered dexterous manipulation. In an extensive evaluation on 18 domains, including **all** 13 OpenAI Gym Robotics tasks, we show that VDS is significantly better than state-of-the-art baselines such as HER on 10/18, while being on par on the remaining 8. Our results are hence not cherry-picked on domains. We are excited to see R1 and R3 both acknowledge the significance of our empirical evaluation. However, R2 and R4 do not see this as a sufficient improvement. Hence, to reiterate the significance of our results, we note that in the three hardest tasks: `HandManipulateEgg`, `HandManipulateBlock`, and `HandManipulateBlockRotateParallel`, we achieve $\approx 3\times$ the performance of HER, our strongest baseline. While, on the three easy `Maze` tasks we report no improvements. This indicates that VDS offers more value in difficult domains, without hurting performance on simpler ones.

Acknowledgement of prior work: We thank all four reviewers for their suggestions of prior work notably in intrinsic motivation RL (R1), curriculum learning (R2), active learning (R3), and uncertainty estimation (R4). Indeed, our work although simple, is connected to large bodies of prior work, which we will cite and discuss thoroughly in our paper. However, R2 believes that some papers we missed citing hurts the novelty of our work. Particularly, R2 cites 5 papers [a-e], and we describe why although important works, most would not be valid baselines for VDS. Graves et al. 2017 [a] and Matiisen et al. 2017 [b] is an overarching framework that explicitly tracks learning progress for every 'goal'. This is only possible to do efficiently with a small discrete goal space, and would not scale to a large continuous goal space that is used in standard goal-conditioned RL. Sukhbaatar et al. 2018 [c] is in fact included in our citations (please see Section 5.2 and [44]) and a baseline for Florensa et al. 2018, which we significantly improve upon in Fig. 4. Pong et al. 2020 [d] and Racaniere and Lampinen et al. 2020 [e] presents exciting approaches for automatic curricula with a focus on pixel-space observation and as such not directly applicable as a baseline. However, we believe both methods can be used in conjunction with VDS for better learning on image-based domains.

Algorithmic / Experimental details

Why value disagreement? (R1, R2, R4). We highlight that value disagreement assigns high probability to goals associated with high learning progress (Fig. 5.) and goals of intermediary difficulty, which are two properties that are explicitly optimized for by some previous methods. However, there are several techniques to estimate epistemic uncertainty such as using Dropout, or using Bayesian neural networks. Studying the effects of such choices and the behavior that emerges would serve as an exciting avenue for future research.

Diversity in ensemble training (R3). The ensemble is trained with random initialization and independent mini-batches. Thereafter, the method also benefits from the agent's exploration strategy or a stochastic RL policy. Even if all Q functions are initialized as the same, since the agent does not behave deterministically, members in the ensemble would see different transitions and will not be identical throughout the training.

Hard goals are down-weighted (R1, R3). Selecting a goal that is particularly difficult to reach provides little to no reward signal in a sparse-reward environment. VDS empirically down-weights these difficult goals early on in training and only begins to sample them once the policy is performative on easier goals (Fig. 5).

Definition of a challenging task (R4). By "challenging tasks" we refer to goals that are far away from the solvable region. The agent is unlikely to obtain reward signals with these challenging training goals given sparse reward. VDS assigns less probability to these challenging training goals since the agent consistently obtains low reward.

Disentangling effects of VDS (R3, R4). We thank the reviewers for their questions on clarity of baselines, which we will improve in the paper. GoalGAN is based on TRPO, while the GMM baselines use SAC. To avoid potential performance downgrade, we replicated the entire codebase of the baselines in a single framework. We note that the performance of baselines in our framework that uses the newer SAC backbone is higher than the original implementations. All our implementations will be publicly released for others to compare with and build on. Additionally, one of the baselines "RandomDDPG" incorporates HER which can bring significant benefits in sparse-reward, goal-conditioned tasks, and hence in certain domains performs better than more recent curriculum learning baselines (Fig. 4).

Number of environments (R4). The 18 tasks include 16 tasks in Fig. 3 and 2 Ant environments in Fig. 4.

We thank the reviewers for their thorough and thoughtful review. Due to constraints in space, we had to exclude discussions on interesting questions, and we will defer detailed analysis to the main paper.