Response to all reviewers: We have significantly updated the results of the paper to show that our meta-gradient method can indeed learn auxiliary questions fast enough to improve learning performance on the main task. In these experiments, as suggested, we use both the actor-critic loss and the (continuously adapting) meta-learned auxiliary question losses to update the state representation, as is usually done in work using hand-crafted auxiliary losses. We compared our approach to a baseline agent that uses only the actor-critic loss, and to agents using both the actor-critic loss and two hand-crafted auxiliary losses introduced in UNREAL: reward-prediction, and pixel-control. Our approach performs significantly better on 7 games, and performs significantly worse on just 1 game of the 10 Atari games tried so far (see Figure 1). This suggests that discovering GVF questions using meta-gradients is a promising approach for auxiliary task discovery that improves learning. As in UNREAL, both our agents and the corresponding baselines were trained for 200M frames, and hyperparameters such as the weighting of the auxiliary loss were tuned per game.

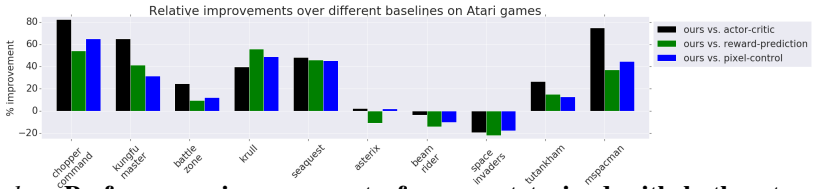

Figure 1: *New results:* **Performance improvement of an agent trained with both actor-critic and (continuously adapting) meta-gradient-learned auxiliary question losses, compared to three baseline agents after 200M frames of training.** The baselines are the standard actor-critic agent, and agents trained with both the actor-critic loss and one of two hand-crafted auxiliary losses: reward-prediction and pixel-control. Our approach performs significantly better on 7 Atari games while performing significantly worse on 1 Atari game of the 10 tried so far.

(**R1**): We extended our representation-learning experiments to include more complex Atari games; as in the experiments in the submitted paper, here only the meta-learned auxiliary question loss is used to update the state representation. Figure 2 shows performance on 5 such games. As in the results in the submitted paper, the meta-gradient approach can discover questions in the form of GVFs that drive the learning of a state representation that is good enough to support the main-task learning, even on these more challenging RL domains. Note that this is not the case for the baseline auxiliary tasks (pixel-control and reward-prediction).

(**R2** & **R3**): We agree that there is a strong relationship between meta-gradient and MAML/L2L. However, we highlight several important differences: a) meta-gradient RL updates meta-parameters of the update function (in this case the auxiliary tasks), whereas MAML (typically) updates the initial parameters, b) meta-gradient RL may be applied to adapt and improve performance during the lifetime of a single agent, whereas MAML learns across many lifetimes (impractical in many applications), c) Our paper demonstrates this idea on far more complex RL tasks (needing millions of training steps as opposed to hundreds). We will revise our claim by stating these additional details and add the necessary citations to clarify this point. Thanks. We do prefer to continue using the questions/answer terminology to highlight the conceptual similarity to Temporal-Difference network, but agree that we should more carefully introduce and motivate both the terminology and the question framing in the revision.

(**R2**): We thank the reviewer for highlighting interesting related work: PowerPlay is indeed similarly motivated as our submission, but uses greedy search over all possible task descriptions for discovering tasks instead of meta-gradients. Curiosity-based RL uses hand-designed intrinsic rewards. We will clarify the relation to our method in our revision and emphasize that our primary contribution is the discovery of GVFs as auxiliary questions through meta-gradients (as pointed out by **R1**).

(**R3**): The algorithm does perform several forward passes per meta-update; we will clarify this in the revision. We will also clarify that performance decreases if the unroll length is increased too much, and we will perform new experiments to confirm the hypothesis that variance increases with unroll length. Thanks for pointing this out.

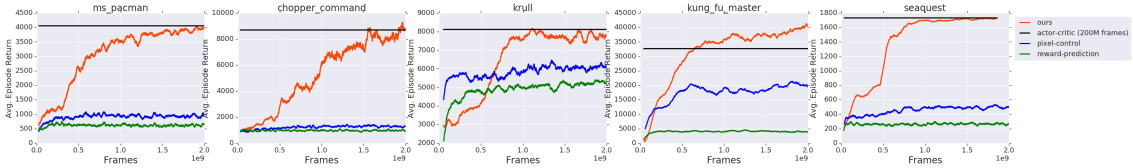

Figure 2: *New results*: **Additional more complex Atari tasks to test representation learning driven only by the meta-learned auxiliary questions, compared to hand-crafted baselines.** The plot includes a horizontal black line that shows the episode return achieved by a standard actor-critic agent after 200M training frames along with the learning curves of the meta-gradient discovery approach, and hand-crafted reward-prediction and pixel-control auxiliary losses. We demonstrate that our approach of discovering auxiliary questions in the form of GVFs can drive representation learning which eventually allows the agent to reach a comparable level of performance to the standard actor-critic agent. These new results show that the approach scales to complex RL domains.