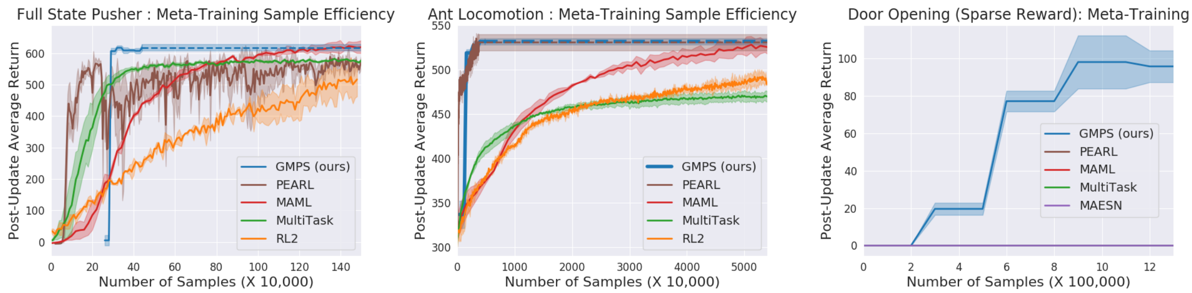

Figure 1: Evaluation with added comparison to PEARL, showing meta-training curves on full state pushing (left), ant locomotion (middle), and sparse reward door opening (right). PEARL is more sample-efficient and achieves similar asymptotic performance on dense reward tasks. However, **GMPS significantly outperforms PEARL on sparse reward tasks.**

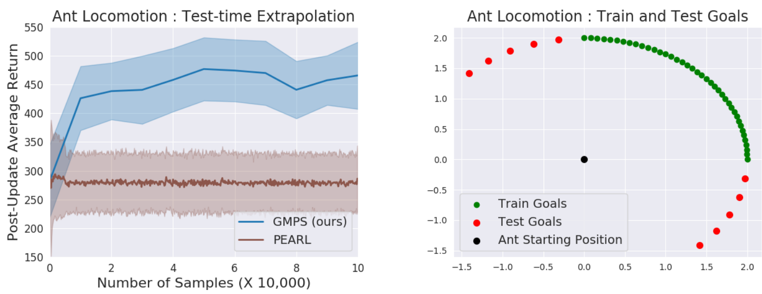

Figure 2: Test-time extrapolation for dense reward ant locomotion Left: Performance comparison. Right: Train and test goals. **GMPS is better able to learn out-of-distribution tasks.**

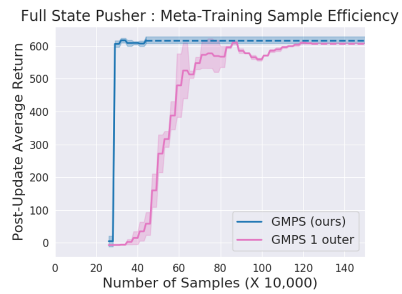

Figure 3: Ablation for number of consecutive outer updates, as requested by reviewer 3. Using 500 imitation steps (blue) results in significantly greater sample efficiency than using only one (pink).

We thank the reviewers for their positive and constructive feedback.

The primary concern from Reviewer 1 was the comparison to PEARL (Rakelly et al.). We have now added this comparison. We performed this comparison for meta-training sample efficiency (Fig 1 left, middle), meta-training on sparse reward tasks (Fig 1 right), and extrapolation to out of distribution tasks at test time (Fig 2).

With dense rewards, we observe PEARL and GMPS require about the same number of samples to meta-train, and achieve similar performance (Fig 1 left, middle). On the sparse reward door task, we train PEARL in a setting that matches ours: PEARL is trained with sparse rewards passed to the encoder during meta-training and meta-testing and shaped rewards are used for meta-training the actor and critic weights. In this setting (Fig 1 right), PEARL is unable to learn a strategy that explores sufficiently for the encoder to detect the sparse rewards. On out of distribution tasks (Fig 2 right), GMPS performs substantially better than PEARL (Fig 2 left). This is because GMPS uses policy gradient to adapt, which enables it to continuously make progress on tasks even if they are out of distribution.

Reviewer 1. See PEARL comparisons above. We will also add a discussion of PEARL to the related work.

Reviewer 2. We will add a discussion of the algorithm's limitations and hyperparameter tuning to the revised paper. One limitation for GMPS and for all current meta-RL methods is the difficulty in meta-training across qualitatively distinct task families. This is due to two factors, the lack of benchmarks containing many different task families and because learning only a few disjoint behaviors is challenging for a single neural network. The most important hyperparameters to tune are the number of imitation steps per sampling step and the dimension of the bias transformation variable. We will discuss alpha and beta in the revised version. Alpha is learned and beta is fixed at 0.01 across all experiments.

Reviewer 3. PEARL uses a different inner update rule than our algorithm (amortized inference instead of policy gradient), and we show how this leads to worse extrapolation for PEARL (Fig. 2)

As requested, we added an ablation to show the effect of consecutive gradient steps between each outer iteration (Fig. 3), where we compare taking 500 imitation steps per sampling step (as in the paper) to taking only one imitation step per sampling step (GMPS 1 outer). This results in poorer sample efficiency, since we no longer perform off-policy gradient updates.

For ant locomotion, in the sparse setting, reward is provided only when the ant is within a certain distance of the goal. Hence even if the ant performs the right behavior, its obtained return will be less than in the dense case since it receives sparse reward for much of its trajectory.