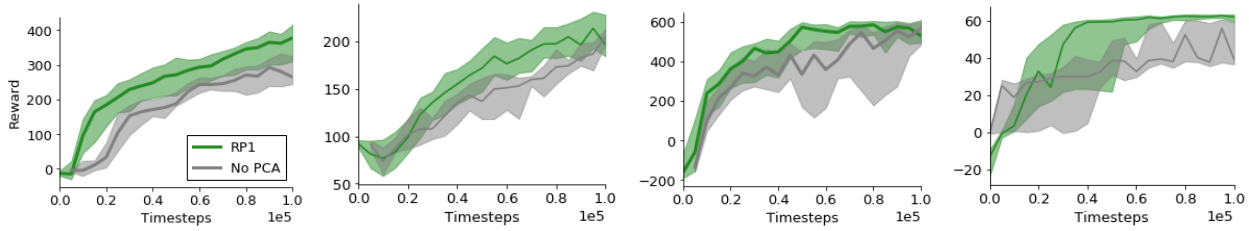
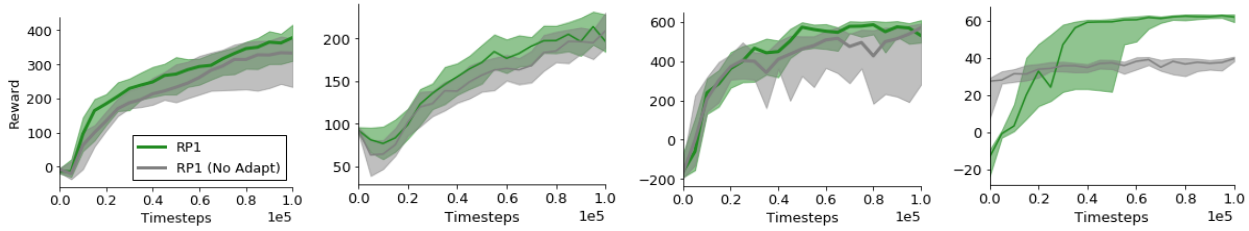
A. Additional Ablation Studies

In this section we show the full results from the ablation study in Table 3.



(a) **HalfCheetah** (b) **Ant** (c) **Hopper** (d) **Swimmer**

Figure 4. Ablation study where we consider removing the early stopping mechanism. All results show the median performance across ten seeds.



(a) **HalfCheetah** (b) **Ant** (c) **Hopper** (d) **Swimmer**

Figure 5. Ablation study where we consider removing the adaptive mechanism. All results show the median performance across ten seeds.

B. Implementation Details

In terms of approach, we follow ME-TRPO (Kurutach et al., 2018) with some adjustments. Instead of using TRPO loss (Schulman et al., 2015), we leverage the first-order approximation loss in PPO (Schulman et al., 2017). We do not apply parameter noise, and we modify policy training slightly by ensuring at least 10 updates are performed before termination is considered, which we find helps to improve convergence. We do not apply GAE nor the overall training approach in (Schulman et al., 2017), as this introduced instabilities. We instead found that generating large batch sizes gave better and more consistent performance, which corroborates the findings in (Ilyas et al., 2018) with respect to true policy gradients.

We augment each environment with an additional state which contains velocity information. This is also done in the 'FixedSwimmer' environment in (Wang et al., 2019), and allows us to infer the reward from the states directly. It must also be noted that in the original 'rllab' (Duan et al., 2016) environments used in (Kurutach et al., 2018), one of the observable states was the velocity state used to calculate rewards, and we therefore mirror this in our OpenAI Gym (Brockman et al., 2016) implementation; we do not anticipate there to be any problem integrating reward prediction with our framework. Furthermore, we provide this state in both the model-free and model-based benchmarks to ensure there is no advantage, and do not notice any noticeable improvement in the model-free setting when this is provided; we hypothesize that some close proxy to the true velocity state already exists in the original state-space. We remove contact information from all environments, and instead of 'Swimmer-v2' we use the aforementioned 'FixedSwimmer', since this can be solved by our policy in a model-free regime. We remove early stopping from Hopper since we found it was necessary for convergence, but left the early stopping in for Ant since it was possible to train performant policies. We train the policy for 100 time steps in HalfCheetah and Ant, and for 200 time steps in Swimmer and Hopper. In experiments without the early stopping mechanism, data collection defaults to 3,000 timesteps per iteration. Full hyperparameter values can be found in Table 5.

We use the following approach to normalize G to produce $\hat{G}$; for convenience, we write $\hat{G}_{\phi_t}(\theta_{t+1})$ as $\hat{G}_t$.

$$\hat{G}_t = \frac{G_t - \frac{1}{5}\sum_{\tau=t-5}^{t-1}(G_\tau)}{l_{\text{val}}} \tag{9}$$

where l_{val} is the final model validation loss from the iteration $t-1$.

Attention should be drawn to the α parameter used to determine early stopping in Algorithm 2. For the tasks we test on, we choose to fix this to 0.0005, and therefore do not tune it to be task specific. We found that at this setting of α, the early stopping mechanism generally collects significantly fewer than the default 3,000 samples (which acts as an upper bound in RP1), but can still expand the batch size collected to the full amount where appropriate (i.e., under policies that provide non-homogenous trajectories).

All experiments were run on a virtual machine with an 8-core Intel Skylake microarchitecture CPU, 32GB of RAM, and an NVIDIA Tesla P100 GPU.

Table 4. Hyperparameters used in the Policy

Hyperparameter Name	Value
Optimizer	Adam
Learning Rate	3e-4
Loss	PPO
Discount Factor	0.99
Batch Size	50,000
Epochs per Batch	10
ϵ-clip	0.2
Default Action σ	0.5
Hidden Layer Size	32
Number of Hidden Layers	2
Activation Function	ReLU

Table 5. Hyperparameters used in the World Model

Hyperparameter Name	Value
Optimizer	Adam
Learning Rate	1e-3
Train/Validation Split	2:1
Number of Models	5
Batch Size	1,024
Hidden Layer Size	1,024
Number of Hidden Layers	2
Activation Function	ReLU
Early Stopping α in PCA	0.0005