A. Detailed Analysis of Algorithm 1 (Proof of Theorem 4.4)

In this section, we provide detailed proof for Theorem 4.4.

Consider a Markov Chain $p : \mathcal{S} \to \Delta(\mathcal{S})$ over horizon H. Denote d_p as the induced state distribution under p and ρ_p as the induced state trajectory distribution under p.

Lemma A.1. *Consider two Markov Chains $p_i : \mathcal{S} \to \Delta(\mathcal{S})$ with $i \in \{1, 2\}$. If*

$$\mathbb{E}_{s\sim d_{p_1}} \left[\|p_1(\cdot|s) - p_2(\cdot|s)\|\right] \le \delta,$$

then for trajectory distributions, we have:

$$\|\rho_{p_1} - \rho_{p_2}\| \le O(H\delta).$$

The above lemma implies that if the Markov chain p_2 can predict the Markov chain p_1 *under the state distribution induced by* p_1, then we can guarantee that the state-wise trajectory distributions from p_1 and p_2 are close as well.

Proof. Denote $\rho_{p,s_1,\dots s_h}$ as the trajectory distribution induced by p conditioned on the first h many states are equal to $s_1, \dots, s_h$. Denote $p(\cdot|s_0)$ as the initial state distribution for s_1 with s_0 being a faked state. By definition, we have:

$$\begin{aligned}
&\|\rho_{p_1} - \rho_{p_2}\| = \sum_{\tau} |\rho_{p_1}(\tau) - \rho_{p_2}(\tau)| \\
&= \sum_{s_1,\dots s_H} \left| \prod_{h=1}^{H} p_1(s_h|s_{h-1}) - \prod_{h=1}^{H} p_2(s_h|s_{h-1}) \right| \\
&= \sum_{s_1,\dots s_H} \left| \prod_{h=1}^{H} p_1(s_h|s_{h-1}) - p_1(s_1|s_0)\prod_{h=2}^{H} p_2(s_h|s_{h-1}) + p_1(s_1|s_0)\prod_{h=2}^{H} p_2(s_h|s_{h-1}) - \prod_{h=1}^{H} p_2(s_h|s_{h-1}) \right| \\
&\le \sum_{s_1} p_1(s_1|s_0) \sum_{s_2,\dots,s_H} \left| \prod_{h=2}^{H} p_1(s_h|s_{h-1}) - \prod_{h=2}^{H} p_2(s_h|s_{h-1}) \right| + \sum_{s_1} |p_1(s_1|s_0) - p_2(s_1|s_0)| \left(\sum_{s_2,\dots,s_H} \prod_{h=2}^{H} p_2(s_h|s_{h-1}) \right) \\
&= \mathbb{E}_{s_1\sim p_1} \left[\|\rho_{p_1,s_1} - \rho_{p_2,s_1}\|\right] + \|p_1(\cdot|s_0) - p_2(\cdot|s_0)\| \\
&\le \mathbb{E}_{s_1,s_2\sim p_1} \left[\|\rho_{p_1,s_1,s_2} - \rho_{p_2,s_1,s_2}\|\right] + \|p_1(\cdot|s_0) - p_2(\cdot|s_0)\| + \mathbb{E}_{s_1\sim d_{\pi_1;1}}[\|p_1(\cdot|s_1) - p_2(\cdot|s_1)\|].
\end{aligned}$$

Recursively applying the same operation on $\|\rho_{p_1,s_1} - \rho_{p_2,s_1}\|$ to time step H, we have:

$$\|\rho_{p_1} - \rho_{p_2}\| \le \sum_{h=1}^{H} \mathbb{E}_{s_h\sim d_{p_1;h}} \left[\|p_1(\cdot|s_h) - p_2(\cdot|s_h)\|\right] \le H\delta,$$

where we recall $d_\pi = \sum_{h=1}^{H} d_{\pi;h}/H$ by definition. Extension to continuous state can be achieved by simply replaying summation by integration.

□

The next lemma shows that by leveraging the no-regret property of FTL, we can learn a locally accurate model.

Lemma A.2 (Local Accuracy of the Learned Model)**.** *Denote the sequence of models learned in Alg. 1 as $\{\hat{f}_1, \dots, \hat{f}_T\}$, there exists a model $\hat{f} \in \{\hat{f}_1, \dots, \hat{f}_T\}$ such that:*

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}} \left[\mathbb{E}_{a\sim U(\mathcal{A}^{(t)})} \left[D_{KL}(f^{(t)}(\cdot|s,a), \hat{f}(\cdot|s,a)) \right] \right] \le O(1/T),$$

where $\pi_f(s) \triangleq \operatorname{argmin}_{a\in\mathcal{A}^{(t)}} \|f(\cdot|s,a) - f^{(s)}(\cdot|s,\pi^{(s)}(s))$ for all $s \in \mathcal{S}$ for any f.

Proof. Denote loss function $\ell_e(f)$ as:

$$\ell_e(f) \triangleq \mathbb{E}_{s\sim d_{\pi_e^{(t)}}, a\sim U(\mathcal{A}^{(t)})}\left[\mathbb{E}_{s'\sim f^{(t)}_{s,a}}\left[-\log(f(s'|s,a))\right]\right].$$

Since Alg. 1 is equivalent to running FTL on the sequence of strongly convex loss functions $\{\ell_e(f)\}_{e=1}^T$, we have (Shalev-Shwartz et al., 2012):

$$\sum_{e=1}^{T}\ell_e(\hat{f}_e) \le \min_{f\in\mathcal{F}}\sum_{e=1}^{T}\ell_e(f) + O\left(\log T\right).$$

Add $\sum_{e=1}^T \mathbb{E}_{s\sim d_{\pi_e^{(t)}}, a\sim U(\mathcal{A}^{(t)})}[\mathbb{E}_{s'\sim f^{(t)}_{s,a}} \log f^{(t)}(s'|s,a)]$ on both sides of the above inequality and using the definition of KL divergence, we have:

$$\begin{aligned}
&\sum_{e=1}^{T}\mathbb{E}_{s\sim d_{\pi_e^{(t)}}, a\sim U(\mathcal{A}^{(t)})}\left[D_{KL}(f^{(t)}(\cdot|s,a), \hat{f}_e(\cdot|s,a))\right] \\
&\le \min_{f\in\mathcal{F}}\sum_{e=1}^{T}\mathbb{E}_{s\sim d_{\pi_e^{(t)}}, a\sim U(\mathcal{A}^{(t)})}\left[D_{KL}(f^{(t)}(\cdot|s,a), f(\cdot|s,a))\right] + O(\log(T)) = O(\log(T)),
\end{aligned}$$

where the last equality comes from the realizability assumption $f^{(t)} \in \mathcal{F}$.

Using the fact that the minimum among a sequence is less than the average of the sequence, we arrive:

$$\min_{\hat{f}\in\{\hat{f}_e\}_{e=1}^T}\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}, a\sim U(\mathcal{A}^{(t)})}\left[D_{KL}(f^{(t)}(\cdot|s,a), \hat{f}(\cdot|s,a))\right] \le \tilde{O}(1/T).$$

□

The above lemma indicates that as $T\to\infty$, we will learn a model $\hat{f}$ which is close to the target true model $f^{(t)}$ under the state distribution induced by $\pi_{\hat{f}}$. But it does not state the difference between the behavior generated by $\pi_{\hat{f}}$ at the target domain and the behavior generated by $\pi^{(s)}$ at the source domain. The next lemma uses the definition $\pi_{\hat{f}}$ to show that when executing $\pi_{\hat{f}}$ in the target domain $\mathcal{M}^{(t)}$, $\pi_{\hat{f}}$ can actually generates behavior that is similar to the behavior generated by $\pi^{(s)}$ in the source domain $\mathcal{M}^{(s)}$.

Lemma A.3 (The Behavior of $\pi_{\hat{f}}$)**.** *Denote $d_{\pi_{\hat{f}}}$ as the state distribution induced by $\pi_{\hat{f}}$ induced at $\mathcal{M}^{(t)}$ (target domain).*

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\pi_{\hat{f}}(s)) - f^{(s)}(\cdot|s,\pi^{(s)}(s))\| \le O(|\mathcal{A}^{(t)}|/\sqrt{T}) + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\left[\epsilon_{s,\pi^{(s)}(s)}\right],$$

where we recall the definition of ϵ in Assumption 4.1.

Proof. Consider the Markov chain that is defined with respect to $f^{(t)}$ and $\pi_{\hat{f}}$, i.e., $f^{(t)}(s'|s,\pi_{\hat{f}}(s))$. Denote the state distri-

bution induced by $f^{(t)}(s'|s, \pi_{\hat{f}}(s))$ at the target domain as $d_{\pi_{\hat{f}}}$. Let us bound $\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s, \pi_{\hat{f}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\|$.

$$
\begin{aligned}
&\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s, \pi_{\hat{f}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\| \\
&\leq \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s, \pi_{\hat{f}}(s)) - \hat{f}(\cdot|s, \pi_{\hat{f}}(s))\| + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|\hat{f}(\cdot|s, \pi_{\hat{f}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\| \\
&\leq \sqrt{\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}} D_{KL}(f^{(t)}(\cdot|s, \pi_{\hat{f}}(s)), \hat{f}(\cdot|s, \pi_{\hat{f}}(s)))} + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|\hat{f}(\cdot|s, \pi_{\hat{f}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\| \\
&\leq O(|\mathcal{A}^{(t)}|/\sqrt{T}) + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|\hat{f}(\cdot|s, \pi_{\hat{f}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\| \\
&\leq O(|\mathcal{A}^{(t)}|/\sqrt{T}) + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|\hat{f}(\cdot|s, \pi_{f^{(t)}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\| \\
&\leq O(|\mathcal{A}^{(t)}|/\sqrt{T}) + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|\hat{f}(\cdot|s, \pi_{f^{(t)}}(s)) - f^{(t)}(\cdot|s, \pi_{f^{(t)}}(s))\| \\
&\quad + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s, \pi_{f^{(t)}}(s)) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\| \\
&\leq O(|\mathcal{A}^{(t)}|/\sqrt{T}) + \sqrt{\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}} D_{KL}(f^{(t)}(\cdot|s, \pi_{f^{(t)}}(s)), \hat{f}(\cdot|s, \pi_{f^{(t)}}(s)))} + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\left[\epsilon_{s,\pi^{(s)}(s)}\right] \\
&= O(|\mathcal{A}^{(t)}|/\sqrt{T}) + \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\left[\epsilon_{s,\pi^{(s)}(s)}\right],
\end{aligned}
$$

where the first inequality uses triangle inequality, the second inequality uses Pinsker's inequality, the third inequality uses the local accuracy of the learned model $\hat{f}$ (Lemma A.2), the fourth inequality uses the definition that $\pi_{\hat{f}}$, and the fifth inequality uses triangle inequality again, and the sixth inequality uses Pinsker's inequality again with the definition of adaptive ability together with the definition of $\pi_{f^{(t)}}(s) \triangleq \text{argmin}_{a\sim\mathcal{A}^{(t)}}\|f^{(t)}(\cdot|s, a) - f^{(s)}(\cdot|s, \pi^{(s)}(s))\|$. □

Proof of Theorem 4.4. Use Lemma A.1 and Lemma A.3 and consider $f^{(s)}(\cdot|s, \pi^{(s)}(s))$ as a Markov chain, we have that:

$$\|\rho^t_{\pi_{\hat{f}}} - \rho^s_{\pi_s}\| \leq O(H|\mathcal{A}^{(t)}|/\sqrt{T}) + H\epsilon,$$

where we denote $\epsilon := \mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\left[\epsilon_{s,\pi^{(s)}(s)}\right]$ □

This concludes the proof of Theorem 4.4.

A.1. Extension to Continuous Action Space (Proof of Corollary 4.6)

For simplicity, we consider $\mathcal{A}^{(t)} = [0, 1]$.[2] We consider Lipschitz continuous transition dynamics with and only with respect to actions, i.e.,

$$\|f(\cdot|s, a) - f(\cdot|s, a')\| \leq L|a - a'|, \tag{4}$$

where L is a Lipschitz constant. We emphasize here that we only assume Lipschitz continuity with respect to action in $\mathcal{M}^{(t)}$. Hence this is a much weaker assumption than the common Lipschitz continuity assumption used in RL community, which requires Lipschitz continuity in both action and state spaces. We also assume that our function class $\mathcal{F}$ only contains function approximators that are Lipschitz continuous with respect to action a (e.g., feedforward fully connected ReLu network is Lipschitz continuous).

Proof of Corollary 4.6. For analysis purpose, let us discretize the action space into bins with size $\delta \in (0, 1)$. Denote the discrete action set $\bar{\mathcal{A}}^{(t)} = \{0.5\delta, 1.5\delta, 2.5\delta, \ldots, 1 - 0.5\delta\}$ (here we assume $1/\delta = \mathbb{N}^+$). Here $|\bar{\mathcal{A}}^{(t)}| = 1/\delta$.

Now consider the following quantity:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s, \hat{a}) - \hat{f}(\cdot|s, \hat{a})\|,$$

for any $\hat{a}$. Without loss of generality, we assume $\hat{a} \in [0, \delta]$. Via Pinsker's inequality and Lemma A.2, we have:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,1])}\|f^{(t)}(\cdot|s, a) - \hat{f}(\cdot|s, a)\| \leq O(1/\sqrt{T}),$$

[2] We can always normalize action to $[0, 1]$.

which implies that:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta])}\|f^{(t)}(\cdot|s,a)-\hat{f}(\cdot|s,a)\|\leq O(1/(\delta\sqrt{T})).$$

We proceed as follows:

$$\begin{aligned}
&\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta])}\|f^{(t)}(\cdot|s,a)-\hat{f}(\cdot|s,a)\|\\
&=\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta])}\|f^{(t)}(\cdot|s,\hat{a}+a-\hat{a})-\hat{f}(\cdot|s,\hat{a}+a-\hat{a})\|\\
&\geq\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta])}\left(\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L|a-\hat{a}|\right)\\
&=\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\mathbb{E}_{a\sim U([0,\delta])}|a-\hat{a}|\\
&\geq\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\mathbb{E}_{a\sim U([0,\delta])}\delta\\
&=\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\delta,
\end{aligned}$$

where the first inequality uses the fact that $\hat{a}\in[0,\delta]$ and the Lipschitz conditions on both $f^{(t)}$ and $\hat{f}\in\mathcal{F}$, the second inequality uses the fact that $|a-\hat{a}|\leq\delta$ for any $a\in[0,\delta]$ as $\hat{a}\in[0,\delta]$.

Hence, we have:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|\leq 2L\delta+O(1/(\delta\sqrt{T}))=O(T^{-1/4}),\forall\hat{a}\in\mathcal{A}^{(t)},$$

where in the last step we set $\delta=\Theta(T^{-1/4})$.

Now, we can simply repeat the process we have for proving Lemma A.3, we will have:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\pi_{\hat{f}}(s))-f^{(s)}(\cdot|s,\pi^{(s)}(s))\|\leq O(T^{-1/4})+\epsilon.$$

Combine with Lemma A.1, we prove the corollary for continuous action setting. □

For general n-dim action space, our bound will scale in the order $O(\sqrt{n}T^{-1/(2n+2)})+\epsilon$. The proof of the n-dim result is similar to the proof of the 1-d case and is included below for completeness:

Proof for n-dim Action Space. For n-dimensional action space, we have:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta]^n)}\|f^{(t)}(\cdot|s,a)-\hat{f}(\cdot|s,a)\|\leq O(1/(\delta^n\sqrt{T})).$$

Using the Lipschitz property:

$$\begin{aligned}
&\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta]^n)}\|f^{(t)}(\cdot|s,a)-\hat{f}(\cdot|s,a)\|\\
&=\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta]^n)}\|f^{(t)}(\cdot|s,\hat{a}+a-\hat{a})-\hat{f}(\cdot|s,\hat{a}+a-\hat{a})\|\\
&\geq\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\mathbb{E}_{a\sim U([0,\delta]^n)}\left(\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\|a-\hat{a}\|\right)\\
&=\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\mathbb{E}_{a\sim U([0,\delta]^n)}\|a-\hat{a}\|\\
&\geq\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\mathbb{E}_{a\sim U([0,\delta]^n)}\sqrt{n}\delta\\
&=\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|-2L\sqrt{n}\delta,
\end{aligned}$$

Combining with above leads to

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\hat{a})-\hat{f}(\cdot|s,\hat{a})\|\leq 2\sqrt{n}L\delta+O(1/(\delta^n\sqrt{T}))=O(\sqrt{n}T^{-1/(2n+2)})$$

where in the last step we set $\delta=\Theta((\frac{T}{n})^{\frac{-1}{2(n+1)}})$. Finally we have:

$$\mathbb{E}_{s\sim d_{\pi_{\hat{f}}}}\|f^{(t)}(\cdot|s,\pi_{\hat{f}}(s))-f^{(s)}(\cdot|s,\pi^{(s)}(s))\|\leq O(\sqrt{n}T^{-1/(2n+2)})+\epsilon.$$

□

B. Detailed description of our experiments

B.1. Environment descriptions.

For each of the four environments (HalfCheetah, Ant, Reacher) we tested on, we include a detailed numerical description of the environments in table 2.

B.2. Descriptions of the perturbations

B.2.1. CHANGING GRAVITY.

We change the gravity in the whole Mujoco locomotion task by a max range from 50% to 200% of the normal gravity. The normal gravity is set to 0 on x and y axis and -9.81 on z axis.

B.2.2. CHANGING MASS.

We change the mass of the agent in the Mujoco locomotion task by a max range from 50% to 200% of its original mass. Note that most agents are composed of links with independent masses, so besides changing the agent's mass as a whole, we also design experiments that change the mass of each individual links of the agent respectively.

B.2.3. CHANGING PLANE ORIENTATION.

In the Reacher task, we tilt the platform of the Reacher so that it forms an angle of 45 degree of the original plane.

B.2.4. CHANGING ARM LENGTH.

In the Reacher task, we change the first link of the Reacher arm (which is composed of two parts) into one tenth of its original length.

B.2.5. CHANGING FRICTION.

We change the frictional coefficient in the environment on an uniform scale. We found if friction is the only change in the target domain, then the adaptation task is relatively simple, so we incorporate it as one of the changes in the Multi-dimensional perturbation task.

B.2.6. MOTOR NOISE.

This task tries to mimic the motor malfunction or inaccuracy in the real world. After the agent outputs an action, we add on a noise from a normal distribution with mean 0 and a fixed standard deviation from 0.2 to 1.

B.2.7. MULTIPLE DOFS OF CHANGES.

In the 3DOF task, we set the gravity to 0.8, mass to 1.2 and friction coefficient to 0.9. In the 15DOF task, we uniformly sample a coefficient from 0.9 to 1.1 for each of the following configuration: gravity, friction coefficient and the mass of each joint of the agent. We record these changes and apply for all comparing methods.

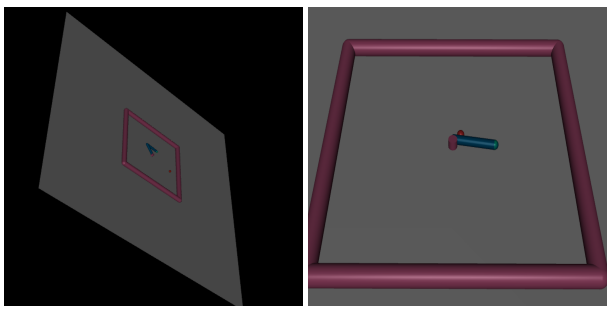


Figure 6. Visual illustration of modified Reacher environment. **left:** 45 degrees of tilted plane. **right:** Reacher with 10% length of its first arm.

Environment name	Full state space size	Model agnostic state size[3]	Action space size	Reward function
HalfCheetah	18	1	6	forward reward - 0.1 $\times$ control cost
Ant	29	2	8	forward reward - 0.5 $\times$ control cost - 0.0005 $\times$ contact cost + survive reward
Reacher	16	4	2	forward distance - control cost

Table 2. Description of the OpenAI gym environments. Note that to enforce safety of the agent in the target environment, we make a modification on HalfCheetah environment such that the episode will be terminated when the cheetah falls off.

B.3. Hyperparameters

	source	PADA-DM	PADA-DM with target	Christiano et al., 2016	Zhu et al., 2018	PPO
# timesteps	2e6	8e4 (5e4,8e4,12e4,15e4)	8e4 (5e4,8e4,1.2e5,1.5e5)	2e5 (1e5,2e5,4e5)	2e6	2e6
learning rate (with linear decay)	7e-4	5e-3	5e-3	5e-3	7e-4	7e-4
soft update rate			every 3000 timesteps (3000,5000,10000)			
explore rate ϵ		0.01 (0.01,0.02)	0.01 (0.01,0.02)			
reward tradeoff λ					0.5	

Table 3. Final hyperparameters we use for our methods and baselines in the experiments. The values in the brackets are the value we considered.

C. Implementation details

C.1. Pretraining of the source dynamics model

In section 5.1, one assumption we make is that we have a pretrained model $\hat{f}^{(s)}$ that well approximates the source dynamics $f^{(s)}$. In practice, we pretrain the model $\hat{f}^{(s)}$ with the (s, a, s') triplets generated during the training of $\pi^{(s)}$. The model $\hat{f}^{(s)}$ is a two-layer neural network with hidden size 128 and ReLU nonlinearity, trained with MSE loss since we assume deterministic transitions. Using these existing samples has two major advantages: first is that we don't need further interaction with the source environment and second is that the trained model $\hat{f}^{(s)}$ especially well approximates the transitions around the actions taken by $\pi^{(s)}$, which is important to our algorithm.

Remark that if we already have the ground truth source dynamics $f^{(s)}$, which is a mild assumption while using a simulator, we can also directly use $f^{(s)}$ to replace $\hat{f}^{(s)}$. During our experiments, we observe that whether using $f^{(s)}$ or $\hat{f}^{(s)}$ won't affect the performance of our method.

C.2. Cross Entropy Method

Here we provide a pseudocode of the Cross Entropy Method that we used in our method, as in Alg. 3. In our implementation, we use $T = 10$ iterations, $N = 200$ actions sampled per iteration, $K = 40$ elites (elite ratio 0.4) and $\sigma_0 = 0.5$. We use the output of target policy as the initial mean, and when we don't have the target policy, we use $\pi^{(s)}(s)$ as the initial mean. To avoid invalid actions, for each action a_i we sample, we clip the action if certain dimension is out of the bound of $\mathcal{A}^{(t)}$ (usually [-1,1] on each dimension).

[3]This means the number of states that won't be passed as inputs to models or policies, e.g., the current coordinate or location of the agent.

Algorithm 3 Cross Entropy Method

Require: Initial mean μ_0, initial standard deviation σ_0, action space $\mathcal{A}^{(t)}$, current model δ_θ, current state s, number of iterations T, sample size N, elite size K.

$\Sigma_0 \leftarrow I_{|\mathcal{A}^{(t)}|}(\sigma_0^2)$
for t = 1, ... T **do**
  Sample $\{a_i\}_{i=1}^N \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$
  $\{a_i\}_{i=1}^N \leftarrow clip(a_i, \min(\mathcal{A}^{(t)}), \max(\mathcal{A}^{(t)}))$
  Sort $\{a_i\}$ with descending $\|\delta_\theta(s, a_i)\|_2^2$ and pick the first K actions $\{a_j\}_{j=1}^K$
  $\mu_t \leftarrow \frac{1}{K}\Sigma_{j=1}^K a_j$
  $\Sigma_t \leftarrow \frac{1}{K}\Sigma_{j=1}^K (a_j - \mu_t)(a_j - \mu_t)^T$
end for
Output: μ_T

D. Supplemental experiments

D.1. Accuracy of Deviation Model

We evaluate the performance of the deviation model by comparing its output with the actual deviation (i.e., $\Delta^{\pi^{(s)}}$) in the target and source environment states. We compare the performance between linear and nonlinear deviation models. We include linear models as they are convenient if we want to use optimal control algorithms. The nonlinear model is the same model we use for our PADA-DM algorithm, which has two 128-neuron layers with ReLU (Nair & Hinton, 2010) nonlinearity. Both of the deviation models are tested on the same initial state distribution after training on 80k samples. In 7(left), we plot the output of the deviation model against the ground truth deviation. We test on 50 trajectories and each data point refers to the average L2 state distance along one trajectory. In 7(right), we plot the ground truth deviation, and the outputs of the linear and nonlinear deviation models over time, on 10 test trajectories.

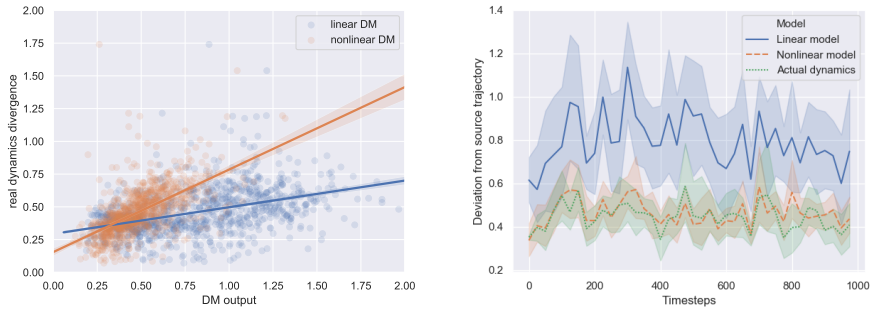


Figure 7. Comparing the performances of the linear and nonlinear deviation models. **left:** This plot depicts the correlation between the predicted deviation and the actual deviation. The nonlinear deviation model is more accurate since its slope is closer to 1. **right:** This plot shows the predicted and actual deviation over the course of 10 trajectories. Here again, the nonlinear model (orange) curve lies very close to the actual deviation curve (green).

D.2. Long-term learning curves

In this section we show a more comprehensive long-term learning curve in Fig. 8. Each task here corresponds to the task in Fig. 2. Note that here again the x-axis is in natural logarithm scale.

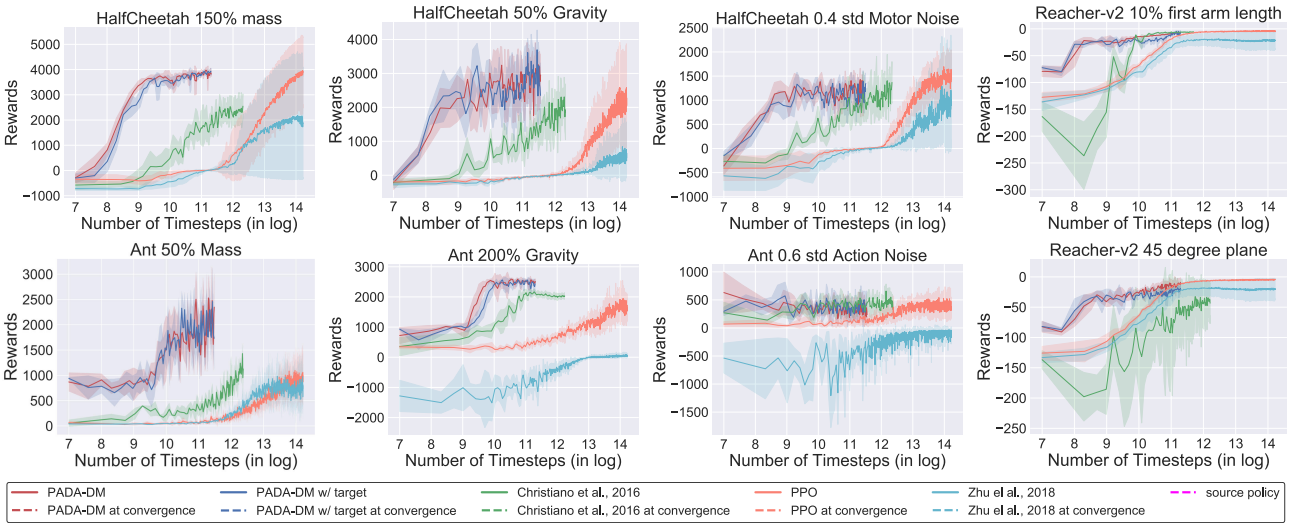


Figure 8. We plot the learning curves across 5 random seeds of our adaptation method (using PADA-DM and PADA-DM with target), and the baseline methods on a number of tasks including perturbation of the mass, gravity, motor noise for the locomotion environments (HalfCheetah and Ant), and the plane angle and arm lengths for the navigation environment (Reacher). The title of each plot corresponds to the perturbation in the target domain, e.g., HalfCheetah Mass 150% means in mass of the agent in the target domain is 150% of that in the source domain. The shaded area denotes 1 standard deviation range from the mean.

D.3. Comparing using true reward

To further verify the efficiency of our choice of reward (the deviations between two environments), we conduct an additional experiments where we use the ground truth reward for CEM. We fix all the hyperparameters and train an additional model to learn to ground truth reward. To ensure the fairness of the comparison, when we use the ground truth reward, we keep doing 1 step look-ahead during CEM. The results verify that the deviation serves as a better reward for conducting policy adaptation, where the ground truth reward leads to a local minimum that results in suboptimal performance.

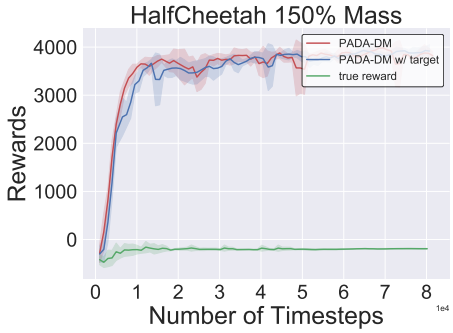


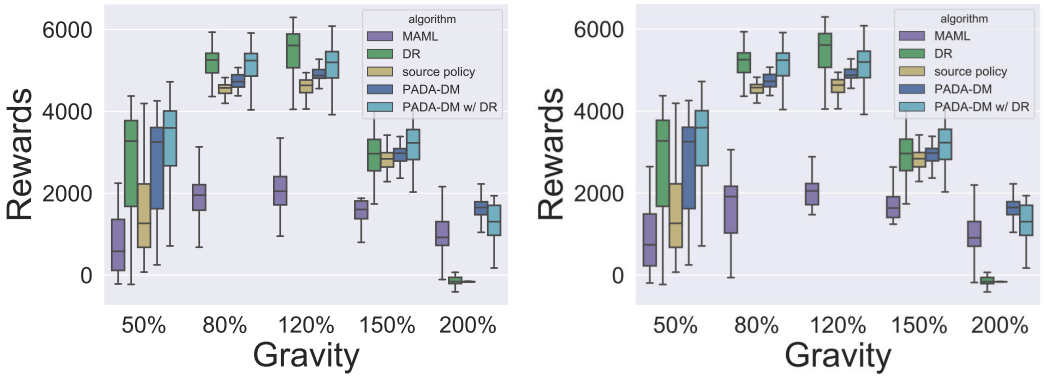
Figure 9. Comparing learning ground truth reward, learning deviations quickly adapts the policy in the target domain, where using ground truth reward may not necessarily leads to optimal performance in the target domain.

D.4. Additional Experiments for Meta-Learning MAML

In this section we conduct an additional experiment to Section 6.3. In section 6.3, we use 80k samples of adaptation for our method and MAML to conduct a fair comparison. However, using so many samples for adaptation contradicts MAML's claim of few-shot adaptation and we also observe that MAMLs test performance does not improve too much as we change the sample size in adaptation phase. Thus here we report the additional experiments to support this claim: we adopt the same experiment setting in Section 6.3, and this time we use 20k samples for MAML during adaptation. The test performance

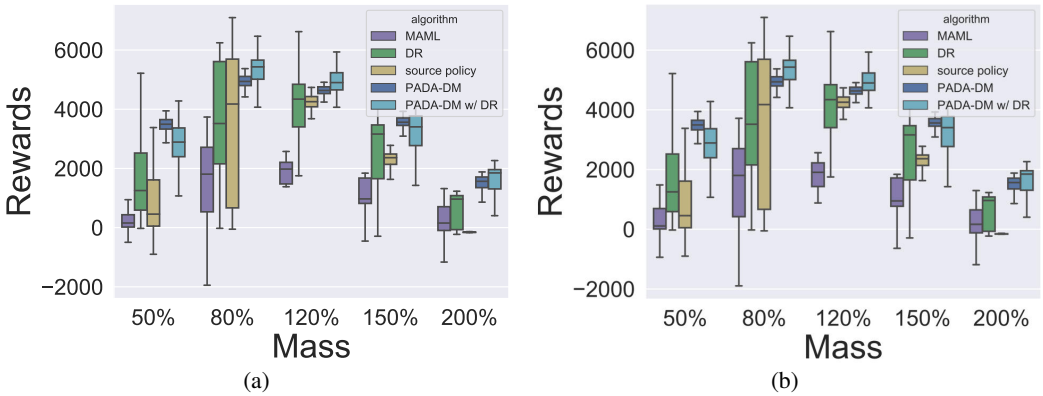
is recorded in Fig. 10(a) and Fig. 11(a). Comparing with the original performance (Fig. 10(b) and Fig. 11(b)), the test performance of MAML does not change that much as the number of adaptation samples decreases and our approach still outperforms MAML consistently.

In addition, we record the mean and the standard deviation of the test performance of each method to deliver a more direct comparison in Table 4 and Table 5. As we can see, our approach outperforms other baselines most of the time. When the perturbation is small (e.g., the 120% columns in both tables), DR also delivers very strong performances. However when perturbation is large (e.g., 200% columns in both tables), DR fails to adapt completely, which indicates that DR has troubles to adapt to out-of-distribution new environments.



(a) (b)

Figure 10. Ablation experiments using domain randomization and meta-learning. (a) MAML with 20k adaptation samples. (b) MAML with 80k adaptation samples. The boxplots show the median of the data while more statistics such as mean and standard deviation are shown in the following tables.



(a) (b)

Figure 11. Ablation experiments using domain randomization and meta-learning. (a) MAML with 20k adaptation samples. (b) MAML with 80k adaptation samples.

	Gravity Perturbation				
	50%	80%	120%	150%	200%
Source policy	1549.74 (1090.73)	4444.86 (637.77)	4592.18 (201.30)	2543.55 (916.16)	-156.51 (25.93)
Domain Randomization	2282.74 (1563.70)	4838.87 (1134.98)	**5236.46 (1179.85)**	2896.03 (554.00)	-43.97 (423.47)
PADA-DM	2694.78 (1166.88)	4739.32 (279.06)	4889.02 164.38)	2998.32 (266.75)	**1531.23 (400.96)**
PADA-DM w/ DR	**3230.29 (1280.54)**	**5036.59 (657.98)**	4934.04 (720.34)	**3200.73 (521.50)**	1431.53 (496.11)
MAML (20k)	854.24 (692.50)	1810.51 (663.72)	1895.86 (650.76)	1575.06 (653.09)	831.07 (717.78)
MAML (80k)	876.37 (711.98)	1778.67 (669.86)	1894.28 (644.55)	1568.60 (646.79)	823.13 (721.15)

Table 4. Mean and standard deviation (in the brackets) of the episodic rewards of each method in the target environment with perturbed gravity across 100 trajectories of 5 random seeds (500 trajectories in total).

	Mass Perturbation				
	50%	80%	120%	150%	200%
Source policy	921.13 (1192.43)	3343.05 (2317.32)	4166.10 (494.94)	2045.26 (665.30)	-149.92 (27.28)
Domain Randomization	1665.05 (1357.31)	3823.45 (1944.70)	3932.86 (1791.18)	2635.72 (1105.15)	944.50 (1134.32)
PADA-DM	**3271.52 (752.62)**	4914.67 (379.73)	4584.95 375.51)	**3557.25 (183.46)**	1398.88 (500.09)
PADA-DM w/ DR	2673.87 (1009.75)	**5348.98 (556.19)**	**4854.30 (591.50)**	3276.70 (874.54)	**1616.75 (490.53)**
MAML (20k)	854.24 (692.50)	1810.51 (663.72)	1895.86 (650.76)	1575.06 (653.09)	831.07 (717.78)
MAML (80k)	876.37 (711.98)	1778.67 (669.86)	1894.28 (644.55)	1568.60 (646.79)	823.13 (721.15)

Table 5. Mean and standard deviation (in the brackets) of the episodic rewards of each method in the target environment with perturbed mass across 100 trajectories of 5 random seeds (500 trajectories in total).