A Proofs

A.1 Proof of Proposition 1

We start by recalling an important Lemma of [Achiam *et al.*, 2017].

Lemma 1. *For any function $f : \mathcal{S} \to \mathbb{R}$, policy π and $\delta_f(s, a, s') = r(s, a, s') + \gamma f(s') - f(s)$:*

$$J_P^\pi = \mathbb{E}_{s\sim\rho_0}\left[f(s)\right] + \frac{1}{1-\gamma}\mathbb{E}_{\substack{s\sim d_P^\pi(\cdot),\\ a\sim\pi(\cdot|s),\\ s'\sim P(\cdot|s,a)}}\left[\delta_f(s,a,s')\right]. \tag{7}$$

Then, we propose this general Lemma that serves as a basis for our Proposition 1.

Lemma 2. *For any function $f : \mathcal{S} \to \mathbb{R}$, let:*

$$L_f^{\pi,P_t,P_s} = \mathbb{E}_{\substack{s\sim d_{P_s}^\pi(\cdot),\\ a\sim\pi(\cdot|s)}}\left[\mathbb{E}_{s'\sim P_t(\cdot|s,a)}\left[\delta_f(s,a,s')\right] - \mathbb{E}_{s'\sim P_t(\cdot|s,a)}\left[\delta_f(s,a,s')\right]\right] \tag{8}$$

$$\epsilon_f^{P_t} = \max_{s\in\mathcal{S}}\left|\mathbb{E}_{a\sim\pi,s'\sim P_t}\left[\delta_f(s,a,s')\right]\right|. \tag{9}$$

The following bound holds:

$$J_{P_t}^\pi \geq J_{P_s}^\pi + \frac{1}{1-\gamma}\left(L_f^{\pi,P_t,P_s} - 2\epsilon_f^{P_t} D_{TV}\big(d_{P_s}^\pi, d_{P_t}^\pi\big)\right). \tag{10}$$

Proof. According to Lemma 1:

$$J_{P_\mathrm{t}}^\pi - J_{P_\mathrm{s}}^\pi = \frac{1}{1-\gamma}\left(\mathbb{E}_{\substack{s\sim d_{P_\mathrm{t}}^\pi(\cdot),\\ a\sim\pi(\cdot|s)\\ s'\sim P_\mathrm{t}(\cdot|s,a)}}\left[\delta_f(s,a,s')\right] - \mathbb{E}_{\substack{s\sim d_{P_\mathrm{s}}^\pi(\cdot),\\ a\sim\pi(\cdot|s),\\ s'\sim P_\mathrm{s}(\cdot|s,a)}}\left[\delta_f(s,a,s')\right]\right). \tag{11}$$

The first term can be written, with $\bar{\delta}_f^{P_\mathrm{t}}(s) = \mathbb{E}_{\substack{a\sim\pi(\cdot|s)\\ s'\sim P_\mathrm{t}(\cdot|s,a)}}\left[\delta_f(s,a,s')\right]$:

$$\mathbb{E}_{\substack{s\sim d_{P_\mathrm{t}}^\pi(\cdot)\\ a\sim\pi(\cdot|s)\\ s'\sim P_\mathrm{t}(\cdot|s,a)}}\left[\delta_f(s,a,s')\right] = \langle d_{P_\mathrm{t}}^\pi\,,\,\bar{\delta}_f^{P_\mathrm{t}}\rangle \tag{12}$$

$$= \langle d_{P_\mathrm{s}}^\pi\,,\,\bar{\delta}_f^{P_\mathrm{t}}\rangle + \langle d_{P_\mathrm{t}}^\pi - d_{P_\mathrm{s}}^\pi\,,\,\bar{\delta}_f^{P_\mathrm{t}}\rangle. \tag{13}$$

We apply Holder's inequality with $p = 1$ and $q = \infty$, and get:

$$\mathbb{E}_{\substack{s\sim d_{P_\mathrm{t}}^\pi(\cdot)\\ a\sim\pi(\cdot|s)\\ s'\sim P_\mathrm{t}(\cdot|s,a)}}\left[\delta_f(s,a,s')\right] \geq \langle d_{P_\mathrm{s}}^\pi, \bar{\delta}_f^{P_\mathrm{t}}\rangle - 2\epsilon_f^{P_\mathrm{t}} D_{\mathrm{TV}}\left(d_{P_\mathrm{s}}^\pi, d_{P_\mathrm{t}}^\pi\right), \tag{14}$$

with $\epsilon_f^{P_\mathrm{t}} = \max_{s\in\mathcal{S}}\left|\mathbb{E}_{a\sim\pi,s'\sim P_\mathrm{t}}\left[\delta_f(s,a,s')\right]\right|$. The Total Variation distance comes from the 1-norm resulting from the application of Holder's inequality. We obtain:

$$\begin{aligned}(1-\gamma)\left(J_{P_\mathrm{t}}^\pi - J_{P_\mathrm{s}}^\pi\right) \geq\ & \mathbb{E}_{\substack{s\sim d_{P_\mathrm{s}}^\pi(\cdot),\\ a\sim\pi(\cdot|s)}}\left[\mathbb{E}_{s'\sim P_\mathrm{t}(\cdot|s,a)}\left[\delta_f(s,a,s')\right] - \mathbb{E}_{s'\sim P_\mathrm{t}(\cdot|s,a)}\left[\delta_f(s,a,s')\right]\right] \\ & - 2\epsilon_f^{P_\mathrm{t}} D_{\mathrm{TV}}\big(d_{P_\mathrm{s}}^\pi, d_{P_\mathrm{t}}^\pi\big).\end{aligned} \tag{15}$$

$\square$

To conclude the proof of Proposition 1, we choose f as the null function $f : \mathcal{S} \to 0$ and upper bound the remaining term by reusing Holder's inequality:

$$\mathbb{E}_{\substack{s\sim d^{\pi}_{P_{\mathrm{s}}}(\cdot),\\ a\sim\pi(\cdot|s),\\ s'\sim P_{\mathrm{t}}(\cdot|s,a)}}[r(s,a,s')] - \mathbb{E}_{\substack{s\sim d^{\pi}_{P_{\mathrm{s}}}(\cdot),\\ a\sim\pi(\cdot|s),\\ s'\sim P_{\mathrm{t}}(\cdot|s,a)}}[r(s,a,s')] \geq -2R_{\max}\mathbb{E}_{\substack{s\sim d^{\pi}_{P_{\mathrm{s}}}(\cdot),\\ a\sim\pi(\cdot|s)}}\left[D_{\mathrm{TV}}(P_{\mathrm{s}}(\cdot|s,a), P_{\mathrm{t}}(\cdot|s,a))\right] \tag{16}$$

$$\geq -2R_{\max}D^{\pi}_{\mathrm{TV}}(P_{\mathrm{s}}, P_{\mathrm{t}}). \tag{17}$$

Other choice for f The function f could also be chosen as the value function associated with the source system $V^{\pi}_{P_{\mathrm{s}}}$. In which case, we get with Lemma 2:

$$J^{\pi}_{P_{\mathrm{t}}} \geq J^{\pi}_{P_{\mathrm{s}}} + \frac{1}{1-\gamma}\left(\mathbb{E}_{\substack{s\sim d^{\pi}_{P_{\mathrm{s}}}(\cdot),\\ a\sim\pi(\cdot|s),\\ s'\sim P_{\mathrm{s}}(\cdot|s,a)}}\left[\frac{P_{\mathrm{t}}(s'|s,a)}{P_{\mathrm{s}}(s'|s,a)}\left(r(s,a,s') + \gamma V^{\pi}_{P_{\mathrm{s}}}(s') - V^{\pi}_{P_{\mathrm{s}}}(s)\right)\right] - 2\epsilon_f^{P_{\mathrm{t}}} D_{\mathrm{TV}}\left(d^{\pi}_{P_{\mathrm{s}}}, d^{\pi}_{P_{\mathrm{t}}}\right)\right). \tag{18}$$

It also introduces an additional term than Proposition 2. Here, it is an importance sampling term between the transition probabilities that is difficulty optimized. In principle, it could be estimated with the classifiers proposed by DARC but would introduce a new level of complexity to the algorithm. Hence, we preferred focusing on proposing the simpler Proposition 2.

A.2 Proof of Proposition 2

We present here the proof of our simpler Proposition 2 that we restate below, as well as its extensions using different discrepancy measures.

Proposition 3. *Let $\nu^{\pi}_{P}(s,a,s')$ the state-action-state visitation distribution, where $\nu^{\pi}_{P}(s,a,s') = (1-\gamma)\mathbb{E}_{\rho_0,\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\left(s_t = s, a_t = a, s_{t+1} = s'\right)\right]$. For any policy π and any transition probabilities P_t and P_s, the following holds:*

$$J^{\pi}_{P_t} \geq J^{\pi}_{P_s} - \frac{2R_{max}}{1-\gamma}D_{TV}\left(\nu^{\pi}_{P_s}, \nu^{\pi}_{P_t}\right), \tag{19}$$

with D_{TV} the Total Variation distance.

Proof. It is known that $J^{\pi}_{P} = \frac{1}{1-\gamma}\mathbb{E}_{(s,a,s')\sim\nu^{\pi}_{P}}\left[r(s,a,s')\right]$. Now:

$$\begin{aligned}
\left|J^{\pi}_{P_{\mathrm{t}}} - J^{\pi}_{P_{\mathrm{s}}}\right| &= \frac{1}{1-\gamma}\left|\left(\mathbb{E}_{(s,a,s')\sim\nu^{\pi}_{P_{\mathrm{t}}}}\left[r(s,a,s')\right] - \mathbb{E}_{(s,a,s')\sim\nu^{\pi}_{P_{\mathrm{s}}}}\left[r(s,a,s')\right]\right)\right| &(20)\\
&= \frac{1}{1-\gamma}\left|\int_{s,a,s'}\left(r(s,a,s')\nu^{\pi}_{P_{\mathrm{t}}}(s,a,s') - r(s,a,s')\nu^{\pi}_{P_{\mathrm{s}}}(s,a,s')\right)\mathrm{d}\{sas'\}\right| &(21)\\
&= \frac{1}{1-\gamma}\left|\int_{s,a,s'} r(s,a,s')\left(\nu^{\pi}_{P_{\mathrm{t}}}(s,a,s') - \nu^{\pi}_{P_{\mathrm{s}}}(s,a,s')\right)\mathrm{d}\{sas'\}\right| &(22)\\
&\leq \frac{2R_{\max}}{1-\gamma}D_{\mathrm{TV}}\left(\nu^{\pi}_{P_{\mathrm{s}}}, \nu^{\pi}_{P_{\mathrm{t}}}\right). &(23)
\end{aligned}$$

The last inequality is an application of Holder's inequality, by setting p to ∞ and q to 1.

$\square$

An application of Pinsker inequality [Csiszar and Körner, 1981] provides a similar upper bound with the Kullback Leibleir divergence.

Corollary 1. *Let $\nu_P^\pi(s,a,s')$ the state-action-state visitation distribution, where $\nu_P^\pi(s,a,s') = (1-\gamma)\mathbb{E}_{\rho_0,\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left(s_t = s, a_t = a, s_{t+1} = s'\right)\right]$. For any policy π and any transition probabilities P_t and P_s such that $\nu_{P_s}^\pi$ is absolutely continuous with respect to $\nu_{P_t}^\pi$, the following holds:*

$$J_{P_t}^\pi \geq J_{P_s}^\pi - \frac{\sqrt{2}R_{max}}{1-\gamma}\sqrt{D_{KL}\left(\nu_{P_s}^\pi \parallel \nu_{P_t}^\pi\right)}, \tag{24}$$

with D_{KL} the Kullback Leibleir divergence.

A lower bound with the Jensen Shannon divergence can also be found thanks to [Corander *et al.*, 2021, Proposition 3.2].

Corollary 2. *We assume the state-action space. Let $\nu_P^\pi(s,a,s')$ the state-action-state visitation distribution, where $\nu_P^\pi(s,a,s') = (1-\gamma)\mathbb{E}_{\rho_0,\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left(s_t = s, a_t = a, s_{t+1} = s'\right)\right]$. We assume the support of $\nu_{P_s}^\pi$ and $\nu_{P_s}^\pi$ is $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Then, for any policy π and any transition probabilities P_t and P_s, the following holds:*

$$J_{P_t}^\pi \geq J_{P_s}^\pi - \frac{4R_{max}}{(1-\gamma)}\sqrt{D_{JS}\left(\nu_{P_s}^\pi \parallel \nu_{P_t}^\pi\right)}, \tag{25}$$

with D_{JS} the Jensen Shannon divergence.

B Algorithms Details

In this section, we further present the different algorithms used in this paper.

B.1 Domain Adaptation with Rewards from Classifiers (DARC)

We introduce our main baseline Domain Adaptation with Rewards from Classifiers (DARC), which is the prominent state-of-the-art algorithm that tackles the off-dynamics task by modifying the RL objective.

DARC takes a variational perspective to this problem. Given a trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$, the target distribution $p(\tau)$ over trajectories is defined as the one inducing trajectories that maximize the exponentiated rewards in the target environment:

$$p(\tau) = \rho(s_0)\left(\prod_t P_{\text{t}}(s_{t+1}|s_t,a_t)\right)\exp\left(\sum_t r(s_t,a_t,s_{t+1})\right). \tag{26}$$

Let the agent's distributions over trajectories in the source environment $q^{\pi_\theta}(\tau)$ be:

$$q^{\pi_\theta}(\tau) = \rho(s_0)\left(\prod_t P_{\text{s}}(s_{t+1}|s_t,a_t)\right)\pi_\theta(a_t|s_t). \tag{27}$$

DARC minimizes the reversed KL-divergence between $q^{\pi_\theta}(\tau)$ and $p(\tau)$, which results in the following objective expression:

$$-D_{\text{KL}}(q^{\pi_\theta}(\tau) \parallel p(\tau)) = \mathbb{E}_{\tau\sim q^{\pi_\theta}(\cdot)}\left[\sum_{t=1}^{T} r(s_t,a_t,s_{t+1}) + \mathcal{H}\left(\pi_\theta(\cdot|s_t)\right) + \Delta r(s_t,a_t,s_{t+1})\right], \tag{28}$$

with $\Delta r(s_t,a_t,s_{t+1}) = \log P_{\text{t}}(s_{t+1}|s_t,a_t) - \log P_{\text{s}}(s_{t+1}|s_t,a_t)$ and $\mathcal{H}(\cdot)$ the entropy.

The additional reward term incentivizes the agent to select transitions from the source that are similar to the target environment. Since the transition probabilities are unknown, DARC uses a pair of binary classifiers to infer whether transitions come from the source or target environment. These classifiers are then used to create a proxy equivalent to Δr.

B.2 Generative Adversarial Imitation Learning Applied for Transition Distributions

Generative Adversarial Imitation Learning (GAIL) [Ho and Ermon, 2016] is a state-of-the-art Imitation Learning algorithm. Its goal is to recover an expert policy π_e by minimizing the Jensen-Shanon divergence between the state-action visitation distributions of the expert and the learning policy. It has been proved that it is able to handle transition visitation distributions in [Desai *et al.*, 2020] as follows. To comply with our previous notations, π_e is now denoted as π_{θ_k} (fixed).

The authors define the general objective to solve by introducing a convex cost function regularizer $\psi : \mathbb{R}^{S\times A\times S} \rightarrow \mathbb{R}$ and its convex conjugate ψ^*:

$$\min_{\theta \in \Theta} \quad \psi^*(\nu_{P_s}^{\pi_\theta} - \nu_{P_t}^{\pi_{\theta_k}}). \tag{29}$$

Following Equation 13 of [Ho and Ermon, 2016] which defines ψ_{GAIL}, the authors establish the following equivalence:

$$\psi^*_{\text{GAIL}}(\nu_{P_s}^{\pi_\theta} - \nu_{P_t}^{\pi_{\theta_k}}) = \sup_{D\in(0,1)^{S\times A\times S}} \mathbb{E}_{(s,a,s')\sim\nu_{P_s}^{\pi_\theta}} [\log(D(s,a,s'))] + \mathbb{E}_{(s,a,s')\sim\nu_{P_t}^{\pi_{\theta_k}}} [\log(1-D(s,a,s'))] \tag{30}$$

where $D : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \rightarrow (0,1)$ is a classifier. Finally, it is demonstrated this specific convex cost function induces the following objective:

$$\min_{\theta\in\Theta} \psi^*_{\text{GAIL}}(\nu_{P_s}^{\pi_\theta} - \nu_{P_t}^{\pi_{\theta_k}}) = \min_{\theta\in\Theta} D_{\text{JS}}(\nu_{P_s}^{\pi_\theta} \parallel \nu_{P_t}^{\pi_{\theta_k}}). \tag{31}$$

In practice, the classifier D is trained to distinguish between samples $(s,a,s') \in (\mathcal{S}\times\mathcal{A}\times\mathcal{S})$ from $\nu_{P_s}^{\pi_\theta}$ and $\nu_{P_t}^{\pi_{\theta_k}}$. The reward used for optimizing the RL agent is given by $r_{\text{imit}} = -\log(D(s,a,s'))$.

B.3 Conservative Q-Learning (CQL)

In the offline setting, agents aim to learn a good policy from a fixed data set of M transitions $\mathcal{D} = \{(s_i, a_i, s_{i+1}\}_{i=0}^{M}$ that was collected with an unknown behavioral policy π_β, which is here π_{θ_k}. Offline RL algorithms have demonstrated impressive results when the data set is gathered with a sufficiently good policy and possesses enough transitions, often outperforming the behavioral policy.

Conservative Q-Learning (CQL) [Kumar *et al.*, 2020] is a state-of-the-art offline RL algorithm. It modifies the learning procedure of the Q-functions to favor transitions appearing in the data set. At iteration k, the Q-values are updated as follows at step j:

$$\min_{\omega \in \Omega} \quad \beta\, \mathbb{E}_{s\sim\mathcal{D}} \left[\left(\log \sum_{a\in\mathcal{A}} \exp\left(Q_\omega^{\pi_{\theta_j}}(s,a)\right) - \mathbb{E}_{a\sim\pi_{\theta_k}(\cdot|s)} \left[Q_\omega^{\pi_{\theta_j}}(s,a)\right]\right)\right] + \mathcal{E}\left(Q_\omega^{\pi_{\theta_j}}\right), \tag{32}$$

where $\mathcal{E}(Q)$ represents the traditional Bellman loss associated with the Q-functions. The regularization, controlled by the hyper-parameter β, penalizes the Q-values associated with state-action pairs not appearing in the data set.

C Experimental Details

In this section, in addition to the values of the hyperparameters necessary to replicate our experiments, we provide further details of the experimental protocol and training. In this section, considering the possible high variance of RL_s, the standard deviation is multiplied by a factor of 0.3. The original variance can be found in Table 2.

C.1 Environment Details

In all the considered environments, one property is modified in the target environment.

Gravity Pendulum Gravity is increased to 14 instead of 10. Since the pendulum requires more time to reach the objective, we also increase the length of each episode to 500 time-steps in the target environment, while keeping the original length of 200 time-steps in the source system.

Broken Joint or Leg environments In these environments, the considered robot - either HalfCheetah or Ant - is crippled in the target domain, where the effect of one or two joints is removed. In practice, this means that it sets one or two dimensions of the action to 0. These environments were extracted from the open source code of [Eysenbach *et al.*, 2020].

Heavy Cheetah The total mass of the HalfCheetah MuJoCo robot is increased from 14 to 20.

Friction Cheetah The friction coefficient of the HalfCheetah MuJoCo robot's feet is increased from 0.4 to 1.

Low Fidelity Minitaur The original Minitaur environment uses a linear torque-current linear relation for the actuator model. It has been improved in [Tan *et al.*, 2018] by introducing non-linearities into this relation where they managed to close the Sim-to-Real gap for a real Minitaur environment. In practice, the Minitaur environment can be found in the PyBullet library [Coumans and Bai, 2016 2021]. The high fidelity is registered as MinitaurBulletEnv-v0. The low fidelity environment can be recovered by calling MinitaurBulletEnv-v0 and by setting the argument `accurate motor model enabled` to False and `pd control enabled` to True.

C.2 Learning Curves

We report in Figure 2 the learning curves of the different agents mentioned in this paper. For clarity purposes, we keep all baselines fixed except for our agent and DARC, our main competitor. Here, FOOD uses the regularization with d_P^π for Gravity Pendulum and ν_P^π for the other environments as GAIL proved to be more stable when FOOD used PPO.

C.3 Global Hyper-parameters

Our experiments are based on the A2C and PPO implementations proposed by the open-source code [Kostrikov, 2018]. We also found that it may be profitable to add a TanH function at the end of the network's policy for the PPO agent to increase the performance of RL_s. We have selected their hyper-parameters according to the source [Raffin, 2020] and included them in Table 3.

Table 3: Chosen hyper-parameters for both A2C and PPO. The PPO hyper-parameters were fixed for the other environments.

Hyperparameters	**A2C**	**PPO**
num-processes	8	8
num-steps	200	1000
lr	$2.5 * 10^{-4}$	$3.0 * 10^{-4}$
γ	0.99	0.99
use-gae	True	True
gae-lambda	0.9	0.95
entropy-coef	0.01	0.001
value-loss-coef	0.4	0.5
use-linear-lr-decay	True	True
ppo-epoch	N/A	5
num-mini-batch	N/A	32
clip-param	N/A	0.1
TanH Squash	False	True

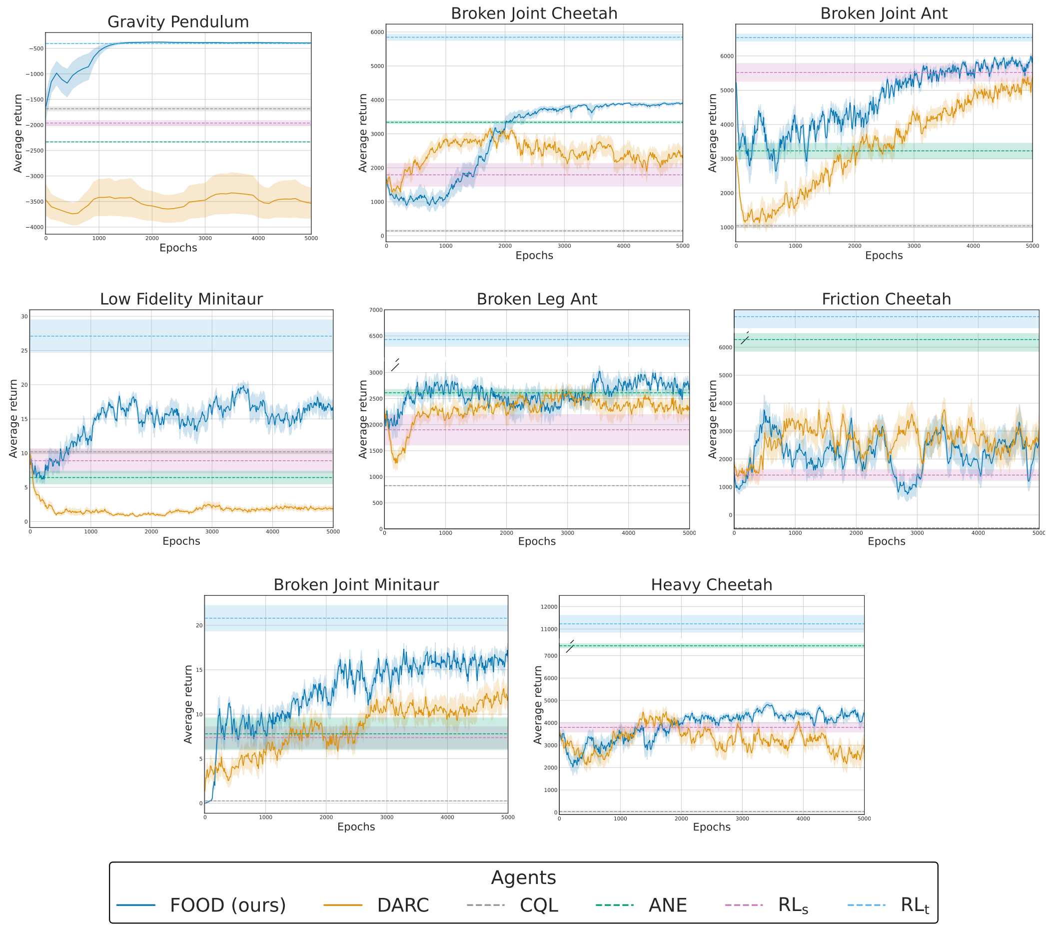

Figure 2: Learning curves of FOOD and DARC for all the proposed environments.

The Minitaur environments As proposed by the PyBullet library [Coumans and Bai, 2016 2021], γ is set to 0.995 for the Minitaur environments. Besides, unlike the Gym and Mujoco environments, they do not use a Tanh squashing function in their policy and the `num-processes` hyper-parameter is set to 1.

Algorithms optimization To allow a fair comparison between the different agents, FOOD, DARC, and ANE use the same underlying agent to optimize their objective. It is A2C for Gravity Pendulum and PPO for the others.

Discriminators training Both FOOD and DARC incorporate classifiers in their objective. At each epoch, 1000 data points are sampled from both source and target transition data sets. The classifiers are then trained with batch sizes of 128 for Pendulum and 256 for the MuJoCo environments. They share the same network structure: a 2 hidden layer MLP with 64 (for Pendulum) or 256 (for MuJoCo) units and ReLU activations. We did not find that the size of the networks play an important role in the results.

C.4 FOOD Hyper-parameters Sensitivity Analysis

This subsection investigates the impact of our main hyper-parameter α, which regulates the strength of regularization that defines a threshold between maximizing the rewards of the source MDP and staying close to the target trajectories. All FOOD results are summarized in Figure 3, where, similar to the previous section, FOOD uses the regularization with $d_{\mathcal{P}}^{\pi}$ in Gravity Pendulum and $\nu_{\mathcal{P}}^{\pi}$ for the other environments. Note that for the Gravity Pendulum environment, $\alpha \in \{0, 1, 5, 10\}$.

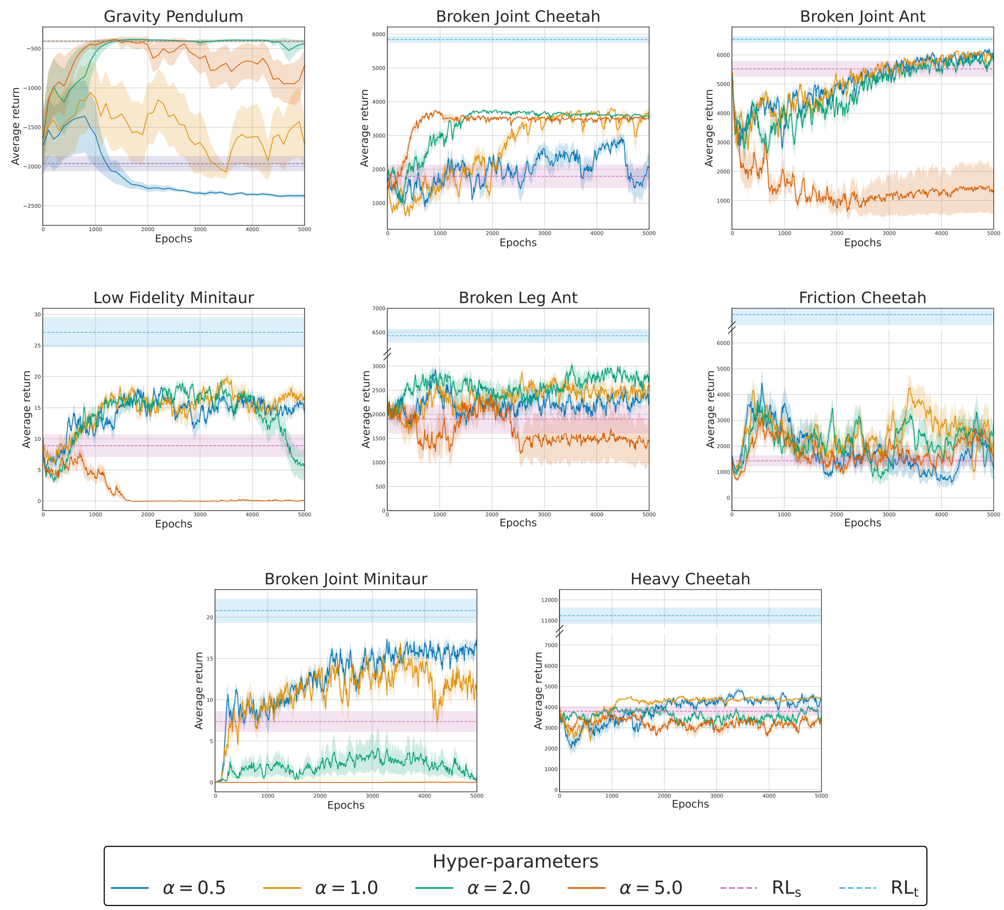

Figure 3: Complete hyperparameter sensitivity analysis for the best FOOD agent on the different off-dynamics environments.

In all the studied environments where PPO was used, we observe that unless for the low or high values of α ($\alpha \in \{0.5, 5\}$), the FOOD agent improves performance compared to RL_s. Both cases can be explained. If the value is too high, it may disrupt the gradients and prevent convergence to a good solution. As mentioned in the main paper, this phenomenon also affects the performance in the source environment, so it would be easy for practitioners to remove such bad hyper-parameters. It may also happen that the strength of the regularization is too low. In that case, FOOD has approximately the same performance as RL_s, as illustrated in Broken Joint HalfCheetah.

Hence, we recommend setting the regularization to have approximately the same weight as the average return. For this, since its advantages are normalized, we recommend using PPO and setting the α parameter to 1.

C.5 Comparison Between the Different IL Algorithms for the FOOD Agent

FOOD is a general algorithm that may use any chosen Imitation Learning algorithm. Each algorithm minimizes a certain type of divergence between state or state-action visitation distributions, as summarized in Table 1. Here, we investigate which IL is better suited for the considered environments.

We compare GAIL-$\mu^\pi_\mathcal{P}$ [Ho and Ermon, 2016], GAIL-$d^\pi_\mathcal{P}$, GAIL-$\nu^\pi_\mathcal{P}$, AIRL-$\mu^\pi_\mathcal{P}$ [Fu *et al.*, 2017], PWIL-$\mu^\pi_\mathcal{P}$ [Dadashi *et al.*, 2020], PWIL-$d^\pi_\mathcal{P}$ and PWIL-$\nu^\pi_\mathcal{P}$ in Table 4. GAIL and its extensions were extracted directly from [Kostrikov, 2018], AIRL from [Gangwani, 2021], and PWIL and its extensions were recoded from scratch.

Environment	GAIL-d	GAIL-μ	GAIL-ν	AIRL-μ	PWIL-d	PWIL-μ	PWIL-ν
Gravity Pendulum	$\mathbf{-485 \pm 54^{*}}$	-2224 ± 43	-2327 ± 14	-1926 ± 572	$\mathbf{-980 \pm 838}$	$\mathbf{-948 \pm 789}$	$\mathbf{-978 \pm 816}$
Broken Joint Cheetah	$\mathbf{3888 \pm 201}$	$\mathbf{3801 \pm 155}$	$\mathbf{3921 \pm 85^{*}}$	3617 ± 225	3537 ± 248	2999 ± 752	$\mathbf{3797 \pm 389}$
Heavy Cheetah	$\mathbf{4828 \pm 553}$	$\mathbf{4876 \pm 181^{*}}$	$\mathbf{4519 \pm 240}$	$\mathbf{4604 \pm 184}$	2945 ± 856	2771 ± 1235	3494 ± 318
Broken Joint Ant	5547 ± 204	$\mathbf{6145 \pm 98^{*}}$	$\mathbf{6135 \pm 122}$	5014 ± 401	3725 ± 988	3483 ± 747	3182 ± 1337
Friction Cheetah	$\mathbf{3212 \pm 2279}$	$\mathbf{3890 \pm 1495}$	$\mathbf{3289 \pm 236}$	$\mathbf{2957 \pm 1526}$	$\mathbf{3451 \pm 361}$	$\mathbf{3926 \pm 735}$	$\mathbf{4227 \pm 740^{*}}$
Broken Joint Minitaur	$\mathbf{13.6 \pm 3.8}$	$\mathbf{14.9 \pm 3}$	$\mathbf{16.9 \pm 4.7^{*}}$	$\mathbf{15.8 \pm 2.3}$	$\mathbf{14.6 \pm 1.9}$	$\mathbf{12.1 \pm 5.2}$	10.5 ± 6.1
Low Fidelity Minitaur	15.7 ± 2.8	$\mathbf{17 \pm 2}$	$\mathbf{17.6 \pm 0.4^{*}}$	7.5 ± 5.7	$\mathbf{13.6 \pm 5.1}$	11.4 ± 3.5	12.1 ± 5.5
Broken Leg Ant	$\mathbf{2345 \pm 806}$	$\mathbf{2652 \pm 356}$	$\mathbf{2977 \pm 85^{*}}$	1634 ± 857	1490 ± 714	1554 ± 886	1697 ± 393

Table 4: FOOD sensitivity analysis with respect to the Imitation Learning agent used. We report the average return over 4 seeds associated with their best hyper-parameter α.

Overall, we observe that all GAIL-associated algorithms have the best results. We attribute this success to the implementation we used, which was optimized for the PPO agent. In addition, FOOD with PWIL has poor results in some environments. This can be attributed to two factors. First, we cannot rule out an error in our code, as we coded it from scratch. Second, this algorithm was introduced in the D4PG agent [Barth-Maron *et al.*, 2018]: it is possible that PPO does not leverage well the PWIL's rewards.

An interesting discussion is about GAIL-d^{π}_{P}, GAIL-μ^{π}_{P} and GAIL-ν^{π}_{P}. Intuitively, the one that focuses on state visitation distributions should give the FOOD agent more freedom to find a better action. This is for example what is observed in the Gravity Pendulum environment. However, in most cases, GAIL-μ^{π}_{P} or GAIL-ν^{π}_{P} provide better results as they provide more information regarding the target trajectories. GAIL-ν^{π}_{P} is the one directly derived from Proposition 2, and it seems GAIL-μ^{π}_{P} is implicitely able to optimize the second term in Proposition 1.

C.6 Data Sensitivity Analysis

In this sub-section, we conduct a comparative analysis between FOOD and DARC across the environments where PPO is used on the number of source trajectories they use. The trained agent RL_{s} samples 5, 10, 25 and 50 trajectories on the source environment. During certain trajectories, the robot directly falls: we exclude them for both FOOD and DARC to avoid misleading regularization.

As depicted in Figure 4, both methods demonstrate relative robustness to the number of source trajectories. Their reliance on a discriminator explains why a small number of trajectories appears to be sufficient for the development of a good agent. Additional insights can be extracted from Figure 4. First, in Friction Cheetah, a larger amount of target data allows DARC to outperform FOOD. Second, in Broken Leg Ant and Heavy Cheetah, an increased number of trajectories decreases FOOD's performance. This decline may result from including trajectories that have medium to poor performance in the target environment, leading to misguided regularization.

C.7 DARC Hyperparameters Sensitivity Analysis

We detail in Figure 5 DARC's sensitivity to its main hyper-parameter σ_{DARC}. We observe a clear dependence on the noise added to the discriminator, although there seems to be no pattern for choosing the right hyper-parameter. For instance, the best hyper-parameter for Broken Joint Cheetah and Broken Joint Ant is $\sigma_{\text{DARC}} = 0.1$, but this value leads to worse performance than RL_{s} on the two other presented environments.

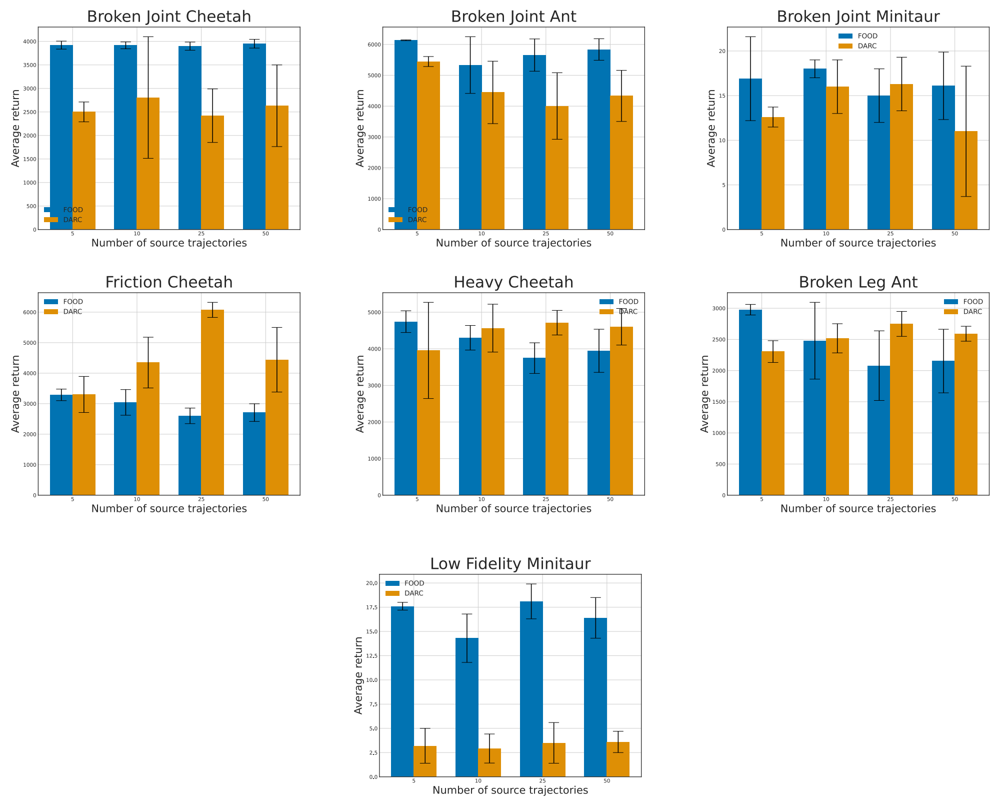

Figure 4: Data sensitivity analysis for both FOOD and DARC agents on the environments where PPO is used.

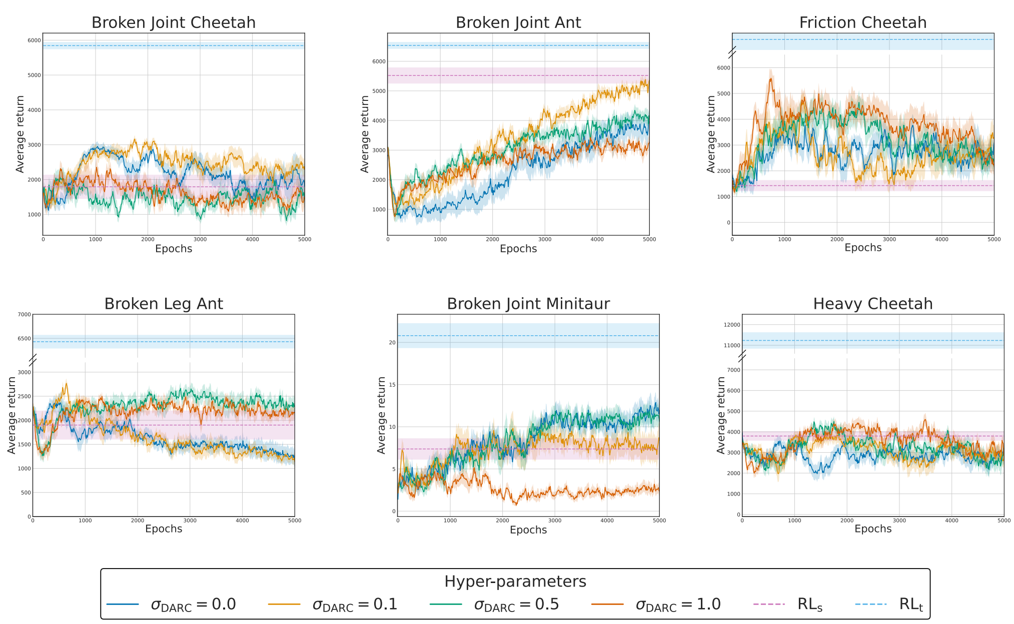

Figure 5: Hyper-parameter sensitivity analysis for the DARC agent on the different environments where DARC works well.

C.8 ANE Hyperparameters Sensitivity Analysis

We also detail the ANE's results for all environments in Figure 6. As a reminder, ANE adds a centered Gaussian noise with std $\sigma_{\text{ANE}} \in \{0.1, 0.2, 0.3, 0.5\}$ to the action during training.

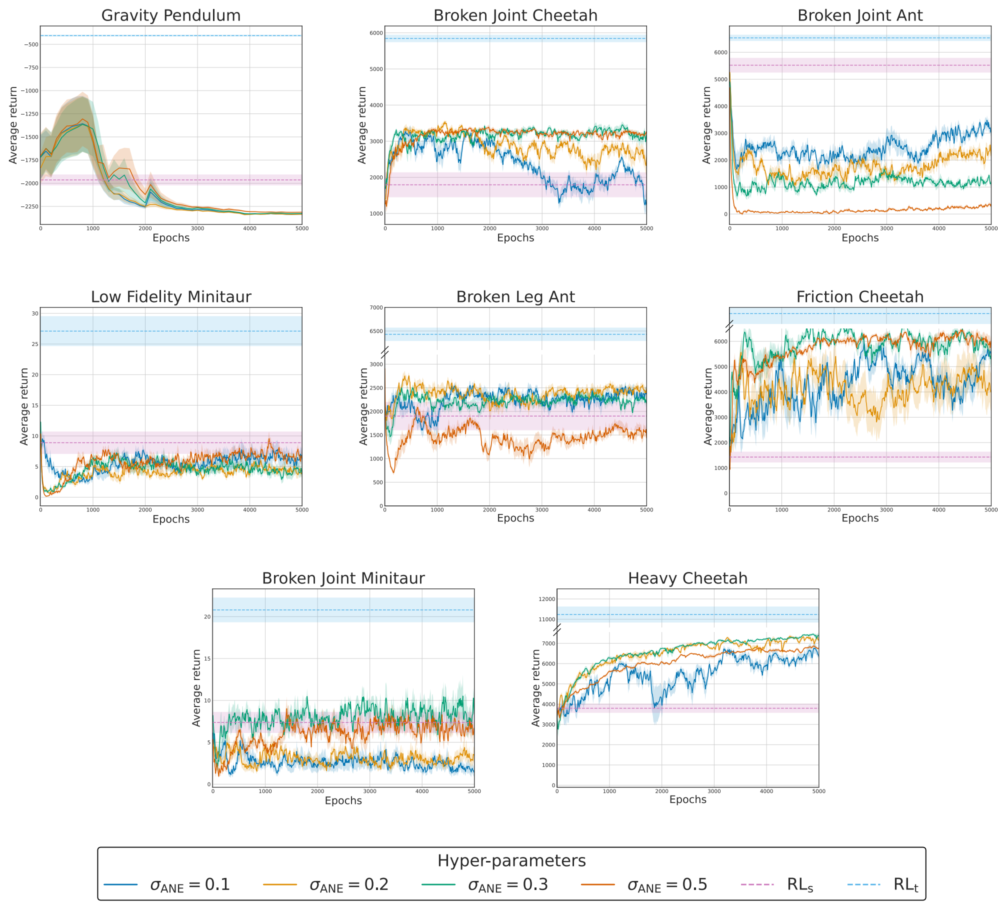

Figure 6: Hyper-parameter sensitivity analysis for the ANE agent on the different environments.

These figures are not easily interpretable. This technique may work very well as observed for Heavy Cheetah, but may fail for other environments such as Broken Joint Ant or Low Fidelity Minitaur.

C.9 H2O Results

Finally, we report H2O results in Figure 7. This method combines the regularization of DARC and CQL in the off-dynamics scenario when the agent has access to a large amount of target data. Since the agent also uses data from the source domain in its learning process, the strength of the regularization is lower than in CQL. It was set to 0.01 in most of the benchmarks in H2O and to 1 for the others. We did a grid search on these 2 values. Given its poor results on the 2 out of 3 environments we tried and the high resources it requires, we did not try it on the other environments.

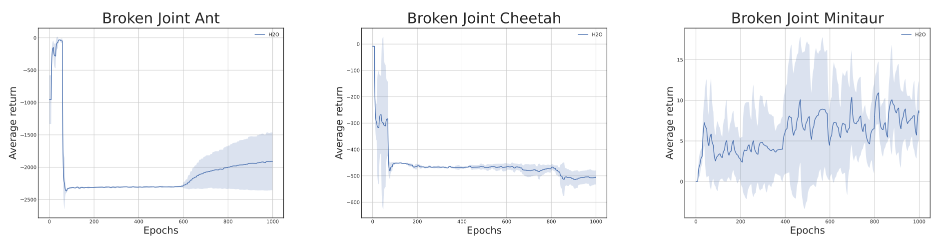

Figure 7: H2O results on 3 environments.