Imitating Cost Constrained Behaviors in Reinforcement Learning

Primary Keywords: *(2) Learning;*

A. Theoretical Analysis

The objective function of the imitation learning problem can be represented using (1),

$$\min_{\pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}), \tag{1}$$

And the distance measure in the GAIL framework is defined as (2).

$$\psi^*(\rho_\pi - \rho_{\pi_E}) = \max_{D} \mathbb{E}_\pi[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] \tag{2}$$

Our proof is based on the GAIL framework, and the objective function of the cost-constrained imitation learning problem is formulated in (3).

$$\begin{aligned} L(\omega,\lambda,\theta) \triangleq \min_{\theta} \max_{\omega,\lambda} \mathbb{E}_{\pi_\theta}[\log D_\omega(s,a)] + \\ \mathbb{E}_{\pi_E}\left[\log(1 - D_\omega(s,a))\right] + \\ \lambda\left(\mathbb{E}_{\pi_\theta}[d(s,a)] - \mathbb{E}_{\pi_E}[d(s,a)]\right) - \beta H(\pi_\theta), \end{aligned} \tag{3}$$

However, it is important to note that the form of the distance measure will differ from that of (2), as will be explained in the following theory.

Theorem 1 *The objective function of the cost-constrained imitation learning problem is:*

$$\min_{\pi \in \Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}), \tag{4}$$

where $\psi^*(\rho_\pi - \rho_{\pi_E}) = \max_{D,\lambda} \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] + \lambda(\mathbb{E}_\pi[d(s,a)] - \mathbb{E}_{\pi_E}[d(s,a)])$

There are broadly two steps to the proof :

Step 1: Typically, optimal policy in an imitation learning setting is obtained by first solving the Inverse Reinforcement Learning (IRL) problem to get the optimal reward function r^* and then running an RL algorithm on the obtained reward function. In GAIL, these two steps were compressed into optimizing a $\psi-$regularized objective. **Our first step is to show this can be also done for Cost Constrained Imitation Learning problems, albeit with an altered $\psi-$ regularized objective.**

Step 2: Our second step is to derive the specific form of ψ^* for CCIL problems.

Step 1

Constrained Markov Decision Process (CMDP) is commonly solved by utilizing the Lagrangian relaxation technique (Tessler, Mankowitz, and Mannor 2018). Then CMDP is transformed into an equivalent unconstrained problem by incorporating the cost constraint into the objective function:

$$\max_{\lambda \geq 0} \min_{\pi \in \Pi} \mathbb{E}_\pi[-r(s,a)] + \lambda(\mathbb{E}_\pi[d(s,a)] - d_0) \tag{5}$$

In the aforementioned equation, our objective is to find the saddle point of the minimax problem. Since the reward function $r(s,a)$ is not provided, our goal is to determine the optimal policy by utilizing the expert policy π_E and the given cost functions $d(s,a)$. To accomplish this, we utilize the maximum casual entropy Inverse Reinforcement Learning (IRL) method (Ziebart, Bagnell, and Dey 2010)(Ziebart et al. 2008) to solve the following optimization problem:

$$\begin{aligned} \max_{\substack{r \in \mathcal{R} \\ \lambda \geq 0}} \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[-r(s,a)] + \lambda(\mathbb{E}_\pi[d(s,a)] - d_0) \right) \\ - \left(\mathbb{E}_{\pi_E}[-r(s,a)] + \lambda(\mathbb{E}_{\pi_E}[d(s,a)] - d_0)\right) \end{aligned} \tag{6}$$

Where $\mathcal{R}$ is a set of reward functions. Maximum casual entropy IRL aims to find a reward function $r \in \mathcal{R}$ that gives low rewards to the learner's policy while giving high rewards to the expert policy. The optimal policy can be found via a reinforcement learning procedure:

$$RL(r,\lambda) = \arg\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[\lambda d(s,a) - r(s,a)] - \lambda d_0 \tag{7}$$

We study policies generated through reinforcement learning, utilizing rewards learned through IRL on the most extensive set of reward functions, denoted as $\mathcal{R}$ in Eq.(6), which encompasses all functions mapping from $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ to $\mathbb{R}$. However, as the use of a large $\mathcal{R}$ can lead to overfitting in the IRL process, we employ a concave reward function regularizer (Finn, Levine, and Abbeel 2016), denoted as ψ, to define the IRL procedure:

$$\begin{aligned} & IRL_\psi(\pi_E, d) = \\ & \arg\max_{\substack{r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \\ \lambda \geq 0}} \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[\lambda d(s,a) - r(s,a)] \right) \\ & - \mathbb{E}_{\pi_E}[\lambda d(s,a) - r(s,a)] + \psi(r) \end{aligned} \tag{8}$$

Given $(\tilde{r}, \tilde{\lambda}) \in IRL_{\psi}(\pi_E, d)$, our objective is to learn a policy defined by $RL(\tilde{r}, \tilde{\lambda})$. To characterize $RL(\tilde{r}, \tilde{\lambda})$, it is commonly beneficial to convert optimization problems involving policies into convex problems. We use occupancy measure ρ_π to accomplish this. After which we express the expected value of the reward and the expected value of the constraint as: $\mathbb{E}_\pi[r(s,a)] = \sum_{s,a} \rho_\pi(s,a) r(s,a)$ and $\mathbb{E}_\pi[d(s,a)] = \sum_{s,a} \rho_\pi(s,a) d(s,a)$ as described in (Altman 1999). IRL can be reformulated as:

$$\begin{aligned} IRL_{\psi}(\pi_E, d) = \arg\max_{\substack{r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \\ \lambda \geq 0}} \min_{\pi \in \Pi} -H(\pi) + \psi(r) + \\ \sum_{s,a} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))[\lambda d(s,a) - r(s,a)] \end{aligned} \tag{9}$$

We then characterize $RL(\tilde{r}, \tilde{\lambda})$, the policy learned by RL on the reward recovered by IRL as the optimal solution of Eq.(4).

Proposition 1 *(Theorem 2 of (Syed, Bowling, and Schapire 2008)) If $\rho \in \mathcal{D}$, then ρ is the occupancy measure for $\pi_\rho(a|s) \triangleq \rho(s,a)/\sum_a' \rho(s,a')$, and π_ρ is the only policy whose occupancy measure is ρ.*

Proposition 2 *(Lemma 3.1 of (Ho and Ermon 2016)) Let $\bar{H}(\rho) = -\sum_{s,a} \rho(s,a) \log(\rho(s,a)/\sum_{a'} \rho(s,a'))$. Then, $\bar{H}$ is strictly concave, and for all $\pi \in \Pi$ and $\rho \in \mathcal{D}$, we have $H(\pi) = \bar{H}(\rho_\pi)$ and $\bar{H}(\rho) = H(\pi_\rho)$*

Proposition 3 *Let $(\tilde{r}, \tilde{\lambda}) \in IRL_{\psi}(\pi_E, d)$, $\tilde{\pi} \in RL(\tilde{r}, \tilde{\lambda})$, and*

$$\begin{aligned} \pi_A \in \arg\min_{\pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}) \\ = \arg\min_{\pi} \max_{r,\lambda} -H(\pi) + \psi(r) + \\ \sum_{s,a} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))[\lambda d(s,a) - r(s,a)] \end{aligned} \tag{10}$$

Then $\pi_A = \tilde{\pi}$.

Proof. Let ρ_A be the occupancy measure of π_A and $\tilde{\rho}$ be the occupancy measure of $\tilde{\pi}$. By using Proposition 1, we define $\bar{L} : \mathcal{D} \times \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \times \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} \bar{L}(\rho, (r,\lambda)) = -\bar{H}(\rho) + \psi(r) + \\ \sum_{s,a} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))[\lambda d(s,a) - r(s,a)] \end{aligned} \tag{11}$$

The following relationship then holds:

$$\rho_A \in \arg\min_{\rho \in \mathcal{D}} \max_{r,\lambda} \bar{L}(\rho, (r,\lambda)) \tag{12}$$

$$(\tilde{r}, \tilde{\lambda}) \in \arg\max_{r,\lambda} \min_{\rho \in \mathcal{D}} \bar{L}(\rho, (r,\lambda)) \tag{13}$$

$$\tilde{\rho} \in \arg\min_{\rho \in \mathcal{D}} \bar{L}(\rho, (\tilde{r}, \tilde{\lambda})) \tag{14}$$

$\mathcal{D}$ is compact and convex, $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is convex. Due to convexity of $-\bar{H}$,it follows that $\bar{L}(\rho, \cdot)$ is convex for all ρ. $\bar{L}(\cdot, (r,\lambda))$ is concave for all (r,λ) (see proof in).

Therefore, we can use minimax duality (Millar 1983):

$$\min_{\rho \in \mathcal{D}} \max_{\substack{r \in \mathcal{R} \\ \lambda}} \bar{L}(\rho, (c,\lambda)) = \max_{\substack{r \in \mathcal{R} \\ \lambda}} \min_{\rho \in \mathcal{D}} \bar{L}(\rho, (c,\lambda)) \tag{15}$$

Hence,from Eqs.(12) and (13), $(\rho_A, (\tilde{r}, \tilde{\lambda}))$ is a saddle point of $\bar{L}$, which implies that:

$$\rho_A \in \arg\min_{\rho \in \mathcal{D}} \bar{L}(\rho, (\tilde{r}, \tilde{\lambda})) \tag{16}$$

Because $\tilde{L}(\cdot, (r,\lambda))$ is strictly concave for all (r,λ), Eqs.(14) and (16) imply $\rho_A = \tilde{\rho}$. Since policies whose corresponding occupancy measure are unique(Proposition 2), finally we get $\pi_A = \tilde{\pi}$

Proposition 3 illustrates the process of IRL in finding the optimal reward function and Lagrangian multiplier, represented by (r^*, λ^*). By utilizing the output of IRL, reinforcement learning can be executed to obtain the optimal policy, represented by π^*. And we prove that π^* is the same as by directly solving the ψ-regularized imitation learning problem $\tilde{L}$. Furthermore, ψ-regularized imitation learning aims to identify a policy whose occupancy measure is similar to that of an expert, as measured by the convex function ψ^*. Subsequently, we deduce the form of ψ^*.

Step 2

In the GAIL paper (Ho and Ermon 2016), the authors present a cost regularizer, ψ_{GA}, that leads to an imitation learning algorithm, as outlined in Eq.(1), which aims to minimize the Jensen-Shannon divergence between the occupancy measures. Specifically, they convert a surrogate loss function, ϕ, which is used for binary classification of state-action pairs drawn from the occupancy measures ρ_π and ρ_{π_E}, into cost function regularizers ϕ, such that $\phi^*(\rho_\pi - \rho_{\pi_E})$ represents the minimum expected risk, $R_\phi(\rho_\pi, \rho_{\pi_E})$, for the function ϕ (Ho and Ermon 2016).

$$R_\phi(\rho_\pi, \rho_{\pi_E}) = \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a)\phi(\gamma) + \rho_{\pi_E}(s,a)\phi(-\gamma) \tag{17}$$

Here we use the same formula of surrogate loss function ϕ as in GAIL paper: $\psi_\phi(c) = \sum_{\rho_{\pi_E}} g_\phi(c(s,a))$, where $g_\phi(x) = -x + \phi(-\phi^{-1}(-x))$, ϕ is a strictly decreasing convex function (Proposition A.1 from (Ho and Ermon 2016)). Noted that in GAIL paper they adopt cost function $c(s,a)$ not reward function $r(s,a)$, then we write in this form: $\psi_\phi(-r) = \sum_{\rho_{\pi_E}} g_\phi(-r(s,a))$.

Then formulation of $\psi^*_\phi(\rho_\pi - \rho_{\pi_E})$ is represented as follows(see proof in):

$$\begin{aligned} &\psi^*_\phi(\rho_\pi - \rho_{\pi_E}) \\ &= -R_\phi(\rho_\pi, \rho_{\pi_E}) + \max_{\lambda} \sum_{s,a} \lambda(\rho_\pi(s,a) - \rho_{\pi_E}(s,a)) d(s,a) \end{aligned}$$

Using the logistic loss $\phi(\gamma) = \log(1 + e^{-\gamma})$, the same form in GAIL paper, then $-R_\phi(\rho_\pi, \rho_{\pi_E}) = \max_{D \in (0,1)^{\mathcal{S}\times\mathcal{A}}} \sum_{s,a} \rho_\pi(s,a) \log D(s,a) + \rho_{\pi_E}(s,a) \log(1 -$

$D(s,a))$. Therefore, we obtain the final form of $\psi^*(\rho_\pi - \rho_{\pi_E})$ as follows:

$$\psi^*(\rho_\pi - \rho_{\pi_E}) = \max_{\substack{D\in(0,1)^{\mathcal{S}\times\mathcal{A}} \\ \lambda}} \sum_{s,a} \rho_\pi(s,a)\log D(s,a) +$$
$$\rho_{\pi_E}(s,a)\log(1-D(s,a)) + \lambda(\rho_\pi(s,a) - \rho_{\pi_E}(s,a))d(s,a)$$
$$= \max_{\substack{D\in(0,1)^{\mathcal{S}\times\mathcal{A}} \\ \lambda}} \mathbb{E}_\pi[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1-D(s,a))]$$
$$+ \lambda(\mathbb{E}_\pi[d(s,a)] - \mathbb{E}_{\pi_E}[d(s,a)])$$

Other Proofs

Prove concavity of $\bar{L}$ $\bar{L}(\cdot,(r,\lambda))$ is concave for all (r,λ). *Proof* We known that $\psi(r)$ is concave, suppose $\alpha \in [0,1]$.

$$\bar{L}(\cdot,(\alpha r_1 + (1-\alpha)r_2, \alpha\lambda_1 + (1-\alpha)\lambda_2)) = -\bar{H}(\rho) +$$
$$\psi(\alpha r_1 + (1-\alpha)r_2) +$$
$$\sum_{s,a}(\rho_\pi - \rho_{\pi_E})[d(\alpha\lambda_1 + (1-\alpha)\lambda_2) - (\alpha r_1 + (1-\alpha)r_2)]$$
$$\geq \alpha\psi(r_1) + (1-\alpha)\psi(r_2) + \alpha\sum_{s,a}(\rho_\pi - \rho_{\pi_E})(\lambda_1 d - r_1)$$
$$+ (1-\alpha)\sum_{s,a}(\rho_\pi - \rho_{\pi_E})(\lambda_2 d - r_2)$$

Therefore, $\bar{L}(\cdot,(\alpha r_1 + (1-\alpha)r_2, \alpha\lambda_1 + (1-\alpha)\lambda_2)) \geq \bar{L}(\cdot,(\alpha r_1, \alpha\lambda_1) + \bar{L}(\cdot,((1-\alpha)r_2,(1-\alpha)\lambda_2)), \bar{L}(\cdot,(r,\lambda))$ is concave for all (r,λ).

Proof of $\psi^*_\phi(\rho_\pi - \rho_{\pi_E})$ We deduce the form of $\psi^*_\phi(\rho_\pi - \rho_{\pi_E})$ as:

$$\psi^*_\phi(\rho_\pi - \rho_{\pi_E}) =$$
$$- R_\phi(\rho_\pi, \rho_{\pi_E}) + \max_\lambda \lambda \sum_{s,a}(\rho_\pi(s,a) - \rho_{\pi_E}(s,a))d(s,a)$$

We will simplify notation by using the symbols ρ_π, ρ_{π_E}, r, and d to represent $\rho_\pi(s,a)$,$\rho_{\pi_E}(s,a)$,$r(s,a)$ and $d(s,a)$, respectively.

$$\psi^*_\phi(\rho_\pi - \rho_{\pi_E}) = \max_{\substack{r\in\mathcal{R} \\ \lambda}} \sum_{s,a}(\rho_\pi - \rho_{\pi_E})(\lambda d - r) - \psi_\phi(-r)$$
$$= \max_{\substack{r\in\mathcal{R} \\ \lambda}} \sum_{s,a}(\rho_\pi - \rho_{\pi_E})(\lambda d - r) - \sum_{s,a}\rho_{\pi_E} g_\phi(-r)$$
$$= \max_{r\in\mathcal{R}} \sum_{s,a}(\rho_\pi - \rho_{\pi_E})(-r) - \sum_{s,a}\rho_{\pi_E}(r + \phi(-\phi^{-1}(r)))$$
$$+ \max_\lambda \sum_{s,a}\lambda(\rho_\pi - \rho_{\pi_E})d$$
$$= \max_{r\in\mathcal{R}} \sum_{s,a}\rho_\pi(-r) - \sum_{s,a}\rho_{\pi_E}\phi(-\phi^{-1}(r))$$
$$+ \max_\lambda \sum_{s,a}\lambda(\rho_\pi - \rho_{\pi_E})d$$

Then we make the change of variables $r \to \phi(\gamma)$, the above equation becomes:

$$\psi^*_\phi(\rho_\pi - \rho_{\pi_E}) =$$
$$\sum_{s,a}\max_{\gamma\in\mathbb{R}} \rho_\pi(-\phi(\gamma)) - \rho_{\pi_E}\phi(-\gamma) + \max_\lambda \lambda\sum_{s,a}(\rho_\pi - \rho_{\pi_E})d$$
$$= -R_\phi(\rho_\pi, \rho_{\pi_E}) + \max_\lambda \lambda\sum_{s,a}(\rho_\pi - \rho_{\pi_E})d$$

Therefore, we prove Theorem 1 and the objective function of cost-constrained imitation learning is Eq.(3).

B. Algorithms for MALM and CVAG

Algorithm 1 and 2 are pseudocodes for Meta-Gradients for Lagrangian multipliers(MALM) and Cost Violation based Alternating Gradient(CVAG) methods.

C. Experiment Figures

C.1 Experiments results

Figure 1 and figure 2 are experiment results of Mujoco tasks and PointButton1 tasks.

Table 1: Hyper-parameters in experiments

hyperparameter	value
Policy and Value network size	(100,100)
Actor and Critic network size (for IQ-Learn)	(256,256)
Activation	Tanh
Batch Size	2000
Generator network update times	3
Discriminator network update times	1
Generalized Advantange Estimation γ	0.995
Generalized Advantange Estimation λ	0.97
Maximum KL	0.01
Learning rate(Value network)	1×10^{-3}
Learning rate(Discriminator network)	3×10^{-4}
Policy entropy	0.0
Discriminator entropy	1×10^{-3}
Initial Lagrangian penalty	0.01
Lagrangian penalty learning rate	0.05
Meta learning rate	0.05

C.2 Experiment Hyperparameters

Table 1 is the illustration of experiment hyper-parameters:

References

Altman, E. 1999. *Constrained Markov decision processes: stochastic modeling*. Routledge.

Finn, C.; Levine, S.; and Abbeel, P. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, 49–58. PMLR.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

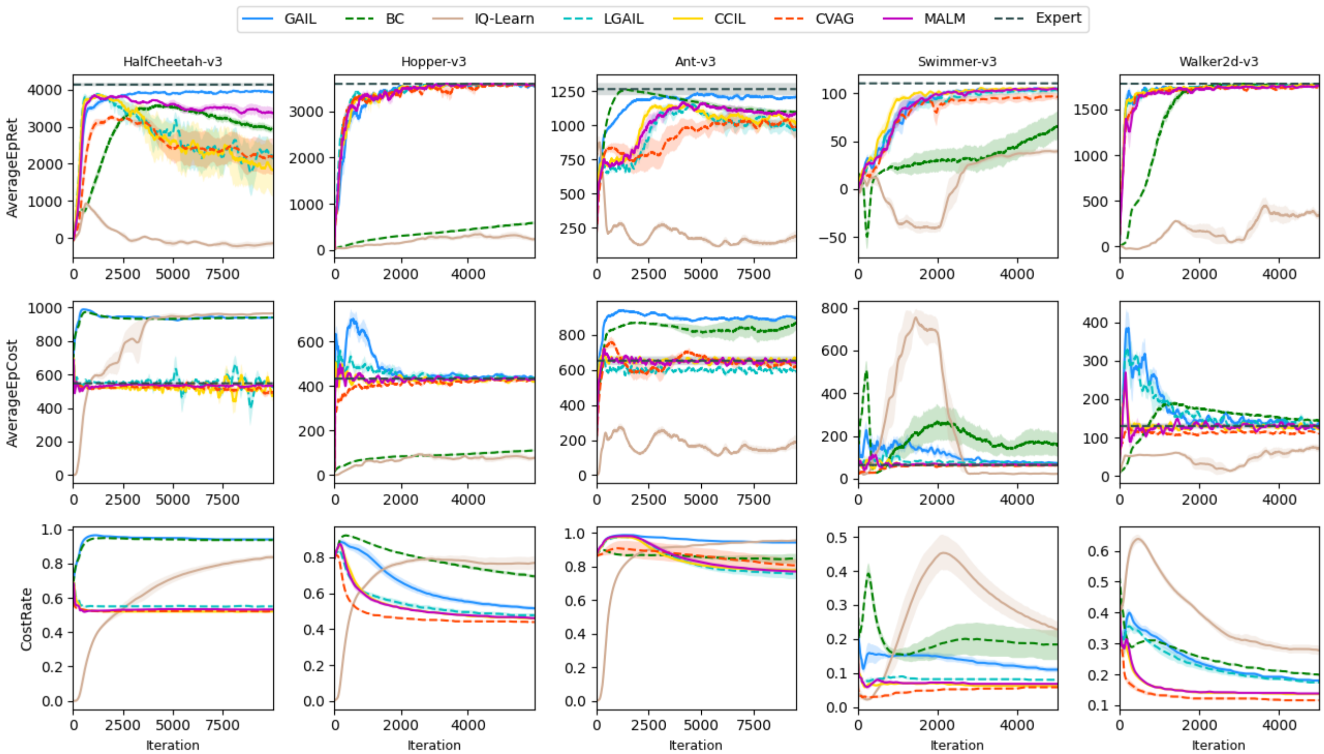

Figure 1: Performance of mujoco environments. The x-axes indicate the number of iterations, and the y-axes indicate the performance of the agent, including average rewards/costs/cost rates with standard deviations.

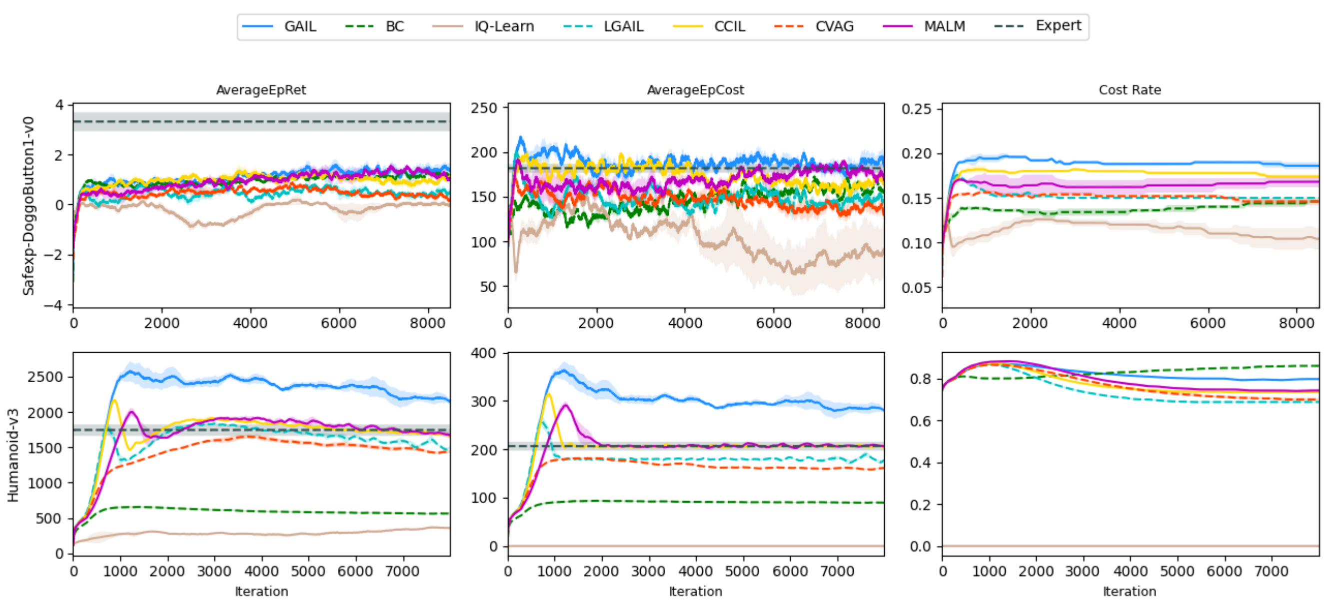

Figure 2: Performance of Humanoid and DoggoButton tasks The x-axes indicate the number of iterations, and the y-axes indicate the performance of the agent, including average rewards/costs/cost rates with standard deviations.

Algorithm 1: Meta-Gradients for Lagrangian Multipliers

Input: initial parameters of policy θ, reward value network ϕ_r, cost value network ϕ_d, discriminator network ω, batch size K, a set of expert trajectories $\Phi_E = \{\tau_E \sim \pi_E\}$, initial Lagrangian multipliers λ, entropy parameter β,learning rates $\alpha_r, \alpha_d, \alpha_\omega, \alpha_\lambda$
Output: Optimal policy π_θ

1: Compute the average cost of expert trajectories: $J_{\mathbb{E}} = \frac{1}{|\Phi_k|} \sum_{\tau \in \Phi_E} \sum_{t=1}^{T} d_t$
2: **for** $k = 1, 2, ...$ **do**
3: Collect set of learner's trajectories $\Phi_k = \{\tau_i\}$ by running policy π_{θ_k} for K time steps.
4: Collect the reward r_t of K time steps by using the discriminator output:$r_t = -\log(D_\omega(s_t, a_t))$
5: Compute $V^r_{\phi_r}(s_t)$ and $V^d_{\phi_d}(s_t)$ of K time steps.
6: Compute the reward and cost advantage $A^r(s_t, a_t)$ and $A^d(s_t, a_t)$, reward to go $\hat{R}^r_t$ and cost to go $\hat{R}^d_t$ of K time steps by using GAE.
7: Compute the average episode cost of learner's trajectories: $J_k = \frac{1}{|\Phi_k|} \sum_{\tau \in \Phi_k} \sum_{t=1}^{T} d_t$
8: Split the data of K time steps into training and validation sets K_{tr}, K_{va}
9: **Inner loss:**
10: Update policy by using TRPO rule:
$\theta' = \arg\max_\theta \sum_{t=1}^{K_{tr}} \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} (A^r(s_t, a_t) - \lambda A^d(s_t, a_t)) + \beta H(\pi_{\theta_k})$
11: Update reward value network:
$\phi'_r \leftarrow \phi_r - \frac{1}{K_{tr}} \sum_{t=1}^{K_{tr}} \alpha_r \bigtriangledown_{\phi_r} (V^r_{\phi_r}(s_t) - \hat{R}^r_t)^2$
12: Update cost value network:
$\phi'_d \leftarrow \phi_d - \frac{1}{K_{tr}} \sum_{t=1}^{K_{tr}} \alpha_d \alpha_r \bigtriangledown_{\phi_d} (V^d_{\phi_d}(s_t) - \hat{R}^d_t)^2$
13: Update discriminator network:
$\omega' \leftarrow \omega + \frac{1}{K} \sum_{t=1}^{K} \alpha_\omega (\bigtriangledown_\omega [\log(D_\omega(s_t, a_t))] + \bigtriangledown_\omega [\log(1 - D_\omega(s_t, a_t))])$
14: Update Lagrangian multipliers:
$\lambda' \leftarrow \lambda + \alpha_\lambda (J_k - J_{\mathbb{E}})$
15: **Outer loss:**
16: Meta-parameter update:
$\lambda'' \leftarrow \lambda' - \frac{1}{K_{va}} \sum_{t=1}^{K_{va}} \bigtriangledown_{\lambda'} (A^r(s_t, a_t) - \lambda' d_t)^2$
17: $\theta \leftarrow \theta', \phi_r \leftarrow \phi'_r, \phi_d \leftarrow \phi'_d, \omega \leftarrow \omega', \lambda \leftarrow \lambda''$.
18: **end for**

Algorithm 2: Cost Violation based Alternating Gradient

Input: initial parameters of policy θ, reward value network ϕ_r, cost value network ϕ_d, discriminator network ω, batch size K, a set of expert trajectories $\Phi_E = \{\tau_E \sim \pi_E\}$, entropy parameter β, learning rates $\alpha_r, \alpha_d, \alpha_\omega$.
Output: Optimal policy π_θ

1: Compute the average cost of expert trajectories: $J_{\mathbb{E}} = \frac{1}{|\Phi_k|} \sum_{\tau \in \Phi_E} \sum_{t=1}^{T} d_t$
2: **for** $k = 1, 2, ...$ **do**
3: Collect set of learner's trajectories $\Phi_k = \{\tau_i\}$ by running policy π_{θ_k} for K time steps.
4: Collect the reward r_t of K time steps by using the discriminator output:$r_t = -\log(D_\omega(s_t, a_t))$
5: Compute $V^r_{\phi_r}(s_t)$ and $V^d_{\phi_d}(s_t)$ of K time steps.
6: Compute the reward and cost advantage $A^r(s_t, a_t)$ and $A^d(s_t, a_t)$, reward to go $\hat{R}^r_t$ and cost to go $\hat{R}^d_t$ of K time steps by using GAE.
7: Compute the average episode cost of learner's trajectories: $J_k = \frac{1}{|\Phi_k|} \sum_{\tau \in \Phi_k} \sum_{t=1}^{T} d_t$
8: **if** $J_k \le J_{\mathbb{E}}$ **then**
9: Update policy towards maximizing the return:
$\theta' = \arg\max_\theta \sum_{t=1}^{K} \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^r(s_t, a_t) + \beta H(\pi_{\theta_k})$
10: **else**
11: Update policy towards minimizing the cost:
$\theta' = \arg\min_\theta \sum_{t=1}^{K} \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^d(s_t, a_t) - \beta H(\pi_{\theta_k})$
12: **end if**
13: Update reward value network:
$\phi'_r \leftarrow \phi_r - \frac{1}{K} \sum_{t=1}^{K} \alpha_r \bigtriangledown_{\phi_r} (V^r_{\phi_r}(s_t) - \hat{R}^r_t)^2$
14: Update cost value network:
$\phi'_d \leftarrow \phi_d - \frac{1}{K} \sum_{t=1}^{K} \alpha_d \bigtriangledown_{\phi_d} (V^d_{\phi_d}(s_t) - \hat{R}^d_t)^2$
15: Update discriminator network:
$\omega' \leftarrow \omega + \frac{1}{K} \sum_{t=1}^{K} \alpha_\omega (\bigtriangledown_\omega [\log(D_\omega(s_t, a_t))] + \bigtriangledown_\omega [\log(1 - D_\omega(s_t, a_t))])$
16: $\theta \leftarrow \theta', \phi_r \leftarrow \phi'_r, \phi_d \leftarrow \phi'_d, \omega \leftarrow \omega'$.
17: **end for**

Millar, P. W. 1983. The minimax principle in asymptotic statistical theory. In *Ecole d'Eté de Probabilités de Saint-Flour XI—1981*, 75–265. Springer.

Syed, U.; Bowling, M.; and Schapire, R. E. 2008. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, 1032–1039.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

Ziebart, B. D.; Bagnell, J. A.; and Dey, A. K. 2010. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 1255–1262.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; Dey, A. K.; et al. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, 1433–1438. Chicago, IL, USA.