PRESCRIBED SAFETY PERFORMANCE IMITATION LEARNING FROM A SINGLE EXPERT DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing safe imitation learning (safe IL) methods mainly focus on learning safe policies that are similar to expert ones, but may fail in applications requiring different safety constraints. In this paper, we propose the Lagrangian Generative Adversarial Imitation Learning (LGAIL) algorithm, which can adaptively learn safe policies from a single expert dataset under diverse prescribed safety constraints. To achieve this, we augment GAIL with safety constraints and then relax it as an unconstrained optimization problem by utilizing a Lagrange multiplier. The Lagrange multiplier enables explicit consideration of the safety and is dynamically adjusted to balance the imitation and safety performance during training. Then, we apply a two-stage optimization framework to solve LGAIL: (1) a discriminator is optimized to measure the similarity between the agent-generated data and the expert ones; (2) forward reinforcement learning is employed to improve the similarity while considering safety concerns enabled by a Lagrange multiplier. Furthermore, theoretical analyses on the convergence and safety of LGAIL demonstrate its capability of adaptively learning a safe policy given prescribed safety constraints. At last, extensive experiments in OpenAI Safety Gym conclude the effectiveness of our approach.

1 INTRODUCTION

Imitation learning (IL), which learns from expert data or expert policies to reproduce an expert policy, has achieved remarkable successes in various applications such as self-driving (Li et al., 2017; Pan et al., 2020), navigation (Hussein et al., 2018), and robot locomotion (Yuan & Kitani, 2020). Most of these algorithms are trained in simulated environments, in which agents are free to make mistakes. However, when deploying IL in real-world applications, the safety of agents is paramount (Amodei et al., 2016; Ray et al., 2019; Arora & Doshi, 2021), and safety requirements can dynamically vary depending on the target application. A policy that is trained without considering safety could generate improper or even harmful actions, and those actions may destroy the safety of agents, which must be avoided in safety-critical scenarios (Sinha et al., 2020).

Nevertheless, little attention has been paid to ensuring the safety of agents in IL. Furthermore, existing several works on safe IL (Zhang & Cho, 2016; Menda et al., 2019; Bhattacharyya et al., 2019; 2020) lack a direct control over the safe level and thus could not generate policies that satisfy a prescribed safety requirement. Conducting safe IL from a single expert dataset with a configurable safety constraint is more realistic because: (1) it is likely that target applications require more stringent safety constraints compared to expert data (Koschuch et al., 2019; Chia et al., 2022), but it is costly and laborious to re-collect new expert data; (2) when the safety standards for new tasks are different from those of expert ones (Phillips & Shikora, 2018), expert data that used to be safe during collection will not work in new applications; (3) an expert dataset could contain some dangerous information because even experts could make mistakes and take dangerous actions (Council et al., 2003; Bickmore et al., 2018; Liu et al., 2020; Lattanzi & Freschi, 2021), resulting in the infeasibility of directly mimicking expert data without the consideration of safety. For example, in autonomous driving, safe speed limits in urban and rural areas are distinct (Warner & Åberg, 2008; Seff & Xiao, 2016); forbidden zones in a robot navigation task can be redesignated under different scenarios (Paternain et al., 2022).

In this paper, we consider the more practical safe IL task, where the agent is required to learn policies given a prescribed safety constraint possibly different from those of experts. In particular, we are given an application-specific safety constraint and the cost signal (or constraint violation signal) from the environment (Achiam et al., 2017; Ray et al., 2019; Stooke et al., 2020; Marchesini et al., 2022)

in addition to an expert dataset. When the safety constraint for the target application is smaller than that of expert data, we regard expert data as unsafe, then the aim of safe IL is to reach super-expert performance regarding the safety. The more realistic task makes it challenging to recover policies that can concurrently achieve expert-level performance and satisfy prescribed safety needs. Unfortunately, to the best of our knowledge, the safe IL task described above is of significance but has not been investigated until now.

To produce policies that can simultaneously achieve expert-level cumulative rewards and satisfy prescribed safety constraints by imitating expert data, we interpret this safe IL task as a constrained optimization problem with Constrained Markov Decision Process (CMDP) (Altman, 1999), *i.e.*, the agent should try to behave as similarly as possible to the expert under safety constraints. Specifically, we introduce an auxiliary cost constraint to restrict the policy generated by Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016), leading to a constrained minimax problem. To tackle the difficult inequality constraint, we adopt a Lagrange multiplier to relax the constrained GAIL problem as an unconstrained one. Then, we propose a new two-stage optimization framework, abbreviated as LGAIL. Specifically, in the first stage, a discriminator is optimized to better measure the similarity between the agent-generated state-action pairs and the expert ones. In the second stage, forward reinforcement learning is employed to improve the similarity while considering safety concerns via a Lagrange multiplier. To the end, we summarize our contributions as three-fold:

- We formalize a new safe IL task with CMDP, where the agent is required to generate policies that satisfy prescribed safety constraints utilizing one expert dataset and the cost signal provided by the environment.
- We develop a safe IL algorithm–LGAIL, a neat yet effective way to tackle the new safe IL task. Theoretical analyses provide nonasymptotic convergence and safety guarantees of LGAIL, which indicates the proposed LGAIL can adaptively learn safe policy with given safety constraints.
- We carry out extensive experiments on various robot tasks in the OpenAI Safety Gym (Ray et al., 2019) to illustrate that LGAIL can work well in the novel safe IL task and can achieve super-expert performance regarding safety.

2 Related Work

Safe Reinforcement Learning (Safe RL). RL with safety-critical constraints, also known as safe RL, has received extensive attention recently (Ray et al., 2019; Liu et al., 2021). The most popular way to deal with safe RL is to convert it into a constrained optimization problem via CMDP (Altman, 1999). There are two major classes of methods to solve safe RL featured by CMDP, *i.e.*, direct approaches (Achiam et al., 2017; Yang et al., 2020; Zhang et al., 2020) and indirect approaches (Ray et al., 2019; Stooke et al., 2020). Constrained Policy Optimization (CPO) (Achiam et al., 2017) is a representative algorithm of direct methods, in which a policy is optimized under performance improvement objective and safety constraints. Yang et al. (2020) split the optimization problem in CPO into two steps: (1) optimize the policy with consideration of only rewards; (2) project the optimized policy into the nearest safe policy. Two milestones of indirect algorithms are TRPO-Lagrangian and PPO-Lagrangian (Ray et al., 2019), which use Lagrange multipliers and show outstanding performance of satisfying constraints. Stooke et al. (2020) improve the Lagrangian methods with PID control to reduce constraint-violating behaviors. However, above methods cannot guarantee the safety of agents during training. To achieve training safety, another spectrum of safe RL algorithms is developed based on Lyapunov functions (Chow et al., 2018; 2019; Jeddi et al., 2021).

Imitation Learning (IL) & Safe Imitation Learning (Safe IL). IL commits to reproducing an expert policy from expert data or expert policies. In general, IL can be divided into behavioral cloning (BC) (Bain & Sammut, 1995; Ross et al., 2011) and inverse reinforcement learning (IRL) (Abbeel & Ng, 2004). The major difference is that the former solves IL in a supervised learning manner, whereas IRL solves IL from the perspective of RL (Torabi et al., 2018). BC enjoys merits of simpleness but suffers from the compounding error and often fails to recover an expert policy compared to IRL (see Hussein et al. (2017) and its reference therein). However, when it comes to safe IL, there is few work. A class of methods are SafeDAgger (Zhang & Cho, 2016) and EnsembleDAgger (Menda et al., 2019), which are built on the framework DAGGER (Ross et al., 2011). They measure the difference between decisions of the learner and the expert while interacting with environments. When the difference goes

beyond a predefined bound, the expert decision will be executed to ensure the safety of the learner. However, both algorithms require a safe expert policy, which is difficult to be satisfied in practice. Based on GAIL, Bhattacharyya et al. (2019; 2020) develop Reward Augmented Imitation Learning (RAIL) via imposing a fixed large penalty on dangerous state-action pairs.

We consider a more practical task where agents are required to conduct safe IL with prescribed safety performance. Compared to Zhang & Cho (2016); Menda et al. (2019), we do not demand safe expert policies to teach the imitator; in contrast to Bhattacharyya et al. (2019; 2020), our algorithm LGAIL can explicitly control the safety performance by dynamically balancing reward and safety needs via a Lagrange multiplier. In addition, the considered safe IL task is different from IL from imperfect demonstration (Wu et al., 2019) that merely considers performance and neglects safety issues, while LGAIL simultaneously focuses on safety and performance issues.

3 PRELIMINARIES

Constrained Markov Decision Process (CMDP). CMDP (Altman, 1999; Achiam et al., 2017; Guan et al., 2021) is modeled by $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, c, d_0, \gamma)$, where $\mathcal{S}$ is state space, $\mathcal{A}$ represents action space, $\mathcal{T} = \mathcal{T}(s'|s,a)$ is the environment transition dynamics, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost function, d_0 is the safety constraint (or cost limit), and γ is the discount factor. The reward function $r(s,a)$ is assumed to be bounded, *i.e.*, $|r(s,a)| \leq R_{max}$, whereas $c(s,a)$ is assumed to be an indicator function (Ray et al., 2019) such that $c(s,a) = 0$ if the agent is safe, $c(s,a) = 1$ otherwise. We consider the tabular setting here, *i.e.*, both space $\mathcal{S}$ and $\mathcal{A}$ are finite. Let $\pi_\theta(a_t|s_t) : \mathcal{S} \times \mathcal{A} \to [0,1]$, whose parameter is θ, be a stochastic policy for the agent. Besides, we assume $\pi_\theta(a|s)$ is directly parameterized by θ, *i.e.*, $\pi_\theta(a|s) = \theta_{s,a}$ and $\theta \in \Theta_p := \{\theta_{s,a} \geq 0, \sum_{a\in\mathcal{A}} \theta_{s,a} = 1, \forall s \in \mathcal{S}\}$. The cost in CMDP refers to safety, *i.e.*, when we talk about "cost" in this paper, it indicates that we are focusing on safety. The expected discounted reward is denoted as $V(\pi, r) = \mathbf{E}_{s_0\sim\zeta}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where $a_t \sim \pi(a_t|s_t)$, $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, a_t)$, and $\zeta(s_0)$ is the probability distribution of the initial state s_0. And we define the expected discounted cost as $J_C(\pi) = \mathbf{E}_{s_0\sim\zeta}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)]$ on the same way. The goal of safe RL defined in Eq. (1) is to find the optimal policy $\pi^*(a_t|s_t)$ which simultaneously satisfies the cost limit,

$$\pi^* = \arg\max_{\pi} V(\pi, r) \quad s.t.\ J_C(\pi) \leq d_0. \tag{1}$$

Generative Adversarial Imitation Learning (GAIL). In GAIL, there are no reward functions as in RL but a discriminator $r_\alpha(s,a)$ that is parameterized by α ($\alpha \in \Lambda \subset \mathbb{R}^q$, and Λ is assumed to be a bounded closed set such that $\forall \alpha_1, \alpha_2 \in \Lambda, \|\alpha_1 - \alpha_2\|_2 \leq C_\alpha$). GAIL is formulated as the following minimax saddle point optimization (Ho & Ermon, 2016; Guan et al., 2021)

$$\min_{\theta\in\Theta_p} \max_{\alpha\in\Lambda} F(\theta, \alpha) := V(\pi_E, r_\alpha) - V(\pi_\theta, r_\alpha) - \psi(\alpha), \tag{2}$$

where π_E stands for expert policies, and $\psi(\alpha)$ is a regularizer.

4 PRESCRIBED SAFETY PERFORMANCE IMITATION LEARNING

In this section, we present the proposed safe IL paradigm, Lagrangian Generative Adversarial Imitation Learning (LGAIL). Below, we first formalize the new task of safe IL with CMDP in Subsection 4.1. Then, the detailed description of LGAIL is presented in Subsections 4.2 and 4.3.

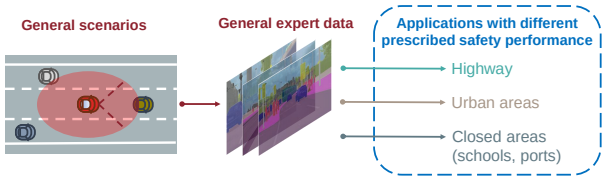


Figure 1: An illustrative example where prescribed safety performance different from expert data is required.

4.1 PROBLEM FORMULATION

Motivations. Three significant factors motivate us to study safe IL with prescribed safety performance. First, ensuring the safety of agents is paramount in most applications. For example, in some safety-critical domains such as human-robot interaction or autonomous driving, robots or vehicles could cause irrevocable human injuries if no special operations are designed for safety. Hence, it

is of significance to pay attention to safety issues in IL. Second, it is likely that target applications have different safety requirements from those of expert data due to: (1) the safety level is generally application-related such that it is probable the target task has a stricter safety level compared to expert data (Koschuch et al., 2019; Chia et al., 2022); (2) even if under the same safety level, safety standards could not be static (Phillips & Shikora, 2018). For example, in autonomous driving, the requirements on safe speed limits in urban and rural areas are distinct (Warner & Åberg, 2008; Seff & Xiao, 2016); the layout of forbidden zones that serve as one of the safety standards in robot navigation could be redesigned under different scenarios (Paternain et al., 2022). We give an illustrative example where prescribed safety performance is required in Figure 1. Last but not least, it is natural that expert data may contain a portion of dangerous data due to the following two reasons: (1) practical expert data often come from various sources with distinct qualities (Tangkaratt et al., 2020); (2) even senior experts could not be immune to mistakes or dangerous decisions (Best, 1992; Culverhouse et al., 2003). Without explicitly considering safety, it is hard to ensure the prescribed safety of generated policies with expert data. Therefore, with the three motivations and the dilemma current safe IL approaches, we aim to solve the safe IL task formalized subsequently.

We characterize the property of expert data with an assumption.

Assumption 1. *We have access to a series of expert trajectories, whose safety constraints could be different from the prescribed one for the imitator.*

The expert trajectories are denoted as $\tau_E = \{\tau_E^1, \tau_E^2, ..., \tau_E^N\}$, and each trajectory τ_E^i where $i \in \{1, ..., N\}$ is composed of chronological states and actions. These expert trajectories can achieve high episodic cumulative rewards, and τ_E^i is sampled from an expert whose safety constraint is $d_i^{\tau_E}$. We denote the minimum safety constraint of τ_E as $d_{min}^{\tau_E} = \min\{d_i^{\tau_E} | i = 1, ..., N\}$. The prescribed safety constraint for safe IL is d_0. As we discussed above, it is probable that d_0 for the target application is smaller than the minimum of expert data such that $d_0 < d_{min}^{\tau_E}$. Under the circumstances, the expert can be regarded as unsafe in terms of target tasks.

Naively conducting IL using the expert data without considering the prescribed safety constraint d_0 could generate unsafe policies. To make it possible to achieve a safe agent, a reasonable assumption on the access to the safety information is made below.

Assumption 2. *The agent can receive cost signals from the environment.*

Assumption 2 is commonly adopted in safe RL Achiam et al. (2017); Ray et al. (2019); Stooke et al. (2020); Marchesini et al. (2022) and makes sense in reality because safety functions are quite possibly simpler than reward functions. For example, in autonomous driving, dangerous conditions such as collisions with pedestrians or cars can be easily identified (Shin & Kim, 2019). Therefore, the task of interest of this paper is presented as follows:

The task of interest: Given expert trajectories τ_E in Assumption 1, a prescribed safety constraint d_0, and the cost signal in Assumption 2, we aim to find a policy that can mimic the expert as much as possible under the prescribed safety constraint.

Although conducting safe IL in this task is arduous, it is worth investigating due to its potential for practical applications compared to former tasks (Bhattacharyya et al., 2019; 2020; Wu et al., 2019).

4.2 Prescribed Safety Performance Imitation Learning

In this new task of safe IL, there are two learning objectives. The first one is that the agent should mimic the expert as much as possible via given expert trajectories when it comes to the episodic cumulative rewards. The second one is that the agent should behave safely to meet prescribed safety constraints utilizing the environment feedback. The safety should be considered as a hard constraint because it represents physical requirement and should not be violated, which motivates us to model safe IL as constrained optimization, *i.e.*, the agent is supposed to mimic the expert as much as possible under prescribed safety constraints. Note that it is not a pure IL problem because the agent should behave unlike the expert in some states due to safety concerns. Thus, we formulate safe IL on the top of GAIL as a constrained minimax optimization problem,

$$\min_{\theta \in \Theta_p} \max_{\alpha \in \Lambda} F(\theta, \alpha) \; s.t. \; J_C(\pi_\theta) \leq d_0. \tag{3}$$

Algorithm 1 Lagrangian Generative Adversarial Imitation Learning (LGAIL). The operation on vectors are element-wise and $h_k^2 = h_k \odot h_k$.

Initialize: Prescribed cost limit d_0, $m_0 = (0, \ldots, 0) \in \mathbb{R}^{q+1}$, $v_0 = (\mu, \ldots, \mu) \in \mathbb{R}_+^{q+1}$, $\theta_0 \in \Theta_p$, and $\alpha_1^0 \in \Lambda$.
for $t = 0, ..., T-1$ **do**
  for $k = 1, ..., K$ **do**
    Sample $(s_i^E, a_i^E) \sim \widetilde{\mathcal{T}}^{\pi_E}$ and $(s_i^{\theta_t}, a_i^{\theta_t}) \sim \widetilde{\mathcal{T}}^{\pi_{\theta_t}}$
    $h_k = -\hat{\nabla}_{\widetilde{\alpha}} F(\theta_t, \widetilde{\alpha}_k^t)$
    $v_k = \iota_k v_{k-1} + (1 - \iota_k) h_k^2$
    $m_k = \beta_k m_{k-1} + (1 - \beta_k) h_k$
    $\widetilde{\alpha}_{k+1}^t = \widetilde{\alpha}_k^t - b_k \cdot \frac{m_k}{\sqrt{v_k}}$ ▷ update $\widetilde{\alpha} = (\alpha, \lambda)$ with the Adam optimizer
  end for
  Randomly choose $\widetilde{\alpha}_t$ from $\{\widetilde{\alpha}_1^t, .., \widetilde{\alpha}_K^t\}$
  $\theta_{t+1} = \mathbf{TRPO}$ in Eq. (5) ▷ update policy π_θ with TRPO
end for

In other words, when optimizing the policy, both similarities calculated by the discriminator $r_\alpha(s, a)$ as well as violations of safety constraints should be considered simultaneously.

4.2.1 Lagrangian Generative Adversarial Imitation Learning

To solve the safe IL problem, we propose a two-stage optimization framework, LGAIL, whose pseudo-code is illustrated in Algorithm 1. Directly solving the safe IL task, which is a constrained optimization problem, is challenging. We employ a Lagrange multiplier to relax the constrained optimization problem into an unconstrained one (Boyd et al., 2004), *i.e.*, prescribed safety constraints are converted into penalties. As a result, we augment the policy improvement stage in GAIL with a Lagrange multiplier. Concretely, the constrained optimization problem in Eq. (3) can be solved by penalizing violations of safety constraints with a Lagrange multiplier when optimizing the policy to mimic the expert,

$$\min_{\theta \in \Theta_p} \max_{\alpha \in \Lambda, \lambda \geq 0} F(\theta, \alpha, \lambda) := V(\pi_E, r_\alpha) - V(\pi_\theta, r_\alpha) + \lambda(J_C(\pi_\theta) - d_0) - \psi(\alpha, \lambda), \tag{4}$$

where λ is the Lagrange multiplier with $\lambda_{max} \geq \lambda \geq 0$, and $\psi(\alpha, \lambda)$ regularizes both α and λ. Let $\widetilde{\alpha} \triangleq (\alpha, \lambda)$, which means adding the 1-dimension scalar λ to the q-dimension vector α. Hence, we obtain a $(q+1)$-dimension vector $\widetilde{\alpha}$. Consequently, we rewrite the object function $F(\theta, \alpha, \lambda)$ as $F(\theta, \widetilde{\alpha})$.

This optimizing target contains both rewards and costs, which can help imitate the expert as well as guarantee safety. There are two stages in LGAIL taking turns to (i) optimize a discriminator to enhance its ability on judging the quality of state-action pairs, and (ii) improve the performance of the agent's policy with the discriminator and safety feedback information enabled by a Lagrange multiplier. The Lagrange multiplier λ helps balance the competition between improving rewards and reducing costs, and it is dynamically updated in stage (i). When the current policy is unsafe, λ will increase so that the penalty on violations of constraints will play a bigger role. On the contrary, λ would decrease so that the optimization concentrates more on mimicking the expert. Parameters of the discriminator $r_\alpha(s, a)$ and Lagrange multiplier λ are updated with an Adam optimizer in Chen et al. (2021). As for the policy, it is updated using Trust Region Policy Optimization (TRPO) (Shani et al., 2020) as follows

$$\pi_{\theta_{t+1}}(\cdot|s) \in \arg\min_{\pi} (\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s, \cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s, \cdot), \pi - \pi_{\theta_t}(\cdot|s) \rangle + \eta_t^{-1} B_w(\pi, \pi_{\theta_t}(\cdot|s))), \tag{5}$$

where $\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s, a)$ and $\hat{Q}_c^{\pi_{\theta_t}}(s, a)$ are estimated Q-value functions for rewards and costs using GAE (Schulman et al., 2015b), respectively, and B_w is the Bregman distance (Wu et al., 2009).

4.3 Theoretical Analyses

In this part, we provide convergence analyses and safety guarantees for LGAIL presented in Algorithm 1. We first give some notations before our proof. Denote $g(\theta)$ as the marginal-maximum

function of $F(\theta, \alpha, \lambda)$, *i.e.*, $g(\theta) := \max_{\alpha \in \Lambda, \lambda \geq 0} F(\theta, \alpha, \lambda) = \max_{\widetilde{\alpha}} F(\theta, \widetilde{\alpha})$. For any fixed θ, we define $\widetilde{\alpha}_\theta^* = \arg\max_{\widetilde{\alpha}} F(\theta, \widetilde{\alpha})$, which means $(\alpha, \lambda)_\theta^* = \arg\max_{\alpha \in \Lambda, \lambda \geq 0} F(\theta, \alpha, \lambda)$. Next we have to define the measure of convergence in our algorithm and present some basic assumptions.

Definition 1. *Guan et al. (2021) The output $\bar{\theta}$ of our algorithm achieves ϵ-global convergence if $g(\bar{\theta}) - g(\theta^*) \leq \epsilon$, where $\theta^* = \arg\min_\theta g(\theta)$ and $\epsilon \in (0, 1)$.*

Assumption 3. *$F(\theta, \widetilde{\alpha})$ is uniformly bounded that $|F(\theta, \widetilde{\alpha})| \leq g^*, \forall \theta$.*

Uniform boundedness means F function is upper bounded no matter how the variables θ and $\widetilde{\alpha}$ change. And the assumption 3 further guarantees the existence of $g(\theta)$.

Assumption 4. *The regularizer $\psi(\widetilde{\alpha})$ is L_ψ-Lipschitz smooth, where L_ψ is the Lipschitz constant.*

Assumption 5. *For any given θ, $F(\theta, \widetilde{\alpha})$ is μ-strongly concave on $\widetilde{\alpha}$, which is usually satisfied by designing the regularizer $\psi(\widetilde{\alpha})$ to be strongly convex in practice.*

Assumption 6. *(Guan et al., 2021) There are some restrictions on the parameterization of reward function.*

(1) $\forall \alpha \in \Lambda$, there exists $C_r \in \mathbb{R}$ such that $\|\nabla_\alpha r_\alpha\|_{\infty,2} := \sqrt{\sum_{i=1}^q \|\frac{\partial r_\alpha}{\partial \alpha_i}\|_\infty^2} \leq C_r$;

(2) $\forall s \in \mathcal{S}, a \in \mathcal{A}, \forall \alpha_1, \alpha_2 \in \Lambda$, there exists $L_r \in \mathbb{R}$ such that $\|\nabla_\alpha r_{\alpha_1}(s,a) - \nabla_\alpha r_{\alpha_2}(s,a)\|_2 \leq L_r \|\alpha_1 - \alpha_2\|_2$.

Assumption 7. *(Bhandari et al., 2018) (Ergodicity) The MDP with policy π_θ and transition kernel $\widetilde{\mathcal{T}}(\cdot|s,a) = \gamma \mathcal{T}(\cdot|s,a) + (1-\gamma)\zeta(\cdot)$ is ergodic, which means for some positive constants $C_M > 0$ and $0 < \rho < 1$,*

$$sup_{s \in \mathcal{S}} d_{TV}(P(s_t \in \cdot | s_0 = s), \chi_\theta) \leq C_M \rho^t, \forall t \geq 0,$$

where $d_{TV}(\cdot, \cdot)$ calculates the total variation distance and χ_θ represents the stationary distribution generated from $\widetilde{\mathcal{T}}(\cdot|s,a)$ or $\mathcal{T}(\cdot|s,a)$ with policy π_θ.

Assumption 8. *Estimated derivatives of $F(\theta, \widetilde{\alpha})$ regarding θ and $\widetilde{\alpha}$ are unbiased. And the estimate of stochastic gradient h_k satisfies uniform boundedness: $\mathbf{E}\|h_k\|^2 \leq G$.*

For the convergence of Adam (Chen et al., 2021), we give some necessary conditions for its parameters:

1. There exists a constant β satisfying $0 \leq \beta_k \leq \beta < 1, \forall k$.
2. The sequence $\{\iota_k\}$ is non-decreasing with $0 < \iota_k < 1$ and $\lim_{k \to \infty} \iota_k \triangleq \iota > \beta^2$.
3. Let $\chi_k := \frac{b_k}{\sqrt{1-\iota_k}}$. There exists a sequence $\{b_k\}$ that is non-increasing and an independent constant C_0 such that $b_k \leq \chi_k \leq C_0 b_k$.

Theorem 1. *(Convergence) When assumptions 1- 8 are satisfied, Adam parameters meet the requirements 1-3, and the update stepsize of θ is $\eta_t = \frac{1-\gamma}{\sqrt{T}}$, we can get:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}[g(\theta_t)] - g(\theta^*) \leq \mathcal{O}\left(\frac{1}{(1-\gamma)^2 \sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^2 \sqrt{K}}\right).$$

The full proof of Theorem 1 is presented in Appendix A and we only give a concise proof sketch below due to space limit:

The proof can be decomposed into two parts. Firstly, $\widetilde{\alpha}_k^t$ is updated to $\widetilde{\alpha}_t$ in K iterations using Adam methods and we can measure the convergence through the term $\|\nabla_{\widetilde{\alpha}} F(\theta_t, \widetilde{\alpha}_t)\|^2$; Secondly, parameter θ is updated by TRPO when $\widetilde{\alpha}_t$ is already chosen. So eventually we can measure the global convergence starting from $\mathbb{E}[g(\theta_t)] - g(\theta^*)$ involving above two parts.

Theorem 1 demonstrates that LGAIL attains ϵ-global convergence with the convergence rate $\mathcal{O}(\frac{1}{(1-\gamma)^2\sqrt{T}})$ and convergence complexity $TK = \widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$ when we set $T = \mathcal{O}(\frac{1}{\epsilon^2})$ and $K = \mathcal{O}(\frac{1}{\epsilon^2})$.

Remark 1. *(Safety) On the one hand, θ^* satisfies the Karush-Kuhn-Tucker (KKT) condition of the Lagarange function in Eq. (4) that $J_C(\pi_{\theta^*}) - d_0 - \frac{\partial \psi(\alpha,\lambda)}{\partial \lambda} \leq 0$. In reality, the term $\frac{\partial \psi(\alpha,\lambda)}{\partial \lambda}$ is relatively small so that it can be omitted into $J_C(\pi_{\theta^*}) \leq d_0$. On the other hand, we have measured the non-asymptotical convergence complexity according to Theorem 1, which means the output policy $\bar{\theta}$ converges to θ^* so that the safety can be assumed to be guaranteed.*

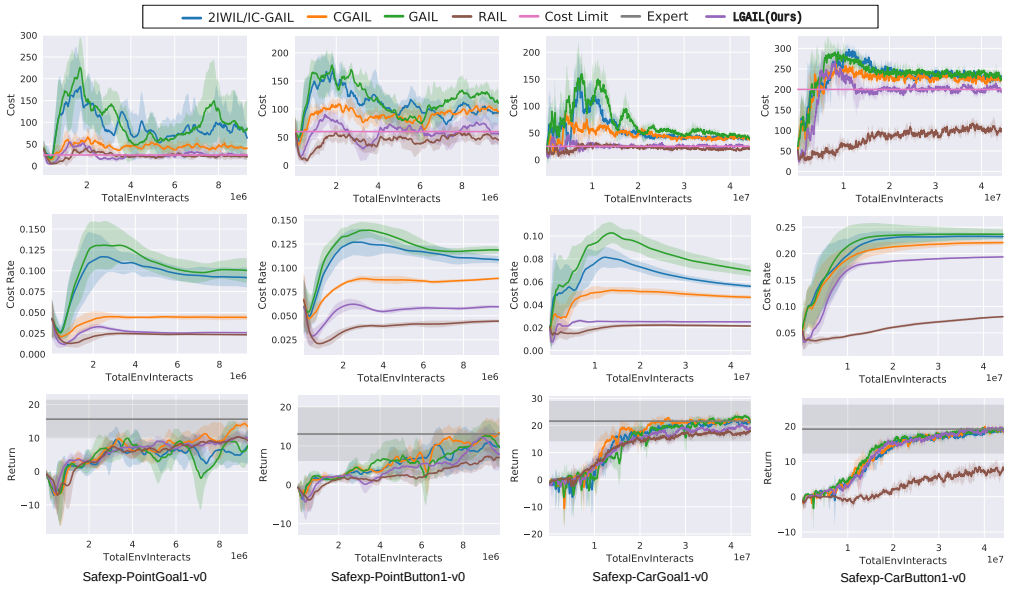


Figure 2: Learning curves of LGAIL and the other baselines on Safety Gym benchmarks. Performance is measured with Cost, Cost Rate, and Return. The x-axis represents time steps of interactions with the environment. Each algorithm is evaluated with 5 random seeds.

5 EXPERIMENTS

We investigate whether our algorithm LGAIL is able to solve the new safe IL task in this paper, *i.e.*, whether LGAIL has the ability to produce safety-prescribed expert behaviors from expert data with the cost signal from the environment. We introduce our experiments from two aspects, setups (Subsection 5.1) and results (Subsection 5.2).

5.1 SETUPS

In the experiments, we adopt six standard Safety Gym environments (Ray et al., 2019) to demonstrate the ability of LGAIL. In terms of robots, we use Point, Car, and Doggo; in terms of tasks, Goal and Button are employed. The level of difficulty of the employed environments is set to 1. More details on environments and expert data are in Appendix B.

Baselines. The safe IL task in this paper is constructed for the first time, so there are few corresponding baselines to compare. We select one representative IL algorithm, GAIL (Ho & Ermon, 2016), to serve as the baseline. The safe IL algorithm RAIL that imposes a fixed large penalty on dangerous state-action pairs is employed (Bhattacharyya et al., 2019; 2020). We also construct a comparable baseline by combing the safe RL method CPO (Achiam et al., 2017) with GAIL and name it as CGAIL. In addition, we relax the exact problem formulation of LGAIL to compare with IL algorithms of learning from imperfect data (2IWIL and IC-GAIL) (Wu et al., 2019). Specifically, we merge 2IWIL and IC-GAIL as 2IWIL/IC-GAIL and conduct experiments with expert trajectories whose cumulative costs are smaller than the prescribed safety constraint, which is equivalent to GAIL with safe expert data). More details on RAIL and 2IWIL/IC-GAIL are in Appendix C.

Metrics. To comprehensively measure the performance of all algorithms, three metrics are employed, *i.e.*, Cost, Cost Rate, and Return. Cost $J_C(\pi_\theta)$ is the average episodic sum of costs, while Return $J_R(\pi_\theta)$ is the average episodic return. Cost Rate is the rate that can be obtained by dividing the total sum of costs of the whole training process by the total number of agent-environment interactions. Cost and Cost Rate are related to the safety of the agent: the smaller they are, the safer the agent is considered. Although both metrics are related to safety, Cost focuses on measuring the current safety of a policy, while Cost Rate emphasizes the safety of whole training process. Hence, Cost Rate could be interpreted as a metric for the training safety to some extent. Return is used to evaluate the performance of mimicking the expert.

Table 1: Summary of quantitative results. The columns represent the algorithms, while the rows represent environments and metrics. Each result is averaged over 30 trails of a policy. Cost limits are presented in brackets under the environment names. Gray color indicates the costs of methods exceed the cost limit.

Environment		**LGAIL(ours)**	GAIL	2IWIL/IC-GAIL	RAIL	CGAIL
PointGoal1-v0 (Cost Limit:25)	Cost	22.5±4.2	86.5±72.1	63.8±18.5	20.6±6.5	40.4±5.2
	Cost Rate	0.026	0.1	0.092	0.023	0.044
	Return	**10.1±2.0**	7.7±4.6	7.6±5.4	9.2±1.2	13.4±2.4
PointButton1-v0 (Cost Limit:60)	Cost	55.2±17.6	110.6±21.7	93.9±18.3	44.8±9.5	96.1±7.5
	Cost Rate	0.059	0.119	0.108	0.044	0.089
	Return	**7.8±3.9**	9.6±5.3	10.0±2.8	7.0±2.5	13.2±3.8
CarGoal1-v0 (Cost Limit:25)	Cost	22.0±3.1	39.0±4.1	40.8±6.3	21.5±2.0	39.4±4.9
	Cost Rate	0.025	0.07	0.056	0.021	0.046
	Return	**19.0±1.8**	21.6±1.2	21.6±1.0	18.1±1.3	21.7±1.6
CarButton1-v0 (Cost Limit:200)	Cost	193.6±9.9	228.0±9.8	228.6±10.0	104.2±16.8	227.0±9.1
	Cost Rate	0.194	0.237	0.232	0.081	0.221
	Return	**18.8±0.6**	19.0±1.1	18.6±1.1	8.1±1.9	18.9±1.0

5.2 RESULTS

In this subsection, we present the experiment results of the proposed algorithm–LGAIL. Learning curves of four environments are presented in Figure 2, while quantitative results are in Table 1. More experiment results are deferred to Appendix D. From Figure 2 and Table 1, it is clear that LGAIL is able to reproduce a safe policy that can satisfy the prescribed safety constraints with comparable performance in imitating the expert.

Safety. It can be seen that LGAIL can achieve much lower Cost and lower Cost Rate compared to other baselines except RAIL, meaning the safety has both been improved during training and at the end of training. Surprisingly, CGAIL, which directly solves the constrained optimization problem of safe IL, performs poorly and even achieve similar costs to GAIL and 2IWIL/IC-GAIL in some environments as in (Ray et al., 2019). In contrast, using a large penalty, RAIL satisfies the cost limit in all environments at the expense of noticeable performance degradation in Return. And it could be laborious to search a suitable penalty scale for different cost limits and environments. LGAIL is able to adaptively drive the learning process to generate an agent that can satisfy the prescribed safety constraint. Compared to GAIL, 2IWIL/IC-GAIL slightly improves the safety measured by Cost and Cost Rate because 2IWIL/IC-GAIL tries to learn from safe expert data. These results verify that a portion of unsafe expert data could cause a negative impact on the safety of IL algorithms, but also indicates that GAIL fails to recover a safe policy from safe expert data, which is further discussed in the Appendix D.6. Besides, we test the performance of the rollout policy of each algorithm with 100 trajectories in Safexp-PointGoal1-v0. The proportion of trajectories that satisfies the cost limit for LGAIL, GAIL, 2IWIL/IC-GAIL, RAIL, and CGAIL is 70%, 30%, 31%, 69%, and 40%, respectively. It is clear that the proportion of safe trajectories of LGAIL during testing is dramatically improved compared against GAIL, 2IWIL/IC-GAIL, and CGAIL, meaning that LGAIL is much safer than the others. Besides, based on learning from an expectation of costs, an individual episodes sampled from the algorithms might exceed the cost limit, which is also observed in experts and Yang et al. (2021).

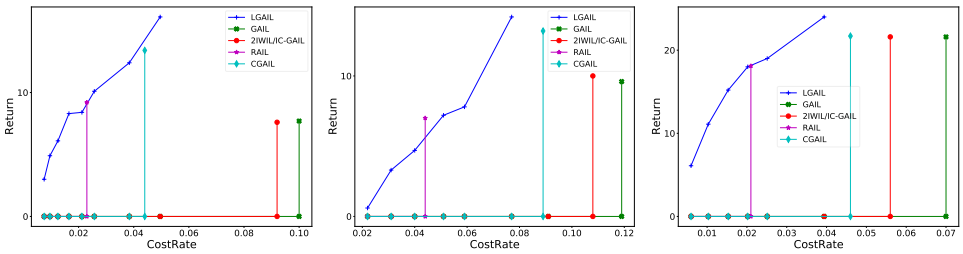


Figure 3: Undiscounted Return vs CostRate. From left to right, the tested environments are Safexp-PointGoal1-v0, Safexp-PointButton1-v0, and Safexp-CarGoal1-v0, respectively.

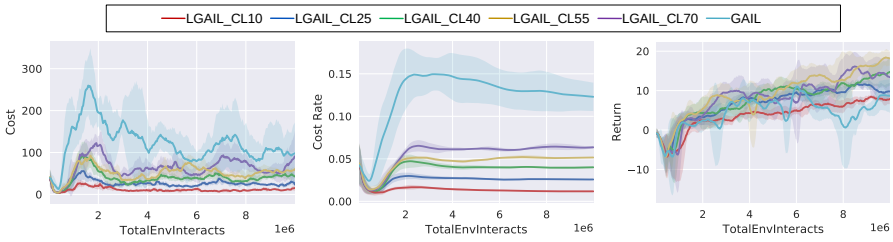


Figure 4: Impact of the cost limit on LGAIL with unsafe expert data. In the legend LGAIL_CL{x}, x represents the cost limit d_0.

Return. It is clear that, with little sacrifice in Return, LGAIL obtains a notably safer agent compared against other baselines except RAIL. Although RAIL achieves lower Cost compared to LGAIL, RAIL obtains the lowest return among all algorithms, meaning that RAIL fails to imitate the expert. It is surprising that CGAIL achieves high Return with lower Cost compared to GAIL. In complex tasks, LGAIL performs slightly worse than GAIL and 2IWIL/IC-GAIL. We think that there are two possible reasons: (1) to keep safe, the agent in LGAIL should try to avoid and to keep away from dangerous areas. This means that the agent should travel the long way around, resulting in a decrease in the rewards of LGAIL in fixed steps. On the contrary, GAIL and 2IWIL/IC-GAIL do not take safety into consideration, so they can walk across dangerous areas to achieve higher rewards; (2) LGAIL adaptively seeks a balance between rewards and costs, and adopts a more conservative exploration strategy, leading to marginal performance degradation. When a policy is unsafe, the Lagrange multiplier will increase and penalize the policy to ensure safety. Therefore, LGAIL would take actions that are more conservative when it explores in the environment. In complex environments, exploration is important for discovering better policies. Although LGAIL might perform marginally worse than GAIL and CGAIL in complex environments regarding Return, the agent of LGAIL can achieve the prescribed safety performance, which is paramount in safety-critical environments when deploying IL algorithms.

Furthermore, we test LGAIL's ability of generating constraint-satisfied policies, whose results are shown in Figures 3 and 4. In Figure 4, an extreme case where expert data are unsafe is tested in environment Safexp-PointGoal1-v0. We employ 15 unsafe expert trajectories, with their Cost 69.5 ± 15.3 and Return 18.1 ± 2.4. In particular, we adjust the cost limit d_0 from 10 to 70 to investigate its impact on safety and reward performance. More results are deferred to Appendix D.5. From Figure 3, we can see that LGAIL can achieve higher Returns with smaller Cost Rates, meaning that the total number of safety violations during training is significantly reduced. From the perspective of Cost, LGAIL is able to obtain a safe agent that satisfies the prescribed safety constraint with even unsafe expert data, whereas traditional safe IL cannot. Namely, given a fixed d_0 no matter it is large or small before training, LGAIL is able to reproduce a policy such that $J_C(\pi_\theta) \leq d_0$. With the decrease in the cost limit, the performance of the agent after training decreases slightly. Even if the safety of the agent in LGAIL has been improved dramatically, the performance of LGAIL is comparable to that of GAIL.

6 Conclusion

In this paper, a new but more practical safe IL task is constructed, in which an agent is required to achieve prescribed safety performance with the cost signal from the environment and a single expert dataset. To conduct safe IL, we develop a two-stage optimization framework, dubbed LGAIL, which can successfully imitate the expert and adaptively produce policies that satisfy the prescribed safety constraint. LGAIL turns the constrained safe IL problem into a corresponding unconstrained one with a Lagrange multiplier. The effectiveness and performance are illustrated in extensive OpenAI Safety Gym benchmarks, meaning that our algorithm is able to deal with the new safe IL task. In addition, the safety of agents during training is also enhanced dramatically compared to the baselines. Although the training safety of LGAIL is significantly enhanced, LGAIL fails to strictly maintain the safety of agents during training. A promising future direction would be achieving the training safety in safe IL.

REPRODUCIBILITY STATEMENT

We acknowledge the importance of reproducibility for research work and try whatever we can to ensure the reproducibility of our work. From the theoretical aspect, we clearly explain the employed assumptions in Subsection 4.3, and the detailed proof of our theorem is presented in Appendix A. From the empirical aspect, we first introduce the environments used in detail in Appendix B. Since we are investigating a new safe imitation learning task, there is no existing data to conduct experiments. Hence, we present how we obtain expert data for this new task in Appendix B. As for the implementation of our algorithm, details such as hyperparameters are provided in Appendix C. Finally, we introduce error bars as well as the computing resources in Appendix D. Our codes and data will be released upon publication.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *arXiv preprint arXiv:1705.10528*, 2017.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, pp. 103500, 2021.

M. Bain and C. Sammut. A framework for behavioural cloning. *In Machine Intelligence*, 15:103–129, 1995.

Chris F Best. Even experts make mistakes. *Risk Management*, 39(1):48–50, 1992.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Raunak Bhattacharyya, Blake Wulfe, Derek Phillips, Alex Kuefler, Jeremy Morton, Ransalu Senanayake, and Mykel Kochenderfer. Modeling human driving behavior through generative adversarial imitation learning. *arXiv preprint arXiv:2006.06412*, 2020.

Raunak P Bhattacharyya, Derek J Phillips, Changliu Liu, Jayesh K Gupta, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Simulating emergent properties of human driving behavior using multi-agent reward augmented imitation learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 789–795. IEEE, 2019.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *Journal of medical Internet research*, 20(9):e11510, 2018.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Congliang Chen, Li Shen, Fangyu Zou, and Wei Liu. Towards practical adam: Non-convexity, convergence theory, and mini-batch acceleration. *arXiv preprint arXiv:2101.05471*, 2021.

Wei Ming Dan Chia, Sye Loong Keoh, Cindy Goh, and Christopher Johnson. Risk assessment methodologies for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in neural information processing systems*, pp. 8092–8101, 2018.

Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

Forrest M Council, David L Harkey, Daniel T Nabors, Asad J Khattak, and Yusuf M Mohamedshah. Examination of fault, unsafe driving acts, and total harm in car-truck collisions. *Transportation research record*, 1830(1):63–71, 2003.

Phil F Culverhouse, Robert Williams, Beatriz Reguera, Vincent Herry, and Sonsoles González-Gil. Do experts make mistakes? a comparison of human and machine indentification of dinoflagellates. *Marine ecology progress series*, 247:17–25, 2003.

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. `https://github.com/openai/baselines`, 2017.

Ziwei Guan, Tengyu Xu, and Yingbin Liang. When will generative adversarial imitation learning algorithms attain global convergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 1117–1125. PMLR, 2021.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pp. 4565–4573, 2016.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Ahmed Hussein, Eyad Elyan, Mohamed Medhat Gaber, and Chrisina Jayne. Deep imitation learning for 3d navigation tasks. *Neural computing and applications*, 29(7):389–404, 2018.

Ashkan B Jeddi, Nariman L Dehghani, and Abdollah Shafieezadeh. Lyapunov-based uncertainty-aware safe reinforcement learning. *arXiv preprint arXiv:2107.13944*, 2021.

Manuel Koschuch, Walter Sebron, Zsolt Szalay, Árpád Török, Hans Tschiürtz, and István Wahl. Safety & security in the context of autonomous driving. In *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*, pp. 1–7. IEEE, 2019.

Emanuele Lattanzi and Valerio Freschi. Machine learning techniques to identify unsafe driving behavior by means of in-vehicle sensor data. *Expert Systems with Applications*, 176:114818, 2021.

Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, pp. 3812–3822, 2017.

Yongshuai Liu, Avishai Halev, and Xin Liu. Policy learning with constraints in model-free reinforcement learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

Zuxin Liu, Hongyi Zhou, Baiming Chen, Sicheng Zhong, Martial Hebert, and Ding Zhao. Safe model-based reinforcement learning with robust cross-entropy method. *arXiv preprint arXiv:2010.07968*, 2020.

Enrico Marchesini, Davide Corsi, and Alessandro Farinelli. Exploring safer behaviors for deep reinforcement learning. AAAI, 2022.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5041–5048, 2019. doi: 10.1109/IROS40897.2019.8968287.

Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos A Theodorou, and Byron Boots. Imitation learning for agile autonomous driving. *The International Journal of Robotics Research*, 39(2-3):286–302, 2020.

Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

Blaine T Phillips and Scott A Shikora. The history of metabolic and bariatric surgery: development of standards for patient safety and efficacy. *Metabolism*, 79:97–107, 2018.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019.

S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, 2011.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

Ari Seff and Jianxiong Xiao. Learning from maps: Visual common sense for autonomous driving. *arXiv preprint arXiv:1611.08583*, 2016.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

MyungJae Shin and Joongheon Kim. Adversarial imitation learning via random search. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Aman Sinha, Matthew O'Kelly, Russ Tedrake, and John C Duchi. Neural bridge sampling for evaluating safety-critical autonomous systems. *Advances in Neural Information Processing Systems*, 33, 2020.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. *arXiv preprint arXiv:2007.03964*, 2020.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Voot Tangkaratt, Bo Han, Mohammad Emtiyaz Khan, and Masashi Sugiyama. Variational imitation learning with diverse-quality demonstrations. In *International Conference on Machine Learning*, pp. 9407–9417. PMLR, 2020.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

Henriette Wallén Warner and Lars Åberg. Drivers' beliefs about exceeding the speed limits. *Transportation research part F: traffic psychology and behaviour*, 11(5):376–389, 2008.

Lei Wu, Rong Jin, Steven Hoi, Jianke Zhu, and Nenghai Yu. Learning bregman distance functions and its application for semi-supervised clustering. *Advances in neural information processing systems*, 22, 2009.

Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. *arXiv preprint arXiv:1901.09387*, 2019.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *AAAI*, pp. 10639–10646, 2021.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *ICLR*, 2020.

Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. *arXiv preprint arXiv:2006.07364*, 2020.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:1605.06450*, 2016.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33:15338–15349, 2020.

A PROOF OF THEOREM

In this part, we present the detailed proof of Theorem 1, which demonstrates LGAIL's convergence and safety. First, some necessary lemmas for the analysis are introduced.

A.1 BASIC CONCEPTS

For better derivation, we introduce some basic concepts. Recall the definition of expected discounted reward:

$$V(\pi, r) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim \zeta, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t)].$$

$V(\pi, r)$ explicitly emphasizes its dependence on the reward function $r(s, a)$, which is important in GAIL. We also name it as average value function. Besides, $V(\pi, r)$ can be calculated in a distribution manner

$$V(\pi, r) = \frac{1}{1-\gamma} \mathbf{E}_{(s,a)\sim\nu_\pi}[r(s, a)],$$

where $\nu_\pi(s, a)$ is referred to as the state-action visitation distribution, and it is defined as $\nu_\pi(s, a) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a)$. To proceed our analysis, we also need to define the average cost function in the same way,

$$J_C(\pi) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 \sim \zeta, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t)] = \frac{1}{1-\gamma} \mathbf{E}_{(s,a)\sim\nu_\pi}[c(s, a)].$$

Note that we do not explicitly stress the cost function $c(s, a)$ in $J_C(\pi)$ because $c(s, a)$ is provided by the environment as in Assumption 2. As for the accumulated reward, starting from a given state s or state-action pair (s, a), respectively, we give notations below:

$$v_{r_\alpha}^{\pi_\theta}(s) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t r_\alpha(s_t, a_t) | s_0 = s, a_t \sim \pi_\theta(\cdot|s_t), s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t)]$$

$$Q_{r_\alpha}^{\pi_\theta}(s, a) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t r_\alpha(s_t, a_t) | s_0 = s, a_0 = a, s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t), a_t \sim \pi_\theta(\cdot|s_t)].$$

In a similar fashion, we specify the accumulated cost starting from a given state or state-action pair as follows,

$$v_c^{\pi_\theta}(s) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 = s, a_t \sim \pi_\theta(\cdot|s_t), s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t)]$$

$$Q_c^{\pi_\theta}(s, a) = \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 = s, a_0 = a, s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t), a_t \sim \pi_\theta(\cdot|s_t)].$$

A.2 PROOF SKETCH

To make our proof easier to comprehend, we first provide a proof sketch before we expand our analysis.

proof sketch. Recalling our two-stage algorithm 1, we can decompose the convergence analysis into two parts: the Adam-maximization on parameter $\widetilde{\alpha}$ and the TRPO-minimization on parameter θ. Starting from the ϵ-global convergence's definition 1:

$$\mathbf{E}[g(\theta_t)] - g(\theta^*) \leq \mathbf{E}[g(\theta_t) - F(\theta_t, \widetilde{\alpha}_t)] + \mathbf{E}[V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})] + \lambda_t \mathbf{E}[J_C(\pi_{\theta_t}) - J_C(\pi_{\theta^*})]. \tag{6}$$

The last two terms in inequality 6 are related to parameter θ, whose update follows:

$$\sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s) \mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s), \pi_{\theta_0}(\cdot|s)) - B_w(\pi_{\theta^*}(\cdot|s), \pi_{\theta_T}(\cdot|s))].$$

Using the optimal condition and making a summation we can get (detailed derivation can refer to Eq. (14)):

$$\begin{aligned}&\frac{\eta_t^2 C_Q^2}{2}+\sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_t}(\cdot|s))-\sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)\mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_{t+1}}(\cdot|s))|\mathcal{F}_t] \\ &\geq \eta_t(1-\gamma)(V(\pi_{\theta^*},r_{\alpha_t})-V(\pi_{\theta_t},r_{\alpha_t}))-\lambda_t\eta_t(1-\gamma)(J_C(\pi_{\theta^*})-J_C(\pi_{\theta_t})).\end{aligned} \tag{7}$$

As for the first term $\mathbf{E}[g(\theta_t)-F(\theta_t,\widetilde{\alpha}_t)]$ in inequality 6, it is associated with parameter α updated by the Adam optimizer. And it can be scaled to $\frac{1}{2\mu}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_t)\|^2$ using the strong concavity of F on parameter (which can refer to 15).

So our focus turns to the term $\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_t)\|^2$. We can bound $\mathbf{E}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_t)\|^2 \leq \frac{C}{\sqrt{K}}$ by Lemma 7, where the constant C is presented specifically in the proof of Lemma 7. Combining above two parts into the original inequality 6 and summing up from $t=0$ to $T-1$, some terms in the left side of inequality 7 can be eliminated into

$$\sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)\mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_0}(\cdot|s))-B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_T}(\cdot|s))].$$

So finally we can get the average convergence result $\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}[g(\theta_t)]-g(\theta^*)\leq\mathcal{O}(\frac{1}{(1-\gamma)^2\sqrt{T}})+\mathcal{O}(\frac{1}{(1-\gamma)^2\sqrt{K}})$ as shown in Theorem 1.

And in the following sections, we will give the whole proof. □

A.3 Continuity Properties of the Objective Function $F(\theta,\widetilde{\alpha})$

We first show the derivatives of $F(\theta,\widetilde{\alpha})$ on each component and represent them in the following lemma.

Lemma 1. *The derivatives of the optimization problem $F(\theta,\widetilde{\alpha})$ can be calculated as follows:*

$$\nabla_\theta F(\theta,\widetilde{\alpha})=\left[-\frac{1}{1-\gamma}d_{\pi_\theta}(s)Q_{r_{\widetilde{\alpha}}}^{\pi_\theta}(s,a)+\frac{\lambda}{1-\gamma}d_{\pi_\theta}(s)Q_c^{\pi_\theta}(s,a)\right]_{|\mathcal{S}|\times|\mathcal{A}|},(s,a)\in\mathcal{S}\times\mathcal{A}$$

$$\nabla_{\widetilde{\alpha}}F(\theta,\widetilde{\alpha})=(\nabla_\alpha F(\theta,\alpha,\lambda),\nabla_\lambda F(\theta,\alpha,\lambda)),$$

where $d_{\pi_\theta}(s)$ is the normalized stationary state distribution, and we specify the detailed expression:

$$\nabla_\alpha F(\theta,\alpha,\lambda)_i=\frac{1}{1-\gamma}[\sum_{s,a}(\nu_{\pi_E}(s,a)-\nu_{\pi_\theta}(s,a))\frac{\partial r_\alpha(s,a)}{\partial\alpha_i}]-\frac{\partial\psi(\alpha,\lambda)}{\partial\alpha_i}$$

$$\nabla_\lambda F(\theta,\alpha,\lambda)=J_C(\pi_\theta)-d_0-\frac{\partial\psi(\alpha,\lambda)}{\partial\lambda}.$$

Proof. The derivative of $F(\theta,\widetilde{\alpha})$ regarding θ is

$$\nabla_\theta F(\theta,\widetilde{\alpha})=-\nabla_\theta V(\pi_\theta,r_\alpha)+\lambda\nabla_\theta J_C(\pi_\theta).$$

We need to calculate two derivatives $\nabla_\theta V(\pi_\theta,r_\alpha)$ and $\nabla_\theta J_C(\pi_\theta)$, respectively. Based on the policy gradient theorem (Sutton et al., 1999), we obtain

$$\nabla_\theta V(\pi_\theta,r_\alpha)=\frac{1}{1-\gamma}\nabla_\theta\mathbf{E}_{(s,a)\sim v_{\pi_\theta}}[r_\alpha(s,a)]=\frac{1}{1-\gamma}\sum_{s\in S}d_{\pi_\theta}(s)\sum_{a\in A}\nabla_\theta\pi_\theta(a|s)Q_{r_\alpha}^{\pi_\theta}(s,a).$$

For any entry $\theta_{s,a}$, $\nabla_\theta V(\pi_\theta,r_\alpha)|_{s,a}=\frac{1}{1-\gamma}\frac{\partial\mathbf{E}_{(s,a)\sim v_{\pi_\theta}}[r_\alpha(s,a)]}{\partial\theta_{s,a}}=\frac{1}{1-\gamma}d_{\pi_\theta}(s)Q_{r_\alpha}^{\pi_\theta}(s,a)$. In a similar manner, we get

$$\nabla_\theta J_C(\pi_\theta)=\frac{1}{1-\gamma}\nabla_\theta\mathbf{E}_{(s,a)\sim v_{\pi_\theta}}[c(s,a)]=\frac{1}{1-\gamma}\sum_{s\in S}d_{\pi_\theta}(s)\sum_{a\in A}\nabla_\theta\pi_\theta(a|s)Q_c^{\pi_\theta}(s,a)$$

$$\nabla_\theta J_C(\pi_\theta)|_{s,a}=\frac{1}{1-\gamma}\frac{\partial\mathbf{E}_{(s,a)\sim v_{\pi_\theta}}[c(s,a)]}{\partial\theta_{s,a}}=\frac{1}{1-\gamma}d_{\pi_\theta}(s)Q_c^{\pi_\theta}(s,a).$$

Hence,

$$\nabla_\theta F(\theta,\widetilde{\alpha}) = \left[-\frac{1}{1-\gamma}d_{\pi_\theta}(s)Q^{\pi_\theta}_{r_{\widetilde{\alpha}}}(s,a) + \frac{\lambda}{1-\gamma}d_{\pi_\theta}(s)Q^{\pi_\theta}_{c}(s,a)\right]_{|\mathcal{S}|\times|\mathcal{A}|}, \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

As for the derivatives of $F(\theta,\widetilde{\alpha})$ on each component of $\widetilde{\alpha}$, we can calculate in a direct way. □

Lemma 2. *(Xu et al. (2020), Lemma 3) The state-action distribution ν_{π_θ} is C_ν-Lipschitz smooth on parameter θ,* i.e.*, given any $\theta_1, \theta_2 \in \Theta_p$:*

$$\|\nu_{\pi_{\theta_1}} - \nu_{\pi_{\theta_2}}\|_{TV} \leq C_\nu \|\theta_1 - \theta_2\|_2,$$

where $C_\nu = \frac{\sqrt{|\mathcal{A}|}}{2}(1 + \lceil \log_\rho C_M^{-1} \rceil + (1-\rho)^{-1})$.

Lemma 3. *$F(\theta,\widetilde{\alpha})$ is Lipschitz smooth on both parameters θ and $\widetilde{\alpha}$:*

$$\begin{aligned}\|\nabla_\theta F(\theta_1,\widetilde{\alpha}_1) - \nabla_\theta F(\theta_2,\widetilde{\alpha}_2)\|_2 &\leq L_{11}\|\theta_1-\theta_2\|_2 + L_{12}\|\widetilde{\alpha}_1-\widetilde{\alpha}_2\|_2\\ \|\nabla_{\widetilde{\alpha}} F(\theta_1,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_2)\|_2 &\leq L_{21}\|\theta_1-\theta_2\|_2 + L_{22}\|\widetilde{\alpha}_1-\widetilde{\alpha}_2\|_2,\end{aligned}$$

where $L_{21} = \frac{2C_\nu\sqrt{C_r^2+1}}{1-\gamma}$ *and* $L_{22} = (\frac{2\sqrt{q}L_r}{1-\gamma} + L_\psi)$.

For the derivation of L_{11} and L_{12}, we refer readers to Guan et al. (2021) with consideration of a new term $\lambda(J_C(\pi_\theta) - d_0)$. During our derivation, only L_{21} and L_{22} are employed. Hence, we present how to derive L_{21} and L_{22}.

Proof. We prove the second inequality in the following way:

$$\begin{aligned}&\|\nabla_{\widetilde{\alpha}} F(\theta_1,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_2)\|_2\\ =\ &\|\nabla_{\widetilde{\alpha}} F(\theta_1,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_1) + \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_2)\|_2\\ \leq\ &\|\nabla_{\widetilde{\alpha}} F(\theta_1,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_1)\|_2 + \|\nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_2)\|_2\end{aligned}$$

We denote $T_1 = \|\nabla_{\widetilde{\alpha}} F(\theta_1,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_1)\|_2$ and $T_2 = \|\nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}} F(\theta_2,\widetilde{\alpha}_2)\|_2$.

Upper-bound of T_1:

We first consider the derivative on the i-th component of α:

$$\begin{aligned}&|(\nabla_\alpha F(\theta_1,\alpha_1,\lambda_1) - \nabla_\alpha F(\theta_2,\alpha_1,\lambda_1))_i|\\ =\ &|(\nabla_\alpha V(\pi_E, r_{\alpha_1}) - \nabla_\alpha V(\pi_{\theta_1}, r_{\alpha_1}) - \frac{\partial\psi(\alpha_1,\lambda_1)}{\partial\alpha} - (\nabla_\alpha V(\pi_E, r_{\alpha_1}) - \nabla_\alpha V(\pi_{\theta_2}, r_{\alpha_1}) - \frac{\partial\psi(\alpha_1,\lambda_1)}{\partial\alpha}))_i|\\ =\ &|(\nabla_\alpha V(\pi_{\theta_2}, r_{\alpha_1}) - \nabla_\alpha V(\pi_{\theta_1}, r_{\alpha_1}))_i|\\ =\ &\frac{1}{1-\gamma}|\sum_{s,a}(\nu_{\pi_{\theta_1}}(s,a) - \nu_{\pi_{\theta_2}}(s,a))(\nabla_\alpha r_{\alpha_1})_i|\\ \leq\ &\frac{1}{1-\gamma}|\sum_{s,a}(\nu_{\pi_{\theta_1}}(s,a) - \nu_{\pi_{\theta_2}}(s,a))|\|(\nabla_\alpha r_{\alpha_1})_i\|_\infty\\ \leq\ &\frac{\|\nu_{\pi_{\theta_1}} - \nu_{\pi_{\theta_2}}\|_1\|\frac{\partial r_\alpha}{\partial\alpha_i}\|_\infty}{1-\gamma}\\ \leq\ &\frac{2C_\nu\|\theta_1-\theta_2\|_2\|\frac{\partial r_\alpha}{\partial\alpha_i}\|_\infty}{1-\gamma},\end{aligned}$$

where the last inequality is due to: $\|\nu_{\pi_{\theta_1}} - \nu_{\pi_{\theta_2}}\|_1 = 2\|\nu_{\pi_{\theta_1}} - \nu_{\pi_{\theta_2}}\|_{TV} \leq 2C_\nu\|\theta_1-\theta_2\|_2$.

Next, the derivative on the Lagrange multiplier λ:

$$\begin{aligned}
&|\nabla_\lambda F(\theta_1,\alpha_1,\lambda_1)-\nabla_\lambda F(\theta_2,\alpha_1,\lambda_1)|\\
&=|J_C(\pi_{\theta_1})-J_C(\pi_{\theta_2})|\\
&=\frac{1}{1-\gamma}|\sum_{s,a}(\nu_{\pi_{\theta_1}}(s,a)-\nu_{\pi_{\theta_2}}(s,a))c(s,a)|\\
&\le\frac{1}{1-\gamma}|\sum_{s,a}(\nu_{\pi_{\theta_1}}-\nu_{\pi_{\theta_2}})|\\
&\le\frac{\|\nu_{\pi_{\theta_1}}-\nu_{\pi_{\theta_2}}\|_1}{1-\gamma}\\
&\le\frac{2C_\nu\|\theta_1-\theta_2\|_2}{1-\gamma}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\|\nabla_{\widetilde{\alpha}}F(\theta_1,\widetilde{\alpha}_1)-\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha_1})\|_2\\
&=\sqrt{\sum_{i=1}^{q}|(\nabla_\alpha F(\theta_1,\alpha_1,\lambda_1)-\nabla_\alpha F(\theta_2,\alpha_1,\lambda_1))_i|^2+|\nabla_\lambda F(\theta_1,\alpha_1,\lambda_1)-\nabla_\lambda F(\theta_2,\alpha_1,\lambda_1)|^2}\\
&\le\frac{2C_\nu\|\theta_1-\theta_2\|_2}{1-\gamma}\sqrt{\sum_{i=1}^{q}\|\frac{\partial r_\alpha}{\partial\alpha_i}\|_\infty^2+1}\\
&\le\frac{2C_\nu\sqrt{C_r^2+1}}{1-\gamma}\|\theta_1-\theta_2\|_2.
\end{aligned}$$

Upper-bound of T_2: Similarly, we start from component-wise derivative of $\widetilde{\alpha}$:

$$\begin{aligned}
&|(\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_1)-\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_2))_i|\\
&=|(\nabla_\alpha F(\theta_2,\alpha_1,\lambda_1)-\nabla_\alpha F(\theta_2,\alpha_2,\lambda_2),\nabla_\lambda F(\theta_2,\alpha_1,\lambda_1)-\nabla_\lambda F(\theta_2,\alpha_2,\lambda_2))_i|\\
&=|(\frac{1}{1-\gamma}\sum_{s,a}\nu_{\pi_E}(s,a)(\nabla_\alpha r_{\alpha_1}(s,a)-\nabla_\alpha r_{\alpha_2}(s,a))-\frac{1}{1-\gamma}\sum_{s,a}\nu_{\pi_{\theta_2}}(s,a)(\nabla_\alpha r_{\alpha_1}(s,a)-\nabla_\alpha r_{\alpha_2}(s,a))\\
&\quad-(\frac{\partial\psi(\alpha_1,\lambda_1)}{\partial\alpha}-\frac{\partial\psi(\alpha_2,\lambda_2)}{\partial\alpha}),-(\frac{\partial\psi(\alpha_1,\lambda_1)}{\partial\lambda}-\frac{\partial\psi(\alpha_2,\lambda_2)}{\partial\lambda}))_i|
\end{aligned}$$

As a result, the 2-norm is bounded as follows:

$$\begin{aligned}
&\|\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_1)-\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_2)\|_2\\
&=\sqrt{\sum_{i=1}^{q+1}|(\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_1)-\nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_2))_i|^2}\\
&\le\frac{1}{1-\gamma}(\sqrt{\sum_{i=1}^{q}|\sum_{s,a}\nu_{\pi_E}(s,a)(\nabla_\alpha r_{\alpha_1}(s,a)-\nabla_\alpha r_{\alpha_2}(s,a))_i|^2}\\
&\quad+\sqrt{\sum_{i=1}^{q}|\sum_{s,a}\nu_{\pi_{\theta_2}}(s,a)(\nabla_\alpha r_{\alpha_1}(s,a)-\nabla_\alpha r_{\alpha_2}(s,a))_i|^2})\\
&\quad+\sqrt{\sum_{i=1}^{q}(\frac{\partial\psi(\alpha_1,\lambda_1)}{\partial\alpha}-\frac{\partial\psi(\alpha_2,\lambda_2)}{\partial\alpha})^2+(\frac{\partial\psi(\alpha_1,\lambda_1)}{\partial\lambda}-\frac{\partial\psi(\alpha_2,\lambda_2)}{\partial\lambda})^2}\\
&\overset{(i)}{\le}\frac{2\sqrt{q}L_r}{1-\gamma}\|\alpha_1-\alpha_2\|_2+L_\psi\|\widetilde{\alpha}_1-\widetilde{\alpha}_2\|_2\\
&\le(\frac{2\sqrt{q}L_r}{1-\gamma}+L_\psi)\|\widetilde{\alpha}_1-\widetilde{\alpha}_2\|_2
\end{aligned}$$

where (i) follows from:$|(\nabla_\alpha r_{\alpha_1}(s,a) - \nabla_\alpha r_{\alpha_2}(s,a))_i| \leq \|\nabla_\alpha r_{\alpha_1}(s,a) - \nabla_\alpha r_{\alpha_2}(s,a)\|_2 \leq L_r\|\alpha_1 - \alpha_2\|_2$ and the last term is the expression of $\|\nabla_{\widetilde{\alpha}}\psi(\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}}\psi(\widetilde{\alpha}_2)\|_2$. With upper-bounds of T_1 and T_2, we have

$$\|\nabla_{\widetilde{\alpha}}F(\theta_1,\widetilde{\alpha}_1) - \nabla_{\widetilde{\alpha}}F(\theta_2,\widetilde{\alpha}_2)\|_2 \leq L_{21}\|\theta_1 - \theta_2\|_2 + L_{22}\|\widetilde{\alpha}_1 - \widetilde{\alpha}_2\|_2,$$

where $L_{21} = \frac{2C_\nu\sqrt{C_r^2+1}}{1-\gamma}$ and $L_{22} = (\frac{2\sqrt{q}L_r}{1-\gamma} + L_\psi)$.

□

A.4 Proof of Convergence of Adam

For better presentation, we define some notations (Chen et al., 2021):

1. Select a constant $\iota' > 0$ that satisfies $\beta^2 < \iota' < \iota$. Define $\delta := \beta^2/\iota' < 1$ and $C_1 = \prod_{j=1}^{N}(\frac{\iota_j}{\iota'})$, in which N stands for the maximal index j that $\iota_j < \iota'$.
2. Let $\Delta_k = \widetilde{\alpha}_{k+1} - \widetilde{\alpha}_k$, $\hat{v}_k = \iota_k v_{k-1} + (1-\iota_k)\sigma_k^2$ where $\sigma_k^2 = \mathbf{E}_k[h_k^2]$, and $\hat{\xi}_k = \frac{b_k}{\sqrt{\hat{v}_k}}$.
3. For the positive vector $\hat{\xi}_k$, we define the weighted norm as $\|v_k\|_{\hat{\xi}_k}^2 = \langle v_k, \hat{\xi}_k v_k\rangle = \sum_{i=1}^{q+1}\hat{\xi}_{k,i}|v_{k,i}|^2$.

Lemma 4. *(Chen et al. (2021), Lemma 33) Let $M_k = \mathbf{E}[-\langle\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_k),\Delta_k\rangle + L_{22}\|\Delta_k\|^2]$, and $\chi_k = b_k/\sqrt{1-\iota_k}$. Then $\forall K \geq 1$, we have:*

$$\sum_{k=1}^{K} M_k \leq C_3\mathbf{E}[\sum_{k=1}^{K}\chi_k\|\frac{\sqrt{1-\iota_k}h_k}{\sqrt{v_k}}\|^2] - \frac{1-\beta}{2}\mathbf{E}[\sum_{k=1}^{K}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_k)\|_{\hat{\xi}_k}^2],$$

where $C_3 = \frac{C_0}{\sqrt{C_1}(1-\sqrt{\delta})}(\frac{C_0^2\chi_1 L_{22}}{C_1(1-\sqrt{\delta})^2} + 2(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\delta)\iota_1}} + 1)^2 G)$.

Lemma 5. *(Chen et al. (2021), Lemma 35)The following estimate is hold:*

$$\mathbf{E}[\sum_{k=1}^{K}\chi_k\left\|\frac{\sqrt{1-\iota_k}h_k}{\sqrt{v_k}}\right\|^2] \leq C_0(q+1)[\chi_1\log(1+\frac{G^2}{\mu(q+1)}) + \frac{1}{\iota_1}\sum_{k=1}^{K}b_k\sqrt{1-\iota_k}].$$

Lemma 6. *(Chen et al. (2021), Lemma 36) We assume α_t is randomly chosen from $\{\alpha_1^t, ..., \alpha_K^t\}$ and add them up with equal probabilities $1/K$. Then, we obtain*

$$\frac{1}{K}\mathbf{E}\sum_{k=1}^{K}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_k^t)\|^2 \leq \frac{C_0\sqrt{G^2+\mu(q+1)}}{Kb_K}\mathbf{E}[\sum_{k=1}^{K}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_k^t)\|_{\hat{\xi}_k}^2].$$

Lemma 7. *When the parameters satisfy all requirements 1, 2, and 3, Adam is convergent and:*

$$\mathbf{E}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_t)\|^2 \leq \frac{C}{\sqrt{K}},$$

where

$$C = \frac{2C_0\sqrt{G^2+\mu(q+1)}}{b(1-\beta)}[2g^* + C_3C_0(q+1)b(\frac{1}{\sqrt{\iota}}\log(1+\frac{G^2}{\mu(q+1)}) + \frac{\sqrt{\iota}}{1-\iota})].$$

Proof. According to the Lipschitz smoothness of $F(\theta,\widetilde{\alpha})$, we have

$$F(\theta_t,\widetilde{\alpha}_k^t) \leq F(\theta_t,\widetilde{\alpha}_{k+1}^t) + \langle -\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_k^t), \widetilde{\alpha}_{k+1}^t - \widetilde{\alpha}_k^t\rangle + \frac{L_{22}}{2}\|\widetilde{\alpha}_{k+1}^t - \widetilde{\alpha}_k^t\|^2.$$

Taking expectation and summing up from $k=1$ to K, we can get $F(\theta_t,\widetilde{\alpha}_1^t) \leq \mathbf{E}[F(\theta_t,\widetilde{\alpha}_{K+1}^t)] + \sum_{k=1}^{K}M_k$. According to the definition of $g(\theta)$, we further have,

$$F(\theta_t,\widetilde{\alpha}_1^t) \leq g(\theta_t) + \sum_{k=1}^{K}M_k.$$

Hence, we can get:

$$
\begin{aligned}
&\mathbf{E}\|\nabla_{\widetilde{\alpha}} F(\theta_t, \widetilde{\alpha}_t)\|^2 \\
&= \frac{1}{K}\mathbf{E}\sum_{k=1}^{K}\|\nabla_{\widetilde{\alpha}} F(\theta_t, \widetilde{\alpha}_k^t)\|^2 \\
&\overset{(i)}{\leq} \frac{C_0\sqrt{G^2+\mu(q+1)}}{Kb_K}\mathbf{E}[\sum_{k=1}^{K}\|\nabla_{\widetilde{\alpha}} F(\theta_t, \widetilde{\alpha}_k^t)\|_{\hat{\xi}_k}^2] \\
&\overset{(ii)}{\leq} \frac{2C_0\sqrt{G^2+\mu(q+1)}}{Kb_K(1-\beta)}(C_3\mathbf{E}[\sum_{k=1}^{K}\chi_k\|\frac{\sqrt{1-\iota_k}h_k}{\sqrt{v_k}}\|^2] - \sum_{k=1}^{K}M_k) \\
&\overset{(iii)}{\leq} \frac{2C_0\sqrt{G^2+\mu(q+1)}}{Kb_K(1-\beta)}\{C_3C_0(q+1)[\chi_1\log(1+\frac{G^2}{\mu(q+1)}) + \frac{1}{\iota_1}\sum_{k=1}^{K}b_k\sqrt{1-\iota_k}] + 2g^*\},
\end{aligned}
$$

where (i) and (ii) follow from Lemmas 6 and 4, respectively, and (iii) is due to Lemma 5. Then we take $b_k = \frac{b}{\sqrt{K}}$, $\beta_k = \beta$, and $\iota_k = 1 - \frac{\iota}{K}$ which satisfies $\delta = \frac{\beta}{1-\frac{\iota}{K}} < 1$ and $\iota_k \geq \frac{1}{4}$. So we can get:

$$
\mathbf{E}\|\nabla_{\widetilde{\alpha}} F(\theta_t, \widetilde{\alpha}_t)\|^2 \leq \frac{C}{\sqrt{K}},
$$

where

$$
C = \frac{2C_0\sqrt{G^2+\mu(q+1)}}{b(1-\beta)}[2g^* + C_3C_0(q+1)b(\frac{1}{\sqrt{\iota}}\log(1+\frac{G^2}{\mu(q+1)}) + \frac{\sqrt{\iota}}{1-\iota})].
$$

□

A.5 Lemmas on Convergence of TRPO

Lemma 8. *For estimated Q-value and Q_c functions, they are upper bounded:*

$$
\|-\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda\hat{Q}_c^{\pi_{\theta_t}}(s,\cdot)\|_\infty \leq C_Q.
$$

Proof. For any state-action pair (s, a), the estimated Q-value for reward is bounded by $\frac{R_{max}}{1-\gamma}$ because of its calculating iteration (Guan et al., 2021) and the estimated Q-value for cost is bounded by $\frac{1}{1-\gamma}$ in the same sense. Hence,

$$
\|-\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda\hat{Q}_c^{\pi_{\theta_t}}(s,\cdot)\|_\infty \leq \frac{R_{max} + \lambda_{\max}}{1-\gamma} \triangleq C_Q.
$$

□

Lemma 9. *For any policy π, π', and any reward function r_α, the following equations hold:*

$$
(1-\gamma)(V(\pi, r_\alpha) - V(\pi', r_\alpha)) = \sum_{s\in\mathcal{S}} d_{\pi'}(s)\langle -Q_{r_\alpha}^{\pi}(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s)\rangle \tag{8}
$$

$$
(1-\gamma)(J_C(\pi') - J_C(\pi)) = \sum_{s\in\mathcal{S}} d_{\pi'}(s)\langle Q_c^{\pi}(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s)\rangle. \tag{9}
$$

Proof. Eq. (8) is a simplified case of Lemma 11 in Guan et al. (2021) when the regularized multiplier is chosen to be zero. Here, we only present the detailed proof of Eq. (9).

For any state $s \in \mathcal{S}$:

$$
\begin{aligned}
&\langle Q_c^{\pi}(s,\cdot), \pi'(\cdot|s)\rangle \\
&\triangleq \sum_a \pi'(a|s)Q_c^{\pi}(s,a) \\
&= \sum_a \pi'(a|s)\left(c(s,a) + \gamma\sum_{s'}\mathcal{T}(s'|s,a)v_c^{\pi}(s')\right) \\
&= T^{\pi'}v_c^{\pi}(s),
\end{aligned} \tag{10}
$$

where $\mathcal{T}^{\pi'}$ is the Bellman operator.

Combing with the fact that $\langle Q_c^\pi(s,\cdot), \pi(\cdot|s)\rangle = \sum_a Q_c^\pi(s,a)\pi(a|s) = v_c^\pi(s)$, we obtain:

$$\langle Q_c^\pi(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s)\rangle = T^{\pi'} v_c^\pi(s) - v_c^\pi(s). \tag{11}$$

Besides, for any state s, $v_c^\pi = c^\pi + \gamma \mathcal{T}^\pi v_c^\pi$, which can be written as $(\mathbf{I} - \gamma\mathcal{T}^\pi)v_c^\pi = c^\pi$. Hence,

$$\begin{aligned} v_c^{\pi'} - v_c^\pi &= (\mathbf{I} - \gamma\mathcal{T}^{\pi'})^{-1} c^{\pi'} - (\mathbf{I} - \gamma\mathcal{T}^{\pi'})^{-1}(\mathbf{I} - \gamma\mathcal{T}^{\pi'})v_c^\pi \\ &= (\mathbf{I} - \gamma\mathcal{T}^{\pi'})^{-1}(c^{\pi'} + \gamma\mathcal{T}^{\pi'} v_c^\pi - v_c^\pi) \\ &= (\mathbf{I} - \gamma\mathcal{T}^{\pi'})^{-1}(T^{\pi'} v_c^\pi - v_c^\pi). \end{aligned}$$

Multiplying both side by the state visitation distribution $d_{\zeta,\pi'} = (1-\gamma)\zeta(\mathbf{I} - \gamma\mathcal{T}^{\pi'})^{-1}$, we get:

$$\zeta(v_c^{\pi'} - v_c^\pi) = \frac{1}{1-\gamma} d_{\zeta,\pi'}(T^{\pi'} v_c^\pi - v_c^\pi). \tag{12}$$

Combining Eq. (10) with (12), we have

$$\begin{aligned} & J_C(\pi') - J_C(\pi) \\ &= \sum_s \zeta(s)(v_c^{\pi'}(s) - v_c^\pi(s)) \\ &= \frac{1}{1-\gamma}\sum_s d_{\zeta,\pi'}(s)(T^{\pi'} v_c^\pi(s) - v_c^\pi(s)) \\ &= \frac{1}{1-\gamma}\sum_s d_{\pi'}(s)\langle Q_c^\pi(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s)\rangle. \end{aligned}$$

Multiplying both side by fractor $1-\gamma$ and we can get the second equality. □

A.6 Proof of Theorem 1

We restate Theorem 1 here for better understanding.

Theorem. ***1*** *(Convergence) When assumptions 1- 8 are satisfied, parameters meet the requirements 1- 3, and the update stepsize of θ is $\eta_t = \frac{1-\gamma}{\sqrt{T}}$, we can get:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}[g(\theta_t)] - g(\theta^*) \le \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{K}}\right).$$

Proof. The parameter θ_{t+1} is updated according to the following law

$$\pi_{\theta_{t+1}}(\cdot|s) \in \arg\min_{\pi\in\Theta_p}(\left\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot), \pi - \pi_{\theta_t}(\cdot|s)\right\rangle + \eta_t^{-1} B_w(\pi, \pi_{\theta_t}(\cdot|s))).$$

Based on the optimal condition, we have

$$\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot) + \eta_t^{-1}(\nabla w(\pi_{\theta_{t+1}}(\cdot|s)) - \nabla w(\pi_{\theta_t}(\cdot|s))), \pi - \pi_{\theta_{t+1}}(\cdot|s)\rangle \ge 0$$

holds for any π.

Let $\pi = \pi_{\theta^*}(\cdot|s)$ in the above inequality, then

$$\begin{aligned} 0 \le\ & \eta_t\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\rangle \\ & + \eta_t\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot), \pi_{\theta_t}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\rangle \\ & + \langle \nabla w(\pi_{\theta_{t+1}}(\cdot|s)) - \nabla w(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\rangle \\ \le\ & \eta_t\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\rangle \\ & + \frac{\eta_t^2 \| -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot)\|_\infty^2}{2} + \frac{\|\pi_{\theta_t}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\|_1^2}{2} \\ & + B_w(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)) - B_w(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s)) - B_w(\pi_{\theta_{t+1}}(\cdot|s), \pi_{\theta_t}(\cdot|s)) \\ \overset{(i)}{\le}\ & \eta_t\langle -\hat{Q}_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_{\theta_t}}(s,\cdot), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\rangle + \frac{\eta_t^2 C_Q^2}{2} \\ & + B_w(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)) - B_w(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s)), \end{aligned} \tag{13}$$

where (i) is due to Lemma 8 and the following relationship:

$$\begin{aligned}\frac{\|\pi_{\theta_t}(\cdot|s)-\pi_{\theta_{t+1}}(\cdot|s)\|_1^2}{2} &= 2\delta(\pi_{\theta_t}(\cdot|s),\pi_{\theta_{t+1}}(\cdot|s))\\ &\le KL(\pi_{\theta_t}(\cdot|s)||\pi_{\theta_{t+1}}(\cdot|s))\\ &= B_w(\pi_{\theta_t}(\cdot|s),\pi_{\theta_{t+1}}(\cdot|s)).\end{aligned}$$

We take a conditional expectation on $\mathcal{F}_t=\sigma(\theta_0,\theta_1,...,\theta_t)$ over inequality 13 and obtain:

$$\begin{aligned}0 \le &\eta_t\langle -Q_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot)+\lambda_t\hat{Q}_c^{\pi_{\theta_t}}(s,\cdot),\pi_{\theta^*}(\cdot|s)-\pi_{\theta_t}(\cdot|s)\rangle+\frac{\eta_t^2C_Q^2}{2}\\ &+B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_t}(\cdot|s))-\mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_{t+1}}(\cdot|s))|\mathcal{F}_t].\end{aligned}$$

For any $s\in\mathcal{S}$, the above inequality holds. Hence, we take a summation on the distribution $s\sim d_{\pi_{\theta^*}}$ and have

$$\begin{aligned}&\frac{\eta_t^2C_Q^2}{2}+\sum_{s\in\mathcal{S}}d_{\pi_{\theta^*}}(s)B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_t}(\cdot|s))-\sum_{s\in\mathcal{S}}d_{\pi_{\theta^*}}(s)\mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_{t+1}}(\cdot|s))|\mathcal{F}_t]\\ \ge &-\eta_t\sum_{s\in\mathcal{S}}d_{\pi_{\theta^*}}(s)\langle -Q_{r_{\alpha_t}}^{\pi_{\theta_t}}(s,\cdot)+\lambda_tQ_c^{\pi_{\theta_t}}(s,\cdot),\pi_{\theta^*}(\cdot|s)-\pi_{\theta_t}(\cdot|s)\rangle \qquad (14)\\ \overset{(i)}{=} &\eta_t(1-\gamma)(V(\pi_{\theta^*},r_{\alpha_t})-V(\pi_{\theta_t},r_{\alpha_t}))-\lambda_t\eta_t(1-\gamma)(J_C(\pi_{\theta^*})-J_C(\pi_{\theta_t})),\end{aligned}$$

where (i) follows from Lemma 9.

Next we proceed our proof as follows:

$$\begin{aligned}&\mathbf{E}[g(\theta_t)]-g(\theta^*)\\ =&\mathbf{E}[g(\theta_t)-F(\theta_t,\widetilde{\alpha}_t)]+\mathbf{E}[F(\theta_t,\widetilde{\alpha}_t)-g(\theta^*)]\\ \le&\mathbf{E}[g(\theta_t)-F(\theta_t,\widetilde{\alpha}_t)]+\mathbf{E}[F(\theta_t,\widetilde{\alpha}_t)-F(\theta^*,\widetilde{\alpha}_t)]\\ =&\mathbf{E}[g(\theta_t)-F(\theta_t,\widetilde{\alpha}_t)]+\mathbf{E}[V(\pi_{\theta^*},r_{\alpha_t})-V(\pi_{\theta_t},r_{\alpha_t})]+\lambda_t\mathbf{E}[J_C(\pi_{\theta_t})-J_C(\pi_{\theta^*})]\\ \le&\mathbf{E}[g(\theta_t)-F(\theta_t,\widetilde{\alpha}_t)]+\frac{1}{\eta_t(1-\gamma)}\sum_{s\in\mathcal{S}}d_{\pi_{\theta^*}}(s)B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_t}(\cdot|s))\\ &-\frac{1}{\eta_t(1-\gamma)}\sum_{s\in\mathcal{S}}d_{\pi_{\theta^*}}(s)\mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_{t+1}}(\cdot|s))|\mathcal{F}_t]+\frac{\eta_tC_Q^2}{2(1-\gamma)}\end{aligned}$$

Considering the fact that $F(\theta,\alpha)$ is μ-strongly concave on $\widetilde{\alpha}$, we have

$$\begin{aligned}&F(\theta_t,\widetilde{\alpha}^*(\theta_t))-F(\theta_t,\widetilde{\alpha}_t)\\ \le&\langle\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}),\widetilde{\alpha}^*(\theta_t)-\widetilde{\alpha}_t\rangle-\frac{\mu}{2}\|\widetilde{\alpha}^*(\theta_t)-\widetilde{\alpha}_t\|^2\\ \le&\frac{1}{2\mu}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_t)\|^2+\frac{\mu}{2}\|\widetilde{\alpha}^*(\theta_t)-\widetilde{\alpha}_t\|^2-\frac{\mu}{2}\|\widetilde{\alpha}^*(\theta_t)-\widetilde{\alpha}_t\|^2 \qquad (15)\\ =&\frac{1}{2\mu}\|\nabla_{\widetilde{\alpha}}F(\theta_t,\widetilde{\alpha}_t)\|^2.\end{aligned}$$

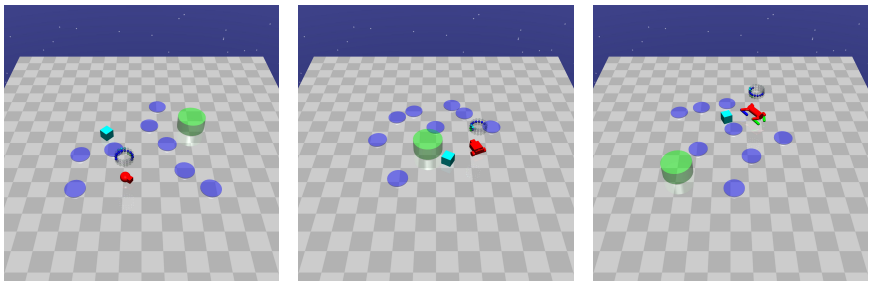
We select $\eta_t=\frac{1-\gamma}{\sqrt{T}}$ and make a summation:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}[g(\theta_t)]-g(\theta^*)\\ \overset{(i)}{\le}&\frac{1}{(1-\gamma)^2\sqrt{T}}\sum_{s\in\mathcal{S}}d_{\pi_{\theta^*}}(s)\mathbf{E}[B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_0}(\cdot|s))-B_w(\pi_{\theta^*}(\cdot|s),\pi_{\theta_T}(\cdot|s))]\\ &+\frac{C}{2\mu\sqrt{K}}+\frac{C_Q^2}{2\sqrt{T}}\\ \overset{(ii)}{\le}&\frac{C}{2\mu\sqrt{K}}+\frac{(1-\gamma)^2C_Q^2+2\log|\mathcal{A}|}{2(1-\gamma)^2\sqrt{T}}\\ \overset{(iii)}{\le}&\mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right)+\mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{K}}\right)\end{aligned}$$

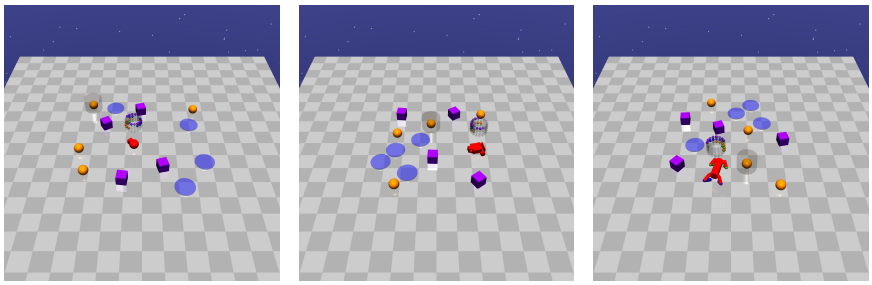
where (i) follows from Lemma 7, (ii) is due to $0 \leq B_w(\pi_1, \pi_2) \leq \log |\mathcal{A}|$, and (iii) is because $C_Q = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$ and the definition of C and C_3. □

B Environment and Expert Data

In this section, we introduce the OpenAI Safety Gym benchmarks (Ray et al., 2019) used in our experiments and give details on how to generate expert data.



(a) Safexp-PointGoal1-v0 (b) Safexp-CarGoal1-v0 (c) Safexp-DoggoGoal1-v0



(d) Safexp-PointButton1-v0 (e) Safexp-CarButton1-v0 (f) Safexp-DoggoButton1-v0

Figure 5: Screenshots of the OpenAI Safety Gym environments. In Safexp-PointGoal1-v0, the red Point should navigate to the green cylinder while avoiding the purple hazards on the floor.

B.1 Environment Overview

OpenAI Safety Gym (Ray et al., 2019) is a highly configurable environment, which supports users to construct desired environments with different robots, tasks, constraints, and observation spaces. In general, tasks in Safety Gym demand the robot to navigate dangerous environments including hazards and vases. There are three optional robots, *i.e.*, Point, Car, and Doggo, while three task options are offered, *i.e.*, Goal, Button, and Push. Constraints such as hazards and vases can be selected and placed into the environment. The information that an agent could receive may come from standard robot sensors, velocity sensors, and lidars. Furthermore, the level of difficulty of the man-made environment can be adjusted by increasing or decreasing the number of constraints. Generally, Safety Gym is such a huge system that it cannot be explained in detail due to the various configurable choices available.

Therefore, to give an intuitive understanding of the environments, we introduce a standard Safety Gym environment–Safexp-PointGoal1-v0, which is shown in Figure 5. As can be interpreted from its name, the robot in this environment is Point (the red object in Figure 5(a)), a small robot with two actuators, one for turning and the other for moving forward/backward; the task is Goal, which means that the robot should move to a goal position as depicted by the green area in Figure 5(a); the number "1" after the task Goal represents the difficulty level of this task. In terms of constraints, there are several hazards (purple circles on the floor in Figure 5(a)) that are randomly placed during the environment initialization. When the robot steps into a hazardous area, the cost indicator c_t will be 1; otherwise, $c_t = 0$ at each step. One episode will end after 1,000 steps. During the 1,000 steps,

if the goal has been achieved, a new goal will be randomly placed on the map. For more details on the Safety Gym, we refer the readers to Ray et al. (2019).

B.2 Environment Specifications

The specifications of the tested environments are listed in Table 2.

Table 2: Specifications of the OpenAI Safety Gym Benchmarks.

Environment	State Space	Action Space	Max-Step
Safexp-PointGoal1-v0	60	2	1000
Safexp-PointButton1-v0	76	2	1000
Safexp-CarGoal1-v0	72	2	1000
Safexp-CarButton1-v0	88	2	1000
Safexp-DoggoGoal1-v0	104	12	1000
Safexp-DoggoButton1-v0	120	12	1000

B.3 Expert Data

Here, we demonstrate how to generate the expert data. First, we use the safe RL algorithm TRPO-Lagrangian implementation (Trust Region Policy Optimization Lagrangian) in Ray et al. (2019) to train an agent with a prescribed cost limit. After training, an agent, which can achieve high cumulative rewards and satisfy the cost limit, is obtained. This agent can be regarded as a safe expert when its prescribed cost limit is the same as that of the target application, and we can get a series of expert data by executing this policy in the Safety Gym environments. Although this agent is considered to be safe in most cases, some trajectories sampled from it could be unsafe, *i.e.*, the Cost of a trajectory is larger than cost limit, due to dynamically changing environments. This means that safe experts may still make mistakes and take dangerous actions, which is consistent with one of the motivations Best (1992); Culverhouse et al. (2003).

As a result, we can sample both safe trajectories and unsafe trajectories using such a safe expert. With consideration of the fact that practical expert data may come from a variety of sources, we also generate the data with multiple expert policies. In particular, for every Safety Gym environment, we use TRPO-Lagrangian to train three safe experts from scratch separately. Default hyper-parameters in Ray et al. (2019) are adopted and the cost limit for each environment are listed in Table 3. After training, we construct 10 safe expert trajectories and 5 unsafe expert trajectories by sampling from each expert. Both states and actions of the expert are recorded sequentially and a trajectory contains $1,000$ states and actions. Since there are three experts, we obtain a total number of 45 expert trajectories for each environment, in which 30 trajectories are safe and the other 15 trajectories are unsafe. The 45 expert trajectories are what we use for IL, and no labels are provided to indicate whether one expert trajectory is safe or not during imitating.

Table 3: Cost limits for training safe experts.

Environment	Cost Limit d_0
Safexp-PointGoal1-v0	25
Safexp-PointButton1-v0	60
Safexp-CarGoal1-v0	25
Safexp-CarButton1-v0	200
Safexp-DoggoGoal1-v0	60
Safexp-DoggoButton1-v0	250

C Implementation Details

We implement LGAIL based on two open source codes, OpenAI Baselines (Dhariwal et al., 2017) and Safety Starter Agents (Ray et al., 2019). Following Dhariwal et al. (2017), we use the RL algorithm Trust Region Policy Optimization (TRPO) (Schulman et al., 2015a) to serve as the generator. We adopt

the discriminator from OpenAI Baselines to replace the reward that is fed back from environments in Safety Starter Agents. Two separate neural networks are constructed to represent the policy $\pi_\theta(s,a)$ and discriminator $r_\alpha(s,a)$. Then the loss function for updating discriminator is as follows

$$\max_{\alpha} \hat{\mathbf{E}}_{\tau_E \sim \pi_E}[\log r_\alpha(s,a)] + \hat{\mathbf{E}}_{\tau_\theta \sim \pi_\theta}[\log(1 - r_\alpha(s,a))], \tag{16}$$

where τ_E and τ_θ are expert trajectories and agent trajectories, respectively. Besides, during training, the Lagrange multiplier λ is dynamically updated to ensure the agent satisfy safety constraints according to,

$$\max_{\lambda} \hat{\mathbf{E}}_{\tau_\theta}(\lambda(J_C(\pi_\theta) - d_0)). \tag{17}$$

For the optimization of the policy, the rapid variation of λ may affect the training stability. Hence, we adopt a technique (Stooke et al., 2020) to regulate the policy improvement process

$$\hat{\mathbf{E}}_{\tau_\theta \sim \pi_\theta}[\frac{1}{1+\lambda}(Q^{\pi_\theta}_{r_\alpha}(s,a) - \lambda Q^{\pi_\theta}_{c}(s,a))] - \beta H(\pi_\theta), \tag{18}$$

in which $Q^{\pi_\theta}_{r_\alpha}(\bar{s},\bar{a}) = \hat{\mathbf{E}}_{\tau_\theta}[-\log(r_\alpha(s,a))|s_0 = \bar{s}, a_0 = \bar{a}]$ and $Q^{\pi_\theta}_{c}(s,a) = \hat{\mathbf{E}}_{\tau_\theta}[c(s,a))|s_0 = \bar{s}, a_0 = \bar{a}]$. Two independent neural networks are adopted to maintain accurate approximation of Q-values for the reward and cost. Finally, we present the number of expert trajectories and the complete hyper-parameters used for imitation learning in Table 4 and Table 5.

We want to discuss a little bit more about one of the baselines, 2IWIL/IC-GAIL (Wu et al., 2019), *i.e.*, algorithms of learning from imperfect demonstration. The basic problem for learning from imperfect demonstration is that expert data could be sampled from experts with different qualities (Wu et al., 2019). Note that the quality here stands for the performance of the expert. In other words, some expert data are sampled from optimal policies while others are sampled from sub-optimal policies. The expert data sampled from sub-optimal policies could mislead the imitator to sub-optimal performance. Besides, only a small portion of expert data is labeled with confidence scores. If the confidence score $conf(s,a) = 1$, then the state-action pair (s,a) is sampled from optimal policies. On the contrary, $conf(s,a) = 0$ means that (s,a) is sampled from sub-optimal policies. Essentially, the aim of their solutions, 2IWIL and IC-GAIL, is to find all the state-action pairs that are sampled from optimal policies and learn from these optimal data without distractions of sub-optimal data (Wu et al., 2019). Therefore, in our experiments, we conduct imitation learning from safe expert data, which is the ultimate form of 2IWIL and IC-GAIL.

Another algorithm CGAIL is a direct extension of CPO (Achiam et al., 2017) to the IL setting, where the reward is replaced with the output of a discriminator. As for RAIL, it bears some resemblance to our algorithm LGAIL. The core difference is that LGAIL explicitly considers the prescribed safety constraint and dynamically adjust the weight of costs while RAIL only imposes a large fixed penalty on dangerous state-action pairs. In addition, LGAIL is guaranteed to achieve a safe policy, while RAIL usually needs to search a proper penalty to balance safety and reward issues. For example, it is necessary to finetune the scale of penalty for RAIL given a new environment or a prescribed cost limit, which could be arduous. In contrast, LGAIL can automatically adapts to different environments and prescribed cost limits. In all experiments, we set the penalty scale for RAIL to 5.

Table 4: Number of expert trajectories. We abbreviate trajectories as Trajs.

Environment	Total Expert Trajs	Safe Expert Trajs	Unsafe Expert Trajs
Safexp-PointGoal1-v0	45	30	15
Safexp-PointButton1-v0	45	30	15
Safexp-CarGoal1-v0	45	30	15
Safexp-CarButton1-v0	45	30	15
Safexp-DoggoGoal1-v0	45	30	15
Safexp-DoggoButton1-v0	45	30	15

D ADDITIONAL EXPERIMENTS

We present more experimental results (including the quantitative results) in different environments with various configurations here to further validate the proposed algorithm–LGAIL.

Table 5: Hyper-parameters in experiments.

Hyper-parameters	Value
Common parameters	
Network size (Except the discriminator network)	(256,256)
Network size (Discriminator network)	(100,100)
Activation	$tanh$
Batch size	3,000
Optimizer	Adam
Generator network update times	1
Discriminator network update times	1
Common parameters for TRPO	
Generalized Advantage Estimation Gamma	0.99
Generalized Advantage Estimation Lambda	0.97
Maximum KL	0.01
Learning rate (Value network)	1×10^{-3}
Value iteration	80
Policy entropy	0.0
Discriminator parameters	
Learning rate (Discriminator network)	3×10^{-4}
Discriminator entropy	1×10^{-3}
Penalty parameters	
Initial penalty	1
Penalty learning rate	5×10^{-2}

D.1 Computing Resources

We use CPUs to run our experiments. The model name of the CPU is Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz. The computation time for each environment is provided in Table 6.

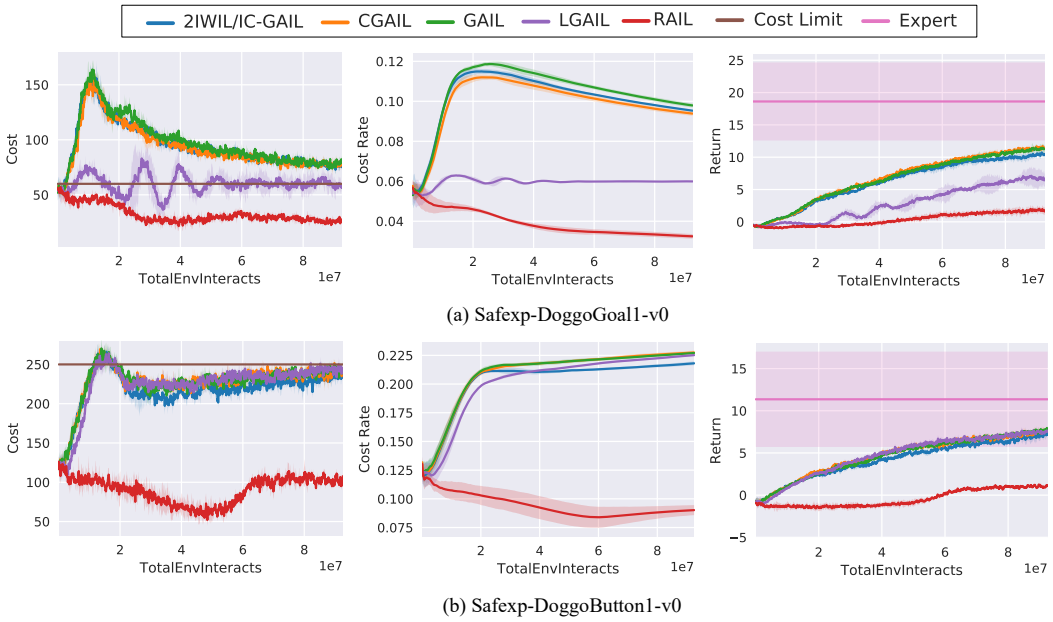
Table 6: Computation time.

Environment	Time
Safexp-PointGoal1-v0	about 4 hours
Safexp-PointButton1-v0	about 4 hours
Safexp-CarGoal1-v0	about 13 hours
Safexp-CarButton1-v0	about 20 hours
Safexp-DoggoGoal1-v0	about 14 hours
Safexp-DoggoButton1-v0	about 20 hours

D.2 Experiments on Doggo Tasks

The learning curves in Safexp-DoggoGoal1-v0 and Safexp-DoggoButton1-v0 are presented in Figure 6, and quantitative results of these two environments are listed in Table 7. Even in these complex environments, the proposed algorithm LGAIL can still mimic the expert under the prescribed safety constraint.

The phenomenon that LGAIL performs slightly worse than the other baselines except RAIL regrading Return has been discussed in the paper. For Safexp-DoggoButton1-v0, the phenomenon that LGAIL did not reduce the cost is because the Cost of LGAIL is lower than the cost limit $d_0 = 250$. According to our algorithm, the Lagrange multiplier will be zero if the current policy satisfies the cost limit. In other words, LGAIL focuses on improving rewards when the policy is safe. Hence, the learning curve of LGAIL in Safexp-DoggoButton1-v0 is reasonable. To demonstrate that LGAIL is able to reduce costs, we also conduct new experiments with lower cost limit $d_0 = 200$. The learning curves of LGAIL in Safexp-DoggoButton1-v0 with cost limit $d_0 = 200$ are presented in Figure 7.



(a) Safexp-DoggoGoal1-v0

(b) Safexp-DoggoButton1-v0

Figure 6: Learning curves in Safexp-DoggoGoal1-v0 and Safexp-DoggoButton1-v0. Performance is measured with Cost, Cost Rate, and Return. The x-axis represents time steps of interactions with the environment. Each algorithm is evaluated with 3 random seeds.

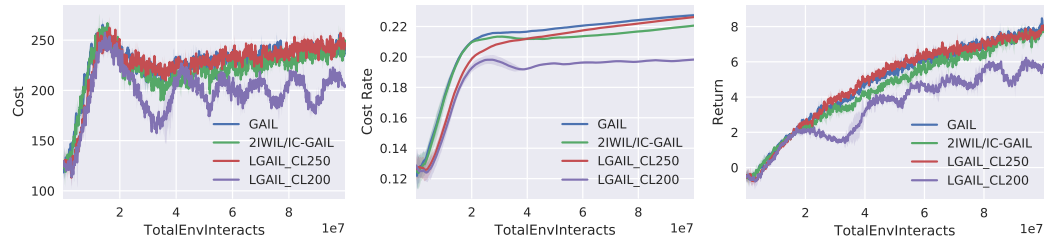


Figure 7: Learning curves in Safexp-DoggoButton1-v0 with different cost limits.

Table 7: Summary of quantitative results. The columns represent the algorithms, while the rows represent environments and metrics. Each result is averaged over 30 trails of a policy.

Environment		LGAIL	GAIL	2IWIL/IC-GAIL	CGAIL	RAIL
DoggoGoal1-v0	Cost	56.3±2.6	79.4±1.9	75.2±3.3	76.9±0.7	25.6±1.4
	Cost Rate	0.06	0.098	0.095	0.094	0.032
	Return	6.6±1.2	11.3±0.5	10.4±0.3	11.3±0.4	1.8±0.5
DoggoButton1-v0	Cost	245.8±6.1	243.6±1.9	241.6±2.4	245.0±4.0	107.7±4.4
	Cost Rate	0.225	0.227	0.218	0.227	0.09
	Return	7.5±0.1	7.8±0.1	7.2±0.5	7.8±0.3	1.1±0.1

D.3 Expert Performance

The performance of the expert data is presented in Table 8. As we discussed above, we sample 10 trajectories from each expert. Besides, during sampling, we select trajectories according to specific reward or safety desires. The performance of expert data in Table 8 is calculated from the sampled data. However, Safety Gym is a dynamically changing environment such that it is not enough to

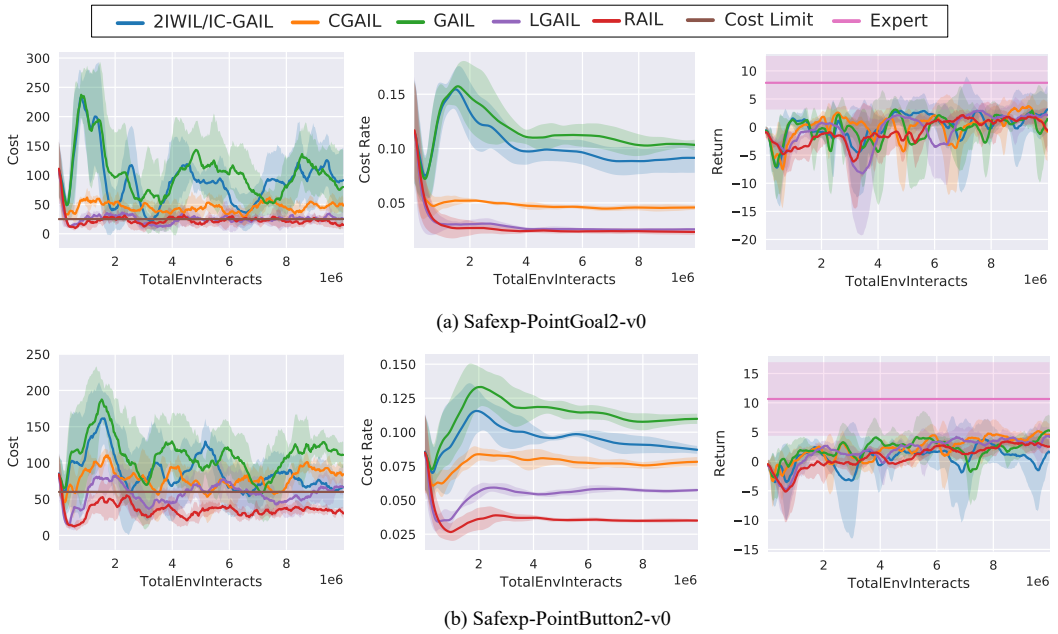
Table 8: Performance of the expert data.

Environment		Cost	Return
Safexp-PointGoal1-v0	Safe	8.0±8.29	19.5±2.8
	Unsafe	64.9±13.3	18.9±2.7
	Mixed	27.0±28.7	19.3±2.8
Safexp-PointButton1-v0	Safe	26.6±15.0	20.1±4.2
	Unsafe	164.8±49.3	19.0±3.0
	Mixed	72.7±72.1	19.8±4.0
Safexp-CarGoal1-v0	Safe	6.7±7.9	25.9±4.0
	Unsafe	82.6±36.90	22.6±3.2
	Mixed	32.0±42.1	24.8±4.1
Safexp-CarButton1-v0	Safe	139.2±41.7	23.8±5.4
	Unsafe	310.5±40.9	24.3±4.0
	Mixed	196.3±90.8	24.0±5.0
Safexp-DoggoGoal1-v0	Safe	24.9±12.4	21.0±3.7
	Unsafe	121.6±26.5	20.4±2.8
	Mixed	57.2±49.1	20.8±3.5
Safexp-DoggoButton1-v0	Safe	172.9±61.6	15.5±5.8
	Unsafe	349.3±35.0	14.4±3.6
	Mixed	231.7±99.3	15.1±5.2

Table 9: Performance of the expert that is evaluated with 100 trajectories.

Environment		Cost	Return
Safexp-PointGoal1-v0	Expert 1	27.8±18.5	15.6±3.2
	Expert 2	12.5±18.9	13.4±7.7
	Expert 3	26.1±27.9	17.9±4.6
Safexp-PointButton1-v0	Expert 1	54.8±40.2	13.4±5.3
	Expert 2	46.7±55.1	11.8±8.5
	Expert 3	70.6±56.7	13.9±6.4
Safexp-CarGoal1-v0	Expert 1	28.4±29.3	21.8±6.3
	Expert 2	20.7±25.9	21.6±7.6
	Expert 3	24.5±26.2	21.4±8.3
Safexp-CarButton1-v0	Expert 1	220.4±87.6	17.6±5.8
	Expert 2	220.0±117.9	19.8±7.1
	Expert 3	197.1±85.0	20.2±7.6
Safexp-DoggoGoal1-v0	Expert 1	54.2±45.2	15.0±5.8
	Expert 2	56.6±46.1	22.8±4.3
	Expert 3	57.2±42.2	18.1±5.1
Safexp-DoggoButton1-v0	Expert 1	218.0±95.1	9.5±5.2
	Expert 2	243.4±115.9	11.5±4.6
	Expert 3	256.8±107.4	13.0±6.4

evaluate an expert with only 10 trajectories. Hence, we also provide the performance of experts in Table 9, which is evaluated with 100 trajectories. As we can see from Table 9, the variance of an expert is relatively high. So even we test an expert with 100 trajectories, the performance of an expert could vary if we retest it. In the learning curves, we plot the expert performance rather than the performance of expert data because the former is fairer.



(a) Safexp-PointGoal2-v0

(b) Safexp-PointButton2-v0

Figure 8: Learning curves of LGAIL and the other baselines on Level 2 tasks. Performance is measured with Cost, Cost Rate, and Return. The x-axis represents time steps of interactions with the environment. Each algorithm is evaluated with 5 random seeds.

D.4 Experiments on "Level 2" Tasks

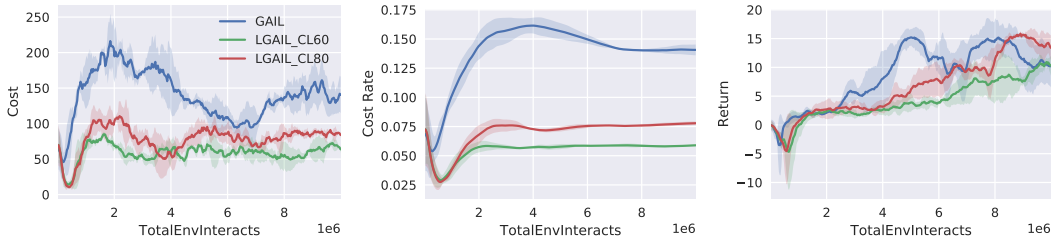
We conduct experiments on "level 2" tasks (Safexp-PointGoal2-v0 and Safexp-PointButton2-v0) to demonstrate the performance of LGAIL against other baselines. The learning curves are presented in Figure 8. In more complex environments, performance degradation is observed for experts and IL algorithms. However, experiment results show that LGAIL can work effectively in these complex environments, i.e., LGAIL is able to achieve the same level of performance regarding Return compared with the other baselines (GAIL, 2IWIL/IC-GAIL, CGAIL) and simultaneously satisfy the cost limit. Although RAIL could also return safe policies, RAIL performs pretty poorly in terms of Return.

D.5 Impact of Cost Limits

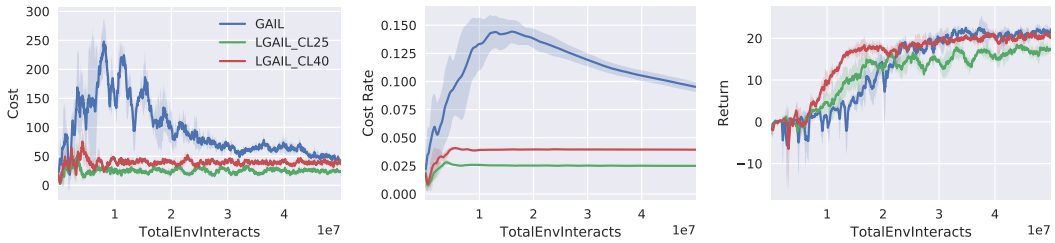
In the paper, we carry out experiments to investigate the impact of cost limit on LGAIL's performance with unsafe expert data. We only present the results using Safexp-PointGoal1-v0 in the paper. Here, more experiments in other environments are given, which are shown in Figure 9. We can see that LGAIL is able to obtain a policy that satisfies the prescribed cost limit.

D.6 GAIL with Purely Safe Expert Data

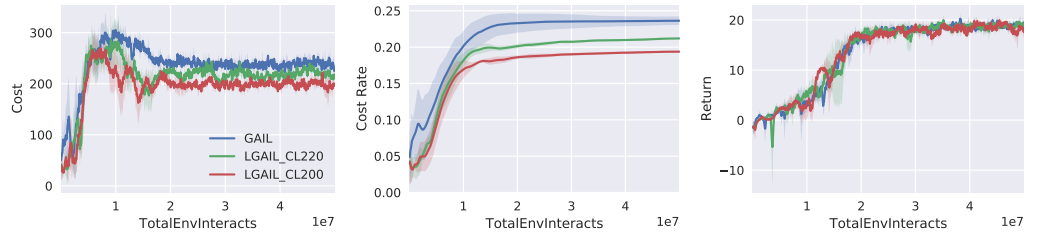
It is worthy of investigating why GAIL could not reproduce safe policies with safe expert data. We conduct experiments to test the impact of the amount of expert data and the diversity of expert data on the safety performance of GAIL. Concretely, in environments Safexp-PointGoal1-v0 and Safexp-PointButton1-v0, we train GAIL with different numbers of expert trajectories (including 10, 30, 100, 300, and 1000 trajectories). These expert data are sampled from a single expert. Each trajectory contains 1,000 state-action pairs. Hence, it means that one million state-action pairs are provided when we use 1000 trajectories to train GAIL, which is a huge amount of data. Besides, we also use safe expert data that are sampled from three independent experts to train an agent. For the experiments that use data sampled from a mixture of expert policies, we evaluate the performance of GAIL against different numbers of expert trajectories (300, 900, and 1500). The learning curves are



(a) Safexp-PointButton1-v0



(b) Safexp-CarGoal1-v0

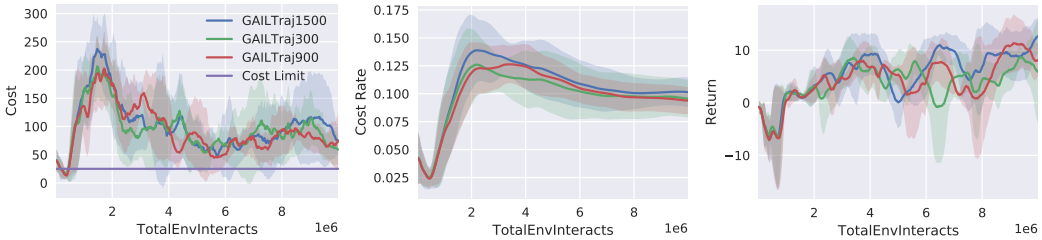


(c) Safexp-CarButton1-v0

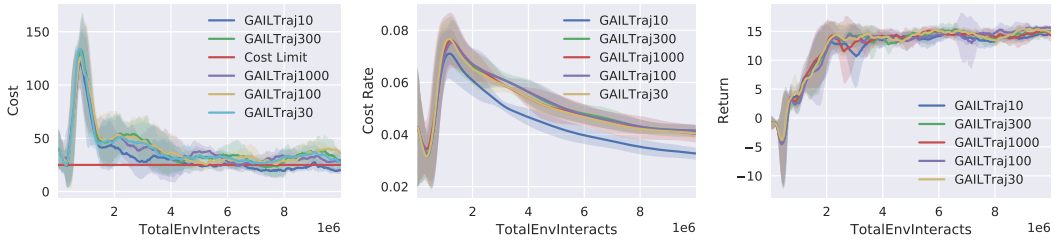
Figure 9: Impact of constraint limit on LGAIL with unsafe expert data. In the legend LGAIL_CL{x}, x represents the cost limit d_0.

presented in Figure 10. Results are shown in Figure 10. From Figure 10, it is clear that: (1) GAIL usually fails to recover a safe policy even with abundant safe expert data; (2) increasing the number of expert data does not help improve the performance of GAIL in terms of both rewards and costs; (3) GAIL performs worse with expert data that are sampled from multiple experts. In practice, expert data are often collected from various sources.

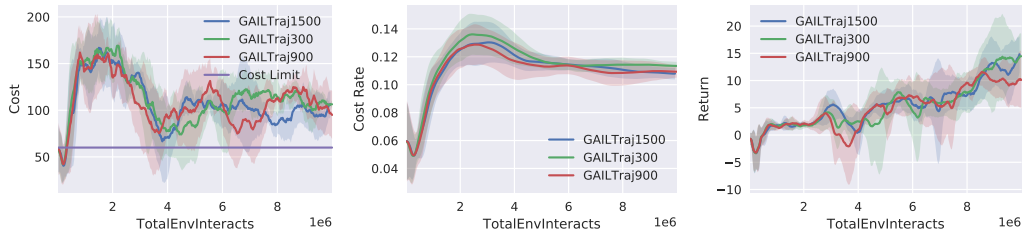
In our opinion, there are three possible reasons: (1) The safe expert data are unbalanced. We think that expert data contain more information on how to achieve rewards compared to the information on how to be safe. Every safe expert trajectory achieves high rewards but low costs, which means that rewards are dense while costs are sparse. As a result, GAIL is likely to mainly develop the ability to accomplish tasks but neglecting the connotative ability to be safe. (2) GAIL could not adapt well to dynamic environments due to the poor generalization ability. GAIL employs RL algorithms to serve as the generator, and RL algorithms often struggle with generalization problems. Hence, GAIL is likely to generalize poorly in dynamic environments such that the recovered policy could be unsafe. (3) Expert data sampled from a mixture of expert policies could provide opposite information about safety. Different experts have their own preferences, which may mislead the agent to dangerous actions. In contrast, our algorithm LGAIL explicitly considers safety issues during imitating and regards them as constraints to regulate the IL process. This explicit modeling enables LGAIL to generate safer policies.



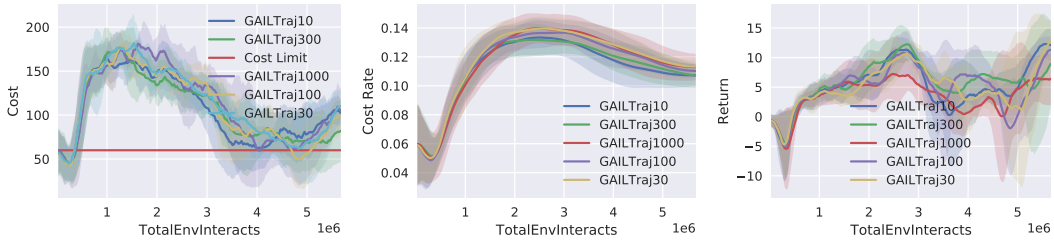
(a) Safexp-PointGoal1-v0 (safe expert data sampled from three experts)



(b) Safexp-PointGoal1-v0 (safe expert data sampled from one expert)



(c) Safexp-PointButton1-v0 (safe expert data sampled from three experts)



(d) Safexp-PointButton1-v0 (safe expert data sampled from one expert)

Figure 10: Impact of the number of expert trajectories on GAIL with safe expert data. In the legend GAILTraj{x}, x represents the number of expert trajectories.