Risk-Sensitive Generative Adversarial Imitation Learning

Jonathan Lacotte
Electrical Engineering
Stanford University

Yinlam Chow
DeepMind

Mohammad Ghavamzadeh
DeepMind

Marco Pavone
Aeronautics & Astronautics
Stanford University

Abstract

We study risk-sensitive imitation learning where the agent's goal is to perform at least as well as the expert in terms of a risk profile. We first formulate our risk-sensitive imitation learning setting. We consider the generative adversarial approach to imitation learning (GAIL) and derive an optimization problem for our formulation, which we call risk-sensitive GAIL (RS-GAIL). We then derive two different versions of our RS-GAIL optimization problem that aim at matching the risk profiles of the agent and the expert w.r.t. Jensen-Shannon (JS) divergence and Wasserstein distance, and develop risk-sensitive generative adversarial imitation learning algorithms based on these optimization problems. We evaluate the performance of our JS-based algorithm and compare it with GAIL and the risk-averse imitation learning (RAIL) algorithm in two MuJoCo tasks.

1 Introduction

We study imitation learning, i.e., the problem of learning to perform a task from the sample trajectories generated by an expert. There are three main approaches to this problem: **1)** behavioral cloning (e.g., Pomerleau 1991) in which the agent learns a policy by solving a supervised learning problem over the state-action pairs of the expert's trajectories, **2)** inverse reinforcement learning (IRL) [Ng and Russell, 2000] followed by reinforcement learning (RL), a process also referred to as RL∘IRL [Ho and Ermon, 2016a], where we first find a cost function under which the expert is optimal (IRL part) and then return the optimal policy w.r.t. this cost function (RL part), and **3)** generative adversarial imitation learning (GAIL) [Ho and Ermon, 2016a] that frames the imitation learning problem as occupancy measure matching w.r.t. either the Jensen-Shannon divergence (GAIL) [Ho and Ermon, 2016a] or the Wasserstein distance (InfoGAIL) [Li et al., 2017]. Behavioral cloning algorithms are simple but often need a large amount of data to be successful. IRL does not suffer from the main problems of behavioral cloning [Ross and Bagnell, 2010, Ross et al., 2011], since it takes entire trajectories into account (instead of single time-step decisions) when learning a cost function. However, IRL algorithms are often expensive to run because they require solving a RL problem in their inner loop. This issue had restricted the use of IRL to small problems for a long while and only recently scalable IRL algorithms have been developed [Levine and Koltun, 2012, Finn et al., 2016]. On the other hand, the nice feature of the GAIL approach to imitation learning is that it bypasses the intermediate IRL step and directly learns a policy from data, as if it were obtained by the RL∘IRL process. The resulting algorithm is closely related to generative adversarial networks (GAN) [Goodfellow et al., 2014] that has recently gained attention in the deep learning community.

In many applications, we may prefer to optimize some measure of risk in addition to the standard optimization criterion, i.e., the expected sum of (discounted) costs. In such cases, we would like to use a criterion that incorporates a penalty for the variability (due to the stochastic nature of the system) induced by a given policy. Several risk-sensitive criteria have been studied in the literature of risk-sensitive Markov decision processes (MDPs) [Howard and Matheson, 1972] including the expected exponential utility [Howard and Matheson, 1972, Borkar, 2001, 2002], a variance-related measure [Sobel, 1982, Filar et al., 1989, Tamar et al., 2012, Prashanth and Ghavamzadeh, 2013], or the tail-related measures like value-at-risk (VaR) and conditional value-at-risk (CVaR) [Filar et al., 1995, Rockafellar and Uryasev, 2002, Chow and Ghavamzadeh, 2014, Tamar et al., 2015b].

In risk-sensitive imitation learning, the agent's goal is to perform at least as well as the expert in terms of one or more risk-sensitive objective(s), e.g., mean + λCVaR$_\alpha$, for one or more values of $\lambda \geq 0$. This goal cannot be satisfied by risk-neutral imitation learning. As we will show in Section 3.3, if we use GAIL to minimize the Wasserstein distance between the occupancy measures of the agent and the expert, the distance between their CVaRs could be still large. Santara et al. [2017a] recently showed empirically that the policy learned by GAIL does not have the desirable tail properties, such as VaR and CVaR, and proposed a modification of GAIL, called risk-averse imitation learning (RAIL), to address this issue. We will discuss about RAIL in more details in Section 5 as it is probably the closest work to us in the literature. Another related work is by Singh et al. [2018] on risk-sensitive IRL in which the proposed algorithm infers not only the expert's cost function but her underlying risk measure, for a rich class of static and dynamic risk measures (coherent risk measures). The agent then learns a policy by optimizing the inferred risk-sensitive objective.

In this paper, we study an imitation learning setting in which the agent's goal is to learn a policy with minimum expected sum of (discounted) costs and with CVaR$_\alpha$ that is at least as well as that of the expert. We first provide a mathematical formulation for this setting and derive a GAIL-like optimization problem for our formulation, which we call risk-sensitive GAIL (RS-GAIL), in Section 3.1. In Sections 3.2 and 3.3, we define cost function regularizers that when we compute their convex conjugates and plug them into our RS-GAIL objective function, the resulting optimization problems aim at learning the expert's policy by matching occupancy measures w.r.t. Jensen-Shannon (JS) divergence and Wasserstein distance, respectively. We call the resulting optimization problems JS-RS-GAIL and W-RS-GAIL and propose our risk-sensitive generative adversarial imitation learning algorithm based on these optimization problems in Section 4. It is important to note that unlike the risk-neutral case in which the occupancy measure of the agent is matched with that of the expert, here in the risk-sensitive case, we match two sets of occupancy measures that encode the risk profile of the agent and the expert. This will become more clear in Section 3. We present our understanding of RAIL and how it is related to our work in Section 5. In Section 6, we evaluate the performance of our JS-RS-GAIL algorithm and compare it with GAIL and RAIL in two MuJoCo [Todorov et al., 2012a] tasks that have also been used by Ho and Ermon [2016a] and Santara et al. [2017a]. Finally in Section 7, we conclude the paper and list a number of future directions.

2 Preliminaries

We consider the scenario in which the agent's interaction with the environment is modeled as a Markov decision process (MDP). A MDP is a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, c, p, p_0, \gamma\}$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces; $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $p : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ are the cost function and transition probability distribution, with $c(s,a)$ and $p(\cdot|s,a)$ being the cost and next state probability when taking action a in state s; $p_0 \in \Delta_{\mathcal{S}}$ is the initial state distribution; and $\gamma \in [0,1)$ is a discounting factor. A stationary stochastic policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a mapping from states to a distribution over actions. We denote by Π the set of all such policies. We denote by $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T) \in \Gamma$, where $a_t \sim \pi(\cdot|s_t),\ \forall t \in \{0, \ldots, T-1\}$, a trajectory of the fixed horizon T generated by policy π, by Γ the set of all such trajectories, and by $C(\tau) = \sum_{t=0}^{T-1} \gamma^t c(s_t, a_t)$ the *loss* of trajectory τ. The probability of trajectory τ is given by $\mathbb{P}(\tau|\pi) = p^\pi(\tau) = p_0(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$. We denote by C^π the random variable of the loss of policy π. Thus, when $\tau \sim p^\pi$, $C(\tau)$ is an instantiation of the random variable C^π. The performance of a policy π is usually measured by a quantity related to the loss of the trajectories it generates, the most common would be its expectation, i.e., $\mathbb{E}[C^\pi] = \mathbb{E}_{\tau \sim p^\pi}[C(\tau)]$. We define the occupancy measure of policy π as $d^\pi(s,a) = \sum_{t=0}^{T} \gamma^t \mathbb{P}(s_t = s, a_t = a|\pi)$, which can be interpreted as the unnormalized distribution of the state-action pairs visited by the agent under policy π. Using occupancy measure, we may write the policy's performance as $\mathbb{E}[C^\pi] = \mathbb{E}_{p^\pi}[C(\tau)] = \mathbb{E}_{d^\pi}[c(s,a)] = \sum_{s,a} d^\pi(s,a) c(s,a)$.

2.1 Risk-sensitive MDPs

In risk-sensitive decision-making, in addition to optimizing the expectation of the loss, it is also important to control the variability of this random variable. This variability is often measured by the variance or tail-related quantities such as value-at-risk (VaR) and conditional value-at-risk (CVaR). Given a policy π and a confidence level $\alpha \in (0,1]$, we define the VaR at level α of the loss random variable C^π as its (left-side) $(1-\alpha)$-quantile, i.e., $\nu_\alpha[C^\pi] := \inf\{t \in \mathbb{R} \mid \mathbb{P}(C^\pi \leq t) \geq 1 - \alpha\}$ and its CVaR at level α as $\rho_\alpha[C^\pi] = \inf_{\nu \in \mathbb{R}} \big\{\nu + \frac{1}{\alpha}\mathbb{E}\big[(C^\pi - \nu)_+\big]\big\}$, where $x_+ = \max(x, 0)$. We also define the *risk envelope* $\mathcal{U}^\pi = \big\{\zeta : \Gamma \to [0, \frac{1}{\alpha}] \mid \sum_{\tau \in \Gamma} \zeta(\tau) \cdot p^\pi(\tau) = 1\big\}$, which is a compact, convex, and bounded set. The quantities $p^\pi_\zeta = \zeta \cdot p^\pi,\ \zeta \in \mathcal{U}^\pi$ are called *distorted probability distributions* and we denote by $\mathcal{P}^\pi_\zeta = \big\{p^\pi_\zeta \mid \zeta \in \mathcal{U}^\pi\big\}$ the set of such distributions. The set $\mathcal{P}^\pi_\zeta$ induces a set of *distorted occupancy measures* $\mathcal{D}^\pi_\zeta$, where each element of $\mathcal{D}^\pi_\zeta$ is the occupancy measure induced by a distorted probability distribution in $\mathcal{P}^\pi_\zeta$. The sets $\mathcal{P}^\pi_\zeta$ and $\mathcal{D}^\pi_\zeta$ characterize the risk of policy π. Given the risk envelope $\mathcal{U}^\pi$, we may define the dual representation of CVaR as $\rho_\alpha[C^\pi] = \sup_{\zeta\in\mathcal{U}^\pi} \mathbb{E}_{\tau\sim p^\pi}\big[\zeta(\tau)C(\tau)\big]$, where the supremum is attained at the density $\zeta^*(\tau) = \frac{1}{\alpha}\mathbf{1}_{\{C(\tau)\geq\nu_\alpha[C^\pi]\}}$. Hence, CVaR can be considered as the expectation of the loss random variable, when the trajectories are generated from the distorted distribution $p^\pi_{\zeta^*} = \zeta^*\cdot p^\pi$, i.e., $\rho_\alpha[C^\pi] = \mathbb{E}_{\tau\sim p^\pi_{\zeta^*}}[C(\tau)]$. If we denote by $d^\pi_{\zeta^*}\in\mathcal{D}^\pi_\zeta$ the distorted occupancy measure induced by $p^\pi_{\zeta^*}$, then we may write the CVaR as $\rho_\alpha[C^\pi] = \mathbb{E}_{p^\pi_{\zeta^*}}[C(\tau)] = \mathbb{E}_{d^\pi_{\zeta^*}}[c(s,a)]$.

2.2 Generative Adversarial Imitation Learning

As discussed in Section 1, generative adversarial imitation learning (GAIL) [Ho and Ermon, 2016a] is a framework for directly extracting a policy from the trajectories generated by an expert policy π_E, as if it were obtained by inverse RL (IRL) followed by RL, i.e., RL$\circ$IRL(π_E). The main idea behind GAIL is to formulate imitation learning as occupancy measure matching w.r.t. the Jensen-Shannon divergence D_{JS}, i.e.,

$$\min_\pi \big(D_{\text{JS}}(d^\pi, d^{\pi_E}) - \lambda H(\pi)\big),$$

where $H(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[-\log\pi(a|s)]$ is the γ-discounted causal entropy of policy π, $\lambda\geq 0$ is a regularization parameter, and $D_{\text{JS}}(d^\pi, d^{\pi_E}) := \sup_{f:\mathcal{S}\times\mathcal{A}\to(0,1)} \mathbb{E}_{d^\pi}[\log f(s,a)] + \mathbb{E}_{d^{\pi_E}}[\log(1-f(s,a))]$. Li et al. [2017] proposed InfoGAIL by reformulating GAIL and replacing the Jensen-Shannon divergence $D_{\text{JS}}(d^\pi, d^{\pi_E})$ with the Wasserstein distance $W(d^\pi, d^{\pi_E}) := \sup_{f\in\mathcal{F}_1}\mathbb{E}_{d^\pi}[f(s,a)] - \mathbb{E}_{d^{\pi_E}}[f(s,a)]$, where $\mathcal{F}_1$ is the set of 1-Lipschitz functions over $\mathcal{S}\times\mathcal{A}$.

3 Risk-sensitive Imitation Learning

In this section, we describe the risk-sensitive imitation learning formulation studied in the paper and derive the optimization problems that our proposed algorithms solve to obtain a risk-sensitive policy from the expert's trajectories.

3.1 Problem Formulation

As described in Section 1, we consider the risk-sensitive imitation learning setting in which the agent's goal is to learn a policy with minimum loss and with CVaR that is at least as well as that of the expert. Thus, the agent solves the optimization problem

$$\min_\pi \ \mathbb{E}[C^\pi], \qquad \text{s.t.}\ \ \rho_\alpha[C^\pi]\leq\rho_\alpha[C^{\pi_E}], \tag{1}$$

where C^π is the loss of policy π w.r.t. the expert's cost function c that is unknown to the agent. The optimization problem (1) without the loss of optimality is equivalent to the unconstrained problem

$$\min_\pi \sup_{\lambda\geq 0} \ \mathbb{E}[C^\pi] - \mathbb{E}[C^{\pi_E}] + \lambda\big(\rho_\alpha[C^\pi] - \rho_\alpha[C^{\pi_E}]\big). \tag{2}$$

Note that π_E is a solution of both (1) and (2). However, since the expert's cost function is unknown, the agent cannot directly solve (2), and thus, considers the surrogate problem

$$\min_\pi \sup_{f\in\mathcal{C}} \sup_{\lambda\geq 0} \ \mathbb{E}[C^\pi_f] - \mathbb{E}[C^{\pi_E}_f] + \lambda\big(\rho_\alpha[C^\pi_f] - \rho_\alpha[C^{\pi_E}_f]\big), \tag{3}$$

where $\mathcal{C} = \{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}\}$ and C^π_f is the loss of policy π w.r.t. the cost function f. We employ the Lagrangian relaxation procedure [Bertsekas, 1999] to swap the inner maximization over λ with the minimization over π and convert (3) to the problem

$$\sup_{\lambda\geq 0} \min_\pi \sup_{f\in\mathcal{C}} \ \mathbb{E}[C^\pi_f] - \mathbb{E}[C^{\pi_E}_f] + \lambda\big(\rho_\alpha[C^\pi_f] - \rho_\alpha[C^{\pi_E}_f]\big). \tag{4}$$

We adopt maximum causal entropy IRL formulation [Ziebart et al., 2008, 2010] and add $-H(\pi)$ to the optimization problem (4). Moreover, since $\mathcal{C}$ is large, to avoid overfitting when we are provided with a finite set of expert's trajectories, we add the negative of a convex regularizer $\psi:\mathcal{C}\to\mathbb{R}\cup\{\infty\}$ to the optimization problem (4). As a result we obtain the following optimization problem for our risk-sensitive imitation learning setting, which we call RS-GAIL:

$$\textbf{(RS-GAIL)} \quad \sup_{\lambda\geq 0} \min_\pi \ -H(\pi) + \mathcal{L}_\lambda(\pi,\pi_E), \tag{5}$$

where $\mathcal{L}_\lambda(\pi,\pi_E) := \sup_{f\in\mathcal{C}} \ (1+\lambda)\big(\rho^\lambda_\alpha[C^\pi_f] - \rho^\lambda_\alpha[C^{\pi_E}_f]\big) - \psi(f)$, with $\rho^\lambda_\alpha[C^\pi_f] := \frac{\mathbb{E}[C^\pi_f]+\lambda\rho_\alpha[C^\pi_f]}{1+\lambda}$ being the coherent risk measure for policy π corresponding to mean-CVaR with the risk parameter λ. The parameter λ can be interpreted as the tradeoff between the mean performance and risk-sensitivity of the policy. The objective function $\mathcal{L}_\lambda(\pi,\pi_E)$ can be decomposed into three terms: **1)** the difference between the agent and expert in terms of mean performance, $\mathbb{E}[C^\pi_f] - \mathbb{E}[C^{\pi_E}_f]$, which corresponds to the standard generative imitation learning objective, **2)** the difference between the agent and the expert in terms of risk $\rho_\alpha[C^\pi_f] - \rho_\alpha[C^{\pi_E}_f]$, and **3)** the convex regularizer $\psi(f)$ that encodes our belief about the expert cost function f. For the risk-sensitive quantity $\rho^\lambda_\alpha[C^\pi]$, we define the distorted probability distributions $p^\pi_\xi = \xi\cdot p^\pi$, where $\xi = \frac{1+\lambda\zeta}{1+\lambda}$, $\zeta\in\mathcal{U}^\pi$. We denote by $\mathcal{P}^\pi_\xi$ the set of such distorted distributions and by $\mathcal{D}^\pi_\xi$ the set of distorted occupancy measures induced by the elements of $\mathcal{P}^\pi_\xi$. Similar to CVaR in Section 2.1, we may write the risk-sensitive quantity $\rho_\alpha^\lambda[C^\pi]$ as the expectation $\rho_\alpha^\lambda[C^\pi] = \mathbb{E}_{p^\pi_{\xi^*}}[C(\tau)] = \mathbb{E}_{d^\pi_{\xi^*}}[c(s,a)]$, where $\xi^* = \frac{1+\lambda\zeta^*}{1+\lambda}$ with ζ^* defined in Section 2.1 and $d^\pi_{\xi^*} \in \mathcal{D}^\pi_\xi$ is the distorted occupancy measure induced by $p^\pi_{\xi^*} \in \mathcal{P}^\pi_\xi$. In Theorem 1, we show that the maximization problem $\mathcal{L}_\lambda(\pi,\pi_E)$ over the cost function $f \in \mathcal{C}$ can be rewritten as a sup-inf problem over the distorted occupancy measures $d \in \mathcal{D}^\pi_\xi$ and $d' \in \mathcal{D}^{\pi_E}_\xi$.

Theorem 1. *Let $\psi : \mathcal{C} \to \mathbb{R} \cup \{\infty\}$ be a convex cost function regularizer. Then,*

$$\begin{aligned}\mathcal{L}_\lambda(\pi,\pi_E) &= \sup_{f\in\mathcal{C}} \,(1+\lambda)\big(\rho_\alpha^\lambda[C_f^\pi] - \rho_\alpha^\lambda[C_f^{\pi_E}]\big) - \psi(f) \\ &= \sup_{d\in\mathcal{D}^\pi_\xi}\ \inf_{d'\in\mathcal{D}^{\pi_E}_\xi} \psi^*\big((1+\lambda)(d-d')\big),\end{aligned} \tag{6}$$

where ψ^* is the convex conjugate function of ψ, i.e., $\psi^*(d) = \sup_{f\in\mathcal{C}} d^\top f - \psi(f)$.

Proof. See Appendix A. □

From Theorem 1, we may write the RS-GAIL optimization problem (5) as

$$\sup_{\lambda\geq 0}\ \min_\pi\ -H(\pi) + \sup_{d\in\mathcal{D}^\pi_\xi}\ \inf_{d'\in\mathcal{D}^{\pi_E}_\xi} \psi^*\big((1+\lambda)(d-d')\big). \tag{7}$$

Comparing the RS-GAIL optimization problem (7) with that of GAIL (see Eq. 4 in Ho and Ermon 2016a), we notice that the main difference is the $\sup_{\mathcal{D}^\pi_\xi} \inf_{\mathcal{D}^{\pi_E}_\xi}$ in RS-GAIL that does not exist in GAIL. In the risk-neutral case, $\lambda = 0$, and thus, the two sets of distorted occupancy measures $\mathcal{D}^\pi_\xi$ and $\mathcal{D}^{\pi_E}_\xi$ are singleton and the RS-GAIL optimization problem is reduced to that of GAIL.

Example 1. *Let $\psi(f) = \begin{cases} 0 & \text{if } ||f||_\infty \leq 1 \\ +\infty & \text{otherwise}\end{cases}$, then $\mathcal{L}_\lambda(\pi,\pi_E) = 2(1+\lambda)\sup_{d\in\mathcal{D}^\pi_\xi} \inf_{d'\in\mathcal{D}^{\pi_E}_\xi} ||d-d'||_{TV}$, where $||d-d'||_{TV}$ is the total variation distance between d and d'. Note that similar to GAIL, our optimization problem aims at learning the expert's policy by matching occupancy measures. However, in order to take risk into account, it now involves matching two sets of occupancy measures (w.r.t. the TV distance) that encode the risk profile of each policy.*

3.2 Risk-sensitive GAIL with Jensen-Shannon Divergence

In this section, we derive RS-GAIL using occupation measure matching via Jensen-Shannon (JS) divergence. We define the difference-of-convex cost function regularizer

$$\psi(f) := \begin{cases}(1+\lambda)\big(-\rho_\alpha^\lambda[C_f^{\pi_E}] + \rho_\alpha^\lambda[G_f^{\pi_E}]\big) & \text{if } f<0 \\ +\infty & \text{otherwise}\end{cases}$$

where $C_f^{\pi_E}$ and $G_f^{\pi_E}$ are the loss random variables of policy π_E w.r.t. the cost functions $c(s,a) = f(s,a)$ and $c(s,a) = g\big(f(s,a)\big)$, respectively, with

$$g(x) := \begin{cases} -\log(1-e^x) & \text{if } x<0 \\ +\infty & \text{otherwise}\end{cases}$$

To clarify, $G_f^{\pi_E}$ is a random variable whose instantiations are $G_f(\tau) = \sum_{t=0}^{T-1}\gamma^t g\big(f(s_t,a_t)\big)$, where $\tau \sim p^{\pi_E}$ is a trajectory generated by the expert policy π_E. Similar to the description in Ho and Ermon [2016a], this regularizer places low penalty on cost functions f that assign negative cost to expert's state-action pairs. However, if f assigns large costs (close to zero, which is the upper-bound of the regularizer) to the expert, then ψ will heavily penalize f. In the following theorems, whose proofs are reported in Appendix B, we derive the optimization problem of the JS version of our RS-GAIL algorithm by computing (6) for the above choice of the cost function regularizer $\psi(f)$. We prove the following results directly from the RS-GAIL optimization problem (5).

Theorem 2. *With the cost function regularizer $\psi(f)$ defined above, we may write*

$$\mathcal{L}_\lambda(\pi,\pi_E) = (1+\lambda) \sup_{f:\mathcal{S}\times\mathcal{A}\to(0,1)} \rho_\alpha^\lambda[F^\pi_{1,f}] - \rho_\alpha^\lambda[-F^{\pi_E}_{2,f}], \tag{8}$$

where F_1^π and $F_2^{\pi_E}$ are the loss random variables of policies π and π_E w.r.t. the cost functions $c(s,a) = \log f(s,a)$ and $c(s,a) = \log\big(1-f(s,a)\big)$, respectively.

Corollary 1. *We may write $\mathcal{L}_\lambda(\pi,\pi_E)$ in terms of the Jensen-Shannon (JS) divergence as*

$$\mathcal{L}_\lambda(\pi,\pi_E) = (1+\lambda) \sup_{d\in\mathcal{D}^\pi_\xi}\ \inf_{d'\in\mathcal{D}^{\pi_E}_\xi} D_{JS}(d,d'). \tag{9}$$

From Corollary 1, we write the optimization problem of the Jensen-Shannon version of our RS-GAIL algorithm (JS-RS-GAIL) as

$$\sup_{\lambda\geq 0}\ \min_\pi\ -H(\pi) + (1+\lambda) \sup_{d\in\mathcal{D}^\pi_\xi}\ \inf_{d'\in\mathcal{D}^{\pi_E}_\xi} D_{\text{JS}}(d,d'). \tag{10}$$

Hence in JS-RS-GAIL, instead of minimizing the original GAIL objective, we solve the optimization problem (10) that aims at matching the sets $\mathcal{D}^\pi_\xi$ and $\mathcal{D}^{\pi_E}_\xi$ w.r.t. the JS divergence.

3.3 Risk-sensitive GAIL with Wasserstein Distance

In this section, we derive RS-GAIL using occupation measure matching via the Wasserstein distance. We define the cost function regularizer $\psi(f) := \begin{cases} 0 & \text{if } f \in \mathcal{F}_1 \\ +\infty & \text{otherwise} \end{cases}$.

Corollary 2. *For the cost function regularizer $\psi(f)$ defined above, we may write*

$$\mathcal{L}_\lambda(\pi, \pi_E) = (1+\lambda) \sup_{d \in \mathcal{D}^\pi_\xi} \inf_{d' \in \mathcal{D}^{\pi_E}_\xi} W(d, d'). \tag{11}$$

Proof. See Appendix C. □

From (6) and the $\psi(f)$ defined above, we have $\mathcal{L}_\lambda(\pi, \pi_E) = \sup_{f \in \mathcal{F}_1} \rho^\lambda_\alpha[C^\pi_f] - \rho^\lambda_\alpha[C^{\pi_E}_f]$, which gives the following optimization problem for the Wasserstein version of our RS-GAIL algorithm (W-RS-GAIL):

$$\sup_{\lambda \geq 0} \min_\pi -H(\pi) + (1+\lambda) \sup_{f \in \mathcal{F}_1} \rho^\lambda_\alpha[C^\pi_f] - \rho^\lambda_\alpha[C^{\pi_E}_f]. \tag{12}$$

We conclude this section by a theorem that shows if we use a risk-neutral imitation learning algorithm to minimize the Wasserstein distance between the occupancy measures of the agent and the expert, the distance between their CVaRs could be still large. Thus, new algorithms, such as those developed in this paper, are needed for risk-sensitive imitation learning.

Theorem 3. *Let Δ be the worst-case risk difference between the agent and expert, given that their occupancy measures are δ-close ($\delta > 0$), i.e.,*

$$\begin{aligned} \Delta = \sup_{\pi, p, p_0} \sup_{f \in \mathcal{F}_1} \; & \rho_\alpha[C^\pi_f] - \rho_\alpha[C^{\pi_E}_f] \\ \textit{s.t.} \; & W(d^\pi, d^{\pi_E}) \leq \delta. \end{aligned} \tag{13}$$

Then, $\Delta \geq \frac{\delta}{\alpha}$.

Theorem 3, whose proof has been reported in Appendix C, indicates that the difference between the risks can be $1/\alpha$-times larger than that between the occupancy measures (in terms of Wasserstein).

4 Risk-sensitive Imitation Learning Algorithms

Algorithm 1 contains the pseudocode of our JS-based and Wasserstein-based risk-sensitive imitation learning algorithms. The algorithms aim at finding a saddle-point (π, f) of the objective function (5). We use the parameterizations for the policy $\theta \mapsto \pi_\theta$ and cost function (discriminator) $w \mapsto f_w$. Similar to GAIL [Ho and Ermon, 2016a], the algorithm is TRPO-based [Schulman et al., 2015] and alternates between an Adam [Kingma and Ba, 2014] gradient ascent step for the cost function parameter w and a KL-constrained gradient descent step w.r.t. a linear approximation of the objective. The details about the algorithm, including the gradients, are reported in Appendix D. In the implementation of our algorithms, we use a grid search and optimize over a finite number of the Lagrangian parameters λ. This can be seen as the agent selecting among a finite number of risk profiles of the form (mean + λCVaR$_\alpha$) when she matches her risk profile to that of the expert.

Algorithm 1: Pseudocode of JS-RS-GAIL and W-RS-GAIL Algorithms

Input: Expert trajectories $\{\tau^E_j\}_{j=1}^{N_E} \sim p^{\pi_E}$, risk level $\alpha \in (0, 1]$, and initial policy and cost function parameters θ_0 and w_0

for $i = 0, 1, 2, \ldots$ **do**

Generate N trajectories using the current policy π_{θ_i}, i.e., $\{\tau_j\}_{j=1}^N \sim p^{\pi_{\theta_i}}$

(JS-RS-GAIL) Estimate the VaRs:

$$\hat{\nu}_\alpha(F^\pi_{1, f_{w_i}}) \quad \text{and} \quad \hat{\nu}_\alpha(-F^{\pi_E}_{2, f_{w_i}})$$

(W-RS-GAIL) Estimate the VaRs:

$$\hat{\nu}_\alpha(C^\pi_{f_{w_i}}) \quad \text{and} \quad \hat{\nu}_\alpha(C^{\pi_E}_{f_{w_i}})$$

Update the discriminator parameter by computing a gradient ascent step w.r.t. the objective

(JS-RS-GAIL)

$$w_{i+1} \mapsto (1+\lambda) \Big(\rho^\lambda_\alpha[F^{\pi_{\theta_i}}_{1, f_{w_i}}] - \rho^\lambda_\alpha[-F^{\pi_E}_{2, f_{w_i}}] \Big)$$

(W-RS-GAIL)

$$w_{i+1} \mapsto (1+\lambda) \Big(\rho^\lambda_\alpha[C^{\pi_{\theta_i}}_{f_{w_i}}] - \rho^\lambda_\alpha[C^{\pi_E}_{f_{w_i}}] \Big)$$

Update the policy parameter using a KL-constrained gradient descent step w.r.t. the objective

(JS-RS-GAIL)

$$\theta_{i+1} \mapsto -H(\pi_{\theta_i}) + (1+\lambda) \rho^\lambda_\alpha[F^{\pi_{\theta_i}}_{1, f_{w_{i+1}}}]$$

(W-RS-GAIL)

$$\theta_{i+1} \mapsto -H(\pi_{\theta_i}) + (1+\lambda) \rho^\lambda_\alpha[C^{\pi_{\theta_i}}_{f_{w_{i+1}}}]$$

end

---

5 Related Work: Discussion about RAIL

We start this section by comparing the RAIL optimization problem (Eq. 9 in Santara et al. 2017a) with that of our JS-RS-GAIL reported in Eq. 10, i.e.,

RAIL:

$$\min_{\pi} -H(\pi) + (1+\lambda) \sup_{f\in(0,1)^{\mathcal{S}\times\mathcal{A}}} \rho_{\alpha}^{\lambda}[F_{1,f}^{\pi}] - \mathbb{E}[-F_{2,f}^{\pi_E}],$$

JS-RS-GAIL:

$$\min_{\pi} -H(\pi) + (1+\lambda) \sup_{f\in(0,1)^{\mathcal{S}\times\mathcal{A}}} \rho_{\alpha}^{\lambda}[F_{1,f}^{\pi}] - \rho_{\alpha}^{\lambda}[-F_{2,f}^{\pi_E}].$$

If we write the above optimization problems in terms of the JS divergence, we obtain

RAIL:

$$\min_{\pi} \; -H(\pi) + (1+\lambda) \sup_{d\in\mathcal{D}_{\xi}^{\pi}} D_{\text{JS}}(d, d^{\pi_E}), \tag{14}$$

JS-RS-GAIL:

$$\min_{\pi} \; -H(\pi) + (1+\lambda) \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} D_{\text{JS}}(d, d'). \tag{15}$$

Note that while JS in (15) matches the distorted occupancy measures (risk profiles) of the agent and expert, the JS in (14) matches the distorted occupancy measure (risk profile) of the agent with the occupancy measure (mean) of the expert. This means that RAIL does not take the expert's risk into account in its optimization.

Moreover, the results reported in Santara et al. [2017a] indicate that GAIL performs poorly in terms of optimizing the risk (VaR and CVaR). By looking at the RAIL's GitHub [Santara et al., 2017b], it seems they used the GAIL implementation from its GitHub [Ho and Ermon, 2016b]. Although we used the same GAIL implementation, we did not observe such a poor performance for GAIL, which is not that surprising since the MuJoCo domains used in the GAIL and RAIL papers are all deterministic and the policies are the only source of randomness there. This is why in our experiments in Section 6, we inject noise to the reward function of the problems. Finally, the gradient of the objective function reported in Eq. (A.3) of Santara et al. [2017a] is a scalar, which does not seem to be correct. We corrected this in our implementation of RAIL in Section 6.

6 Experimental Results

We evaluated Algorithm 1 against GAIL [Ho and Ermon, 2016a] and RAIL [Santara et al., 2017a] on 2 physics-based high-dimensional continuous control tasks (Hopper-v1 and Walker2d-v1), solved efficiently by model-free reinforcement learning [Schulman et al., 2015], [Duan et al., 2016]. All environments were simulated with MuJoCo [Todorov et al., 2012b]. Each task comes with a deterministic cost function $c(s,a)$ and a deterministic dynamics function, defined in the OpenAI Gym [Brockman et al., 2016]. For each task, an expert has been trained [Ho and Ermon, 2016a] on these true cost functions to minimize the expected cumulative cost. Because those environments are deterministic and we need a relevant framework for assessing performance in terms of risk criteria, we transform the environments such that (i) the costs and/or transitions are random and (ii) the expert (trained with respect to the original deterministic environment) is risk-sensitive with respect to the transformed stochastic environment.

For simplicity, we opted for a cost transformation, as follows. For each task, we sample trajectories from the experts to get a set of state-action pairs $D = \{(s_i, a_i)\}_{i=1}^{N}$. We run the K-Means clustering algorithm over D to obtain $K = 15$ different clusters. Let w_j be the relative proportion of observed state-action pairs in the j-th cluster. The weights w_j give a heuristic estimate of the expert occupancy measure d_{π_E}. For any state-action pair (s,a), we compute the closest pair $(\hat{s},\hat{a}) \in D$ (w.r.t. the Euclidean distance). Let j be the index of the cluster to which $(\hat{s},\hat{a})$ belongs, and let Z be a random variable independently drawn from the standard Gaussian distribution $\mathcal{N}(0,1)$. Then,

- for Hopper-v1, we use the cost transformation $c_T(s,a) = \frac{1}{0.2+\sqrt{w_j}}|Z(\omega)|c(s,a)$.
- for Walker-v1, we use the cost transformation $c_T(s,a) = \frac{0.4}{\sqrt{w_j - 0.02}}|Z(\omega)|c(s,a)$.

Importantly, the original cost function $c(s,a)$ is negative. For a given task (let's say Hopper-v1), let $E_c < 0$ be the average cost per state-action pair encountered by the expert. Qualitatively, when w_j is small, it means that the expert avoids the state-action pair (s,a) and hence, $c(s,a)$ should be much larger than E_c. Then, we chose the transformation such that $\frac{1}{0.2+\sqrt{w_j}}|Z(\omega)|$ takes values over some range $[m_s, M_s]$ with high probability, and, such that $m_s c(s,a) \geq c(s,a) \gg E_c$ and $M_s c(s,a) \ll E_c$. On the other hand, when w_j is large, it means that $c(s,a)$ is smaller or close to E_c. Moreover, the random variable $\frac{1}{0.2+\sqrt{w_j}}|Z(\omega)|$ takes values over some range $[m_L, M_L]$ more concentrated than $[m_s, M_s]$, so that the transformed cost $c_T(s,a)$ is concentrated around E_c with high probability.

We argue that this cost transformation makes the expert risk-sensitive with respect to the transformed environment. Indeed, even if the transformed cost at the state-action pair (s,a) associated with a small w_j can take a small value $c_T(s,a) \simeq M_s c(s,a) \ll E_c$, there is a non-negligible probability that $c_T(s,a)$ takes a value very large compared to E_c. Hence, this randomness makes the state-action pair 'risky', which is consistent with a small value of w_j for a risk-sensitive expert. On the other hand, if w_j is large, then the transformed cost takes values that are concentrated around E_c. It makes the state-action pair 'safe', which is consistent with a large value of w_j for a risk-sensitive expert. In particular, a risk-neutral expert would look for state-action pairs associated with small values of w_j, since the cost can be very small with non-negligible probability, and hence, yield a better mean return, on the contrary to a risk-sensitive expert who cares about the tail of the cumulative cost distribution. Before running our algorithms, we observed experimentally that those desired qualitative statements are satisfied and, thus, that the cost transformations introduced stochasticity in the cumulative cost of the expert, in a way that she is (i) risk-sensitive and (ii) there is enough variability in the cumulative cost to make the risk-sensitive imitation setting relevant.

For each task, we used JS-RS-GAIL, GAIL and RAIL to train policies of the same neural network architecture, with two layers and tanh nonlinearities in between. The first, respectively second, layer contains a number of neurons on the order of the observation space dimension, respectively action space dimension. Hence, we have a faster training procedure compared to [Ho and Ermon, 2016a] that uses 100 neurons for each layer. An additional reason to motivate our choice of a smaller neural network architecture is the recent work of [Mania et al., 2018] who have shown that policies parameterized in spaces of such dimensions can be trained to achieve state-of-the-art performance on MuJoCo tasks. The discriminator networks of Algorithm 1 also used the same architecture. For each task, we gave to all algorithms the same amount of environment interaction for training. The imitation policy is trained over 500 iterations, similarly to [Ho and Ermon, 2016a] and [Santara et al., 2017a]. At each iteration, 50000 state-action pairs are sampled to evaluate the mean, VaR_α, CVaR_α and gradients with sample averages. Table 1 shows the exact experimental performance with respect to the mean, VaR_α, CVaR_α and ρ_α^λ of the random cumulative (transformed) cost. Due to the increasing amount of samples required to estimate VaR_α and CVaR_α when α decreases, we chose $\alpha = 0.3$, meaning that we are interested in the performance for the 30% worst-case outcomes. For each task, each algorithm is run using 5 different random seeds. For each run, we sample 1000 trajectories using the trained policy. We report the average estimates of each criteria.

On the two high-dimensional control tasks, our algorithm produced policies that (i) perform at least as well as GAIL w.r.t. mean criteria and (ii) outperform GAIL w.r.t. the risk criteria ρ_α^λ. The risk performance of JS-RS-GAIL is actually much closer to the expert's one than GAIL. We also observed slight improvements over RAIL with Hopper-v1 and significant improvements over RAIL with Walker-v1.

Table 1: Learned policy performance, $\alpha = 0.3$, $\lambda = 0.05$.

Transformed cost	Expert	GAIL	RAIL	Ours
Hopper-v1				
Mean	- 6096	-5853	-6064	-6105
VaR_α	-6129	-6019	-6125	-6124
CVaR_α	-5590	-4958	-5493	-5657
ρ_α^λ	-6375	-6100	-6338	-6387

Transformed cost	Expert	GAIL	RAIL	Ours
Walker-v1				
Mean	-7651	-7231	-7363	-7572
VaR_α	-7875	-7274	-7773	-7909
CVaR_α	-6440	-5353	-5505	-5926
ρ_α^λ	-7973	-7498	-7638	-7868

7 Conclusions and Future Work

In this paper, we first formulated a risk-sensitive imitation learning setting in which the agent's goal is to have a risk profile as good as the expert's. We then derived a GAIL-like optimization problem for our formulation, which we called risk-sensitive GAIL (RS-GAIL). We proposed two risk-sensitive generative adversarial imitation learning algorithms based on two variations of RS-GAIL that match the agent and expert's risk profiles w.r.t. Jensen-Shannon (JS) divergence and Wasserstein distance. We experimented with our JS-based algorithm and compared its performance with that of GAIL [Ho and Ermon, 2016a] and RAIL [Santara et al., 2017a] in two MuJoCo tasks.

Future directions include **1)** extending our results to other popular risk measures, such as expected exponential utility and the more general class of coherent risk measures, **2)** investigating other risk-sensitive imitation learning settings, especially those in which the agent can tune its risk profile w.r.t. the expert, e.g., being a more risk averse/seeking version of the expert, and **3)** more experiments, particularly with our Wasserstein-based algorithm and in problems with higher intrinsic stochasticity.

References

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

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

V. Borkar. A sensitivity formula for the risk-sensitive cost and the actor-critic algorithm. *Systems & Control Letters*, 44:339–346, 2001.

V. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27:294–311, 2002.

J. Borwein. A very complicated proof of the minimax theorem. *Minimax Theory and its Applications*, 1(1): 21–27, 2016.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, pages 3509–3517, 2014.

Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

J. Filar, L. Kallenberg, and H. Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

J. Filar, D. Krass, and K. Ross. Percentile performance criteria for limiting average Markov decision processes. *IEEE Transaction of Automatic Control*, 40 (1):2–10, 1995.

C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 49–58, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of Advances in neural information processing systems*, pages 2672–2680, 2014.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Proceedings of Advances in Neural Information Processing Systems*, pages 4565–4573, 2016a.

J. Ho and S. Ermon. Gail github. `https://github.com/openai/imitation`, 2016b.

R. Howard and J. Matheson. Risk sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Levine and V. Koltun. Continuous inverse optimal control with locally optimal examples. In *Proceedings of the 29th International Conference on Machine Learning*, pages 41–48, 2012.

Y. Li, J. Song, and S. Ermon. InfoGAIL: Interpretable imitation learning from visual demonstrations. In *Proceedings of Advances in Neural Information Processing Systems*, 2017.

H. Mania, A. Guy, and B. Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670, 2000.

X. Nguyen, M. Wainwright, and M. Jordan. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876–904, 2009.

A. Pichler. Evaluations of risk measures for different probability measures. *SIAM Journal on Optimization*, 23(1):530–551, 2013.

D. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Proceedings of Advances in Neural Information Processing Systems 26*, pages 252–260, 2013.

R. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 26:1443–1471, 2002.

S. Ross and D. Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 661–668, 2010.

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*, pages 627–635, 2011.

A. Santara, A. Naik, B. Ravindran, D. Das, D. Mudigere, S. Avancha, and B. Kaul. Rail: Risk-averse imitation learning. *arXiv:1707.06658*, 2017a.

A. Santara, A. Naik, B. Ravindran, D. Das, D. Mudigere, S. Avancha, and B. Kaul. Rail github. `https://github.com/Santara/RAIL`, 2017b.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory,* *Second Edition*. Society for Industrial and Applied Mathematics, 2014.

S. Singh, J. Lacotte, A. Majumdar, and M. Pavone. Risk-sensitive inverse reinforcement learning via semi- and non-parametric methods. *Int. Journal of Robotics Research*, 2018.

M. Sobel. The variance of discounted Markov decision processes. *Applied Probability*, pages 794–802, 1982.

A. Tamar, D. D. Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, pages 387–396, 2012.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, pages 1468–1476, 2015a.

A. Tamar, Y. Glassner, and S. Mannor. Optimizing the cvar via sampling. In *AAAI Conference on Artificial Intelligence*, 2015b.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *Proceedings of IROS*, pages 5026–5033, 2012a.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5026–5033. IEEE, 2012b.

B. Ziebart, A. Maas, J. Bagnell, and A. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of AAAI*, pages 1433–1438, 2008.

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

A Proof of Theorem 1

Before proving the main result, we first provide the following two technical lemmas that we will later use in the main analysis.

Lemma 1 (Minimax)**.** *For any fixed policy π and any member of the risk envelop $\zeta \in \mathcal{U}^\pi$ such that $\xi = \frac{1+\lambda\zeta}{1+\lambda}$, denote by $\Lambda(f, \xi) = \mathbb{E}_\pi[\xi F_f] - \mathbb{E}_{\pi_E}[\xi F_f]$ the difference of the expected cumulative cost. Then, the following equality holds:*

$$\sup_{f\in\mathcal{C}} \inf_{\zeta\in\mathcal{U}^\pi} \Lambda\left(f, \frac{1+\lambda\zeta}{1+\lambda}\right) = \inf_{\zeta\in\mathcal{U}^\pi} \sup_{f\in\mathcal{C}} \Lambda\left(f, \frac{1+\lambda\zeta}{1+\lambda}\right) \tag{16}$$

Proof. The function $(f, \xi) \mapsto \Lambda(f, \xi)$ is linear and continuous over $\mathcal{C}$, and ζ is a linear function ξ, and linear and continuous over $\mathcal{U}^\pi$. Since $\mathcal{C}$ is convex, and $\mathcal{U}^\pi$ is nonempty, convex and weakly compact, the result follows from the Von Neumann-Fan minimax theorem [Borwein, 2016]. □

This technical result allows us to swap the $\min$ and the $\max$ operator between the cost and the risk envelops. Next we also have the following technical result justifying the duality between distorted occupation measure and the risk-sensitive probability distribution $p^\pi_\xi = \xi \cdot p^\pi$ over trajectories, for any element in risk envelop $\zeta \in \mathcal{U}^\pi$ such that $\xi = \frac{1+\lambda\zeta}{1+\lambda}$.

Lemma 2. *For any arbitrary pair of (f, ξ) such that $\zeta \in \mathcal{U}^\pi$, $\xi = \frac{1+\lambda\zeta}{1+\lambda}$, and $f \in \mathcal{C}$, the following equality holds:*

$$\mathbb{E}_\pi[\xi(\tau) C^\pi_f(\tau)] = \int_\Gamma d^\pi_\xi(s, a) f(s, a) ds\, da, \tag{17}$$

where d^π_ξ is the $\gamma-$discounted, $\xi-$distorted occupation measure.

Proof. See Theorem 3.1 in Altman [1999]. □

Utilizing the result from Lemma 1, for any arbitrary policy π, the following chain of equalities holds w.r.t. the loss function of RS-GAIL:

$$\begin{aligned}
\mathcal{L}_\lambda(\pi, \pi_E) &= (1+\lambda) \sup_{f\in\mathcal{C}} \rho^\lambda_\alpha[C^\pi_f] - \rho^\lambda_\alpha[C^{\pi_E}_f] - \psi(f) \\
&= (1+\lambda) \sup_{f\in\mathcal{C}} \sup_{\zeta\in\mathcal{U}^\pi} \mathbb{E}\left[\frac{1+\lambda\zeta}{1+\lambda} C^\pi_f\right] - \sup_{\zeta'\in\mathcal{U}^{\pi_E}} \mathbb{E}\left[\frac{1+\lambda\zeta'}{1+\lambda} C^{\pi_E}_f\right] - \psi(f) \\
&= (1+\lambda) \sup_{f\in\mathcal{C}} \sup_{\zeta\in\mathcal{U}^\pi} \inf_{\zeta'\in\mathcal{U}^{\pi_E}} \mathbb{E}\left[\frac{1+\lambda\zeta}{1+\lambda} C^\pi_f\right] - \mathbb{E}\left[\frac{1+\lambda\zeta'}{1+\lambda} C^{\pi_E}_f\right] - \psi(f) \\
&= (1+\lambda) \sup_{\zeta\in\mathcal{U}^\pi} \sup_{f\in\mathcal{C}} \inf_{\zeta'\in\mathcal{U}^{\pi_E}} \mathbb{E}\left[\frac{1+\lambda\zeta}{1+\lambda} C^\pi_f\right] - \mathbb{E}\left[\frac{1+\lambda\zeta'}{1+\lambda} C^{\pi_E}_f\right] - \psi(f)
\end{aligned} \tag{18}$$

Again by applying the result of Lemma 1 to the last expression, the loss function in RS-GAIL can be expressed as:

$$\mathcal{L}_\lambda(\pi, \pi_E) = \sup_{\zeta\in\mathcal{U}^\pi} \inf_{\zeta'\in\mathcal{U}^{\pi_E}} \sup_{f\in\mathcal{C}} (1+\lambda) \cdot \left(\mathbb{E}\left[\frac{1+\lambda\zeta}{1+\lambda} C^\pi_f\right] - \mathbb{E}\left[\frac{1+\lambda\zeta'}{1+\lambda} C^\pi_f\right]\right) - \psi(f). \tag{19}$$

Furthermore, from Lemma 2 we deduce that for any $\zeta \in \mathcal{U}^\pi$, $\zeta' \in \mathcal{U}^{\pi_E}$, and $\xi = \frac{1+\lambda\zeta}{1+\lambda}$, $\xi' = \frac{1+\lambda\zeta'}{1+\lambda}$, the following equality holds:

$$\mathbb{E}\left[\frac{1+\lambda\zeta}{1+\lambda} C^\pi_f\right] - \mathbb{E}\left[\frac{1+\lambda\zeta'}{1+\lambda} C^\pi_f\right] = \int_\Gamma (d^\pi_\xi(s, a) - d^{\pi_E}_{\xi'}(s, a)) f(s, a) ds\, da \tag{20}$$

Combining the above results with the definitions of distorted occupation measure w.r.t. radon-nikodem derivative ξ and policies π, π_E, i.e., $\mathcal{D}^\pi_\xi$ and $\mathcal{D}^{\pi_E}_\xi$, we finally obtain the following desired result:

$$\mathcal{L}_\lambda(\pi, \pi_E) = \sup_{d\in\mathcal{D}^\pi_\xi} \inf_{d'\in\mathcal{D}^{\pi_E}_\xi} \psi^*_\mathcal{C}((1+\lambda)(d - d')), \tag{21}$$

where the convex conjugate function with respect $\psi_{\mathcal{C}}^{*} : \mathbb{R}_{S \times A} \to \mathbb{R}$ is defined as

$$\psi_{\mathcal{C}}^{*}(d) = \sup_{f \in \mathcal{C}} \langle d, f \rangle - \psi(f).$$

B Proofs of RS-GAIL with Jensen Shannon Divergence

In this section, we aim to derive RS-GAIL using occupation measure matching via Jensen Shannon divergence. Consider the original RS-GAIL formulation in (4) with fixed $\lambda \geq 0$, i.e.,

$$(1+\lambda) \min_{\pi} \sup_{f \in \mathcal{C}} \rho_\alpha^\lambda[C_f^\pi] - \rho_\alpha^\lambda[C_f^{\pi_E}].$$

Following the derivation of the GAIL paper, we instead consider the following formulation

$$\min_{\pi} -H(\pi) + \sup_{f \in \mathcal{C}} \rho_\alpha^\lambda[C_f^\pi] - \rho_\alpha^\lambda[C_f^{\pi_E}] - \psi(f) \tag{22}$$

where the entropy regularizer term $H(\pi)$ in the cost incentivizes exploration in policy learning, and the reward regularizer $\psi(f)$ regularizes the inverse reinforcement learning problem.

First we want to find the cost regularizer $\psi(\cdot)$ that leads to the Jensen Shannon divergence loss function between occupation measures. To proceed, we first re-visit the following technical lemma from Ho and Ermon [2016a] about reformulating occupation measure matching as a general $f-$divergence minimization problem, where the corresponding $f-$divergence is induced by a given strictly decreasing convex surrogate function ϕ.

Lemma 3. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is a strictly decreasing convex function. Let Φ be the range of $-\phi$, and define $\psi_\phi : \mathbb{R}_{S\times A} \to \mathbb{R}$ by:*

$$\psi_\phi(f) = \begin{cases} (1+\lambda)\left(-\rho_\alpha^\lambda[C_f^{\pi_E}] + \rho_\alpha^\lambda[G_{\phi,f}^{\pi_E}]\right) & \text{if } f(s,a) \in \Phi, \forall s,a \\ \infty & \text{otherwise} \end{cases}, \tag{23}$$

where $G_{\phi,f}^{\pi_E}$ is the γ-discounted cumulative cost function $G_{\phi,f}^{\pi_E} = \sum_t^\infty \gamma^t - \phi(-\phi^{-1}(-f(s_t,a_t)))$ that is induced by policy π_E. Then, ψ_ϕ is closed, proper, and convex, and by using $\psi = \psi_\phi$ as the cost regularizer, the optimization problem in (5) *is equivalent to*

$$\sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} -R_{\lambda,\phi}(d,d'),$$

where $R_{\lambda,\phi}$ is the minimum expected risk induced by surrogate loss function ϕ, i.e., $R_{\lambda,\phi}(d,d') = (1+\lambda)\sum_{s,a}\min_{\gamma\in\mathbb{R}} d(s,a)\phi(\gamma) + d'(s,a)\phi(-\gamma)$.

Proof. To start with, recall from (5) the following inner objective function of RS-GAIL:

$$\mathcal{L}_\lambda(\pi,\pi_E) = \sup_{f\in\mathcal{C}} (1+\lambda)\big(\rho_\alpha^\lambda[C_f^\pi] - \rho_\alpha^\lambda[C_f^{\pi_E}] - \psi(f)\big).$$

Using the definition of the above regularizer (which is a difference of convex function in f), one has the following chain of inequalities

$$\begin{aligned}
\sup_{d\in\mathcal{D}_\xi^\pi} (1+\lambda)\big(\rho_\alpha^\lambda[C_f^\pi] - \rho_\alpha^\lambda[C_f^{\pi_E}]\big) - \psi_\phi(f) =& (1+\lambda) \sup_{f\in\Phi} \rho_\alpha^\lambda[C_f^\pi] - \rho_\alpha^\lambda[G_{\phi,f}^{\pi_E}] \\
=& (1+\lambda) \sup_{d\in\mathcal{D}_\xi^\pi} \sup_{f\in\Phi} \langle d,f\rangle - \rho_\alpha^\lambda[G_{\phi,f}^{\pi_E}] \\
\overset{(a)}{=}& (1+\lambda) \sup_{d\in\mathcal{D}_\xi^\pi} \sup_{f\in\Phi} \inf_{d'\in\mathcal{D}_\xi^{\pi_E}} \langle d,f\rangle - \langle d', \phi(-\phi^{-1}(-f))\rangle \\
\overset{(b)}{=}& (1+\lambda) \sup_{d\in\mathcal{D}_\xi^\pi} \inf_{d'\in\mathcal{D}_\xi^{\pi_E}} \sup_{f\in\Phi} \langle d,f\rangle - \langle d', \phi(-\phi^{-1}(-f))\rangle,
\end{aligned}$$

where the first equality and the second equality follow from the definitions of ψ_ϕ and the dual representation theorem of coherent risk measure $\rho_\alpha^\lambda[C_{\phi,f}^{\pi_E}]$, the equality in (a) is based on the dual representation theorem of coherent risk $\rho_\alpha^\lambda[G_{\phi,f}^{\pi_E}] = \sup_{d'\in\mathcal{D}_\xi^{\pi_E}} \langle d', -\phi(-\phi^{-1}(-f))\rangle$, and the equality in (b) is based on strong duality, i.e., $\kappa_d(d',f) =$

$\langle d, f\rangle - \langle d', \phi(-\phi^{-1}(-f)\rangle$ is concave in f and is convex in d', and both $\mathcal{D}_{\xi}^{\pi_E}$ and Φ are convex sets. Utilizing the arguments from Proposition A.1 in Ho and Ermon [2016a], the above expression can further be re-written as

$$
\begin{aligned}
&(1+\lambda) \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} \sup_{f\in\Phi} \langle d, f\rangle - \langle d', \phi(-\phi^{-1}(-f))\rangle \\
\overset{(a)}{=}&(1+\lambda) \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} \sum_{s,a} \sup_{\tilde{f}\in\Phi} \left[d(s,a)\tilde{f} - d'(s,a)\phi(-\phi^{-1}(-\tilde{f}))\right] \\
=&(1+\lambda) \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} \sum_{s,a} \sup_{\gamma\in\mathbb{R}} \left[d(s,a)(-\phi(\gamma)) - d'(s,a)\phi(-\gamma)\right], \text{ where } f = -\phi(\gamma) \\
=& \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} -R_{\lambda,\phi}(d,d').
\end{aligned}
$$

The equality in (a) is due to the fact that the outer maximization in the first line is with respect to the cost function f, and the inner maximization in the second line is with respect to an element of the cost function (which is denoted by $\tilde{f}$). The second equality is due to a one-to-one mapping of $f = -\phi(\gamma)$, and the third equality follows from the definition of $R_{\lambda,\phi}(d,d')$. This completes the proof of this result. □

B.1 Proof of Theorem 2

Now we turn to the main result of this section. The following theorem transform the loss function of RS-GAIL into a Jensen Shannon divergence loss function, using the cost regularizer in (23), with the logistic loss $\phi(x) = \log(1+\exp(-x))$, as suggested by the discussions in Section 2.1.4 of Nguyen et al. [2009].

Recall from Lemma 3 that the inner problem of RS-GAIL, i.e., problem (6), can be rewritten as

$$
\sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} -R_{\lambda,\phi}(d,d').
$$

Therefore, one can reformulate the objective function $-R_{\phi}(d,d')$ in this problem as follows:

$$
\begin{aligned}
-R_{\lambda,\phi}(d,d') &= (1+\lambda)\sum_{s,a} \max_{\gamma\in\mathbb{R}} d(s,a)\log\left(\frac{1}{1+\exp(-\gamma)}\right) + d'(s,a)\log\left(\frac{1}{1+\exp(\gamma)}\right) \\
&= (1+\lambda)\sum_{s,a} \max_{\gamma\in\mathbb{R}} d(s,a)\log\left(\sigma(\gamma)\right) + d'(s,a)\log\left(1-\sigma(\gamma)\right) \\
&= (1+\lambda) \sup_{f:S\times A\to(0,1)} \sum_{s,a} d(s,a)\log\left(f(s,a)\right) + d'(s,a)\log\left(1-f(s,a)\right),
\end{aligned}
$$

where $\sigma(\gamma) = 1/(1+\exp(-\gamma))$ is a sigmoid function, and because its range is $(0,1)$, one can further express the inner optimization problem using the discriminator form, given in the third equality. Now consider the objective function $\sum_{s,a} d(s,a)\log\left(f(s,a)\right) + d'(s,a)\log\left(1-f(s,a)\right)$.

Notice that the objective function $(1+\lambda)\sum_{s,a} d(s,a)\log\left(f(s,a)\right) + d'(s,a)\log\left(1-f(s,a)\right)$ is concave in f, and is linear in d' and d. Using the Minimax theorem in Lemma 1, one can swap the $\inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}}$ operator and the $\sup_{f:S\times A\to(0,1)}$ operator in problem (6), i.e.,

$$
\begin{aligned}
&\sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} -R_{\phi}(d,d') \\
=&(1+\lambda)\cdot \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \sup_{f:S\times A\to(0,1)} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} \sum_{s,a} d(s,a)\log\left(f(s,a)\right) + d'(s,a)\log\left(1-f(s,a)\right) \\
=&(1+\lambda)\cdot \sup_{f:S\times A\to(0,1)} \sup_{d\in\mathcal{D}_{\xi}^{\pi}} \inf_{d'\in\mathcal{D}_{\xi}^{\pi_E}} \sum_{s,a} d(s,a)\log\left(f(s,a)\right) + d'(s,a)\log\left(1-f(s,a)\right).
\end{aligned}
$$

Furthermore, by using the equivalence of supremum (or infimum) between the set of distorted occupation measures of

$\mathcal{D}_\xi^\pi$ (or $\mathcal{D}_\xi^{\pi_E}$) and the set of risk envelop $\mathcal{U}^\pi$ (or $\mathcal{U}^{\pi_E}$), one has the following chain of equalities:

$$\begin{aligned}
&\frac{1}{1+\lambda}\cdot \sup_{\zeta\in\mathcal{U}^\pi:\xi=\frac{1+\lambda\zeta}{1+\lambda}} \inf_{\zeta'\in\mathcal{U}^{\pi_E}:\xi'=\frac{1+\lambda\zeta'}{1+\lambda}} -R_\phi(d_\xi^\pi, d_{\xi'}^{\pi_E})\\
=&\sup_{f:S\times A\to(0,1)} \sup_{\zeta\in\mathcal{U}^\pi:\xi=\frac{1+\lambda\zeta}{1+\lambda}} \inf_{\zeta'\in\mathcal{U}^{\pi_E}:\xi'=\frac{1+\lambda\zeta'}{1+\lambda}} \sum_{s,a} d_\xi^\pi(s,a)\log\left(f(s,a)\right) + d_{\xi'}^{\pi_E}(s,a)\log\left(1-f(s,a)\right)\\
=&\sup_{f:S\times A\to(0,1)} \sup_{\zeta\in\mathcal{U}^\pi:\xi=\frac{1+\lambda\zeta}{1+\lambda}} \sum_{s,a} d_\xi^\pi(s,a)\log\left(f(s,a)\right) - \sup_{\zeta'\in\mathcal{U}^{\pi_E}:\xi'=\frac{1+\lambda\zeta'}{1+\lambda}} \sum_{s,a} d_{\xi'}^{\pi_E}(s,a)\left(-\log\left(1-f(s,a)\right)\right)\\
=&\sup_{f:S\times A\to(0,1)} \rho_\alpha^\lambda[F_{1,f}^\pi] - \rho_\alpha^\lambda[-F_{2,f}^{\pi_E}],
\end{aligned}$$

where the first equality and second equality follow from basic arguments in optimization theory, and the third equality follows from the dual representation theory of coherent risk measures of $\rho_\alpha^\lambda[F_{1,f}^\pi]$ and $\rho_\alpha^\lambda[-F_{2,f}^{\pi_E}]$. This completes the proof.

Combining this result to the original problem formulation in (22), one completes the proof of this theorem.

B.2 Proof of Corollary 1

In order to show the following equality:

$$(1+\lambda)\sup_{f:S\times A\to(0,1)} \rho_\alpha^\lambda[F_{1,f}^\pi] - \rho_\alpha^\lambda[-F_{2,f}^{\pi_E}] = (1+\lambda)\sup_{d\in\mathcal{D}_\xi^\pi}\inf_{d'\in\mathcal{D}_\xi^{\pi_E}} D_{\text{JS}}(d,d'),$$

we utilize the fact that the left side is equal to $\sup_{d\in\mathcal{D}_\xi^\pi}\inf_{d'\in\mathcal{D}_\xi^{\pi_E}} -R_\phi(d,d')$, and in the following proof we instead show that the following equality holds:

$$\sup_{d\in\mathcal{D}_\xi^\pi}\inf_{d'\in\mathcal{D}_\xi^{\pi_E}} -R_\phi(d,d') = (1+\lambda)\sup_{d\in\mathcal{D}_\xi^\pi}\inf_{d'\in\mathcal{D}_\xi^{\pi_E}} D_{\text{JS}}(d,d'). \tag{24}$$

For any $d\in\mathcal{D}_\xi^\pi$ and $d'\in\mathcal{D}_\xi^{\pi_E}$, consider the optimization problem:

$$\sum_{s,a}\max_{\tilde{f}\in(0,1)} d(s,a)\log\left(\tilde{f}\right) + d'(s,a)\log\left(1-\tilde{f}\right) \tag{25}$$

For each state-action pair (s,a), since the optimization problem has a concave objective function, by the first order optimality, $\tilde{f}^*$ can be found by:

$$(1-\tilde{f}^*)d(s,a) - \tilde{f}^* d'(s,a) = 0 \quad\implies \tilde{f}^* = \frac{d(s,a)}{d(s,a)+d'(s,a)} \in (0,1).$$

By putting the optimizer back to the problem, one can show that

$$(25) = \sum_{s,a} d(s,a)\log\left(\frac{d(s,a)}{d(s,a)+d'(s,a)}\right) + d'(s,a)\log\left(\frac{d'(s,a)}{d(s,a)+d'(s,a)}\right).$$

Then by putting this result back to (25), one in turn shows that

$$\sup_{d\in\mathcal{D}_\xi^\pi}\inf_{d'\in\mathcal{D}_\xi^{\pi_E}} -R_\phi(d,d') = (1+\lambda)(-\log(4) + \sup_{d\in\mathcal{D}_\xi^\pi}\inf_{d'\in\mathcal{D}_\xi^{\pi_E}} D_{\text{JS}}(d,d')),$$

which completes the proof of this corollary.

C Proofs of RS-GAIL with Wasserstein Distance

Proof of Corollary 2

Corollary 2. *For the cost function regularizer $\psi(f)$ defined above, we may write*

$$\mathcal{L}_\lambda(\pi, \pi_E) = (1+\lambda) \sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} W(d, d').$$

Proof. From Eq. 6, we may write

$$\begin{aligned}
\mathcal{L}_\lambda(\pi, \pi_E) &= \sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} \psi^*\big((1+\lambda)(d-d')\big) \\
&\overset{\textbf{(a)}}{=} (1+\lambda) \sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} \sup_{f \in \mathcal{C}} (d-d')^\top f - \psi(f) \\
&\overset{\textbf{(b)}}{=} (1+\lambda) \sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} \sup_{f \in \mathcal{F}_1} (d-d')^\top f \\
&= (1+\lambda) \sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} \sup_{f \in \mathcal{F}_1} \mathbb{E}_d[f(s,a)] - \mathbb{E}_{d'}[f(s,a)] \\
&\overset{\textbf{(c)}}{=} (1+\lambda) \sup_{d \in \mathcal{D}_\xi^\pi} \inf_{d' \in \mathcal{D}_\xi^{\pi_E}} W(d, d'),
\end{aligned}$$

(a) is from the definition of ψ^*, **(b)** is from the definition of $\psi(f)$, and **(c)** is from the definition of the Wasserstein distance. ☐

Proof of Theorem 3

Theorem 3. *Let Δ be the worst-case risk difference between the agent and expert, given that their occupancy measures are δ-close ($\delta > 0$), i.e.,*

$$\Delta = \sup_{p, p_0, \pi} \sup_{f \in \mathcal{F}_1} \rho_\alpha[C_f^\pi] - \rho_\alpha[C_f^{\pi_E}], \qquad \textit{s.t.} \;\; W(d^\pi, d^{\pi_E}) \le \delta. \tag{26}$$

Then, $\Delta \ge \frac{\delta}{\alpha}$.

Proof. Let $\|\cdot\|$ be a norm on the state-action space $\mathcal{S} \times \mathcal{A}$ and denote by Γ the set of trajectories with horizon T. For a trajectory $\tau = (s_0, a_0, s_1, \ldots, s_T, a_T) \in \Gamma$, we define $\|\tau\|_\Gamma = \sum_{t=0}^T \gamma^t \|(s_t, a_t)\|$. The function $\|\cdot\|_\Gamma$ defines a norm on the trajectory space Γ. Let $\mathcal{G}_1$ be the space of 1-Lipschitz functions over Γ with respect to $\|\cdot\|_\Gamma$. In particular, for $f \in \mathcal{F}_1$ and trajectories τ, τ', we have

$$\begin{aligned}
|C_f(\tau) - C_f(\tau')| &= \Big| \sum_{t=0}^T \gamma^t \left(f(s_t, a_t) - f(s'_t, a'_t) \right) \Big| \\
&\le \sum_{t=0}^T \gamma^t |f(s_t, a_t) - f(s'_t, a'_t)| \\
&\underset{(a)}{\le} \sum_{t=0}^T \gamma^t \|(s_t, a_t) - (s'_t, a'_t)\| \\
&= \|\tau - \tau'\|_\Gamma
\end{aligned}$$

where (a) holds because f is 1-Lipschitz over $\mathcal{S} \times \mathcal{A}$. Hence, for $f \in \mathcal{F}_1$, we have $C_f \in \mathcal{G}_1$. Then, it implies

$$\{(\pi, p, p_0) \mid W(p^\pi, p^{\pi_E}) \le \delta\} \subseteq \{(\pi, p, p_0) \mid W(d^\pi, d^{\pi_E}) \le \delta\} \tag{27}$$

where, p^π (resp. p^{π_E}) denotes the distribution over Γ induced by (π, p, p_0) (resp. (π_E, p, p_0)). Indeed, if $(\pi, p, p_0) \in \{(\pi, p, p_0) \mid W(p^\pi, p^{\pi_E}) \le \delta\}$, then, for any $G \in \mathcal{G}_1$, we have that

$$\mathbb{E}_{p^\pi}[G(\tau)] - \mathbb{E}_{p^{\pi_E}}[G(\tau)] \le \delta$$

For $f \in \mathcal{F}_1$, since $C_f \in \mathcal{G}_1$, we get that $\mathbb{E}_{p^\pi}[C_f(\tau)] - \mathbb{E}_{p^{\pi_E}}[C_f(\tau)] \leq \delta$, which proves (27).

Therefore, we can lower bound Δ as follows

$$\Delta \geq \tilde{\Delta} := \sup_{f \in \mathcal{F}_1} \sup_{(\pi, p, p_0); W(p^\pi, p^{\pi_E}) \leq \delta} \rho_\alpha[C_f^\pi] - \rho_\alpha[C_f^{\pi_E}]. \tag{28}$$

By Theorem 15 in [Pichler, 2013], we have that $\tilde{\Delta} \geq \frac{\delta}{\alpha}$, which concludes the proof.

$\square$

D Gradient formulas

In order to derive the expression of the gradients for JS-RS-GAIL, we first make the following assumption regarding the uniqueness of the quantiles of the random cumulative cost with respect to any cost and policy parameters.

Assumption 1. *For any $\alpha \in (0,1)$, $\theta \in \Theta$ and $w \in \mathcal{W}$, there exists a unique $z_\alpha^\theta \in \mathbb{R}$ (respectively $z_\alpha^{\pi_E} \in \mathbb{R}$) such that $\mathbb{P}[F_{1,f_w}^{\pi_\theta} \leq z_\alpha^\theta] = 1-\alpha$ (respectively $\mathbb{P}[-F_{2,f_w}^{\pi_E} \leq z_\alpha^{\pi_E}] = 1-\alpha$).*

Lemma 4. *Let $\theta \in \Theta$ and $w \in \mathcal{W}$. Then,*

1. *$\rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \inf_{\nu \in \mathbb{R}} \left(\nu + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu]_+\right)$, where $x_+ = \max(x,0)$.*
2. *There exists a unique $\nu^* \in \mathbb{R}$ such that $\rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \nu^* + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu^*]_+$.*
3. *$\nu^* = \nu_\alpha(F_{1,f_w}^{\pi_\theta})$*

Proof. The first point is a standard result about the Conditional-Value-at-Risk (see Shapiro et al. 2014). The second and third points stem from Assumption 1, and Theorem 6.2 in Shapiro et al. [2014]. $\square$

We are now ready to prove the expression of the gradient of $(1+\lambda)\left(\rho_\alpha^\lambda[F_{1,f_w}^{\pi_\theta}] - \rho_\alpha^\lambda[-F_{2,f_w}^{\pi_E}]\right)$ with respect to w. First, we tackle the difficult term of the objective corresponding to the CVaR.

Theorem 4. *For any $\theta \in \Theta$ and any $w \in \mathcal{W}$, we have:*

$$\begin{cases} \nabla_w \rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \frac{1}{\alpha}\mathbb{E}\left[\mathbf{1}_{\{F_{1,w}^{\pi_\theta}(\tau) \geq \nu_\alpha(F_{1,w}^{\pi_\theta})\}} \nabla_w F_{1,w}^{\pi_\theta}(\tau)\right] \\ \nabla_w \rho_\alpha[-F_{2,f_w}^{\pi_E}] = -\frac{1}{\alpha}\mathbb{E}\left[\mathbf{1}_{\{-F_{2,w}^{\pi_E}(\tau) \geq \nu_\alpha(-F_{2,w}^{\pi_E})\}} \nabla_w F_{2,w}^{\pi_E}(\tau)\right] \end{cases} \tag{29}$$

Proof. From Lemma 4, we have that for any $\epsilon > 0$:

$$\rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \inf_{\nu \in [\nu_w^* - \epsilon, \nu_w^* + \epsilon]} \left(\nu + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu]_+\right) \tag{30}$$

where $\nu^* = \nu_\alpha(F_{1,f_w}^{\pi_\theta})$. The set of minimizers Λ of the RHS in Eq. 30 is the singleton $\{\nu_w^*\}$. The interval $[\nu^* - \epsilon, \nu^* + \epsilon]$ is nonempty and compact. By Assumption 1, for any $\nu \in \mathbb{R}$, the function $w \mapsto \nu + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu]_+$is differentiable. The function $(w,\nu) \mapsto \nabla_w\left(\nu + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu]_+\right)$ is continuous. Therefore, we can apply Danskin's theorem [Shapiro et al., 2014] to deduce that $w \mapsto \rho_\alpha[F_{1,f_w}^{\pi_\theta}]$ is differentiable and $\nabla_w \rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \nabla_w\left(\nu^* + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu^*]_+\right)$. It is immediately observed that $\nabla_w\left(\nu^* + \frac{1}{\alpha}\mathbb{E}[F_{f_w} - \nu^*]_+\right) = \mathbb{E}_\theta\left[\frac{1}{\alpha}\mathbf{1}_{\{F_{1,f_w}^{\pi_\theta}(\tau) \geq \nu_\alpha(F_{1,f_w}^{\pi_\theta})\}} \nabla_w F_{1,f_w}^{\pi_\theta}(\tau)\right]$. Similar steps can be carried out to show that $\nabla_w \rho_\alpha[-F_{2,f_w}^{\pi_E}] = -\frac{1}{\alpha}\mathbb{E}\left[\mathbf{1}_{\{-F_{2,f_w}^{\pi_E}(\tau) \geq \nu_\alpha(-F_{2,f_w}^{\pi_E})\}} \nabla_w F_{2,f_w}^{\pi_E}(\tau)\right]$. $\square$

We are now ready to give the sample average estimator expressions for the gradient of the whole objective with respect to the discriminator parameter $w \in \mathcal{W}$.

Corollary 3. *Given trajectories $\{\tau_j\}_{j=1}^N$ sampled from π_θ, trajectories $\{\tau_j^E\}_{j=1}^{N_E}$ sampled from π_E and a cost function parameter $w \in \mathcal{W}$, an estimator of the gradient of $(1+\lambda)\left(\rho_\alpha^\lambda[F_{1,f_w}^{\pi_\theta}] - \rho_\alpha^\lambda[-F_{2,f_w}^{\pi_E}]\right)$ with respect to w is given by*

$$\frac{1}{\alpha N}\sum_{j=1}^{N}\left(\alpha + \lambda\mathbf{1}_{\{F_{1,f_w}^{\pi_\theta}(\tau_j) \geq \hat{\nu}_\alpha(F_{1,f_w}^{\pi_\theta})\}}\right)\nabla_w F_{1,f_w}^{\pi_\theta}(\tau_j) + \frac{1}{\alpha N_E}\sum_{j=1}^{N_E}\left(\alpha + \lambda\mathbf{1}_{\{-F_{2,f_w}^{\pi_E}(\tau_j^E) \geq \hat{\nu}_\alpha(-F_{2,f_w}^{\pi_E})\}}\right)\nabla_w F_{2,f_w}^{\pi_E}(\tau_j^E) \tag{31}$$

Lemma 5. *For any $\theta \in \Theta$, the causal entropy gradient is given by*

$$\nabla_\theta H(\pi_\theta) = \mathbb{E}_{d_{\pi_\theta}}[\nabla_\theta \log \pi_\theta(a \mid s) Q_{log}(s,a)] \tag{32}$$

where $Q_{log}(\bar{s},\bar{a}) = \mathbb{E}_{d_{\pi_\theta}}[-\log \pi_\theta(a \mid s) \mid s_0 = \bar{s}, a_0 = \bar{a}]$

Proof. We refer to the proof of Lemma A.1 in Ho and Ermon [2016a]. □

Lemma 6. *For any $\theta \in \Theta$ and $w \in \mathcal{W}$, we have*

$$\nabla_\theta \rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \frac{1}{\alpha}\mathbb{E}\left[\nabla_\theta \log \pi_\theta(\tau)\left(F_{1,f_w}^{\pi_\theta}(\tau) - \nu_\alpha(F_{1,f_w}^{\pi_\theta})\right)_+\right] \tag{33}$$

and $\nabla_\theta \log \pi_\theta(\tau) = \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t \mid s_t)$*, with* $\tau = (s_0, a_0, \ldots, s_T, a_T)$.

Proof. We refer the reader to the proof in Tamar et al. [2015a]. □

In order to carry out a policy step, we heuristically adapt the TRPO algorithm [Schulman et al., 2015] to our objective function. In particular, instead of considering the linear approximation of the standard risk-neutral objective, we consider the first-order approximation of the mean + λCVaR$_\alpha$ objective. In order to linearize the CVaR$_\alpha$ term, we use an empirical estimator of the gradient expression given by (33). Then, we minimize this first-order approximation with the additional trust region constraint of the TRPO algorithm.

Using similar assumptions and arguments, we get the following expressions of the gradients for W-RS-GAIL.

Theorem 5 (W-RS-GAIL, gradient with respect to cost function parameter)**.**

$$\begin{aligned}\nabla_w(1+\lambda)\left(\rho_\alpha^\lambda[C_{f_w}^{\pi_\theta}] - \rho_\alpha^\lambda[C_{f_w}^{\pi_E}]\right) =& \frac{1}{\alpha}\mathbb{E}\left[(\alpha + \lambda \mathbf{1}_{\{C_{f_w}^{\pi_\theta}(\tau) \geq \nu_\alpha(C_{f_w}^{\pi_\theta})\}})\nabla_w C_{f_w}^{\pi_\theta}(\tau)\right] \\ & - \frac{1}{\alpha}\mathbb{E}\left[(\alpha + \lambda \mathbf{1}_{\{C_{f_w}^{\pi_E}(\tau) \geq \nu_\alpha(C_{f_w}^{\pi_E})\}})\nabla_w C_{f_w}^{\pi_E}(\tau)\right]\end{aligned} \tag{34}$$

Theorem 6 (W-RS-GAIL, gradient with respect to policy parameter)**.**

$$\nabla_\theta \rho_\alpha^\lambda[C_{f_w}^{\pi_\theta}] = \frac{1}{\alpha}\mathbb{E}\left[\nabla_\theta \log \pi_\theta(\tau)\left(C_{f_w}^{\pi_\theta}(\tau) - \nu_\alpha(C_{f_w}^{\pi_\theta})\right)_+\right] \tag{35}$$

E Additional experimental details

Table 2 describes the neural network architectures of the policy, reward and value functions for each task. We used a discount factor $\gamma = .995$.

Table 2: Neural network architectures

Task	Policy	Reward	Value function
Hopper-v1	11 - tanh - 3 - tanh	32 - tanh - 16 - tanh	11 - tanh - 3 - tanh
Walker-v1	32 - tanh - 16 - tanh	32 - tanh - 16 - tanh	32 - tanh - 16 - tanh