A Proof of Theorem 1

Before proving the theorem, we first state and prove the following two technical lemmas that we will later use them in the proof of Theoren 1.

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

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

Proof. The function $(f, \xi) \mapsto \Lambda(f, \xi)$ is linear and continuous over $\mathcal{C}$, and ξ is a linear function of ζ and is 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]. □

The result of Lemma 1 allows us to swap the min and max operators between the cost and risk envelops.

We now prove the following technical lemma that justifies the duality between the distorted occupation measure and the risk-sensitive probability distribution $p_\xi^\pi = \xi \cdot p^\pi$ over trajectories, for $\xi = \frac{1+\lambda\zeta}{1+\lambda}$, when $\zeta \in \mathcal{U}^\pi$ is any element of the risk envelop.

Lemma 2. *For any arbitrary pair (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:*

$$\mathbb{E}_\pi[\xi(\tau) C_f^\pi(\tau)] = \int_\Gamma d_\xi^\pi(s,a) f(s,a) ds\, da,$$

where d_ξ^π is the γ-discounted ξ-distorted occupation measure of policy π.

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

Using Lemma 1, for any arbitrary policy π, the following chain of equalities holds for the loss function of RS-GAIL:

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

By applying Lemma 1 to the last expression, the loss function in RS-GAIL can be expressed as

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

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:

$$\mathbb{E}\left[\frac{1+\lambda\zeta}{1+\lambda} C_f^\pi\right] - \mathbb{E}\left[\frac{1+\lambda\zeta'}{1+\lambda} C_f^\pi\right] = \int_\Gamma \left(d_\xi^\pi(s,a) - d_{\xi'}^{\pi_E}(s,a)\right) f(s,a)\; ds\; da.$$

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

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

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 of Eq. 4 with fixed $\lambda \geq 0$, i.e.,

$$(1+\lambda)\min_{\pi}\sup_{f\in\mathcal{C}}\rho_\alpha^\lambda[C_f^\pi]-\rho_\alpha^\lambda[C_f^{\pi_E}]. \tag{15}$$

Following the derivation of the GAIL paper, we replace (15) with the following formulation:

$$\min_{\pi} -H(\pi)+\sup_{f\in\mathcal{C}}\rho_\alpha^\lambda[C_f^\pi]-\rho_\alpha^\lambda[C_f^{\pi_E}]-\psi(f), \tag{16}$$

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

We first want to find the cost regularizer $\psi(\cdot)$ that leads to the Jensen Shannon divergence loss function between the occupation measures. To proceed, we revisit 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. *Let $\phi:\mathbb{R}\rightarrow\mathbb{R}$ be a strictly decreasing convex function and Φ be the range of $-\phi$. We define $\psi_\phi:\mathbb{R}_{S\times A}\rightarrow\mathbb{R}$ as*

$$\psi_\phi(f)=\begin{cases}(1+\lambda)\left(-\rho_\alpha^\lambda[C_f^{\pi_E}]+\rho_\alpha^\lambda[G_{\phi,f}^{\pi_E}]\right) & \text{if } f(s,a)\in\Phi,\ \forall s,a\\ \infty & \text{otherwise}\end{cases}, \tag{17}$$

where $G_{\phi,f}^{\pi_E}$ is the γ-discounted cumulative cost function $G_{\phi,f}^{\pi_E}=-\sum_{t=0}^{\infty}\gamma^t\phi\Big(-\phi^{-1}\big(-f(s_t,a_t)\big)\Big)$ that is induced by policy π_E. Then, ψ_ϕ is closed, proper, and convex. Using $\psi=\psi_\phi$ as the cost regularizer, the optimization problem (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 the 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. From (5), recall 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), we have the following chain of inequalities:

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

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

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

(a) is due to the fact that the outer maximization on the first line is w.r.t. the cost function f, and the inner maximization on the second line is w.r.t. an element of the cost function (which is denoted by $\tilde f$). The second equality is due to the one-to-one mapping of $f=-\phi(\gamma)$. The third equality follows from the definition of $R_{\lambda,\phi}(d,d')$. This completes the proof. □

B.1 Proof of Theorem 2

We now turn to the main result of this section. The following theorem transforms the loss function of RS-GAIL into a Jensen Shannon divergence loss function using the cost regularizer in (17), 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., the problem in Eq. 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, we can reformulate the objective function $-R_\phi(d,d')$ in this problem as

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

where $\sigma(\gamma)=1/\big(1+\exp(-\gamma)\big)$ is a sigmoid function, and since 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 $\sum_{s,a}d(s,a)\log\big(f(s,a)\big)+d'(s,a)\log\big(1-f(s,a)\big)$. This objective function is concave in f and linear in d and d'. Using the minimax theorem in Lemma 1, one can swap the $\inf_{d'\in\mathcal{D}_\xi^{\pi_E}}$ and $\sup_{f:S\times A\to(0,1)}$ operators in (6), i.e.,

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

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

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

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

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

Combining this result with the original problem formulation in (16) completes the proof.

B.2 Proof of Corollary 1

In order to show the following equality:

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

we utilize the fact that the left-hand-side is equal to

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

In proving Corollary 1, we instead show that the following equality holds:

$$\sup_{d\in\mathcal{D}_\xi^\pi}\ \inf_{d'\in\mathcal{D}_\xi^{\pi_E}} -R_\phi(d,d') = (1+\lambda)\sup_{d\in\mathcal{D}_\xi^\pi}\ \inf_{d'\in\mathcal{D}_\xi^{\pi_E}} D_{\mathrm{JS}}(d,d'). \tag{18}$$

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

$$\sum_{s,a}\max_{\tilde f\in(0,1)} d(s,a)\log(\tilde f) + d'(s,a)\log(1-\tilde f). \tag{19}$$

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 as

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

By putting the optimizer back to the problem, one can show that (19) may be rewritten as

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

Then by putting this result back to (18), one may show that

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

which completes the proof.

C Proofs of RS-GAIL with Wasserstein Distance

C.1 Proof of Corollary 2

Corollary 2. *For the cost function regularizer* $\psi(f) := \begin{cases} 0 & \text{if } f \in \mathcal{F}_1 \\ +\infty & \text{otherwise} \end{cases}$*, 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. □

C.2 Proof of Theorem 3

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

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

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

Proof. Let $\|\cdot\|$ be a norm on the state-action space $\mathcal{S} \times \mathcal{A}$ and denote by Γ the set of all trajectories with horizon T. For a trajectory $\tau = (s_0, a_0, s_1, \dots, s_T, a_T) \in \Gamma$, 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 w.r.t. $\|\cdot\|_\Gamma$. In particular, for $f \in \mathcal{F}_1$ and trajectories τ and τ', we have

$$\begin{aligned}
|C_f(\tau) - C_f(\tau')| &= |\sum_{t=0}^T \gamma^t \left(f(s_t, a_t) - f(s_t', a_t')\right)| \\
&\leq \sum_{t=0}^T \gamma^t |f(s_t, a_t) - f(s_t', a_t')| \\
&\overset{\text{(a)}}{\leq} \sum_{t=0}^T \gamma^t \|(s_t, a_t) - (s_t', a_t')\| \\
&= \|\tau - \tau'\|_\Gamma,
\end{aligned}$$

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

$$\big\{(\pi, p, p_0) \mid W(p^\pi, p^{\pi_E}) \leq \delta\big\} \subseteq \big\{(\pi, p, p_0) \mid W(d^\pi, d^{\pi_E}) \leq \delta\big\}, \tag{20}$$

where p^{π} and p^{π_E} denote the distributions over Γ induced by (π, p, p_0) and (π_E, p, p_0), respectively. Indeed, if $(\pi, p, p_0) \in \big\{(\pi, p, p_0) \mid W(p^{\pi}, p^{\pi_E}) \leq \delta\big\}$, then for any $G \in \mathcal{G}_1$, we have

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

For $f \in \mathcal{F}_1$, since $C_f \in \mathcal{G}_1$, we obtain $\mathbb{E}_{p^{\pi}}\big[C_f(\tau)\big] - \mathbb{E}_{p^{\pi_E}}\big[C_f(\tau)\big] \leq \delta$, which proves (20). Therefore, we can lower-bound Δ as

$$\Delta \geq \tilde{\Delta} := \sup_{f \in \mathcal{F}_1} \sup_{(\pi,p,p_0);W(p^{\pi},p^{\pi_E}) \leq \delta} \rho_\alpha[C_f^{\pi}] - \rho_\alpha[C_f^{\pi_E}]. \tag{21}$$

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

D Algorithmic Details

D.1 JS-RS-GAIL

D.1.1 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 w.r.t. 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 item is a standard result about the Conditional-Value-at-Risk (see Shapiro et al. [2014]). The second and third items stem from Assumption 1 and Theorem 6.2 in Shapiro et al. [2014]. □

Lemma 5. *For any $\theta \in \Theta$ and any $w \in \mathcal{W}$, we have*

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

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

$$\rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \inf_{\nu \in [\nu_w^* - \epsilon, \nu_w^* + \epsilon]} \left(\nu + \frac{1}{\alpha}\mathbb{E}[F_{1,f_w}^{\pi_\theta} - \nu]_+\right), \tag{22}$$

where $\nu^* = \nu_\alpha(F_{1,f_w}^{\pi_\theta})$. The set of minimizers Λ of the RHS of (22) is the singleton $\{\nu_w^*\}$. The interval $[\nu^* - \epsilon, \nu^* + \epsilon]$ is nonempty and compact. Using 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 and 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 $\nabla_w \rho_\alpha[-F_{2,f_w}^{\pi_E}] = -\frac{1}{\alpha}\mathbb{E}\left[\mathbf{1}_{\{-F_{2,f_w}^{\pi_E}(\tau) \geq \nu_\alpha(-F_{2,f_w}^{\pi_E})\}} \nabla_w F_{2,f_w}^{\pi_E}(\tau)\right]$. □

Lemma 6. *For any $\theta \in \Theta$, the causal entropy gradient is given by*

$$\nabla_\theta H(\pi_\theta) = \mathbb{E}_{d_{\pi_\theta}}\big[\nabla_\theta \log \pi_\theta(a \,|\, s) Q_{log}(s,a)\big], \tag{23}$$

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

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

Lemma 7. *For any $\theta \in \Theta$ and $w \in \mathcal{W}$, we have*

$$\nabla_\theta \rho_\alpha[F_{1,f_w}^{\pi_\theta}] = \frac{1}{\alpha}\mathbb{E}\left[\nabla_\theta \log \pi_\theta(\tau) \left(F_{1,f_w}^{\pi_\theta}(\tau) - \nu_\alpha(F_{1,f_w}^{\pi_\theta})\right)_+\right], \tag{24}$$

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

Proof. We refer the reader to the proof in Tamar et al. [2015a, 2017]. □

D.1.2 Estimation of the VaR and Gradients

Estimation of the VaR

Corollary 3. *Let $\{\tau_j\}_{j=1}^N$ be trajectories sampled independently from π_θ. Given a cost function parameter $w \in \mathcal{W}$, let $\left(F_{1,f_w}^{\pi_\theta}(\tau_{(1)}), \ldots, F_{1,f_w}^{\pi_\theta}(\tau_{(N)})\right)$ be the order statistic of the sampled trajectories, i.e.,*

$$F_{1,f_w}^{\pi_\theta}(\tau_{(1)}) \leq \ldots \leq F_{1,f_w}^{\pi_\theta}(\tau_{(N)}).$$

Then, a consistent estimator $\widehat{\nu}_\alpha(F_{1,f_w}^{\pi_\theta})$ of the Value-at-Risk $\nu_\alpha(F_{1,f_w}^{\pi_\theta})$ is given by the $(1-\alpha)$-quantile of the order statistic $\left(F_{1,f_w}^{\pi_\theta}(\tau_{(1)}), \ldots, F_{1,f_w}^{\pi_\theta}(\tau_{(N)})\right)$.

Similarly, let $\{\tau_j^E\}_{j=1}^N$ be trajectories sampled independently from π_E. If $\left(-F_{2,f_w}^{\pi_E}(\tau_{(1)}^E), \ldots, -F_{2,f_w}^{\pi_E}(\tau_{(N)}^E)\right)$ is the order statistic of the sampled expert trajectories, then a consistent estimator $\widehat{\nu}_\alpha(-F_{2,f_w}^{\pi_E})$ of the Value-at-Risk $\nu_\alpha(-F_{2,f_w}^{\pi_E})$ is given by the $(1-\alpha)$-quantile of the order statistic $\left(-F_{2,f_w}^{\pi_E}(\tau_{(1)}^E), \ldots, -F_{2,f_w}^{\pi_E}(\tau_{(N)}^E)\right)$.

Estimation of the Gradients

Corollary 4. *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 $w \in \mathcal{W}$, a consistent estimator of the gradient of $(1+\lambda)\left(\rho_\alpha^\lambda[F_{1,f_w}^{\pi_\theta}] - \rho_\alpha^\lambda[-F_{2,f_w}^{\pi_E}]\right)$ w.r.t. w is given by*

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

The two other gradients are estimated using standard Monte-Carlo techniques from the reinforcement learning literature [Sutton et al., 2000, Ziebart et al., 2008].

D.2 W-RS-GAIL

D.2.1 Gradient Formulas

Using similar assumptions and technical arguments as for JS-RS-GAIL, we obtain the following expressions for the gradients of W-RS-GAIL.

Theorem 4 (W-RS-GAIL, gradient w.r.t. the cost function parameter)**.**

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

Theorem 5 (W-RS-GAIL, gradient w.r.t. the policy parameter)**.**

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

D.2.2 Estimation of the VaR and Gradients

Estimation of the VaR

Corollary 5. *Let $\{\tau_j\}_{j=1}^N$ be trajectories sampled independently from π_θ. Given a cost function parameter $w \in \mathcal{W}$, let $\left(C_{f_w}^{\pi_\theta}(\tau_{(1)}), \ldots, C_{f_w}^{\pi_\theta}(\tau_{(N)})\right)$ be the order statistic of the sampled trajectories, i.e.,*

$$C_{f_w}^{\pi_\theta}(\tau_{(1)}) \leq \ldots \leq C_{f_w}^{\pi_\theta}(\tau_{(N)}).$$

Then, a consistent estimator $\widehat{\nu}_\alpha(C_{f_w}^{\pi_\theta})$ of the Value-at-Risk $\nu_\alpha(C_{f_w}^{\pi_\theta})$ is given by the $(1-\alpha)$-quantile of the order statistic $\left(C_{f_w}^{\pi_\theta}(\tau_{(1)}), \ldots, C_{f_w}^{\pi_\theta}(\tau_{(N)})\right)$.

Similarly, let $\{\tau_j^E\}_{j=1}^N$ *be trajectories sampled independently from* π_E*. If* $\left(C_{f_w}^{\pi_E}(\tau_{(1)}^E), \ldots, C_{f_w}^{\pi_E}(\tau_{(N)}^E)\right)$ *is the order statistic of the sampled expert trajectories, then a consistent estimator* $\widehat{\nu}_\alpha(C_{f_w}^{\pi_E})$ *of the Value-at-Risk* $\nu_\alpha(C_{f_w}^{\pi_E})$ *is given by the* $(1-\alpha)$*-quantile of the order statistic* $\left(C_{f_w}^{\pi_E}(\tau_{(1)}^E), \ldots, C_{f_w}^{\pi_E}(\tau_{(N)}^E)\right)$.

Estimation of the Gradients

Corollary 6. *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* $w \in \mathcal{W}$*, a consistent estimator of* $\nabla_w (1+\lambda)\left(\rho_\alpha^\lambda[C_{f_w}^{\pi_\theta}] - \rho_\alpha^\lambda[C_{f_w}^{\pi_E}]\right)$ *is given by*

$$\frac{1}{\alpha N}\sum_{j=1}^{N}\left(\alpha + \lambda \mathbf{1}_{\left\{C_{f_w}^{\pi_\theta}(\tau_j) \geq \widehat{\nu}_\alpha(C_{f_w}^{\pi_\theta})\right\}}\right)\nabla_w C_{f_w}^{\pi_\theta}(\tau_j) - \frac{1}{\alpha N_E}\sum_{j=1}^{N_E}\left(\alpha + \lambda \mathbf{1}_{\left\{C_{f_w}^{\pi_E}(\tau_j^E) \geq \widehat{\nu}_\alpha(C_{f_w}^{\pi_E})\right\}}\right)\nabla_w C_{f_w}^{\pi_E}(\tau_j^E).$$

The two other gradients are estimated using standard Monte-Carlo techniques from the reinforcement learning literature [Sutton et al., 2000, Ziebart et al., 2008].

E Adding Noise to the Cost Function of Hopper and Walker

For each (deterministic) environment, we pre-train an expert's policy π_E using TRPO. We introduce stochasticity in the cost function in a way that (i) increases the risk-sensitivity of the expert policy π_E w.r.t. the modified cost function and (ii) makes the environment stochastic enough to have a meaningful assessment of risk in terms of tail performance.

Hopper: Given the deterministic cost function $c(s,a)$ of the original implementation, we introduce randomness into $c(s,a)$ as follows: We generate 500 trajectories from the expert's policy π_E. Then, we run a K-Means clustering algorithm with $K = 15$ over the set of collected state-action pairs $\mathcal{D}$. We set $\{w_i\}_{i=1}^{K=15}$ to be the relative proportion of the expert's state-action pairs in the i-th cluster. These weights give us a rough estimate of the occupancy measure of the expert's policy. For any other (unobserved in the expert's trajectories) state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we compute the closest observed state-action pair w.r.t. the Euclidean distance, i.e., $(\hat{s},\hat{a}) \in \mathcal{D}$. Let j be the index of the cluster that $(\hat{s},\hat{a})$ belongs to. We define the noisy cost function $c_{\text{random}}(s,a)$ to be

$$c_{\text{random}}(s,a) := \frac{1}{0.2 + \sqrt{w_j}}\,|Z|\,c(s,a),$$

where $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable truncated between -10 and 10. The lower the value of w_j (i.e., the less 'time' the expert spends in the region of the state-action space around (s,a)), the higher the random gain $\frac{1}{0.2+\sqrt{w_j}}|Z|$. Therefore, a low value of w_j, combined with the random cost $c_{\text{random}}(s,a)$, corresponds, a posteriori, to a region the expert considers as risky.

Walker: We use the exact same procedure as in Hopper for Walker, with the cost function c_{random} defined as

$$c_{\text{random}}(s,a) := \frac{0.4}{\sqrt{\max(0.01, w_j - 0.02)}}\,|Z|\,c(s,a),$$

where $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable truncated between -10 and 10.

The numerical values defined the modified cost functions $c_N(s,a)$ were chosen before running any imitation learning algorithm.