8 APPENDIX

8.1 PROOFS

Proof for Proposition 1:

Proof. We show that $\max_s D_{JS}(\pi^B(s)\|\pi^A(s))$ is well-defined for an MDP $\mathcal{M}$ with two representations $\mathcal{M}^A$ and $\mathcal{M}^B$. From Theorem 1, we know the distribution $\pi(s)$ can be written with respect to its occupancy measure ρ_π. It is sufficient to show that we can map occupancy measures of π^A and π^B to a common MDP. By the definition of an occupancy measure,

$$\begin{aligned}\rho_\pi(s,a) &= \mathbb{P}(\pi(s)=a)\sum_{i=0}^{\infty}\gamma^i\mathbb{P}(s_i=s|\pi)\\ &= \mathbb{E}_{\tau=(s_0,a_0,\cdots,s_n)\sim\pi}[\sum_{i=0}^{n}\gamma^i\mathbf{1}((s_i,a_i)=(s,a))]\end{aligned}$$

that is to say, the occupancy measure is the expected discounted count of a state-action pair to appear in all possible trajectories. Since we have trajectory mappings between $\mathcal{M}^A$ and $\mathcal{M}^B$, we can convert an occupancy measure in $\mathcal{M}^A$ to one in $\mathcal{M}^B$ by mapping each trajectory and perform the count in the new MDP representation. Formally, the occupancy measure $\rho^B_{\pi^B}$ of π^B in $\mathcal{M}^B$ can be mapped to an occupancy measure in $\mathcal{M}^A$ by

$$\begin{aligned}&\rho^A_{\pi^B}(s,a)\\ &= \mathbb{E}_{\substack{\tau^B\sim\pi^B,\\ f_{B\to A}(\tau^B)=(s_0,a_0,\cdots,s_n)}}[\sum_{i=0}^{n}\gamma^i\mathbf{1}((s_i,a_i)=(s,a))]\end{aligned}$$

Following from this, we can compute $D_{JS}(\pi^B(s)\|\pi^A(s))$ using any s in $\mathcal{M}^A$. And the maximum is defined. In the definition, there is a choice whether to map π^A's occupancy measure to $\mathcal{M}^B$ or π^B's to $\mathcal{M}^A$. Though both approaches lead to a valid definition, we use the definition that for $D_{JS}(\cdot\|\cdot)$, we always map the representation in the first argument to that of the second argument. It is preferable to the other one because in Theorem 2, we want to optimize

$$J(\pi'^A) \geq J_{\pi^A}(\pi'^A) - \frac{2\gamma^A(4\beta^B_{\mathcal{D}_2}\epsilon^B_{\mathcal{D}_2}+\alpha^A_{\mathcal{D}}\epsilon^A_{\mathcal{D}})}{(1-\gamma^A)^2}+\delta_2$$

by optimizing

$$\beta^B_{\mathcal{D}_2} = \mathbb{E}_{\mathcal{M}\sim\mathcal{D}_2}[\max_s D_{JS}(\pi^B(s)\|\pi^A(s))]$$

usually via computing the gradient of $\beta^B_{\mathcal{D}_2}$ w.r.t. π^A. If we use $f_{A\to B}$ to map from $\mathcal{M}^A$ to $\mathcal{M}^B$, the gradient will involve a complex composition of $f_{A\to B}$ and π^A, which is undesirable. □

To prove Theorem 2, we need to use a policy improvement result for a single MDP (a modified version of Theorem 1 in [29]).

Theorem 4. *Assume for an MDP $\mathcal{M}$, an expert policy π_E have a higer advantage of over a policy π with a margin, i.e., $\eta(\pi_E,\mathcal{M})-\eta(\pi,\mathcal{M})\geq\delta$ Define*

$$\begin{aligned}\alpha &= \max_s D_{KL}(\pi'(s)\|\pi(s))\\ \beta &= \max_s D_{JS}(\pi'(s)\|\pi_E(s))\\ \epsilon_{\pi_E} &= \max_{s,a}|A_{\pi_E}(s,a)|\\ \epsilon_\pi &= \max_{s,a}|A_\pi(s,a)|\end{aligned}$$

then $\eta(\pi',\mathcal{M}) \geq \eta_\pi(\pi',\mathcal{M}) - \frac{2\gamma(4\beta\epsilon_{\pi_E}+\alpha\epsilon_\pi)}{(1-\gamma)^2}+\delta$

Proof. The only difference from the original theorem is that the original assumes $\mathbb{E}_{a_E\sim\pi_E(s),a\sim\pi(s)}[A_\pi(s,a_E)-A_\pi(s,a)]\geq\delta'>0$ for every state s. It is a stronger assumption which is not needed in their analysis. Notice that the advantage of a policy over itself is zero, i.e., $\mathbb{E}_{a\sim\pi(s)}[A_\pi(s,a)]=0$ for every s, so the margin assumption simplifies to $\mathbb{E}_{a_E\sim\pi_E(s)}[A_\pi(s,a_E)]\geq\delta'$.

By the policy advantage formula,

$$\begin{aligned}\eta(\pi_E,\mathcal{M})-\eta(\pi,\mathcal{M}) &= \mathbb{E}_{\tau\sim\pi_E}[\sum_{i=0}^{\infty}\gamma^iA_\pi(s_i,a_i)]\\ &= \mathbb{E}_{s_i\sim\rho_{\pi_E}}\mathbb{E}_{a_i\sim\pi_E(s_i)}[\sum_{i=0}^{\infty}\gamma^iA_\pi(s_i,a_i)]\\ &\geq \mathbb{E}_{s_i\sim\rho_{\pi_E}}[\delta'\sum_{i=0}^{\infty}\gamma^i]\\ &= \frac{\delta'}{1-\gamma}\end{aligned}$$

So an assumption on per-state advantage translates to a overall advantage. Thus we can make this weaker assumption which is also more intuitive and the original statement still holds with a different δ term. □

Proof of Theorem 2:

Proof. Theorem 2 is a distributional extension to the theorem above. For $\mathcal{M}\sim\mathcal{D}_2$, let $\delta_{\mathcal{M}}=\eta(\pi^B,\mathcal{M}^B)-$

$\eta(\pi^A, \mathcal{M}^A)$.

$$
\begin{aligned}
&J(\pi'^A) \\
&= \mathbb{E}_{\mathcal{M}\sim\mathcal{D}}[\eta(\pi'^A, \mathcal{M}^A)] \\
&= \mathbb{E}_{\mathcal{M}\sim\mathcal{D}_1}[\eta(\pi'^A, \mathcal{M}^A)] + \mathbb{E}_{\mathcal{M}\sim\mathcal{D}_2}[\eta(\pi'^A, \mathcal{M}^A)] \\
&\geq \mathbb{E}_{\mathcal{M}\sim\mathcal{D}_1}[\eta(\pi'^A, \mathcal{M}^A)] + \\
&\mathbb{E}_{\mathcal{M}\sim\mathcal{D}_2}[\eta_{\pi^A}(\pi'^A, \mathcal{M}^A) - \frac{2\gamma^A(4\beta\epsilon_{\pi^B} + \alpha\epsilon_{\pi^A})}{(1-\gamma^A)^2} + \delta_{\mathcal{M}}] \\
&\geq \mathbb{E}_{\mathcal{M}\sim\mathcal{D}_1}[\eta_{\pi^A(}(\pi'^A, \mathcal{M}^A) - \frac{2\gamma^A\alpha\epsilon_{\pi^A}}{(1-\gamma^A)^2}] + \\
&\mathbb{E}_{\mathcal{M}\sim\mathcal{D}_2}[\eta_{\pi^A}(\pi'^A, \mathcal{M}^A) - \frac{2\gamma^A(4\beta\epsilon_{\pi^B} + \alpha\epsilon_{\pi^A})}{(1-\gamma^A)^2} + \delta_{\mathcal{M}}] \\
&= \mathbb{E}_{\mathcal{M}\sim\mathcal{D}}[\eta_{\pi^A}(\pi'^A, \mathcal{M}^A)] - \mathbb{E}_{\mathcal{M}\sim\mathcal{D}}[\frac{2\gamma^A\alpha\epsilon_{\pi^A}}{(1-\gamma^A)^2}] - \\
&\mathbb{E}_{\mathcal{M}\sim\mathcal{D}_2}[\frac{2\gamma^A\cdot 4\beta\epsilon_{\pi^B}}{(1-\gamma^A)^2}] + \mathbb{E}_{\mathcal{M}\sim\mathcal{D}_2}[\delta_{\mathcal{M}}] \\
&\geq J_{\pi^A}(\pi'^A) - \frac{2\gamma^A(4\beta^B_{\mathcal{D}_2}\epsilon^B_{\mathcal{D}_2} + \alpha^A_{\mathcal{D}}\epsilon^A_{\mathcal{D}})}{(1-\gamma^A)^2} + \delta_2
\end{aligned}
$$

The derivation for $J(\pi'^B)$ is the same. □

Finally, we provide the proof for Theorem 3. We first quantify the performance gap between a policy π and an optimal policy π^*. For a policy that is able to achieve ϵ $0-1$ loss, $\ell(s,\pi) = \mathbf{1}(\pi(s) \neq \pi^*(s))$, measured against π^*'s action choices under its own state distributions, then we can bound the performance gap. Let $Q_t^{\pi'}(s,\pi)$ denote the t-step cost of executing π in initial state s and then following policy π'

Theorem 5. *(Theorem 2.2 from [51], adpated to our notations) Let π be such taht $\mathbb{E}_{s\sim\rho_\pi}[\ell(s,\pi)] = \epsilon$, and $Q^{\pi^*}_{T-t+1}(s,\pi^*) - Q^{\pi^*}_{T-t+1}(s,a) \leq u$ for all action $a, t \in \{1,2,\cdots,T\}$, then $\eta(\pi,\mathcal{M}) \geq \eta(\pi^*,\mathcal{M}) - uT\epsilon$.*

Thus the important quantity to measure is ϵ, and by measuring the disagreements between two policies in two views, we can upper bound ϵ. The result is originally stated in the context of classification, and the above theorem justifies the learning reduction approach of reducing policy learning to classification.

Theorem 6. *(Corollary 5 in [16] applied to full classifiers) Using the definitions in Theorem 3, with probability $1-\sigma$ over the choice of a sample set N, for all pairs of classifiers h_1, h_2 such that for all i we have $\zeta_i(h_1,h_2,\sigma) > 0$ and $b_i(h_1,h_2,\sigma) \leq 1$.*

$$\epsilon \leq \max_{j\in\{1,\cdots,k\}} b_j(h_1,h_2,\sigma)$$

Proof. The only change from the original proof is that instead of a partial classifier which can output $\perp$, we consider a full classifier. Then we could eliminate the estimates for $\mathbb{P}(h_1 \neq \perp)$ and the error introduced by converting a partial classifier to a full classifier via random labelling when the output is $\perp$. □

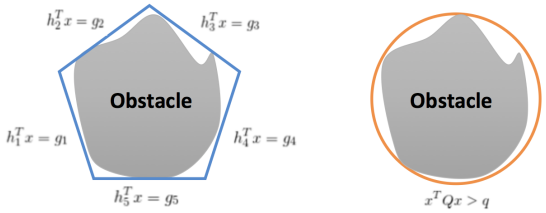

Figure 7: Two views for Risk-Aware Path Planning. On the left, the obstacle is enclosed by a polytope (MILP view) and on the right the obstacle is enclosed by an ellipse (QCQP view).

Proof of Theorem 3:

Proof. For the bound for π^A, we are measuring ϵ_A on its sampled paths. Then directly apply Theorem 6 gives an upper bound on ϵ_A. Apply Theorem 5 gives the result of Theorem 3. □

8.2 PICTORIAL REPRESENTATION OF THE TWO-VIEWS IN RISK-AWARE PATH PLANNING:

We present a pictorial representation of the two different views used in the experiments in Fig 7. In the MILP view, the constraint space is represented using additional auxiliary binary variables to choose the active side of the polytope, whereas in the QCQP view, the constraint space can be encoded in a quadratic constraint.

8.3 RISK-AWARE PLANNING DATASET GENERATION:

We generate 150 obstacle maps. Each map contains 10 rectangle obstacles, with the center of each obstacle chosen from a uniform random distribution over the space $0 \leq y \leq 1$, $0 \leq x \leq 1$. The side length of each obstacle was chosen from a uniform distribution in range [0.01, 0.02] and the orientation was chosen from a uniform distribution between 0° and 360°. In order to avoid trivial infeasible maps, any obstacles centered close to the destination are removed. For MILP view, we directly use the randomly generated rectangles for defining the constraint space. However, for the QCQP view, we enclose the rectangle obstacles with a circle for defining the quadratic constraint.

8.4 DISCRETE/CONTINUOUS CONTROL RESULTS IN TABULAR FORM

	Acrobot	Swimmer	Hopper
A (CoPiEr)	-86.44 ± 10.80	106.35 ± 23.11	217.83 ± 30.03
A (PG)	-169.57 ± 10.48	109.09 ± 21.58	278.66 ± 32.87
A (All)	-252.42 ± 8.73	100.36 ± 22.37	49.39 ± 10.35
B (CoPiEr)	-88.48 ± 15.13	104.16 ± 19.32	168.88 ± 18.21
B (PG)	-257.16 ± 10.93	103.48 ± 21.89	89.34 ± 4.89
B (All)	-251.74 ± 9.65	96.74 ± 19.57	22.59 ± 5.55
A + B	-86.42 ± 3.48	108.71 ± 5.03	346.53 ± 5.91