A Gradient Derivation

This section derives the gradient $\frac{\partial \mathcal{L}}{\partial \theta}$ in equation (12) in full details. Recall $\mathcal{L}$ is defined as:

$$\mathcal{L} = \int p_E(V) \log \frac{p_E(V)}{p_Q(V|U_B, \Sigma; \theta)} dV$$

Firstly, we substitute $\mathcal{L}$ in $\frac{\partial \mathcal{L}}{\partial \theta}$ and rewrite $\log \frac{p_E(V)}{p_Q(V|U_B,\Sigma;\theta)}$ as the difference between $\log\ p_E(V)$ and $p_Q(V|U_B, \Sigma; \theta)$.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} =& \frac{\partial}{\partial \theta} \int p_E(V) \log \frac{p_E(V)}{p_Q(V|U_B, \Sigma; \theta)} dV \\
=& \int p_E(V) \frac{\partial}{\partial \theta} \log\ p_E(V) dV - \int p_E(V) \frac{\partial}{\partial \theta} \log\ p_Q(V|U_B, \Sigma; \theta) dV \\
=& - \int p_E(V) \frac{\partial}{\partial \theta} \log\ p_Q(V|U_B, \Sigma; \theta) dV
\end{aligned} \tag{14}$$

Since $p_E(V)$ is independent of θ, the first derivative in the second line of equation (14) evaluates to 0. Next, we substitute $p_Q(V|U_B, \Sigma; \theta)$ using the optimal control sequence distribution expression in equation (4):

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} =& - \int p_E(V) \frac{\partial}{\partial \theta} \log\ \frac{1}{Z} p_B(V|U_B, \Sigma) \exp(-\frac{1}{\lambda} S(V;\theta)) dV \\
=& - \int p_E(V) \frac{\partial}{\partial \theta} \log p_B(V|U_B, \Sigma)\ dV - \int p_E(V) \frac{\partial}{\partial \theta} \log \exp(-\frac{1}{\lambda} S(V;\theta)) dV \\
&+ \int p_E(V) \frac{\partial}{\partial \theta} \log Z dV \\
=& - \int p_E(V) \frac{\partial}{\partial \theta} (-\frac{1}{\lambda} S(V;\theta)) dV + \int p_E(V) \frac{\partial}{\partial \theta} \log Z dV
\end{aligned} \tag{15}$$

Since $p_B(V|U_B, \Sigma)$ is independent of θ, the first derivative in the second line of equation (14) evaluates to 0. We are left with only two integrals in equation (15).

Next, we factorize out $\frac{\partial}{\partial \theta} \log Z$ from the integral since the partition function Z is constant to all V:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} =& - \int p_E(V) (-\frac{1}{\lambda} \frac{\partial}{\partial \theta} S(V;\theta)) dV + (\frac{\partial}{\partial \theta} \log Z) \int p_E(V) dV \\
=& - \int p_E(V) (-\frac{1}{\lambda} \frac{\partial}{\partial \theta} S(V;\theta)) dV + \frac{1}{Z} \frac{\partial Z}{\partial \theta}
\end{aligned} \tag{16}$$

The second line in equation (16) follows as $\int p_E(V) dV = 1$.

Next, we substitute $Z = \int p_B(V|U_B, \Sigma) \exp(-\frac{1}{\lambda} S(V;\theta)) dV$ in $\frac{\partial Z}{\partial \theta}$ and simplify it:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} =& - \int p_E(V) (-\frac{1}{\lambda} \frac{\partial}{\partial \theta} S(V;\theta)) dV + \frac{1}{Z} \frac{\partial}{\partial \theta} \int p_B(V|U_B, \Sigma) \exp(-\frac{1}{\lambda} S(V;\theta)) dV \\
=& - \int p_E(V) (-\frac{1}{\lambda} \frac{\partial}{\partial \theta} S(V;\theta)) dV + \frac{1}{Z} \int p_B(V|U_B, \Sigma) \frac{\partial}{\partial \theta} \exp(-\frac{1}{\lambda} S(V;\theta)) dV \\
=& - \int p_E(V) (-\frac{1}{\lambda} \frac{\partial}{\partial \theta} S(V;\theta)) dV \\
&+ \frac{1}{Z} \int p_B(V|U_B, \Sigma) \exp(-\frac{1}{\lambda} S(V;\theta)) \frac{\partial}{\partial \theta} (-\frac{1}{\lambda} S(V;\theta)) dV \\
=& - \int p_E(V) (-\frac{1}{\lambda} \frac{\partial}{\partial \theta} S(V;\theta)) dV \\
&+ \int \frac{p_B(V|U_B, \Sigma) \exp(-\frac{1}{\lambda} S(V;\theta))}{Z} \frac{\partial}{\partial \theta} (-\frac{1}{\lambda} S(V;\theta)) dV
\end{aligned} \tag{17}$$

We rewrite $\frac{1}{Z}p_B(V|U_B,\Sigma)\exp(-\frac{1}{\lambda}S(V;\theta))$ in the last line of equation (17) as $p_Q(V|U_B,\Sigma;\theta)$ using equation (4) and finally we have:

$$\frac{\partial\mathcal{L}}{\partial\theta}=\int p_E(V)(\frac{1}{\lambda}\frac{\partial}{\partial\theta}S(V;\theta))dV-\int p_Q(V|U_B,\Sigma;\theta)(\frac{1}{\lambda}\frac{\partial}{\partial\theta}S(V;\theta))dV \tag{18}$$

B Theorem 3.1 and Proof

We use T to denote the task horizon, and K to denote the receding time horizon for local optimization. Let us simplify notations in the optimal control sequence distribution $p_Q(V|U_B,\Sigma,x_t;\theta)$ and remove the explicit dependency on U_B and Σ. We assume that all control sequences are applied with U_B as the base distribution and Σ as the covariate matrix. We assume the expert's underlying cost function is parameterized by θ_E, so we have $p_E(.)=p(.;\theta_E)$.

B.1 Sketch

We present a theoretical analysis on the convergence of RHIRL. Our main theorem 3.1 states that given that the Kullback–Leibler (KL) divergence over local control sequence distribution for each time step $t=0,1,...T-1$ is bounded by ϵ, though we do not query expert during the learning, using the cached expert demonstrations alone allows us to bound the error over global state marginal distribution *linear* in the task horizon T under total variance measure.

First, we show that if the KL-divergence over the local control sequence distribution is bounded by ϵ, so is the KL-divergence over the resulting state distribution. At each time step t, the optimal control sequences distribution $p_Q(V_t|x_t;\theta_t)$ at the initial state x_t contains the full information to generate the corresponding state trajectories $p(\tau_t|x_t;\theta)$, and consequently the state distributions $p_t(x|x_t;\theta)$ (by neglecting the temporal information). Upon applying the information loss (lemma B.1), we prove in lemma B.2 that, given initial state x_t, if the KL-divergence over V_t is bound by ϵ, so is the corresponding state distribution $p_t(x)$, i.e. $D_{\text{KL}}(p_t(x|x_t;\theta_E)\parallel p_t(x|x_t;\theta))\leq D_{\text{KL}}(p(V_t|x_t;\theta_E)\parallel p_Q(V_t|x_t;\theta))\leq\epsilon$.

Next, we show the KL-divergence over global state marginal distribution between two consecutive time steps is also bounded by ϵ. For each optimal control sequence we compute, we only execute the first control and use the rest to warm start the re-planning for the next time step. Therefore, for each time step, we only change a small region of the global state distribution, i.e. reachable space of the current time step. We use $p^t_{\text{RHC}}(x;\theta)$ to denote the global state marginal distribution by recursively applying RHC from time step $i=0,...,t$ under θ and switching to θ_E thereafter until $T-1$. Using generalized log sum inequality, we prove in lemma B.3 that if the KL-divergence over V_t is bounded by ϵ for all $t=0,...,T-1$, the KL-divergence over the global state marginal distribution between each of the two consecutive time steps is bounded by ϵ, i.e. $D_{\text{KL}}(p^t_{\text{RHC}}(x;\theta)\parallel p^{t+1}_{\text{RHC}}(x;\theta))\leq\frac{K+1}{T}\epsilon$.

Finally, we use Pinsker's inequality to upper bound the total variation (TV) distance by KL-divergence over state marginal distribution. Then we use the triangle inequality to show that the TV distance between expert and the actual visited state distribution over the task horizon T using RHIRL is bounded by an error linearly in T, i.e. $D_{\text{TV}}(p_E(x)\parallel p_{\text{RHC}}(x;\theta))<T\sqrt{\epsilon/2}$.

B.2 Proofs

First, we use **Lemma B.1, B.2** to prove that the KL-divergence over control sequence space upper bounds the KL-divergence over the resulting state distribution. We define the control sequence starts at task time step t as $V_t=\{v_t,v_{t+1},...,v_{t+K-1}\}$. Moreover, its corresponding trajectory segment $\tau_t=\{x_t,x_{t+1},...,x_{t+K}\}$ is computed uniquely from V_t and initial state x_t by iteratively applying the dynamic model $x_{t+1}=f(x_t,v_t)$. We use $p(x_t)$ to denote the state density at a single time step t. Assume V_t is optimized based on the cost parameterization θ, then the corresponding state distribution is defined as the summation of all state density over horizon K, i.e. $p_t(x;\theta)=\frac{1}{K+1}\sum_{i=t}^{t+K}p(x_i;\theta)$.

Lemma B.1 (Information loss[16]). *Let a and b be two random variables and $f(.)$ be a convex function. Let $P(a,b)$ be a joint probability distribution. The marginal distributions are $P(a)=\sum_b P(a,b)$ and $P(b)=\sum_a P(a,b)$. Assume that a can explain away b. This is expressed as follows – given any two probability distribution $P(.)$, $Q(.)$, assume the following equality holds for all a, b:*

$$P(b|a)=Q(b|a) \tag{19}$$

Under these conditions, the following inequality holds:

$$\sum_a Q(a) f(\frac{P(a)}{Q(a)}) > \sum_b Q(b) f(\frac{P(b)}{Q(b)}) \tag{20}$$

Lemma B.2. *Given the initial state x_t and the two control sequence distributions $p(V_t|x_t;\theta_E)$ and $p_Q(V_t|x_t;\theta)$, the KL-divergence between the resulting state distribution is upper bounded by KL-divergence between the control sequence distribution.*

$$D_{KL}(p_t(x|x_t;\theta_E) \parallel p_t(x|x_t;\theta)) \leq D_{KL}(p(V_t|x_t;\theta_E) \parallel p_Q(V_t|x_t;\theta)) \tag{21}$$

Proof. Firstly, we prove that the KL-divergence over state trajectory distribution $p(\tau_t)$ is upper bounded by the KL-divergence between $p(V_t)$. Given dynamical model f, the control sequence V_t and the initial state x_t contains all information to generate the corresponding τ_t. Therefore, for any joint distribution $P(\tau_t, V_t|x_t)$ and $Q(\tau_t, V_t|x_t)$, the following is true

$$P(\tau_t|V_t, x_t) = Q(\tau_t|V_t, x_t)$$

Upon applying the information loss Lemma B.1, we have the inequality:

$$D_{\text{KL}}(p(\tau_t|x_t;\theta_E) \parallel p(\tau_t|x_t;\theta)) \leq D_{\text{KL}}(p(V_t|x_t;\theta_E) \parallel p_Q(V_t|x_t;\theta)) \tag{22}$$

Next we prove that the KL-divergence between state distribution is upper bounded by the trajectory distribution. Since a trajectory $\tau_t = \{x_t, x_{t+1}, ..., x_{t+K}\}$ contains full information of the resulting states (by neglecting the temporal information), for any joint distribution $P(x|\tau_t)$ and $Q(x|\tau_t)$, the following is true

$$P(x|\tau_t) = Q(x|\tau_t) \tag{23}$$

Upon applying Lemma B.1 we have the inequality:

$$D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel p_t(x|x_t;\theta)) \leq D_{\text{KL}}(p(\tau_t|x_t;\theta_E) \parallel p(\tau_t|x_t;\theta)) \tag{24}$$

Therefore, given the KL-divergence between the control sequence distribution is upper bounded by ϵ, we use the equality in equation (22) and (24) to show that the KL-divergence between the resulting state distribution is also upper bounded by ϵ.

$$D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel p_t(x|x_t;\theta)) \leq D_{\text{KL}}(p(V_t|x_t;\theta_E) \parallel p_Q(V_t|x_t;\theta)) \leq \epsilon \tag{25}$$

□

Definition 1 (One-step Recoverability [32])**.** Assume that the state distribution of the learner and expert are different at time t, that is $D_{\text{KL}}(p(x_t;\theta_E) \parallel p(x_t;\theta)) \neq 0$, there exists a policy π_{re} that when used for the learner, can bound:

$$D_{\text{KL}}(p(x_{t+1};\theta_E) \parallel p(x_{t+1};\pi_{re})) \leq \epsilon_1 \tag{26}$$

where the current initial state distribution of the student follows $p(x_t;\theta)$.

Intuitively, this condition requires that, no matter what is the current state distribution, the learner can recover to the expert demonstrated distribution in a single time-step. In our case, this is a natural condition since the difference in the initial state distribution $p(x_t;\theta)$ and $p(x_t;\theta_E)$ is not arbitrarily large: we use re-planning to ensure the receding state sequences is always bounded below ϵ, hence this recoverability condition can be easily satisfied. We emphasize that this recoverable policy is **never** executed in our algorithm, it is only used for the theoretical analysis.

Next, we derive a bound over the global state marginal distribution between two consecutive time steps. At each time step $t = 0, ..., T-1$, we re-optimize the local control sequence distribution and only execute the first control, hence we only change state density over a small reachable space. We define $p^t_{\text{RHC}}(x,\theta)$ as the global marginal state distribution by applying RHC from $i = 0, 1, ..., t$ under θ and then using the recoverable policy π_{re} to switch to the expert θ_E thereafter until $T-1$, i.e., $p^t_{\text{RHC}}(x;\theta) = \frac{1}{T}(\sum_{i=0}^{t+1} p(x_i;\theta) + p(x_{t+2};\pi_{re}) + \sum_{i=t+3}^{T} p(x_i;\theta_E))$. According to the definition of the recoverable policy, we have $D_{\text{KL}}(p(x_{t+2};\theta_E) \parallel p(x_{t+2};\pi_{re})) \leq \epsilon_1$, therefore, $p^t_{\text{RHC}}(x;\theta) \approx \frac{1}{T}(\sum_{i=0}^{t+1} p(x_i;\theta) + \sum_{i=t+2}^{T} p(x_i;\theta_E))$. To quantify the change in global state marginal distribution, we derive a bound for the KL-divergence between two consecutive time steps, i.e. $D_{\text{KL}}(p^{t-1}_{\text{RHC}}(x;\theta) \parallel p^t_{\text{RHC}}(x;\theta))$.

Lemma B.3. *If the KL-divergence over resulting state density from the control sequence distribution of length K are bounded by ϵ, i.e. $D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel p_t(x|x_t;\theta)) < \epsilon$, where x_t is the state encountered by our policy at $t = 0, 1, ..., T-1$ and is one-step recoverable, then KL-divergence over the global state marginal distribution between two consecutive control executions are bounded by ϵ,*

$$D_{KL}(p^{t-1}_{\text{RHC}}(x;\theta) \parallel p^{t}_{\text{RHC}}(x;\theta)) \leq \frac{K+1}{T}\epsilon \tag{27}$$

for $t = 1, ..., T-1$.

We state the generalized log sum inequality below in lemma B.4, the proof can be found in the Appendix of [16]. Lemma B.4 and B.5 will be used in the proof for Lemma B.3.

Lemma B.4 (Generalized log sum inequality[16])**.** *Let $p_1, ..., p_n$ and $q_1, ..., q_n$ be non-negative numbers. Let $p = \sum_{i=1}^{n} p_i$ and $q = \sum_{i=1}^{n} q_i$. Let $f(.)$ be a convex function. We have the following:*

$$\sum_{i=1}^{n} q_i f(\frac{p_i}{q_i}) \geq q f(\frac{p}{q}) \tag{28}$$

Lemma B.5. *Let $p(x)$ and $q(x)$ be non-negative functions, and c is a constant factor. We have the following:*

$$\int cp(x) \log \frac{cp(x)}{cq(x)} = c \int p(x) \log \frac{p(x)}{q(x)} \tag{29}$$

Proof.

$$\int cp(x) \log \frac{cp(x)}{cq(x)} dx = \int cp(x) \log \frac{p(x)}{q(x)} dx = c \int p(x) \log \frac{p(x)}{q(x)} dx \tag{30}$$

□

Now, we are ready to prove lemma B.3.

Proof. For each $t = 0, ..., T-1$, we re-plan for the optimal local control sequence start at x_t so that the resulting state distribution over horizon K is bounded, i.e. $D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel p_t(x|x_t;\theta)) < \epsilon$, where $p_t(x|x_t;\theta) = \frac{1}{K+1}\sum_{i=t}^{t+K} p(x_i;\theta)$. However, instead of executing all K controls in the sequence, we only execute the first control at the current time step t and change the state distribution $p(x_{t+1};\theta)$ reachable for that single time step, then we use the remaining control sequence to warm start the local control sequence optimization for the next time step. To account the effect of replanning, for each time step t, since we do not change the state distribution after $p(x_{t+1};\theta)$, we can think of the change in the global state distribution as if we follow the optimal control under θ at time step t and then use the recoverable policy π_{re} to switch to θ_E afterwards over the control sequence horizon K. Hence, the actual state density is $\frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\theta) + p(x_{t+2};\pi_{re}) + \sum_{i=t+3}^{t+K} p(x_i;\theta_E))$. Theoretically, since we do not query the expert online, the initial state x_t distribution in theorem 3.1 and lemma B.2 should follow the expert demonstration at time t, i.e. $p(x_t;\theta_E)$. However, our MPC controller cannot jump to this distribution and we replan from our current state distribution $p(x_t;\theta)$. To resolve this mismatch, we require the recoverability condition in our optimization procedure such that, for each resulting state from the controller, there always exists a one-step recoverable policy π_{re} that can correct the current state distribution $p(x_t;\theta)$ to $p(x_{t+1};\theta_E)$ in one step. Therefore, with this one-step recoverable condition on every state x_t induced by the cost function parameterized by θ, x_t in theorem 3.1 and lemma B.2 now follows the state distribution of our controller, i.e. $p(x_t;\theta)$.

That is,

$$\begin{aligned}
&D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel \frac{1}{K+1}(p(x_t) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E)) \\
=&D_{\text{KL}}(\frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\pi_{re}) + \sum_{i=t+2}^{t+K} p(x;\theta_E)) \parallel \frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E)) \\
\leq&D_{\text{KL}}(\frac{1}{K+1}(p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x;\theta_E)) \parallel \frac{1}{K+1}(p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta)) \\
=&D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel p_t(x|x_t;\theta)) \leq \epsilon
\end{aligned} \tag{31}$$

Recall that we use $p^t_{\text{RHC}}(x;\theta)$ to account for the global state marginal distribution resulted from executing a single optimal control at time step t. More specifically, $p^t_{\text{RHC}}(x;\theta)$ is defined as the state marginal distribution by executing only the first optimal control from the replanned optimal control sequence at each time step from $i = 0, ..., t$ under θ and switching to θ_E thereafter by using the recoverable policy π_{re} until $T-1$, i.e. $p^t_{\text{RHC}}(x;\theta) \approx \frac{1}{T}(\sum_{i=0}^{t+1} p(x_i;\theta) + \sum_{i=t+2}^{T} p(x_i;\theta_E))$. We bound the KL-divergence over global state distribution between two consecutive time step as follow:

$$
\begin{aligned}
&D_{\text{KL}}(p^{t-1}_{\text{RHC}}(x;\theta) \parallel p^t_{\text{RHC}}(x;\theta)) \\
=&D_{\text{KL}}(\frac{1}{T}(\sum_{i=0}^{t} p(x_i;\theta) + p(x_{t+1};\pi_{re}) + \sum_{i=t+2}^{T} p(x_i;\theta_E)) \parallel \frac{1}{T}(\sum_{i=0}^{t+1} p(x_i;\theta) + p(x_{t+2};\pi_{re}) + \sum_{i=t+3}^{T} p(x_i;\theta_E))) \\
\approx&D_{\text{KL}}(\frac{1}{T}(\sum_{i=0}^{t} p(x_i;\theta) + \sum_{i=t+1}^{T} p(x_i;\theta_E)) \parallel \frac{1}{T}(\sum_{i=0}^{t+1} p(x_i;\theta) + \sum_{i=t+2}^{T} p(x_i;\theta_E))) \\
=&\frac{1}{T}\int(\sum_{i=0}^{t} p(x_i;\theta) + \sum_{i=t+1}^{T} p(x_i;\theta_E)) \log \frac{\sum_{i=0}^{t} p(x_i;\theta) + \sum_{i=t+1} p(x_i;\theta_E)}{\sum_{i=0}^{t+1} p(x_i;\theta) + \sum_{i=t+2}^{T} p(x_i;\theta_E)} dx \\
\leq&\frac{1}{T}(\int \sum_{i=0}^{t-1} p(x_i;\theta) \log \frac{\sum_{i=0}^{t-1} p(x_i;\theta)}{\sum_{i=0}^{t-1} p(x_i;\theta)} dx \\
&+ \int (p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E)) \log \frac{p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E)}{p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E)} dx \\
&+ \int \sum_{i=t+K+1}^{T} p(x_i;\theta_E) \log \frac{\sum_{i=t+K+1}^{T} p(x_i;\theta_E)}{\sum_{i=t+K+1}^{T} p(x_i;\theta_E)} dx) \\
=&\frac{1}{T}\int (p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E)) \log \frac{p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E)}{p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E)} dx \\
=&\frac{K+1}{T}\int \frac{1}{K+1}(p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E)) \log \frac{\frac{1}{K+1}(p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E))}{\frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E))} dx \\
=&\frac{K+1}{T} D_{\text{KL}}(\frac{1}{K+1}(p(x_t;\theta) + \sum_{i=t+1}^{t+K} p(x_i;\theta_E)) \parallel \frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E))) \\
=&\frac{K+1}{T} D_{\text{KL}}(p_t(x|x_t;\theta_E) \parallel \frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E))) \\
\leq&\frac{K+1}{T}\epsilon
\end{aligned}
\tag{32}
$$

The first equality in equation (32) follows the definition of $p^t_{\text{RHC}}(x;\theta)$ and the second line follows the definition of the recoverable policy π_{re}. Then, we use lemma B.5 to factor out $\frac{1}{T}$ in the third line. The next inequality follows from the generalized log sum inequality stated in lemma B.4, and we have the first and third terms reduce to 0 and are left with the second term in the next line. We apply lemma B.5 again to the integral using the constant factor $\frac{1}{K+1}$. In addition, to make the equality hold, we multiply the inverse of the constant factor $K+1$ outside the integral. We observe the integral is now the KL divergence between the expert $p_t(x|x_t;\theta_E)$ and one-step-execution of our policy $\frac{1}{K+1}(p(x_t;\theta) + p(x_{t+1};\theta) + \sum_{i=t+2}^{t+K} p(x_i;\theta_E))$. The final inequality follows from the bound derived in equation (31).

For $t = 0$, we have $p^0_{\text{RHC}}(x;\theta) = \frac{1}{T}(p(x_0;\theta) + \sum_{i=1}^{T-1} p(x_i;\theta_E))$. Since the initial state x_0 for expert and our policy are sampled from the same initial state distribution μ, $p(x_0)$ is independent of θ, i.e. $p(x_0;\theta) = p(x_0;\theta_E)$. Therefore, $p_{\text{RHC}}(x;\theta_E) = p(x_0) + \sum_{i=1}^{T-1} p(x_i;\theta_E) = p^0_{\text{RHC}}(x;\theta)$. Moreover, the final global state marginal distribution $p_{\text{RHC}}(x;\theta)$ is the same as the $p^{T-1}_{\text{RHC}}(x;\theta)$,

i.e. $p_{\text{RHC}}(x;\theta) = \sum_{i=0}^{T-1} p(x_i;\theta) = p_{\text{RHC}}^{T-1}(x;\theta)$. For any $t = 1, 2, .., T-1$, we have proved $D_{\text{KL}}(p_{\text{RHC}}^{t-1}(x;\theta) \parallel p_{\text{RHC}}^{t}(x;\theta)) \leq \frac{K+1}{T}\epsilon$.

□

Finally, we are prepared to prove theorem 3.1.

Proof. We evaluate the TV distance over the state marginal distribution between the expert policy and our control law.

$$\begin{aligned} D_{\text{TV}}(p_{\text{RHC}}(x;\theta_E) \parallel p_{\text{RHC}}(x;\theta)) &= D_{\text{TV}}(p_{\text{RHC}}^{0}(x;\theta) \parallel p_{\text{RHC}}^{T-1}(x;\theta)) \\ &\leq \sum_{t=1}^{T-1} D_{\text{TV}}(p_{\text{RHC}}^{t-1}(x;\theta) \parallel p_{\text{RHC}}^{t}(x;\theta)) \end{aligned} \tag{33}$$

The first equality in equation (33) follows from the fact that $p_{\text{RHC}}(x;\theta_E) = p_{\text{RHC}}^{0}(x;\theta)$ and $p_{\text{RHC}}(x;\theta) = p_{\text{RHC}}^{T-1}(x;\theta)$. We use triangle inequality of the TV distance measures to obtain the inequality in the second line.

Recall that by Pinsker's inequality, the total variation (TV) distance is related to Kullback–Leibler (KL) divergence by the following inequality: $D_{\text{TV}}(P \parallel Q) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P \parallel Q)}$. We apply Pinsker's inequality to each of the TV terms in the second line of equation (33) to bound them by a summation of KL-divergence as shown in the first line of equation (34). Next, given the control sequence distribution for every time step is bounded by ϵ, we apply lemma B.2 to show that the resulting state distribution from the optimal control sequences for each time step t is also bounded by ϵ. Next, we use this result and apply Lemma B.3 to bound the KL-divergence over the global state marginal distribution between two consecutive time steps by $\frac{K+1}{T}\epsilon$. Second line in equation (34) follows from this result and finally we derive the final bound linear in T.

$$\begin{aligned} D_{\text{TV}}(p_{\text{RHC}}(x;\theta_E) \parallel p_{\text{RHC}}(x;\theta)) &\leq \sum_{t=1}^{T-1} \sqrt{\frac{1}{2} D_{\text{KL}}(p_{\text{RHC}}^{t-1}(x;\theta) \parallel p_{\text{RHC}}^{t}(x;\theta))} \\ &\leq \sum_{t=1}^{T-1} \sqrt{\frac{(K+1)\epsilon}{2T}} \\ &= (T-1)\sqrt{\frac{K+1}{T}}\sqrt{\epsilon/2} \\ &\leq T\sqrt{\epsilon/2} \end{aligned} \tag{34}$$

The last line follows from the fact that $K << T$, so $\sqrt{\frac{K+1}{T}} < 1$. □

C Extension to Stochastic Dynamics

RHIRL optimizes the trajectories in the space of control sequences $p(V)$, whereas $V = \{u_0, u_1, u_2, ..., u_{K-1}\}$ is a sequence of controls. If the system is deterministic, we can apply V to the dynamical system $x_{t+1} = f(x_t, v_t)$ with start state x_0 and obtain a state sequence $\tau = (x_0, x_1, \ldots, x_K)$. We recall that total state trajectory cost of V defined in equation (3) as follows:

$$S(V, x_0; \theta) = \sum_{k=0}^{K} g(x_k; \theta)$$

We use the information-theoretic MPC (MPPI) [37] to solve for an optimal control sequence distribution at the current start state x_t in iteration t. The main result of MPPI suggests that, under a deterministic system, the optimal control sequence distribution Q minimizes the "free energy" of the dynamical system and this free energy can be calculated from the cost of the state trajectory under Q. Mathematically, the probability density $p_Q(V^*)$, as shown in equation (4) can be expressed as a

function of the state cost $S(V, x_t; \theta)$, with respect to a Gaussian "base" distribution $B(V_B, \Sigma)$ that depends on the control cost:

$$p_Q(V^*|U_B, \Sigma, x_t; \theta) = \frac{1}{Z} p_B(V^*|U_B, \Sigma) \exp(-\frac{1}{\lambda} S(V^*, x_t; \theta)),$$

where Z is the partition function. Intuitively, this result shows that the control sequence V that results in lower state-trajectory cost $S(V)$ are exponentially more likely to be chosen.

In this section, we extend RHIRL to stochastic dynamics where $x_{t+1} = f(x_t, v_t, \omega_t)$, where ω_t is a random variable that models the independent system noise. More specifically, we assume that $x_{t+1} \sim p(x_{t+1}|x_t, v_t)$. Due to the stochasticity of the dynamics, the state trajectory cost in equation (3) and the optimal control distribution in equation (4) are affected. Therefore, we first redefine the trajectory state cost under stochastic dynamics, then derive the counterpart of equation (4) for the optimal control sequence distribution under the stochastic dynamics, finally we adapt our existing RHIRL algorithm to stochastic dynamics.

C.1 State Trajectory Cost

Due to the stochasticity of the dynamics, given the initial state x_0, we no longer have a one-to-one mapping from the control sequence V to the resulting state trajectory $\tau = (x_0, x_1, ..., x_K)$. Instead, we have a distribution of state trajectories:

$$p(\tau|x_0, V) = \prod_{t=0}^{K-1} p(x_{t+1}|x_t, v_t) \tag{35}$$

To accommodate this change, the trajectory state cost of a control sequence $\tilde{S}(V, x_0; \theta)$ is defined over the distribution of the resulting state trajectories, instead of single trajectory:

$$\tilde{S}(V, x_0; \theta) = \int p(\tau|x_0, V) S(\tau|x_0; \theta) d\tau \tag{36}$$

$$= \int \prod_{t=0}^{K-1} p(x_{t+1}|x_t, v_t) \sum_{t=0}^{K} g(x_t; \theta) d\tau \tag{37}$$

We always measure the preferences over the control sequence V by their resulting state trajectories τ. Hence, when the resulting trajectories changes from a single deterministic sequence of states to a distribution of state trajectories, we adapt our measure of the resulting cost: under the deterministic dynamics where each V uniquely maps to the same state trajectory τ, the state trajectory cost is the cost of that specific trajectory; while under the stochastic dynamics where the same control sequence V maps to a distribution of τ, the state trajectory cost of a control sequence is now defined as the expected state cost of the distribution of trajectory. We measure the state trajectory cost of a control sequence $S(V, x_0; \theta)$, instead of simply a state trajectory cost on the states itself $S(\tau, x_0; \theta)$, because we want to use this measure to directly optimize the control sequence.

C.2 Optimal Control Sequence Distribution

Next, we derive the optimal control sequence distribution under the stochastic dynamics. Our derivation is based on MPPI [37], which uses the "free-energy" principle to derive the optimal control sequence distribution under deterministic dynamics.

Definition 2 (Free-energy ([33], Definition 1)**.** Let $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ and the function $\mathcal{J}(x)$: $\mathcal{Z} \to \mathbb{R}$ be a measurable function. The the term:

$$\mathbb{E}(\mathcal{J}(x)) = \log \int \exp(\rho \mathcal{J}(x)) d\mathbb{P} \tag{38}$$

is called free energy of $\mathcal{J}(x)$ with respect to $\mathbb{P}$, ρ is a constant.

Now we have the free-energy of a control system under stochastic dynamics as stated below. It has a "Gaussian" base control sequence distribution $B(U_B, \Sigma)$ such that its control sequence distribution

follows $p_B(V^*|U_B, \Sigma)$ whereas Σ is the Gaussian control noise covariance matrix. $\tilde{S}(V;\theta)$ denotes the state trajectory cost function.

$$\mathcal{F}(\tilde{S}, p_B, x_0, \lambda; \theta) = \log(\mathbb{E}_{p_B}[\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta))]), \tag{39}$$

$\lambda \in \mathbb{R}^+$ is the inverse temperature of the control system.

Suppose now we have another control sequence distribution with probability measure $p(V)$ and these two distributions are absolutely continuous, then we can rewrite the free-energy w.r.t $p_B(V)$ using the expectation over the density of $p(V)$ use the standard importance sampling trick:

$$\begin{aligned}
\mathcal{F}(\tilde{S}, p_B, x_0, \lambda; \theta) &= \log(\mathbb{E}_{p_B}[\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta))]) && (40)\\
&= \log(\mathbb{E}_{p_B}[\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta)\frac{p_B(V^*|U_B, \Sigma)}{p(V)})]) && (41)\\
&\geq \mathbb{E}_p[\log(\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta)\frac{p_B(V^*|U_B, \Sigma)}{p(V)}))] && (42)\\
&= -\frac{1}{\lambda}\mathbb{E}_p[\tilde{S}(V, x_0; \theta) + \lambda\log(\frac{p(V)}{p_B(V^*|U_B, \Sigma)})] && (43)\\
&= -\frac{1}{\lambda}(\mathbb{E}_p[\tilde{S}(V, x_0; \theta)] + \lambda\mathbb{E}_p[\log(\frac{p(V)}{p_B(V^*|U_B, \Sigma)})]) && (44)\\
&= -\frac{1}{\lambda}(\mathbb{E}_p[\tilde{S}(V, x_0; \theta)] + \lambda D_{\text{KL}}(p(V)||p_B(V^*|U_B, \Sigma))) && (45)
\end{aligned}$$

Therefore, we have

$$\mathcal{F}(S, p_B, x_0, \lambda; \theta) \geq -\frac{1}{\lambda}(\mathbb{E}_p[\tilde{S}(V, x_0; \theta)] + \lambda D_{\text{KL}}(p(V)||p_B(V^*|U_B, \Sigma))) \tag{46}$$

The right-hand side is the lower bound of the free-energy of the control system. We use $p_Q(V)$ to denote the optimal control sequence distribution. This distribution is only optimal if and only if the bound in the equation above is tight, i.e. $\mathcal{F}(\tilde{S}, p_B, x_0, \lambda; \theta) = -\frac{1}{\lambda}\mathbb{E}_{p_Q}[\tilde{S}(V, x_0; \theta)] + \lambda D_{\text{KL}}(p_Q(V)||p_B(V^*|U_B, \Sigma))$.

We claim that the optimal control sequence distribution $p_{\tilde{Q}}$ under the stochastic dynamics is as follows:

$$p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta) = \frac{1}{Z}\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta))p_B(V^*|U_B, \Sigma), \tag{47}$$

whereas $Z = \int \exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta))p_B(V^*|U_B, \Sigma)dV$ is the partition function.

We prove that equation (47) is the optimal control sequence distribution under stochastic dynamics by showing that this $p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta)$ tightens the bound of free-energy in equation (46). We substitute $p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta)$ into the RHS of equation (46) and simplify the expression of the KL-divergence:

$$\begin{aligned}
\mathcal{F}(\tilde{S}, p_B, x_0, \lambda; \theta) &\geq -\frac{1}{\lambda}(\mathbb{E}_{p_{\tilde{Q}}}[\tilde{S}(V, x_0; \theta)] + \lambda D_{\text{KL}}(p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta)||p_B(V^*|U_B, \Sigma))) && (48)\\
&= -\frac{1}{\lambda}(\mathbb{E}_{p_{\tilde{Q}}}[\tilde{S}(V, x_0; \theta)] + \lambda\mathbb{E}_{p_{\tilde{Q}}}[\log(\frac{p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta)}{p_B(V^*|U_B, \Sigma)})]) && (49)\\
&= -\frac{1}{\lambda}\mathbb{E}_{p_{\tilde{Q}}}[\tilde{S}(V, x_0; \theta)] - \mathbb{E}_{p_{\tilde{Q}}}[\log(\frac{\frac{1}{Z}\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta))p_B(V^*|U_B, \Sigma)}{p_B(V^*|U_B, \Sigma)})] && (50)\\
&= -\frac{1}{\lambda}\mathbb{E}_{p_{\tilde{Q}}}[\tilde{S}(V, x_0; \theta)] - (\frac{1}{\lambda}\mathbb{E}_{p_{\tilde{Q}}}[\tilde{S}(V, x_0; \theta)] - \log(Z)) && (51)
\end{aligned}$$

Next we substitute the expression for the partition function Z and we found that the RHS is exactly the definition of the the free-energy of the control system with base distribution $B(U_B, \Sigma)$:

$$\mathcal{F}(\tilde{S}, p_B, x_0, \lambda; \theta) \geq \log(Z) \tag{52}$$

$$= \log(\int \exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta)) p_B(V^*|U_B, \Sigma) dV) \tag{53}$$

$$= \log(\mathbb{E}_{p_B}[\exp(-\frac{1}{\lambda}\tilde{S}(V, x_0; \theta))]) \tag{54}$$

$$= \mathcal{F}(\tilde{S}, p_B, x_0, \lambda; \theta) \tag{55}$$

The final equality forces the inequality to be tight. Therefore, $p_Q(V^*|U_B, \Sigma, x_0; \theta)$ in equation (47) is the optimal control sequence distribution under the stochastic dynamics.

We observe that the optimal control sequence distribution under deterministic system $p_Q(V^*|U_B, \Sigma, x_0; \theta)$ in equation (4) and that under the stochastic dynamics $\tilde{p_Q}(V^*|U_B, \Sigma, x_0; \theta)$ in equation (47) only differs in the calculation of the state trajectory cost of the control sequences. Intuitively, it means that under deterministic dynamics, we choose the control sequence V that will, for sure, leads to a state trajectory with lower cost; while when we extend to stochastic dynamics, the control sequence V that results in lower state-trajectory cost $S(V)$ in expectation are exponentially more likely to be chosen. Practically, now we need more samples for a single control sequence to compute the expectation in equation (37).

C.3 RHIRL under Stochastic Dynamics

Next, we adapt our RHIRL algorithm to this new state trajectory cost measure $\tilde{S}(V, x_0; \theta)$ and the optimal control sequences $\tilde{p_Q}(V^*|U_B, \Sigma, x_0; \theta)$ under stochastic dynamics. We recall that under the deterministic dynamics, RHIRL uses importance sampling in equation (12) to estimate the $\frac{\partial \mathcal{L}}{\partial \theta}$ so as to update the cost function parameter θ:

$$\frac{\partial}{\partial \theta}\mathcal{L}(\theta; D, x_0) \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda}\frac{\partial}{\partial \theta}S(V_i, x_0; \theta) - \frac{1}{M}\sum_{j=1}^{M}\frac{1}{\lambda}w(V_j)\frac{\partial}{\partial \theta}S(V_j, x_0; \theta),$$

whereas the N control sequences in the first term are from the expert demonstration D_t and the M control sequences in the second term are from our approximated optimal control sequence distribution $p_Q(V^*)$, and $w(V_j)$ is the importance sampling weight.

To estimate $\frac{\partial \mathcal{L}}{\partial \theta}$, we need to calculate/approximate the importance sampling weight $w(V)$, and the derivative of the state trajectory cost $\frac{\partial}{\partial \theta}S(V, x_0; \theta)$ w.r.t θ. We recall that the importance sampling weight $w(V)$ depends on the state trajectory cost $S(V, x_0; \theta)$ in equation (11):

$$w(V) \propto \exp(-\frac{1}{\lambda}\left(S(V, x_0; \theta) + \lambda\sum_{k=0}^{K-1} u_k^\intercal \Sigma^{-1} v_k\right))$$

Under the deterministic dynamics, the importance sampling weight $w(V)$ is estimated using Monte-Carlo approximation with M state trajectory samples as follows:

$$w(V) \approx \frac{\exp\left(-\frac{1}{\lambda}(S(V, x_0; \theta) + \lambda\sum_{k=0}^{K-1} u_k^\intercal \Sigma^{-1} v_k)\right)}{\sum_{j=0}^{M-1}\exp\left(-\frac{1}{\lambda}(S(V_j, x_0; \theta) + \lambda\sum_{k=0}^{K-1} u_k^{j\intercal} \Sigma^{-1} v_k^j)\right)} \tag{56}$$

$$= \frac{\exp\left(-\frac{1}{\lambda}(\sum_{t=0}^{K} g(x_t; \theta) + \lambda\sum_{k=0}^{K-1} u_k^\intercal \Sigma^{-1} v_k)\right)}{\sum_{j=0}^{M-1}\exp\left(-\frac{1}{\lambda}(\sum_{t=0}^{K} g(x_t^j; \theta) + \lambda\sum_{k=0}^{K-1} u_k^{j\intercal} \Sigma^{-1} v_k^j)\right)} \tag{57}$$

Since the state trajectory cost $S(V, x_0; \theta)$ is a linear sum of the cost of all states, $\frac{\partial S}{\partial \theta}$ can be directly computed as follows:

$$\frac{\partial}{\partial \theta}S(V, x_0; \theta) = \frac{\partial}{\partial \theta}\sum_{t=0}^{K} g(x_t; \theta) = \sum_{t=0}^{K}\frac{\partial}{\partial \theta}g(x_t; \theta) \tag{58}$$

To extend RHIRL to stochastic dynamics, when the state trajectory cost function is now $\tilde{S}(V, x_0; \theta)$, we need to redefine how to estimate $\tilde{w}(V)$ and consequently $\frac{\partial \tilde{S}}{\partial \theta}$.

When extend to stochastic dynamic, we have the following:

$$\tilde{w}(V) \propto \exp(-\frac{1}{\lambda}\Big(\tilde{S}(V, x_0; \theta) + \lambda \sum_{k=0}^{K-1} u_k^\intercal \Sigma^{-1} v_k)\Big) \tag{59}$$

and we still adopts Monte-carlo sampling to approximate $\tilde{w}(V)$. However, since $\tilde{S}(V, x_0; \theta)$ now measures the expected state trajectory cost under the stochastic dynamics, we need to go one step further and use sampling to estimate $\tilde{S}(V, x_0; \theta)$ using M^s number of state trajectories $\tau_h = (x_0, x_1^h, ..., x_K^h)$ per (x_0, V) pair:

$$\tilde{S}(V, x_0; \theta) \approx \frac{1}{M^s} \sum_{h=0}^{M^s-1} \sum_{t=0}^{K} g(x_t^h; \theta) \tag{60}$$

Therefore, now the importance sampling weight $\tilde{w}(V)$ is approximated from $M \times M^s$ state trajectories, with M^s trajectories from each (x_0, V_j) pair as follows:

$$\tilde{w}(V) \approx \frac{\exp\Big(-\frac{1}{\lambda}(\tilde{S}(V, x_0; \theta) + \lambda \sum_{k=0}^{K-1} u_k^\intercal \Sigma^{-1} v_k)\Big)}{\sum_{j=0}^{M-1} \exp\Big(-\frac{1}{\lambda}(\tilde{S}(V_j, x_0; \theta) + \lambda \sum_{k=0}^{K-1} u_k^{j\intercal} \Sigma^{-1} v_k^j)\Big)} \tag{61}$$

$$\approx \frac{\exp\Big(-\frac{1}{\lambda}(\frac{1}{M^s} \sum_{h=0}^{M^s-1} \sum_{t=0}^{K} g(x_t^h; \theta) + \lambda \sum_{k=0}^{K-1} u_k^\intercal \Sigma^{-1} v_k)\Big)}{\sum_{j=0}^{M-1} \exp\Big(-\frac{1}{\lambda}(\frac{1}{M^s} \sum_{h=0}^{M^s-1} \sum_{t=0}^{K} g(x_t^{j^h}; \theta) + \lambda \sum_{k=0}^{K-1} u_k^{j\intercal} \Sigma^{-1} v_k^j)\Big)} \tag{62}$$

We emphasize that under the stochastic dynamics, we use the state trajectory samples to estimate both the state trajectory cost $\tilde{S}(V, x_0; \theta)$ and the importance sampling weight $\tilde{w}(V)$. Since the $\tilde{S}$ now measures the expected cost over a distribution of trajectories, we need more samples to estimate $\tilde{w}(V)$ compared to the deterministic setting.

Moreover, in the final $\frac{\partial \mathcal{L}}{\partial \theta}$, we need to differentiate $\tilde{S}(V, x_0; \theta)$ w.r.t. θ. Since $\tilde{S}$ is estimated from sampling, we have:

$$\frac{\partial}{\partial \theta} \tilde{S}(V, x_0; \theta) \approx \frac{\partial}{\partial \theta} \frac{1}{M^s} \sum_{h=0}^{M^s-1} \sum_{t=0}^{K} g(x_t^h; \theta) \approx \frac{1}{M^s} \sum_{h=0}^{M^s-1} \sum_{t=0}^{K} \frac{\partial}{\partial \theta} g(x_t^h; \theta) \tag{63}$$

Finally, we summarize how to extend RHIRL to stochastic dynamics. In stochastic dynamics, each control sequence V will map to a distribution of state trajectory $p(\tau|V, x_0)$. Hence, we adapt our measure of state trajectory cost $S(V, x_0; \theta)$ from a single trajectory to be the expected cost over a distribution of state trajectories $\tilde{S}(V, x_0; \theta)$ in equation (37). Next, we revise the optimal control sequence distribution to a stochastic setting $p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta)$ in equation (47). More specifically, we show under the stochastic dynamics, the optimal control sequences V^* is chosen based on the expected cost of its resulting state trajectories. Finally, under this new state trajectory cost $\tilde{S}(V, x_0; \theta)$ and the optimal control sequence distribution $p_{\tilde{Q}}(V^*|U_B, \Sigma, x_0; \theta)$, we adapt the approximation of the importance sampling weight $\tilde{w}(V)$ and consequently the gradient of the overall loss w.r.t θ by adding one more sampling process to estimate the new state trajectory cost. Moreover, in practice, we can use the same set of samples to estimate both state trajectory cost $\tilde{S}(V, x_0; \theta)$ and the importance sampling weights $\tilde{w}(V)$.

D Experimental Details

In this section, we list down the implementation details of RHIRL and the baselines. The code is included in the supplementary material. We also report the hyperparameters used in the experiments, the detailed network architectures, training procedures and evaluation procedures used for our experiments.

D.1 Practical Issues of RHIRL

Control Noise Covariance Approximation

The actual control noise covariance Σ is unknown to RHIRL and the baselines. However, RHIRL uses the noise covariance matrix Σ to sample the controls around the nominal control (Algorithm 1, line 6) and calculate the quadratic control cost in Equation (37). Since we have no access to the true Σ, RHIRL approximates Σ as a constant factor of the identity matrix βI, whereas β is the hyperparameter we optimize using grid search and I is the identify matrix with its width equals to the dimension of the action space. Therefore, instead of sampling the controls from $\mathcal{N}(V|U, \Sigma)$, we sample from $\mathcal{N}(V|U, \beta I)$ in practice. We also use βI in Equation (37) to replace the unknown Σ. Even in the noise-free environment, we set β to a non-zero value to foster exploration; otherwise, the importance sampling degenerates to the single nominal control.

Our experiment shows that RHIRL is robust to the choice of β: the cost learning performs well even if $\beta I \neq \Sigma$. This may be attributed to the fact that we jointly optimize the state cost function and β. Therefore the learned state cost function may compensate for the inaccurate approximation for Σ.

Numerical Stability

Equation(11) forms the basis of importance sampling and estimation. However, the learned cost can be a huge negative number, which causes numerical instability in estimating the importance weights. To mitigate this issue, we subtract the minimum trajectory cost S_{min} from all rollouts to improve the numeric stability. Since subtracting the same number from all rollouts does not change the order of the preference, this operation does not affect the optimality of our derivation.

Nominal Control Initialization and Local Optimality

RHIRL samples around the nominal control sequence to collect the samples for importance sampling. However, if the initial nominal control sequence performs poorly, it is not easy to generate any good samples to improve the current control sequences. To mitigate this problem, we add an exploration strategy to the sampling process in (Algorithm 1, line 6): with probability α, we continue with the standard sampling strategy to sample around the nominal control; with probability $1 - \alpha$ we sample uniformly from the entire action space. This helps RHIRL to correct from the unsuitable nominal control initialization and also helps RHIRL to escape the local optimal solution. We set $\alpha = 0.5$ for all tasks in our experiments.

Control Smoothness

Updating the optimal control by importance sampling might cause some jerk in the control space. In order to make the control change smoothly in its local space, we apply a Savitzky–Golay filter over the time horizon dimension to constrain the control that does not change too much over the time horizon.

D.2 Training

We list the hyper-parameters of RHIRL for different tasks. These hyper-parameters were selected via grid search.

Task	K	β	batch size	λ	lr	weight decay
Pendulum-v0	20	0.8	50	0.10	1e-4	8e-5
LunarLanderContinuous-v2	40	0.6	200	0.10	1e-4	8e-5
Hopper-v2	20	0.8	100	0.10	1e-4	8e-5
Walker2d-v2.	30	0.6	150	0.10	1e-4	8e-5
Ant-v2	15	1.2	200	0.10	1e-4	8e-5
CarRacing-v0	15	1.0	200	0.10	1e-4	8e-5

The implementation of the baselines (f-IRL, AIRL and GAIL) are adapted from f-IRL's [27] official repository. We use the hyperparameters reported in f-IRL for the MuJoCo tasks and performed grid search on the hyperparameters for the rest of the tasks. SAC[12] is used as the base MaxEnt RL algorithm for both expert policy and the baselines optimization algorithm. We use a tanh squashed Gaussian as the policy network for Pendulum-v0, LunarLander-v2, and the MuJoCo tasks; and we

use a Gaussian Convolutional policy as the policy network for CarRacing-v0. The mean and std of the Gaussian are parameterized by a ReLU MLP of size (64, 64). Adam is used as the optimizer. We use the reported SAC temperature,$\alpha = 0.2$, reward scale $c = 0.2$, and gradient penalty coefficient $\lambda = 4.0$. The rest of the hyperparameters for f-IRL, GAIL and AIRL are listed below.

Task	SAC learning rate	SAC replay buffer size	Reward/Value model learning rate	l2 weight decay
Pendulum-v0	1e-4	100000	1e-5	1e-3
LunarlanderContinuous-v2	1e-3	100000	1e-5	1e-3
Hopper-v2	1e-5	1000000	1e-5	1e-3
Walker2d-v2	1e-5	1000000	1e-5	1e-3
Ant-v2	3e-4	1000000	1e-4	1e-3
CarRacing-v0	4e-4	10000000	1e-4	1e-3

D.3 Reward Function and Discriminator Network Architectures

We use the same neural network architecture to parameterize the cost-function/reward-function/discriminator for all methods. For continuous control task with raw state input, i.e. pendulum, lunarlander and the MuJoCo tasks, we use two-layer of MLP with ReLU activation function to parameterized the cost function/discriminator. The hidden size for Pendulum-v0 is (32, 32), and (64, 64) for the rest of the tasks.

For continuous control task with image input, i.e. carracing, we use a four convolutional layer with kernel size 3×3 as the feature extractor. The output of the CNN layer is vector with size (128,) and is fed into the same reward network as describe above.

D.4 Additional Experiments Results

We report the average returns and the standard deviation for Table 1 and Table 2 in Table 3 and Table 4 respectively. The mean and standard deviation computed from 3 trials for each entry of the tables.

Table 3: Performance of RHIRL, f-IRL, GAIL, and AIRL. We report the mean and the standard deviation of the policy returns using the ground-truth task reward. Higher values indicate better performance.

		No Noise $\Sigma = 0$	**Mild Noise** $\Sigma = 0.2$	**High Noise** $\Sigma = 0.5$
Pendulum	Expert	-154.69 ± 50.05	-156.50 ± 70.72	-168.54 ± 80.89
	RHIRL	-125.95 ± 1.21	-122.33 ± 3.44	-132.39 ± 10.36
	f-IRL	-121.94 ± 97.21	-127.51 ± 104.55	-197.36 ± 106.92
	AIRL	-131.64 ± 1.16	-184.62 ± 88.16	-203.12 ± 80.57
	GAIL	-207.05 ± 57.41	-207.14 ± 57.52	-253.85 ± 181.84
LunarLander	Expert	235.13 ± 43.59	222.65 ± 56.35	164.52 ± 36.79
	RHIRL	246.39 ± 10.96	233.73 ± 23.75	198.23 ± 47.8
	f-IRL	179.03 ± 9.19	141.73 ± 11.81	121.67 ± 22.77
	AIRL	174.49 ± 35.17	132.76 ± 85.59	95.61 ± 19.25
	GAIL	169.98 ± 15.43	125.5 ± 16.78	100.24 ± 79.04
Hopper	Expert	3222.48 ± 390.65	3159.32 ± 520.00	2887.72 ± 483.93
	RHIRL	3071.63 ± 122.03	3121.72 ± 278.98	2776.2 ± 345.90
	f-IRL	3080.34 ± 458.96	2580.19 ± 637.21	1270.24 ± 539.84
	AIRL	18.9 ± 0.79	33.52 ± 3.86	18.38 ± 7.84
	GAIL	2642.59 ± 187.33	1576.25 ± 1051.98	702.33 ± 151.37
Walker2d	Expert	4999.47 ± 55.99	4500.43 ± 114.48	3624.48 ± 95.05
	RHIRL	4939.44 ± 100.28	4473.332 ± 324.34	3446.55 ± 507.89
	f-IRL	4927.92 ± 529.95	3697.36 ± 711.56	2831.91 ± 993.76
	AIRL	-2.51 ± 0.69	22.24 ± 10.74	6.5 ± 5.03
	GAIL	2489.04 ± 813.31	2884.35 ± 59.88	1840.62 ± 778.3
Ant	Expert	5759.22 ± 173.57	2557.37 ± 501.95	252.62 ± 91.44
	RHIRL	4987.67 ± 149.2	2373.32 ± 529.3	230.8 ± 253.39
	f-IRL	5022.42 ± 108.07	2034.87 ± 262.29	197.2 ± 200.45
	AIRL	1000.4 ± 0.79	849.05 ± 30.15	-7.43 ± 6.01
	GAIL	2784.87 ± 301.66	1022.04 ± 580.49	-416.69 ± 292.23
CarRacing	Expert	903.25 ± 0.23	702.01 ± 0.3	281.12 ± 0.34
	RHIRL	359.61 ± 40.32	206.21 ± 19.87	53.97 ± 3.24
	f-IRL	85.45 ± 47.4	18.32 ± 27.89	2.04 ± 13.8
	AIRL	-21.97 ± 2.67	-25.25 ± 5.98	-32.31 ± 7.43
	GAIL	2.62 ± 3.41	-7.65 ± 4.77	-15.88 ± 5.89

Table 4: Generalization of learned cost functions over different noise levels.

		Noise-free for learning	**Noise Level Σ for Testing** 0.2	0.5
Pendulum	RHIRL	-125.95 ± 1.21	-125.01 ± 4.53	-126.4 ± 7.73
	f-IRL	-121.94 ± 97.21	-199.44 ± 96.99	-220.74±79.75
	AIRL	-131.64 ± 1.16	-247.86±11.44	-304.48±20.78
	GAIL	-207.05 ± 57.41	-220.6± 69.82	-270.81± 79.68
LunarLander	RHIRL	246.39 ± 10.96	205.66±24.67	175.82±52.12
	f-IRL	179.03 ± 9.19	121.80±20.94	102.06±22.31
	AIRL	174.49 ± 35.17	31.46±9.68	22.29±14.01
	GAIL	169.98 ± 15.43	101.80 ± 23.12	78.33± 24.15
Hopper	RHIRL	3071.63 ± 122.03	2577.28 ±409.33	2152.08 ±342.21
	f-IRL	3080.34 ± 458.96	2110.52±26.71	1984.29±31.88
	AIRL	18.9 ± 0.79	18.86±4.80	8.78±10.89
	GAIL	2642.59 ± 187.33	215.29± 27.76	132.15± 30.20
Walker2d	RHIRL	4939.44 ± 100.28	4039.44±39.2	3440.23±531.08
	f-IRL	4927.92 ± 529.95	2976.66±396.57	1090.11±1389.56
	AIRL	-2.51 ± 0.69	1380.84±364.95	1787.15±230.94
	GAIL	2489.04 ± 813.31	103.15± 121.84	124.15±82.84
Ant	RHIRL	4987.67 ± 149.2	3192.82±162.12	867.08±204.28
	f-IRL	5022.42 ± 108.07	2042.41±129.89	472.77±110.2
	AIRL	1000.4 ± 0.79	845.69 ±29.01	0.69 ±20.49
	GAIL	2784.87 ± 301.66	-6.41± 21.17	-79.89± 142.43
CarRacing	RHIRL	359.61 ± 40.32	261.78±54.44	110.12±58.90
	f-IRL	85.45 ± 47.4	16.12±67.82	-24.78±2.12
	AIRL	-21.97 ± 2.67	-27.09±6.65	-23.96±4.11
	GAIL	2.62 ± 3.41	-6.41±3.22	-49.89±7.98