Supplementary Material for "*On Using Hamiltonian Monte Carlo Sampling for Reinforcement Learning Problems in High-dimension*"

A Convergence and Boundedness Results

We proceed to prove theorem by stating convergence properties for HMC as follows. In the initial sampling stage, starting from the initial position Markov chain converges towards to the typical set. In the next stage Markov chain quickly traverse the typical set and improves the estimate by removing the bias. In the last stage Markov chain refine the exploration of typical the typical set provide improved estimates. The number of samples taken during the last stage is referred as effective sample size.

A.1 Proof of Theorem 1

Theorem 1. *Let $\mathcal{T}$ be an optimality operator under HMC given as $(\mathcal{T}Q)(s',a') = r(s',a') + \frac{\gamma}{|\mathcal{H}|}\sum_{s\in\mathcal{H}}\max_a Q(s,a),\ \forall(s',a')\in\mathcal{S}\times\mathcal{A}$, where $\mathcal{H}$ is a subset of next states sampled using HMC from the target distribution given in (6). Then, under update rule (4) and for any given $\epsilon \geq 0$, there exists $n_{\mathcal{H}}, t' > 0$ such that $\|Q^t - Q^*\|_\infty \leq \epsilon\ \forall t \geq t'$.*

Proof of Theorem 1. Let $\bar{Q}^t(s,a) = \frac{1}{n_{\mathcal{H}}}\max_a Q^t(s,a), \forall(s,a)\in\mathcal{S}\times\mathcal{A}$. Here we consider $n_{\mathcal{H}}$ to be the effective number of samples. Let $\mathbb{E}_{\mathcal{P}}Q^t, \mathbf{Var}_{\mathcal{P}}Q^t$ be the expectation and covariance of Q^t with respect to the target distribution. From Central Limit Theorem for HMC we have

$$\bar{Q}^t \sim \mathcal{N}\left(\mathbb{E}_{\mathcal{P}}Q^t, \sqrt{\frac{\mathbf{Var}_{\mathcal{P}}Q^t}{n_{\mathcal{H}}}}\right).$$

Since Q function does not decay fast we provide a proof for the case where Q^t is C-Lipschitz. From Theorem 6.5 in [41] we have that, there exists a $c_0 > 0$ such that

$$||\bar{Q}^t - \mathbb{E}_{\mathcal{P}}Q^t|| \leq c_0. \tag{S.1}$$

Recall that Bellman optimality operator $\mathcal{T}$ is a contraction mapping. Thus from triangle inequality we have

$$\begin{aligned}\Big|\Big|\mathcal{T}Q_1 - \mathcal{T}Q_2\Big|\Big|_\infty &\leq \max_{s',a'}\Big|\Big|r(s',a') + \frac{\gamma}{|\mathcal{H}_1|}\sum_{s\in\mathcal{S}}\max_a Q_1(s,a)\\ &\qquad -r(s',a') - \frac{\gamma}{|\mathcal{H}_2|}\sum_{s\in\mathcal{S}}\max_a Q_2(s,a)\Big|\Big|\\ &\leq \max_{s',a'}\Big|\Big|\frac{\gamma}{|\mathcal{H}_1|}\sum_{s\in\mathcal{S}}\max_a Q_1(s,a) - \frac{\gamma}{|\mathcal{H}_2|}\sum_{s\in\mathcal{S}}\max_a Q_2(s,a)\Big|\Big|\end{aligned}$$

Let $|\mathcal{H}_1| = |\mathcal{H}_2| = n_{\mathcal{H}}$. Then using triangle inequality we have

$$\Big|\Big|\mathcal{T}Q_1 - \mathcal{T}Q_2\Big|\Big|_\infty \leq \max_{s',a'}\gamma\left[\Big|\Big|\bar{Q}_1 - \mathbb{E}_{\mathcal{P}}Q_1\Big|\Big| + \Big|\Big|\bar{Q}_2 - \mathbb{E}_{\mathcal{P}}Q_2\Big|\Big|\right] + \max_{s',a'}\gamma\Big|\Big|\mathbb{E}_{\mathcal{P}}Q_1 - \mathbb{E}_{\mathcal{P}}Q_2\Big|\Big|$$

Since Q function almost surely converge under exhaustive sampling we have

$$\max_{s',a'}\gamma\Big|\Big|\mathbb{E}_{\mathcal{P}}Q_1 - \mathbb{E}_{\mathcal{P}}Q_2\Big|\Big| \leq \gamma\Big|\Big|Q_1 - Q_2\Big|\Big|_\infty \tag{S.2}$$

From (S.1) and (S.2) we have after t time steps

$$\Big|\Big|\mathcal{T}Q_1 - \mathcal{T}Q_2\Big|\Big|_\infty \leq 2c_0 + \gamma\Big|\Big|Q_1 - Q_2\Big|\Big|_\infty$$

Let R_{max} and R_{min} be the maximum and minimum reward values. Then we have that

$$\Big|\Big|Q_1 - Q_2\Big|\Big|_\infty \leq \frac{\gamma}{1-\gamma}R_{max} - R_{min}.$$

Thus for any $\epsilon \geq$ by choosing a γ such there exists a t' such that $\forall t \geq t'$

$$\|Q^t - Q^*\|_\infty \leq \epsilon$$

This concludes the proof of Theorem 1. □

A.2 Proof of Theorem 2

Theorem 2. *Let $Q_{\mathcal{E}}^{t+1}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s \in \mathcal{S}} \mathbb{P}(s|s_t, a_t) \max_a Q_{\mathcal{E}}^t(s, a), \forall (s_t, a_t) \in \mathcal{S} \times \mathcal{A}$ be the update rule under exhaustive sampling, and Q^t be the Q function updated according to Hamiltonian Q-Learning, i.e. by (9)-(10). Then, for any given $\tilde{\epsilon} \geq 0$, there exists $n_{\mathcal{H}}, t' > 0$, such that $\|Q^t - Q_{\mathcal{E}}^t\|_\infty \leq \tilde{\epsilon}\ \forall t \geq t'$.*

Proof of Theorem 2. Note that at each time step we attempt to recover the matrix $Q_{\mathcal{E}}^t$, i.e., Q function time t under exhaustive sampling though a matrix completion method starting from $\widehat{Q}^t$, which is the Q updated function at time t using Hamiltonian Q-Learning. From Theorem 4 in [24] we have that $\forall t \geq t'$ there exists some constant $\delta > 0$ such that when the updated Q function a $\widehat{Q}^t$ satisfy

$$\left|\left|\widehat{Q}^t - Q_{\mathcal{E}}^t\right|\right|_\infty \leq c$$

where c is some positive constant then reconstructed (completed) matrix Q^t satiesfies

$$\left|\left|Q^t - Q_{\mathcal{E}}^t\right|\right|_\infty \leq \delta \left|\left|\widehat{Q}^t - Q_{\mathcal{E}}^t\right|\right|_\infty \tag{S.3}$$

for some $\delta > 0$. This implies that when the initial matrix used for matrix completion is sufficiently close to the matrix we are trying to recover matrix completion iterations converge to a global optimum. From the result of Theorem 1 we have for any given $\epsilon \geq 0$, there exists $n_{\mathcal{H}}, t' > 0$ such that $\forall t \geq t'$

$$\left|\left|\widehat{Q}^t - Q^*\right|\right| \leq \epsilon \tag{S.4}$$

Recall that under the update equation $Q_{\mathcal{E}}^{t+1}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s \in \mathcal{S}} \max_a Q_{\mathcal{E}}^t(s, a), \forall (s_t, a_t) \in \mathcal{S} \times \mathcal{A}$ we have that $Q_{\mathcal{E}}$ almost surely converge to the optimal Q^*. Thus there exists a $t^\dagger$ such that $\forall t \geq t^\dagger$

$$\left|\left|Q_{\mathcal{E}}^t - Q^*\right|\right| \leq \epsilon$$

Let $t^\ddagger = \max\{t^\dagger, t'\}$. Then from triangle inequality we have that

$$\left|\left|\widehat{Q}^t - Q_{\mathcal{E}}^t\right|\right| \leq \left|\left|\widehat{Q}^t - Q^*\right|\right| + \left|\left|Q_{\mathcal{E}}^t - Q^*\right|\right| \leq 2\epsilon.$$

Thus from (S.3) we have that

$$\left|\left|Q^t - Q_{\mathcal{E}}^t\right|\right|_\infty \leq 2\delta\epsilon$$

This concludes the proof of Theorem 2. □

B Sampling Complexity

In this section we provide theoretical results on sampling complexity of Hamiltonian Q-Learning. For brevity of notation we define $\mathcal{M}Q(s) = \max_a Q(s, a)$. Note that we have the following regularity conditions on the MDP studied in this paper.

Regularity Conditions

1. Spaces $\mathcal{S}$ and $\mathcal{A}$ (state space and action space) are compact subsets of $\mathbb{R}^{\mathcal{D}_s}$ and $\mathbb{R}^{\mathcal{D}_a}$ respectively.
2. All the rewards are bounded such that $r(s, a) \in [R_{\min}, R_{\max}]$, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
3. The optimal Q^* is C-Lipschitz such that

$$\left|Q^*(s, a) - Q^*(s', a')\right| \leq C\left(||s - s'||_F + ||a - a'||_F\right)$$

Now we prove some useful lemmas for proving sampling complexity of Hamiltonian Q-Learning

Lemma 1. *For some constant c_1, if*

$$|\Omega_t| \geq c_1 \frac{\max\left\{|\mathcal{S}|^2, |\mathcal{A}|^2\right\}|\mathcal{S}||\mathcal{A}|\mathcal{D}_s\mathcal{D}_a}{\log\left(\mathcal{D}_s + \mathcal{D}_a\right)}$$

with $\left|\left|\widehat{Q}^t(s,a) - Q^*(s,a)\right|\right|_\infty \leq \epsilon$ then there exists a constant c_2 such that

$$\left|\left|Q^t(s,a) - Q^*(s,a)\right|\right|_\infty \leq c_2\epsilon$$

Proof of Lemma 1. Recall that in order to complete a low rank matrix using matrix estimation methods, the matrix can not be sparse. This condition can be formulated using the notion of incoherence. Let Q be a matrix of rank r_Q with the singular value decomposition $Q = U\Sigma V^T$. Let T_Q be the orthogonal projection of $Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to its column space. Then incoherence parameter of $\phi(Q)$ can be give as

$$\phi(Q) = \max\left\{\frac{|\mathcal{S}|}{r_Q}\max_{1\leq i\leq|\mathcal{S}|}||T_U\mathbf{e}_i||_F^2, \frac{|\mathcal{A}|}{r_Q}\max_{1\leq i\leq|\mathcal{A}|}||T_U\mathbf{e}_i||_F^2\right\}$$

where $\mathbf{e}_i$ are the standard basis vectors. Recall that Q^t is the matrix generated in matrix completion phase from $\widehat{Q}$. From Theorem 4 in [24] we have that for some constant C_1 if a fraction of p elements are observed from the matrix such that

$$p \geq C_1 \frac{\phi_t^2 r_Q^2 \mathcal{D}_s\mathcal{D}_a}{\log\left(\mathcal{D}_s + \mathcal{D}_a\right)}$$

where ϕ_t is the coherence parameter of Q^t then with probability at least $1 - C_2(\mathcal{D}_s + \mathcal{D}_a)^{-1}$ for some constant C_2 with $\left|\left|\widehat{Q}^t(s,a) - Q^*(s,a)\right|\right|_\infty \leq \epsilon$ there exists a constant c_2 such that

$$\left|\left|Q^t(s,a) - Q^*(s,a)\right|\right|_\infty \leq c_2\epsilon$$

Note that $p \approx \frac{|\Omega_t|}{|\mathcal{S}||\mathcal{A}|}$. Further we have for some constant c_3

$$\frac{\phi_t^2 r_Q^2 \mathcal{D}_s\mathcal{D}_a}{\log\left(\mathcal{D}_s + \mathcal{D}_a\right)} = c_3 \frac{\max\left\{|\mathcal{S}|^2, |\mathcal{A}|^2\right\}\mathcal{D}_s\mathcal{D}_a}{\log\left(\mathcal{D}_s + \mathcal{D}_a\right)}$$

Thus it follows that for some constant c_1 if

$$|\Omega_t| = c_1 \frac{\max\left\{|\mathcal{S}|^2, |\mathcal{A}|^2\right\}|\mathcal{S}||\mathcal{A}|\mathcal{D}_s\mathcal{D}_a}{\log\left(\mathcal{D}_s + \mathcal{D}_a\right)}$$

with $\left|\left|\widehat{Q}^t(s,a) - Q^*(s,a)\right|\right|_\infty \leq \epsilon$ then there exists a constant c_2 such that

$$\left|\left|Q^t(s,a) - Q^*(s,a)\right|\right|_\infty \leq c_2\epsilon$$

This concludes the proof of Lemma 1. □

Lemma 2. *Let $1 - \xi$ be the spectral gap of Markov chain under Hamiltonian sampling where $\xi \in [0,1]$. Let $\Delta R = R_{\max} - R_{\min}$ be the maximum reward gap. Then $\forall(s',a') \in \mathcal{S}\times\mathcal{A}$ we have that*

$$|\widehat{Q}(s',a') - Q^*(s',a')| \leq \frac{\gamma^2}{1-\gamma}\Delta R + \sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2 \log\left(\frac{2}{\delta}\right)}.$$

with at least probability $1 - \delta$.

Proof of Lemma 2. Let $\widehat{Q}(s',a') = r(s',a') + \frac{\gamma}{|\mathcal{H}|}\sum_{s\in\mathcal{H}} \max_a Q(s,a)$. Recall that $\mathcal{M}Q(s) = \max_a Q(s,a)$. Then we have that $\widehat{Q}(s',a') = r(s',a') + \frac{\gamma}{|\mathcal{H}|}\sum_{s\in\mathcal{H}} \mathcal{M}Q(s)$. Then it follows that

$$\begin{aligned}
|\widehat{Q}(s',a') - Q^*(s',a')\Big| &= \Big|r(s',a') + \frac{\gamma}{|\mathcal{H}|}\sum_{s\in\mathcal{H}} \mathcal{M}Q(s) - r(s',a') - \gamma\mathbb{E}_{\mathcal{P}}\mathcal{M}Q^*(s)\Big| \\
&= \Big|\frac{\gamma}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q(s_i) - \gamma\mathbb{E}_{\mathcal{P}}\mathcal{M}Q^*(s)\Big| \\
&= \Big|\frac{\gamma}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q(s_i) - \frac{\gamma}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q^*(s_i)\Big| \\
&\quad + \Big|\frac{\gamma}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q^*(s_i) - \gamma\mathbb{E}_{\mathcal{P}}\mathcal{M}Q^*(s)\Big| \qquad \text{(S.5)}
\end{aligned}$$

Recall that all the rewards are bounded such that $r(s,a) \in [R_{\min}, R_{\max}]$, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Thus for all s,a we have that $\mathcal{M}Q(s) \leq \frac{\gamma}{1-\gamma}R_{\max}$. Let $\Delta R = R_{\max} - R_{\min}$. Then we have that

$$\Big|\frac{\gamma}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q(s_i) - \frac{\gamma}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q^*(s_i)\Big| \leq \frac{\gamma^2}{1-\gamma}\Delta R. \qquad \text{(S.6)}$$

Let $\xi \in [0,1]$ be a constant such that $1-\xi$ is the spectral gap of the Markov chain under Hamiltonian sampling. Then from [42] we have that

$$\mathbb{P}\left(\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q^*(s_i) - \mathbb{E}_{\mathcal{P}}\mathcal{M}Q^*(s) \geq \vartheta\right) \leq \exp\left(-\frac{1-\xi}{1+\xi}\frac{|\mathcal{H}|\vartheta^2}{2R_{\max}^2}\left(\frac{1-\gamma}{\gamma}\right)^2\right)$$

Let $\delta = \exp\left(-\frac{1-\xi}{1+\xi}\frac{|\mathcal{H}|\vartheta^2}{2R_{\max}^2}\left(\frac{1-\gamma}{\gamma}\right)^2\right)$. Then we have that

$$\vartheta = \sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2 \log\left(\frac{2}{\delta}\right)}.$$

Thus we see that

$$\Big|\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathcal{M}Q^*(s_i) - \mathbb{E}_{\mathcal{P}}\mathcal{M}Q^*(s)\Big| \leq \sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2 \log\left(\frac{2}{\delta}\right)} \qquad \text{(S.7)}$$

with at least probability $1-\delta$. Thus it follows from equations (S.5), (S.6) and (S.7) that

$$|\widehat{Q}(s',a') - Q^*(s',a')\Big| \leq \frac{\gamma^2}{1-\gamma}\Delta R + \sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2 \log\left(\frac{2}{\delta}\right)}.$$

with at least probability $1-\delta$. This concludes the proof of Lemma 2. □

Lemma 3. *For all $(s,a) \in \mathcal{S}\times\mathcal{A}$ we have that*

$$|Q^t(s,a) - Q^*(s,a)\Big| \leq 2c_1\frac{\gamma^2 R_{\max}}{1-\gamma}$$

with probability at least $1-\delta$

Proof of Lemma 3. From Lemma 2 and [14] we have that for all $(s,a) \in \Omega_t$

$$|\widehat{Q}^t(s,a) - Q^*(s,a)\Big| \leq \frac{\gamma^2}{1-\gamma}\Delta R + \sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}_t|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2 \log\left(\frac{2|\Omega_t|T}{\delta}\right)}. \qquad \text{(S.8)}$$

with probability at least $1-\frac{\delta}{T}$. Thus we have that

$$|Q^t(s,a) - Q^*(s,a)\Big| \leq c_1\frac{\gamma^2}{1-\gamma}\Delta R + c_1\sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}_t|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2 \log\left(\frac{2|\Omega_t|T}{\delta}\right)}.$$

with probability at least $1 - \frac{\delta}{T}$. Fro all $1 \leq t \leq T$ letting

$$|\mathcal{H}_t| = \frac{1+\xi}{1-\xi}\frac{2}{\gamma^2}\log\left(\frac{2|\Omega_t|T}{\delta}\right)$$

we obtain

$$\frac{\gamma^2}{1-\gamma}R_{\max} \geq \sqrt{\frac{1+\xi}{1-\xi}\frac{2}{|\mathcal{H}_t|}\left(\frac{\gamma R_{\max}}{1-\gamma}\right)^2\log\left(\frac{2|\Omega_t|T}{\delta}\right)}.$$

Thus we have,

$$\left|Q^t(s,a) - Q^*(s,a)\right| \leq 2c_1\frac{\gamma^2 R_{\max}}{1-\gamma}$$

with probability at least $1-\delta$. Recall that $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$ we have $\mathcal{M}Q(s,a) \leq \frac{\gamma\Delta R}{1-\gamma}$. Thus this also proves that

$$\left|Q^t(s,a) - Q^*(s,a)\right| \leq 2c_1\gamma\left|Q^{t-1}(s,a) - Q^*(s,a)\right|$$

This concludes the proof of Lemma 3. □

Now we proceed to prove the main theorem for sampling complexity as follows.

Theorem 3. *Let $\mathcal{D}_s, \mathcal{D}_a$ be the dimension of state space and action space respectively. Consider the Hamiltonian Q-Learning algorithm presented in Algorithm 1. Under a suitable matrix completion method sampling complexity of the algorithm, Q function converge to the family of ϵ-optimal Q functions with $\widetilde{O}\left(\epsilon^{-(\mathcal{D}_s+\mathcal{D}_a+2)}\right)$ number of samples.*

Proof of Theorem 3. Note that sample complexity of Hamiltonian Q-Learning can be given as

$$\sum_{t=1}^{T_\epsilon}|\Omega_t||\mathcal{H}_t| \leq T_\epsilon|\Omega_{T_\epsilon}||\mathcal{H}_{T_\epsilon}|$$

Let β^t be the discretization parameter at time t and $T_\epsilon = \frac{\log\left(\frac{\gamma R_{\max}}{(1-\gamma)\epsilon}\right)}{\log\left(\frac{1}{2\gamma c_1}\right)}$. Then from Lemmas 1, 2 and 3 it follows that

$$\sum_{t=1}^{T_\epsilon}|\Omega_t||\mathcal{H}_t| = \widetilde{O}\left(\frac{1}{\epsilon^{\mathcal{D}_s+\mathcal{D}_a+2}}\right)$$

This concludes the proof of Theorem 3. □

C Additional Experimental Details for Benchmark Control Tasks

In this section we provide additional details related to the experimental results presented in this paper.

C.1 Experimental Setup

We consider the case that state transition is stochastic due to system noise arise from model uncertainties. Following the conventional approach we model these parameter uncertainties and external disturbances using a multivariate Gaussian perturbation [43, 44, 45]. For all the control tasks we consider the dynamic equations given in [14]. For all simulations we take 100 HMC samples during the update phase. We use trajectory length $L = 100$ and step size $\delta l = 0.02$. We randomly initialize the Q matrix using values between 0 and 1.

Inverted Pendulum Let θ, $\dot{\theta}$ be the angle of the pendulum, respectively. Then, by letting a denote the input torque applied to the pendulum, its dynamics can be expressed as

$$\ddot{\theta} - \sin\theta + \dot{\theta} - a = 0. \tag{S.9}$$

The state space associated with the pendulum is 2-dimensional ($\mathcal{D}_s = 2$) and any state $s \in \mathcal{S}$ is given by $s = (\theta, \dot{\theta})$. We define the range of state space as $\theta \in [-\pi, \pi]$ and $\dot{\theta} \in [-10, 10]$. We consider action space to be a 1-dimensional ($\mathcal{D}_a = 1$) space such that $a \in [-1, 1]$. We discretize each dimension in state space into 25 values and the action space into 10 values. This forms a Q matrix of dimension 625×10.

Also, we consider the noise co-variance of the Gaussian perturbation to be $\Sigma = \text{diag}[0.868, 1.550]$.

Let $s_t = (\theta_t, \dot{\theta}_t)$ and a_t be the state and the action at time t. Then the state transition probability kernel and corresponding target distribution can be given using (7) and (8), respectively, with mean $\mu(s_t, a_t) = (\theta_t + \dot{\theta}_t\tau, \dot{\theta}_t + \ddot{\theta}_t\tau)$, where τ is the discretizated time interval and $\ddot{\theta}_t$ can be obtained from (S.9) by substituting $\theta_t, \dot{\theta}_t, a_t$, and co-variance $\Sigma(s_t, a_t) = \Sigma$.

As our goal is to stabilize the pendulum to the upright position (i.e. to $\theta = 0$) while minimizing the amount of applied torque, we consider the reward function as follows

$$r(\theta, \dot{\theta}, a) = -0.1a^2 + \exp(\cos\theta - 1).$$

Double Integrator By letting x, $\dot{x}$, and a denote the position, velocity, and input torque, respectively, we can express the system dynamics as

$$\ddot{x} = a. \tag{S.10}$$

State space of the double integrator is 2-dimensional ($\mathcal{D}_s = 2$) and any state $s \in \mathcal{S}$ is given as $s = (x, \dot{x})$. We define the range of state space as $x \in [-3, 3]$ and $\dot{\theta} \in [-3, 3]$. We consider action space to be a 1-dimensional ($\mathcal{D}_a = 1$) space such that $a \in [-1, 1]$. We discretize each dimension in state space into 25 values and action space into 10 values. This forms a Q matrix of dimension 625×10.

Here we consider the noise co-variance of the Gaussian perturbation to be $\Sigma = \text{diag}[0.848, 0.848]$.

Let $s_t = (x_t, \dot{x}_t)$ and a_t be the state and the action at time t. Then the state transition probability kernel and corresponding target distribution can be given using (7) and (8), respectively, with mean $\mu(s_t, a_t) = (x_t + \dot{x}_t\tau, \dot{x}_t + \ddot{x}_t\tau)$, where τ is the discretizated time interval and $\ddot{x}_t$ can be obtained from (S.10) by substituting $x_t, \dot{x}_t, a_t$, and co-variance $\Sigma(s_t, a_t) = \Sigma$.

We define the reward function as the quadratic cost

$$r(x, \dot{x}, a) = -\frac{1}{2}\left(x^2 + \dot{x}^2\right).$$

Cartpole Let θ, $\dot{\theta}$ be the angle and angular velocity of the pole, respectively. Similarly, let x, $\dot{x}$ be the position and linear velocity of the cart, respectively. Then, by letting a denote the control force applied to the cart, the dynamics of cart-pole system [**?**] can be expressed as

$$\begin{aligned} l\left(\frac{4}{3}(m+M) - m\cos^2\theta\right)\ddot{\theta} + (a + ml\dot{\theta}^2\sin\theta)\cos\theta - (m+M)g\sin\theta = 0 \\ (m+M)\ddot{x} - ml\left(\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta\right) - a = 0 \end{aligned} \tag{S.11}$$

where, m, M, l and g represent the mass of the pole, mass of the cart, length of the pole and the gravitational acceleration, respectively.

State space of the cart-pole system is 4-dimensional ($\mathcal{D}_s = 4$) and any state $s \in \mathcal{S}$ is given by $s = (\theta, \dot{\theta}, x, \dot{x})$. We define the range of state space as $\theta \in [-pi/2, \pi/2], \dot{\theta} \in [-3.0, 3.0], x \in [-2.4, 2.4]$ and $\dot{x} \in [-3.5, 3.5]$. We consider action space to be a 1-dimensional ($\mathcal{D}_a = 1$) space such that $a \in [-10, 10]$. We discretize each dimension in state space into 5 values and action space into 10 values. This forms a Q matrix of dimensions 625×10.

Although the differential equations (S.11) governing the dynamics of the pendulum on a cart system are deterministic, uncertainty of the parameters and external disturbances to the system causes the cart pole to deviate from the defined dynamics leading to a stochastic state transition. Here we consider the co-variance of the Gaussian perturbation to be $\Sigma = \text{diag}[0.641, 0.848, 0.759, 0.917]$.

Let $s_t = (\theta_t, \dot{\theta}_t, x_t, \dot{x}_t)$ and a_t be the state and the action at time t. Then the state transition probability kernel and corresponding target distribution can be given using (7) and (8), respectively, with mean $\mu(s_t, a_t) = (\theta_t + \dot{\theta}_t\tau, \dot{\theta}_t + \ddot{\theta}_t\tau, x_t + \dot{x}_t\tau, \dot{x}_t + \ddot{x}_t\tau)$, where τ is the discretizated time interval and $\ddot{\theta}_t, \ddot{x}_t$ can be obtained from (S.11) by substituting $\theta_t, \dot{\theta}_t, a_t$, and co-variance $\Sigma(s_t, a_t) = \Sigma$.

Our simulation results use the following value for the system parameters - $m = 0.1kg$, $M = 1kg$, $l = 0.5m$ and $g = 9.8ms^{-2}$. The goal is stabilizing the pole in upright position. Thus we consider the reward function

$$r(\theta, \dot{\theta}, x, \dot{x}, a) = \cos^4(15\theta)$$

Acrobot Let $\theta_1, \dot{\theta}_1$ be the angle and angular velocity of the first pole, respectively. Similarly, let $\theta_2, \dot{\theta}_2$ be the angle and angular velocity of the first pole, respectively. Then, by letting a denote the control torque applied to the second joint, dynamics of the acrobot can be expressed as

$$\begin{aligned} \ddot{\theta}_2 &= \frac{a + \frac{D_2}{D_1}\phi_1 - m_2 l_1 l_{c2}\dot{\theta}_1^2 \sin\theta_2 - \phi_2}{m_2(l_2^2 + l_{c2}^2) - \frac{D_2^2}{D_1}} \\ \ddot{\theta}_1 &= -\frac{D_2\ddot{\theta}_2 + \phi_1}{D_1}, \end{aligned} \tag{S.12}$$

where,

$$\begin{aligned} D_1 &= m_1(l_1^2 + l_{c1}^2) + m_2(l_1^2 + l_2^2 + l_{c2}^2 + 2l_1 l_{c2}\cos\theta_2) \\ D_2 &= m_2(l_2^2 + l_{c2}^2 + l_1 l_{c2}\cos\theta_2) \\ \phi_2 &= m_2 l_{c2} g\sin(\theta_1 + \theta_2) \\ \phi_1 &= -m_2 l_{c2}\dot{\theta}_2(\dot{\theta}_2 + 2\dot{\theta}_1)\sin\theta_2 + (m_1 l_{c1} + m_2 l_1)g\sin\theta_1 + phi_2, \end{aligned}$$

and m_1, m_2, l_1 l_2 and g represent the mass of the poles, length of the poles, and the gravitational acceleration, respectively. We have used $l_{c1} = l_1/2$ and $l_{c2} = l_2/2$. Moreover, our simulation results use the following value for the system parameters: $m_1 = m_2 = 0.1kg$, $l_1 = l_2 = 0.1m$ and $g = 9.8ms^{-2}$.

State space of the acrobot is 4-dimensional ($\mathcal{D}_s = 4$) and any state $s \in \mathcal{S}$ is given by $s = (\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2)$. We define the range of state space as $\theta_i \in [-\pi, \pi]$ and $\dot{\theta}_i \in [-10.0, 10.0]$, $i = 1, 2$. We consider action space to be a 1-dimensional ($\mathcal{D}_a = 1$) space such that $a \in [-1, 1]$. We discretize each dimension in state space into 5 values and action space into 10 values. This forms a Q matrix of dimensions 625×10.

To incorporate the effects from uncertainty in the parameters and external disturbances to the system, we consider the co-variance of the Gaussian perturbation to the system be $\Sigma = \text{diag}[0.686, 1.550, 0.686, 1.550]$.

Let $s_t = (\theta_{1t}, \dot{\theta_{1t}}, \theta_{2t}, \dot{\theta_{2t}})$ and a_t be the state and the action at time t. Then the state transition probability kernel and corresponding target distribution can be given by (7) and (8), respectively, with mean $\mu(s_t, a_t) = (\theta_{1t} + \dot{\theta}_{1t}\tau, \dot{\theta}_{1t} + \ddot{\theta}_{1t}\tau, \theta_{2t} + \dot{\theta}_{2t}\tau, \dot{\theta}_{2t} + \ddot{\theta}_{2t}\tau,)$, where τ is the discretizated time interval and $\ddot{\theta}_{1t}, \ddot{\theta}_{2t}$ can be obtained from (S.12) by substituting $\theta_t, \dot{\theta}_t, a_t$, and co-variance $\Sigma(s_t, a_t) = \Sigma$.

As the objective is to stabilize the acrobot to the upright position, we define the reward function as

$$r(\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2, a) = \exp(-\cos\theta_1 - 1) + \exp(-\cos(\theta_1 + \theta_2) - 1).$$

C.2 Comparison with Deep RL Algorithms

We provided results combining HMC sampling with benchmark Deep RL algorithms DQN and DDPG. We used the same network architecture of DQN and DDPG presented in the original papers [1, 46]. To train the networks, we used the Adam optimizer [47] with learning rate $1e^{-5}$, discount coefficient $\gamma = 0.99$, and batchsize 32. For all results provided in this paper we used following hyper parameters. Also, we set the number of steps between target network update to 10,000.