Policy Optimization with Demonstrations — Supplementary Material

Anonymous Authors[1]

A. Implementation Details

Our algorithm is implemented with Theano (Theano Development Team, 2016) based on the code provided in (Ho & Ermon, 2016). For all experiments, we use a three-layer fully connected neural network with Tanh nonlinear activation functions as our policy network. The number of hidden units is 64 for MountainCar and CartPole, and 100 for the others. The value network share the same architecture with the policy network and are optimized with Adam (Kingma & Ba, 2014). The other hyperparameters are provided in supplement. We use the DQfD[1] code for experiments on MountainCar and CartPole and implement DDPGfD based on the public DDPG code[2]. The hyperparameter setting in the experiments, is given in Table 1.

Table 1. Hyperparameters

Environment	λ_1	λ_2	Iterations	Batch Size
MountainCar-v1	0.1	0.0	300	5,000
CartPole-v0	0.1	0.0	300	5,000
Hopper-v1	0.1	0.0	500	20,000
HalfCheetah-v1	0.01	0.0	1000	20,000
Walker2d-v1	0.01	0.0	500	20,000
DoublePendulum-v1	0.1	0.0	500	20,000
Humanoid-v1	0.1	0.0	1500	50,000
Reacher-v1	0.01	0.001	500	10,000

B. Proofs

B.1. Proof of Policy Improvement Bound

Lemma 1. *Given two distributions $p(x), q(x)$ over random variable x, we have $D_{TV}^2(p,q) \leq 4D_{JS}(p,q)$.*

Proof. By the definition of JS-divergence, we have

$$\begin{aligned} & D_{JS}(p,q) \\ =& \frac{1}{2}D_{KL}(p, \frac{p+q}{2}) + \frac{1}{2}D_{KL}(q, \frac{p+q}{2}) \\ \geq& \frac{1}{2}D_{TV}^2(p, \frac{p+q}{2}) + \frac{1}{2}D_{TV}^2(q, \frac{p+q}{2}) \\ =& \frac{1}{2}\left(\frac{1}{2}\int\left|\frac{p-q}{2}\right|dx\right)^2 + \frac{1}{2}\left(\frac{1}{2}\int\left|\frac{q-p}{2}\right|dx\right)^2 \\ =& \frac{1}{4}\left(\frac{1}{2}\int|p-q|dx\right)^2 = \frac{1}{4}D_{TV}^2(p,q). \end{aligned} \tag{1}$$

Rearranging, the result follows. □

[1] https://github.com/go2sea/DQfD/

[2] https://github.com/shaneshixiang/rllabplusplus

Definition 1. *(coupled policies (Schulman et al., 2015)) $\pi, \tilde{\pi}$ are α-coupled policies if it defines a joint distribution $(a, \tilde{a})|s$, such that $P(a \neq \tilde{a}|s) \leq a$ for all s. π and $\tilde{\pi}$ will denote the marginal distributions of a and $\tilde{a}$, respectively.*

Theorem 1. *(Proposition 4.7 in (Levin & Peres, 2017)) Suppose p_X and p_Y are distributions with $D_{TV}(p_X, p_Y) = \alpha$, Then there exists a joint distribution (X, Y) whose marginals are p_X,p_Y, for which $X = Y$ with probability $1 - \alpha$.*

Corollary 1. *Combining Definition 1 and Theorem 1, we have two policies π and $\tilde{\pi}$, if $\max_s D_{TV}(\pi(a|s), \tilde{\pi}(a|s)) \leq \alpha$, then $\pi, \tilde{\pi}$ are α-coupled policies.*

Lemma 2. *(Adopted from (Schulman et al., 2015)) Suppose π_1, π_2 are two stochastic policies defined on $\mathcal{S} \times \mathcal{A}$, we have*

$$\eta(\pi_1) = \eta(\pi_2) + \mathbb{E}_{\tau \sim \pi_1} \left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_2}(s, a) \right] \tag{2}$$

where the expectation $\mathbb{E}_{\tau \sim \pi_1}$ indicates that trajectory τ are generated by executing $\pi_1(a|s)$.

Proof. Note that $A_{\pi_2}(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s, a) + \gamma V_{\pi_2}(s') - V_{\pi_2}(s) \right]$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\tau \sim \pi_1} \left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_2}(s, a) \right] \\
=&\mathbb{E}_{\tau \sim \pi_1} \left[\sum_{t=0}^{\infty} \gamma^t (r(s, a) + \gamma V_{\pi_2}(s') - V_{\pi_2}(s)) \right] \\
=&\mathbb{E}_{\tau \sim \pi_1} \left[-V_{\pi_2} + \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \\
=&- \mathbb{E}_{s_0} \left[V_{\pi_2}(s_0) \right] + \mathbb{E}_{\tau \sim \pi_1} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \\
=&- \eta(\pi_2) + \eta(\pi_1)
\end{aligned} \tag{3}$$

Rearranging, the result follows. □

Given two arbitrary stochastic policies π_1, π_2, define the expected advantage of π_1 over π_2 at state s as $A^{\pi_1|\pi_2}(s)$:

$$A^{\pi_1|\pi_2}(s) = \mathbb{E}_{a \sim \pi_1(\cdot|s)} \left[A_{\pi_2}(s, a) \right] \tag{4}$$

Lemma 3. *(Adopted from (Schulman et al., 2015)) Given that π_1, π_2 are α-coupled policies, for all s,*

$$\left| A^{\pi_1|\pi_2}(s) \right| \leq 2\alpha \max_a |A_{\pi_2}(s, a)| \tag{5}$$

Proof.

$$\begin{aligned}
\left| A^{\pi_1|\pi_2}(s) \right| &= \left| \mathbb{E}_{a \sim \pi_1(\cdot|s)} \left[A_{\pi_2}(s, a) \right] \right| \\
&= \left| \mathbb{E}_{(a_1, a_2) \sim (\pi_1, \pi_2)} \left[A_{\pi_2}(s, a_1) - A_{\pi_2}(s, a_2) \right] \right| \\
&= \left| P(a_1 \neq a_2|s) \mathbb{E}_{(a_1, a_2) \sim (\pi_1, \pi_2)|a_1 \neq a_2} \left[A_{\pi_2}(s, a_1) - A_{\pi_2}(s, a_2) \right] \right| \\
&\leq \alpha \cdot 2 \max_a |A_{\pi_2}(s, a)|
\end{aligned} \tag{6}$$

□

Lemma 4. *Given that π_1, π_2 are α-coupled policies, for all s, Then*

$$\left| \mathbb{E}_{s_t \sim \pi_1} \left[A^{\pi_1|\pi_2}(s) \right] \right| \leq 2\alpha(1 - (1 - \alpha)^t) \cdot \max_{s,a} |A_{\pi_2}(s, a)| \tag{7}$$

Proof. Consider generating a trajectory using π_1. Specifically, at each time step i we sample $(a_1^i, a_2^i)|s_t$ following π_2 and π_2 respectively, then a_1^i is executed to generated the trajectory, while a_2^i is ignored. Let n_t denote the number of times that $a_1^i \neq a_2^i$ for $i \leq t$, *i.e.*, the number of times π_1 and π_2 disagree before time step t.

$$\begin{aligned}\mathbb{E}_{s_t\sim\pi_1}\left[A^{\pi_1|\pi_2}(s)\right] =& P(n_t=0)\mathbb{E}_{s_t\sim\pi_1|n_t=0}\left[A^{\pi_1|\pi_2}(s)\right] + P(n_t\neq 0)\mathbb{E}_{s_t\sim\pi_1|n_t\neq 0}\left[A^{\pi_1|\pi_2}(s)\right]\\ =& P(n_t=0)\mathbb{E}_{s_t\sim\pi_2}\left[A^{\pi_1|\pi_2}(s)\right] + P(n_t\neq 0)\mathbb{E}_{s_t\sim\pi_1|n_t\neq 0}\left[A^{\pi_1|\pi_2}(s)\right]\\ =& P(n_t\neq 0)\mathbb{E}_{s_t\sim\pi_1|n_t\neq 0}\left[A^{\pi_1|\pi_2}(s)\right]\end{aligned} \tag{8}$$

Since π_1, π_2 are α-coupled, $P(n_t = 0) = (1-\alpha)^t$, thus $P(n_t \neq 0) = 1-(1-\alpha)^t$, and

$$\begin{aligned}\left|\mathbb{E}_{s_t\sim\pi_1}\left[A^{\pi_1|\pi_2}(s)\right]\right| =& (1-(1-\alpha)^t)\left|\mathbb{E}_{s_t\sim\pi_1|n_t\neq 0}\left[A^{\pi_1|\pi_2}(s)\right]\right|\\ \leq& (1-(1-\alpha)^t)\max_s\left|A^{\pi_1|\pi_2}(s)\right|\\ \leq& (1-(1-\alpha)^t)\cdot 2\alpha\max_{s,a}|A_{\pi_2}(s,a)| \qquad \text{(from Lemma 3)}\end{aligned} \tag{9}$$

□

Lemma 5. *Given three arbitrary stochastic policies $\pi, \tilde{\pi}, \pi_E$, the following equation hold.*

$$\eta(\tilde{\pi}) - \eta(\pi_E) = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi_E}(s_t)\right] = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi}(s_t)\right] - \mathbb{E}_{\pi_E}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_E|\pi}(s_t)\right] \tag{10}$$

Proof. Applying Lemma 2 to policy $\tilde{\pi}$ and π_E, we have

$$\eta(\tilde{\pi}) = \eta(\pi_E) + \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi_E}(s_t)\right] \tag{11}$$

Rearranging, the first equality holds. Similarly, writing $\eta(\pi_E)$ and $\eta(\tilde{\pi})$ in terms of π respectively gives

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi}(s_t)\right], \quad \eta(\pi_E) = \eta(\pi) + \mathbb{E}_{\tau\sim\pi_E}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_E|\pi}(s_t)\right] \tag{12}$$

Subtracting $\eta(\tilde{\pi})$ by $\eta(\pi_E)$ in Eqn.(12) gives

$$\eta(\tilde{\pi}) - \eta(\pi_E) = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi}(s_t)\right] - \mathbb{E}_{\pi_E}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_E|\pi}(s_t)\right] \tag{13}$$

Thus the second equality holds. □

Now we are ready to derive the bound given in Theorem 1. Since $\alpha = D_{\text{TV}}^{\max}(\pi, \tilde{\pi})$, and $\beta = D_{\text{TV}}^{\max}(\pi_E, \tilde{\pi})$, thus we know $\pi, \tilde{\pi}$ are α-coupled and $\pi_E, \tilde{\pi}$ are β-coupled through Corollary 1. Furthermore, we suppose expert policyπ_E satisfy Assumption 1, which means

$$\mathbb{E}_{a_E\sim\pi_E(\cdot|s)}[A_\pi(s, a_E)] \geq \delta \tag{14}$$

Applying Lemma 2 to π and $\tilde{\pi}$, then substituting Eqn. (4) into $\tilde{\pi}$ gives

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi}(s_t)\right], \quad J_\pi(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}|\pi}(s_t)\right] \tag{15}$$

Then, subtracting $\eta(\tilde{\pi})$ by $J_\pi(\tilde{\pi})$ gives

$$\eta(\tilde{\pi}) - J_\pi(\tilde{\pi}) \tag{16}$$

$$= \sum_{t=0}^{\infty} \gamma^t \left(\mathbb{E}_{\tilde{\pi}} \left[A^{\tilde{\pi}|\pi}(s_t) \right] - \mathbb{E}_{\pi} \left[A^{\tilde{\pi}|\pi}(s_t) \right] \right) \tag{17}$$

$$= \sum_{t=0}^{\infty} \gamma^t \left(\mathbb{E}_{\tilde{\pi}} \left[A^{\tilde{\pi}|\pi}(s_t) \right] - \mathbb{E}_{\pi_E} \left[A^{\pi_E|\pi}(s_t) \right] + \mathbb{E}_{\pi_E} \left[A^{\pi_E|\pi}(s_t) \right] - \mathbb{E}_{\pi} \left[A^{\tilde{\pi}|\pi}(s_t) \right] \right) \tag{18}$$

$$= \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tilde{\pi}} \left[A^{\tilde{\pi}|\pi}(s_t) \right] - \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_E} \left[A^{\pi_E|\pi}(s_t) \right] \right) + \sum_{t=0}^{\infty} \gamma^t \left(\mathbb{E}_{\pi_E} \left[A^{\pi_E|\pi}(s_t) \right] - \mathbb{E}_{\pi} \left[A^{\tilde{\pi}|\pi}(s_t) \right] \right) \tag{19}$$

$$= \underbrace{\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_E} \left[A^{\tilde{\pi}|\pi_E}(s_t) \right]}_{\text{by Lemma 5}} + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_E} \left[A^{\pi_E|\pi}(s_t) \right] - \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi} \left[A^{\tilde{\pi}|\pi}(s_t) \right] \tag{20}$$

$$\geq - \sum_{t=0}^{\infty} \gamma^t \left| \mathbb{E}_{\pi_E} \left[A^{\tilde{\pi}|\pi_E}(s_t) \right] \right| - \sum_{t=0}^{\infty} \gamma^t \left| \mathbb{E}_{\pi} \left[A^{\tilde{\pi}|\pi}(s_t) \right] \right| + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_E} \left[A^{\pi_E|\pi}(s_t) \right] \tag{21}$$

$$\geq - \sum_{t=0}^{\infty} \gamma^t \left(\underbrace{2\beta(1-(1-\beta)^t) \cdot \max_{s,a} |A_{\pi_E}(s,a)|}_{\text{by Lemma 4}} + \underbrace{2\alpha(1-(1-\alpha)^t) \cdot \max_{s,a} |A_{\pi}(s,a)|}_{\text{by Lemma 4}} - \underbrace{\delta}_{\text{by Eqn. (14)}} \right) \tag{22}$$

$$= - \frac{2\beta^2\gamma\epsilon_E}{(1-\gamma)(1-\gamma(1-\beta))} - \frac{2\alpha^2\gamma\epsilon_\pi}{(1-\gamma)(1-\gamma(1-\alpha))} + \frac{\delta}{1-\gamma} \tag{23}$$

$$\geq - \frac{2\gamma(\beta^2\epsilon_E + \alpha^2\epsilon_\pi)}{(1-\gamma)^2} + \frac{\delta}{1-\gamma} \tag{24}$$

where $\epsilon_E = \max_{s,a} |A_{\pi_E}(s,a)|, \epsilon_\pi = \max_{s,a} |A_\pi(s,a)|$. Applying $D_{TV}^2(p,q) \leq D_{KL}(p,q)$ (Pollard, 2013) and $D_{TV}^2(p,q) \leq 4 D_{JS}(p,q)$ (Lemma 1), Theorem 1 is proved.

B.2. Proof of Theorem 2

Proof. Let $f(v) = v\log(v) - (v+1)\log(v+1)$, f^* be the conjugate function given by $f^*(t) = \sup_{v \in \text{dom}_f} \{vt - f(v)\}$. It is obvious that $f(v)$ is a continuous function on $(0, +\infty)$, the second derivative of $f(v)$ is given by $f'' = \frac{1}{v(v+1)} \geq 0$,which means $f(v)$ is a convex function. Therefore, we can rewrite f in terms of its convex conjugate function f^* as $f(v) = f^{**}(v) = \sup_{t \in \text{dom}_{f^*}} \{tv - f^*(t)\}$. Substituting f^{**} into D_{JS}, we have

$$\begin{aligned}
& D_{JS}(\rho_\pi, \rho_E) \\
\triangleq & \int_{\mathcal{S}\times\mathcal{A}} \rho_\pi \log \frac{2\rho_\pi}{\rho_\pi + \rho_E} + \rho_E \log \frac{2\rho_E}{\rho_\pi + \rho_E} dsda \\
= & \int_{\mathcal{S}\times\mathcal{A}} \rho_\pi \log \frac{\rho_\pi}{\rho_\pi + \rho_E} + \rho_E \log \frac{\rho_E}{\rho_\pi + \rho_E} dsda + \log 4 \\
= & \int_{\mathcal{S}\times\mathcal{A}} \rho_E f(\frac{\rho_\pi}{\rho_E}) dsda + \log 4 \\
= & \int_{\mathcal{S}\times\mathcal{A}} \rho_E \sup_{t \in \text{dom}_{f^*}} \left\{ t\frac{\rho_\pi}{\rho_E} - f^*(t) \right\} dsda + \log 4 \\
\geq & \sup_{T \in \mathcal{T}} \left(\int_{\mathcal{S}\times\mathcal{A}} \rho_\pi T(s,a) - \rho_E f^*(T(s,a)) dsda \right) + \log 4 \\
= & \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{(s,a)\sim\rho_\pi}[T(s,a)] + \mathbb{E}_{(s,a)\sim\rho_E}[-f^*(T(s,a))] \right) + \log 4.
\end{aligned}$$

where the inequality line is given by the Jensen's inequality and replacing t by $T(s,a)$, $\mathcal{T} = \{T(s,a) : \mathcal{S}\times\mathcal{A} \to \text{dom}_{f^*}\}$, *i.e.*, function $T(s,a)$ is valid if and only if $\text{range}_T = \text{dom}_{f^*}$.

Now we show that $T(s,a)$ can be formed with $h(U(s,a))$. We first derive the specific form and domain of f^*.

$$\begin{aligned} f^*(t) &= \sup_{v \in \text{dom}_f} \{vt - f(v)\} \\ &= \sup_{v \in \text{dom}_f} \{vt - v\log(v) + (v+1)\log(v+1)\}. \end{aligned}$$

Let $g(v) = vt - v\log(v) + (v+1)\log(v+1)$, then the supremum of $g(v)$ is achieved when $g' = 0$, which gives $t = \log\frac{v}{v+1} \in (-\infty, 0)$. Substituting this into f^* results in $f^* = \log\frac{1}{1-e^t}$. On the other hand, $U(s,a) \in \mathbb{R}$, $h(u) \in (-\infty, 0)$, thus $\text{range}_T = \text{dom}_{f^*}$.

Then we show that $-f^*(T(s,a)) = \bar{h}(U(s,a))$.

$$\begin{aligned} -f^*(T(s,a)) &= -f^*(h(U(s,a))) \\ &= \log(1 - e^{-\log(1+e^{-U(s,a)})}) \\ &= \log\frac{e^{-U(s,a)}}{1+e^{-U(s,a)}} = \bar{h}(U(s,a)). \end{aligned}$$

□

References

Ho, Jonathan and Ermon, Stefano. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

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

Levin, David A and Peres, Yuval. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Pollard, D. Asymptopia: an exposition of statistical asymptotic theory, 2013.

Schulman, John, Levine, Sergey, Abbeel, Pieter, Jordan, Michael, and Moritz, Philipp. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1889–1897, 2015.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688.