Monotonic Robust Policy Optimization with Model Discrepancy

A. Appendix

A.1. Proof of Lemma 1

Proof. First, we define $\eta(\pi|p_w) - \max_{p\in\mathcal{P}} \eta(\pi|p) \triangleq -C_0$, where $C_0 \geq 0$ depends on π and $\mathcal{P}$. Then, given a policy π and any environment $p \in \mathcal{P}$, and for any non-negative constant $C \geq C_0 \geq 0$, we thus have:

$$\eta(\pi|p_w) - \mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\pi|p)\right] \geq \eta(\pi|p_w) - \max_{p\in\mathcal{P}} \eta(\pi|p) = -C_0 \geq -C. \tag{A.1}$$

Therefore, the second inequality $\eta(\pi|p_w) \geq \mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\pi|p)\right] - C$ is proved, while the first inequality $\mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\pi|p)\right] \geq \eta(\pi|p_w)$ always holds according to the definition of the worst-case environment p_w. □

A.2. Proof of Theorem 1

Lemma A.1. *For any two joint distribution $P_1(\mathrm{x},\mathrm{y}) = P_1(\mathrm{x})P_1(\mathrm{y}|\mathrm{x})$ and $P_2(\mathrm{x},\mathrm{y}) = P_2(\mathrm{x})P_2(\mathrm{y}|\mathrm{x})$ over* x *and* y*, we can bound the total variation distance of them by:*

$$D_{TV}(P_1(\mathrm{x},\mathrm{y})\|P_2(\mathrm{x},\mathrm{y})) \leq D_{TV}(P_1(\mathrm{x})|P_2(\mathrm{x})) + \max_x D_{TV}(P_1(\mathrm{y}|x)\|P_2(\mathrm{y}|x)) \tag{A.2}$$

Proof.

$$\begin{aligned}
D_{TV}(P_1(\mathrm{x},\mathrm{y})\|P_2(\mathrm{x},\mathrm{y})) =& \frac{1}{2}\sum_{x,y}|P_1(x,y) - P_2(x,y)| && \text{(A.3)}\\
=& \frac{1}{2}\sum_{x,y}|P_1(x)P_1(y|x) - P_2(x)P_2(y|x)| && \text{(A.4)}\\
=& \frac{1}{2}\sum_{x,y}|P_1(x)P_1(y|x) - P_1(x)P_2(y|x) + P_1(x)P_2(y|x) - P_2(x)P_2(y|x)| && \text{(A.5)}\\
\leq& \frac{1}{2}\sum_{x,y}P_1(x)|P_1(y|x) - P_2(y|x)| + \frac{1}{2}\sum_x|P_1(x) - P_2(x)| && \text{(A.6)}\\
=& \mathbb{E}_{x\sim P_1} D_{TV}(P_1(\mathrm{y}|x)\|P_2(\mathrm{y}|x)) + D_{TV}(P_1(\mathrm{x})|P_2(\mathrm{x})). && \text{(A.7)}
\end{aligned}$$

□

Lemma A.2. *Suppose the initial state distributions $P_1^0(\mathrm{s})$ and $P_2^0(\mathrm{s})$ are the same. Then the distance in the state marginal at time step t is bounded as:*

$$D_{TV}(P_1^t(\mathrm{s})\|P_2^t(\mathrm{s})) \leq t\max_t \mathbb{E}_{s'\sim P_1^t} D_{TV}(P_1(\mathrm{s}|s')\|P_2(\mathrm{s}|s')). \tag{A.8}$$

Proof. We can first prove the following inequality:

$$|P_1^t(s) - P_2^t(s)| = \left| \sum_{s'} P_1(\mathrm{s}_t = s|s') P_1^{t-1}(s') - \sum_{s'} P_2(\mathrm{s}_t = s|s') P_2^{t-1}(s') \right| \tag{A.9}$$

$$\leq \sum_{s'} |P_1(\mathrm{s}_t = s|s') P_1^{t-1}(s') - P_2(\mathrm{s}_t = s|s') P_2^{t-1}(s')| \tag{A.10}$$

$$= \sum_{s'} |P_1(\mathrm{s}_t = s|s') P_1^{t-1}(s') - P_2(\mathrm{s}_t = s|s') P_1^{t-1}(s') \tag{A.11}$$

$$+ P_2(\mathrm{s}_t = s|s') P_1^{t-1}(s') - P_2(\mathrm{s}_t = s|s') P_2^{t-1}(s')| \tag{A.12}$$

$$\leq \mathbb{E}_{s' \sim P_1^{t-1}} |P_1(s|s') - P_2(s|s')| + \sum_{s'} P_2(s|s') |P_1^{t-1}(s') - P_2^{t-1}(s')|. \tag{A.13}$$

Based on (A.13), we have:

$$D_{TV}(P_1^t(\mathrm{s}) \| P_2^t(\mathrm{s})) \leq \frac{1}{2} \sum_s |P_1^t(s) - P_2^t(s)| \tag{A.14}$$

$$\leq \frac{1}{2} \sum_s \left(\mathbb{E}_{s' \sim P_1^{t-1}} |P_1(s|s') - P_2(s|s')| + \sum_{s'} P_2(s|s') |P_1^{t-1}(s') - P_2^{t-1}(s')| \right) \tag{A.15}$$

$$= \mathbb{E}_{s' \sim P_1^{t-1}} D_{TV}(P_1(\mathrm{s}|s') \| P_2(\mathrm{s}|s')) + D_{TV}(P_1^{t-1}(\mathrm{s}') \| P_2^{t-1}(\mathrm{s}')) \tag{A.16}$$

$$\leq \sum_{i=1}^{t} \mathbb{E}_{s' \sim P_1^{i-1}} D_{TV}(P_1(\mathrm{s}|s') \| P_2(\mathrm{s}|s')) \tag{A.17}$$

$$\leq t \max_t \mathbb{E}_{s' \sim P_1^t} D_{TV}(P_1(\mathrm{s}|s') \| P_2(\mathrm{s}|s')), \tag{A.18}$$

where (A.17) is obtained by recursively applying (A.14)-(A.16) to the second term $D_{TV}(P_1^{t-1}(\mathrm{s}') \| P_2^{t-1}(\mathrm{s}'))$ in (A.16). □

Theorem A.1. *(Theorem 1 in the main text.) In MDPs where the reward function is bounded, and for any distribution P over $\mathcal{P}$, by updating the current policy π to a new policy $\tilde{\pi}$, the following bound holds:*

$$\eta(\tilde{\pi}|p_w) - \mathbb{E}_{\mathrm{p} \sim P} [\eta(\tilde{\pi}|p)] \geq -2|r|_{\max} \frac{\gamma \mathbb{E}_{\mathrm{p} \sim P} [\epsilon(p_w \| p)]}{(1-\gamma)^2} - \frac{4|r|_{\max} d(\pi, \tilde{\pi})}{(1-\gamma)^2}, \tag{A.19}$$

where p_w denotes the environment that corresponds to the worst-case performance under policy π, and we define $\epsilon(p_w \| p) \triangleq \max_t \mathbb{E}_{s' \sim P_\pi^t(\cdot|p_w)} \mathbb{E}_{\mathrm{a} \sim \pi(\cdot|\mathrm{s}')} D_{TV}(\mathcal{T}(\mathrm{s}|\mathrm{s}', \mathrm{a}, p_w) \| \mathcal{T}(\mathrm{s}|\mathrm{s}', \mathrm{a}, p))$, $d(\pi, \tilde{\pi}) \triangleq \max_t \mathbb{E}_{s' \sim P_\pi^t(\cdot|p_w)} D_{TV}(\pi(\mathrm{a}|\mathrm{s}') \| \tilde{\pi}(\mathrm{a}|\mathrm{s}'))$.

Proof. We can rewrite the LHS of (A.19) as:

$$\eta(\tilde{\pi}|p_w) - \mathbb{E}_{\mathrm{p} \sim P} [\eta(\tilde{\pi}|p)] = \eta(\tilde{\pi}|p_w) - \eta(\pi|p_w) + \eta(\pi|p_w) - \mathbb{E}_{\mathrm{p} \sim P} [\eta(\tilde{\pi}|p)] .$$

For the last two terms, we have:

$$\mathbb{E}_{\mathrm{p} \sim P} |\eta(\pi|p_w) - \eta(\tilde{\pi}|p)| = \mathbb{E}_{\mathrm{p} \sim P} \left| \sum_t \gamma^t \sum_{s,a} (P^t(s,a|p_w) - P^t(s,a|p)) R(s,a) \right|$$

$$\leq \mathbb{E}_{\mathrm{p} \sim P} \sum_t \gamma^t \sum_{s,a} |P^t(s,a|p_w) - P^t(s,a|p)| \cdot |R(s,a)|$$

$$\leq 2|r|_{\max} \sum_t \gamma^t \mathbb{E}_{\mathrm{p} \sim P} \left[D_{TV}(P^t(\mathrm{s}, \mathrm{a}|p_w) \| P^t(\mathrm{s}, \mathrm{a}|p)) \right], \tag{A.20}$$

where $|r|_{\max}$ denotes the upper bound of the absolute value of reward function $|R(s,a)|$, and $P^t(\mathrm{s},\mathrm{a}|p_w)=\pi(\mathrm{a}|\mathrm{s})P^t_\pi(\mathrm{s}|p_w)$ and $P^t(\mathrm{s},\mathrm{a}|p)=\tilde{\pi}(\mathrm{a}|\mathrm{s})P^t_{\tilde{\pi}}(\mathrm{s}|p)$. Further referring to Lemma A.1, we have:

$$\begin{aligned}&\mathbb{E}_{\mathrm{p}\sim P}\left[D_{TV}(P^t(\mathrm{s},\mathrm{a}|p_w)|P^t(\mathrm{s},\mathrm{a}|p))\right]\\ \leq&\mathbb{E}_{\mathrm{s}\sim P^t_\pi(\cdot|p_w)}D_{TV}(\pi(\mathrm{a}|\mathrm{s})\|\tilde{\pi}(\mathrm{a}|\mathrm{s}))+\mathbb{E}_{\mathrm{p}\sim P}\left[D_{TV}(P^t_\pi(\mathrm{s}|p_w)\|P^t_{\tilde{\pi}}(\mathrm{s}|p))\right].\end{aligned}\tag{A.21}$$

Note that

$$P(\mathrm{s}|\mathrm{s}',p_w)=\sum_a\mathcal{T}(\mathrm{s}|\mathrm{s}',a,p_w)\pi(a|\mathrm{s}'),\tag{A.22}$$

$$P(\mathrm{s}|\mathrm{s}',\mathrm{p})=\sum_a\mathcal{T}(\mathrm{s}|\mathrm{s}',a,\mathrm{p})\tilde{\pi}(a|\mathrm{s}').\tag{A.23}$$

Similar to Lemma A.1, we have:

$$D_{TV}(P(\mathrm{s}|\mathrm{s}',p_w)\|P(\mathrm{s}|\mathrm{s}',\mathrm{p}))\tag{A.24}$$

$$=\frac{1}{2}\sum_s\sum_a|\mathcal{T}(s|\mathrm{s}',a,p_w)\pi(\mathrm{a}|\mathrm{s}')-\mathcal{T}(s|\mathrm{s}',a,\mathrm{p})\tilde{\pi}(a|\mathrm{s}')|\tag{A.25}$$

$$\leq\frac{1}{2}\sum_s\sum_a|\mathcal{T}(s|\mathrm{s}',a,p_w)-\mathcal{T}(s|\mathrm{s}',a,\mathrm{p})|\pi(a|\mathrm{s}')+\frac{1}{2}\sum_s\sum_a\mathcal{T}(s|\mathrm{s}',a,\mathrm{p})|\pi(a|\mathrm{s}')-\tilde{\pi}(a|\mathrm{s}')|\tag{A.26}$$

$$=\mathbb{E}_{\mathrm{a}\sim\pi(\cdot|\mathrm{s}')}D_{TV}(\mathcal{T}(\mathrm{s}|\mathrm{s}',a,p_w)\|\mathcal{T}(\mathrm{s}|\mathrm{s}',a,\mathrm{p}))+D_{TV}(\pi(\mathrm{a}|\mathrm{s}')\|\tilde{\pi}(\mathrm{a}|\mathrm{s}')).\tag{A.27}$$

Referring to Lemma A.2, we have:

$$\begin{aligned}&\mathbb{E}_{\mathrm{p}\sim P}\left[D_{TV}(P^t_\pi(\mathrm{s}|p_w)\|P^t_{\tilde{\pi}}(\mathrm{s}|p))\right]\\ \leq&t\mathbb{E}_{\mathrm{p}\sim P}\max_t\mathbb{E}_{\mathrm{s}'\sim P^t_\pi(\cdot|p_w)}D_{TV}(P(\mathrm{s}|s',p_w)\|P(\mathrm{s}|s',p))\\ \leq&t\mathbb{E}_{\mathrm{p}\sim P}\max_t\mathbb{E}_{\mathrm{s}'\sim P^t_\pi(\cdot|p_w)}\mathbb{E}_{\mathrm{a}\sim\pi(\cdot|\mathrm{s}')}D_{TV}(\mathcal{T}(\mathrm{s}|s',a,p_w)\|\mathcal{T}(\mathrm{s}|s',a,p))\\ &+t\max_t\mathbb{E}_{\mathrm{s}'\sim P^t_\pi(\cdot|p_w)}D_{TV}(\pi(\mathrm{a}|s')\|\tilde{\pi}(\mathrm{a}|s')).\end{aligned}\tag{A.28}$$

Since $\epsilon(p\|p_w)=\max_t\mathbb{E}_{s'\sim P^t_\pi(\cdot|p_w)}\mathbb{E}_{\mathrm{a}\sim\pi(\cdot|\mathrm{s}')}D_{TV}(\mathcal{T}(\mathrm{s}|\mathrm{s}',\mathrm{a},p_w)\|\mathcal{T}(\mathrm{s}|\mathrm{s}',\mathrm{a},p))$ and $d(\pi,\tilde{\pi})=\max_t\mathbb{E}_{s'\sim P^t_\pi(\cdot|p_w)}D_{TV}(\pi(\mathrm{a}|\mathrm{s}')\|\tilde{\pi}(\mathrm{a}|\mathrm{s}'))$, combining (A.20), (A.21) and (A.28), and referring to Jensen's inequality, we have:

$$\begin{aligned}&|\eta(\pi|p_w)-\mathbb{E}_{\mathrm{p}\sim P}\eta(\tilde{\pi}|p)|\\ \leq&\mathbb{E}_{\mathrm{p}\sim P}|\eta(\pi|p_w)-\eta(\tilde{\pi}|p)|\\ \leq&2|r|_{\max}\sum_t\gamma^t\mathbb{E}_{\mathrm{p}\sim P}\Big[(t+1)\max_t\mathbb{E}_{\mathrm{s}'\sim P^t_\pi(\cdot|p_w)}D_{TV}(\pi(\mathrm{a}|s')\|\tilde{\pi}(\mathrm{a}|s'))\\ &+t\max_t\mathbb{E}_{\mathrm{s}'\sim P^t_\pi(\cdot|p_w)}\mathbb{E}_{\mathrm{a}\sim\pi(\cdot|\mathrm{s}')}D_{TV}(\mathcal{T}(\mathrm{s}|s',a,p_w)\|\mathcal{T}(\mathrm{s}|s',a,p))\Big]\\ =&2|r|_{\max}\sum_t\gamma^t\left[t(\mathbb{E}_{\mathrm{p}\sim P}\left[\epsilon(p_w\|p)\right]+d(\pi,\tilde{\pi}))+d(\pi,\tilde{\pi})\right]\\ =&2|r|_{\max}\left[\frac{\gamma\mathbb{E}_{\mathrm{p}\sim P}\left[\epsilon(p_w\|p)\right]}{(1-\gamma)^2}+\frac{d(\pi,\tilde{\pi})}{(1-\gamma)^2}\right].\end{aligned}\tag{A.29}$$

With policy π be updated to $\tilde{\pi}$, $\eta(\pi|p_w)\leq\mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\tilde{\pi}|p)\right]$. Then, we have:

$$\eta(\pi|p_w)-\mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\tilde{\pi}|p)\right]\geq-2|r|_{\max}\left[\frac{\gamma\mathbb{E}_{\mathrm{p}\sim P}\left[\epsilon(p_w\|p)\right]}{(1-\gamma)^2}+\frac{d(\pi,\tilde{\pi})}{(1-\gamma)^2}\right].\tag{A.30}$$

Similar to the derivation of (A.29) and referring to Janner et al. (2019), we have:

$$\eta(\tilde{\pi}|p_w)-\eta(\pi|p_w)\geq-\frac{2|r|_{\max}d(\pi,\tilde{\pi})}{(1-\gamma)^2}.\tag{A.31}$$

Combining the above results, we end up with the proof, as follows:

$$\eta(\tilde{\pi}|p_w) - \mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\tilde{\pi}|p)\right] \geq -2|r|_{\max}\frac{\gamma\mathbb{E}_{\mathrm{p}\sim P}\left[\epsilon(p_w\|p)\right]}{(1-\gamma)^2} - \frac{4|r|_{\max}d(\pi,\tilde{\pi})}{(1-\gamma)^2}. \tag{A.32}$$

□

A.3. Derivation of Policy Optimization Step

In the policy optimization step, we aim to solve the following optimization problem:

$$\max_{\tilde{\pi}} \mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\tilde{\pi}|p)\right] \quad \text{s.t.} \quad d(\pi,\tilde{\pi}) \leq \delta_1. \tag{A.33}$$

Referring to (1), we have:

$$\mathbb{E}_{\mathrm{p}\sim P}\left[\eta(\tilde{\pi}|p)\right] \geq \mathbb{E}_{\mathrm{p}\sim P}\left[L_\pi(\tilde{\pi}|p)\right] - \frac{2\lambda\gamma}{(1-\gamma)^2}\beta^2. \tag{A.34}$$

We now turn to optimize the RHS of (A.34) to maximize the objective in (6) (or (A.33)) under the constraint $d(\pi,\tilde{\pi}) \leq \delta_1$:

$$\max_{\tilde{\pi}} \mathbb{E}_{\mathrm{p}\sim P}\left[L_\pi(\tilde{\pi}|p)\right] - \frac{2\lambda\gamma}{(1-\gamma)^2}\beta^2 \quad \text{s.t.} \quad d(\pi,\tilde{\pi}) \leq \delta_1. \tag{A.35}$$

Note that we have:

$$d(\pi,\tilde{\pi}) = \max_t \mathbb{E}_{s'\sim P_\pi^t(\cdot|p_w)} D_{TV}(\pi(\mathrm{a}|\mathrm{s}')\|\tilde{\pi}(\mathrm{a}|\mathrm{s}')) \leq \max_{s'} D_{TV}(\pi(\mathrm{a}|s')\|\tilde{\pi}(\mathrm{a}|s')) = \beta. \tag{A.36}$$

Following the approximation in Schulman et al. (2015), (A.35) can be equivalently transformed to:

$$\max_{\tilde{\pi}} \mathbb{E}_{\mathrm{p}\sim P}\left[\mathbb{E}_{s\sim P_\pi(\cdot|p),a\sim\pi(\cdot|s)}\left[\frac{\tilde{\pi}(a|s)}{\pi(a|s)}A_\pi(s,a)\right]\right] \quad \text{s.t.} \quad \beta \leq \delta, \tag{A.37}$$

which can be solved by using the PPO (Schulman et al., 2017).

A.4. Proof of Theorem 2

Proof. Denote $H(\pi^k\|\pi^{k+1}) \triangleq \max_t \mathbb{E}_{s'\sim P_\pi^t(\cdot|p_w)} D_{TV}(\pi^k(\mathrm{a}|\mathrm{s}')\|\pi^{k+1}(\mathrm{a}|\mathrm{s}'))$. Updating π_k to π_{k+1} at each iteration k and following Theorem 1, we have

$$\eta(\pi_{k+1}|p_w^k) \geq \mathbb{E}_{\mathrm{p}\sim P^{k+1}}\left[\eta(\pi_{k+1}|p) - \frac{2|r|_{\max}\gamma\epsilon(p\|p_w^k)}{(1-\gamma)^2}\right] - \frac{4|r|_{max}H(\pi_k\|\pi_{k+1})}{(1-\gamma)^2}. \tag{A.38}$$

Since P^{k+1} and π_{k+1} are obtained by maximizing the RHS of (3), we have

$$\mathbb{E}_{\mathrm{p}\sim P^{k+1}}\left[\eta(\pi_{k+1}|p) - \frac{2|r|_{\max}\gamma\epsilon(p\|p_w^k)}{(1-\gamma)^2}\right] - \frac{4|r|_{max}H(\pi_k\|\pi_{k+1})}{(1-\gamma)^2}. \tag{A.39}$$

$$\geq \mathbb{E}_{\mathrm{p}\sim P^{k+1}}\left[\eta(\pi_k|p) - \frac{2r_{max}\gamma\epsilon(p\|p_w^k)}{(1-\gamma)^2}\right] - \frac{4|r|_{\max}H(\pi_k\|\pi_k)}{(1-\gamma)^2} \tag{A.40}$$

$$= \mathbb{E}_{\mathrm{p}\sim P^{k+1}}\left[\eta(\pi_k|p) - \frac{2r_{max}\gamma\epsilon(p\|p_w^k)}{(1-\gamma)^2}\right] \tag{A.41}$$

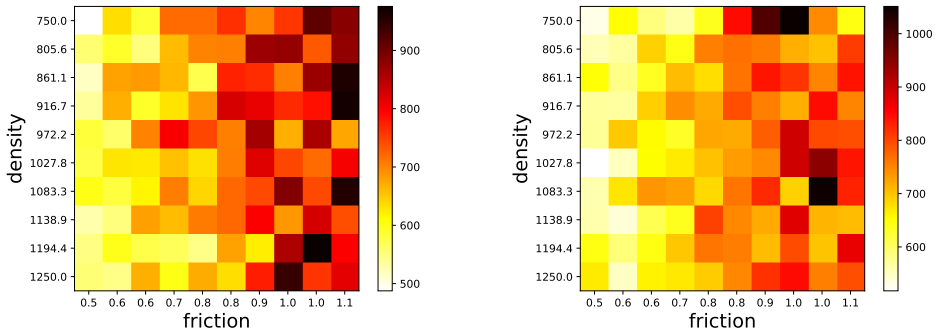
From Line 5 in Algorithm 1, the environment selected for training satisfies:

$$\eta(\pi_k|p) - \frac{2|r|_{\max}\gamma\epsilon(p\|p_w^k)}{(1-\gamma)^2} \geq \eta(\pi_k|p_w^k) - \frac{2|r|_{\max}\gamma\epsilon(p_w^k\|p_w^k)}{(1-\gamma)^2} = \eta(\pi_k|p_w^k). \tag{A.42}$$

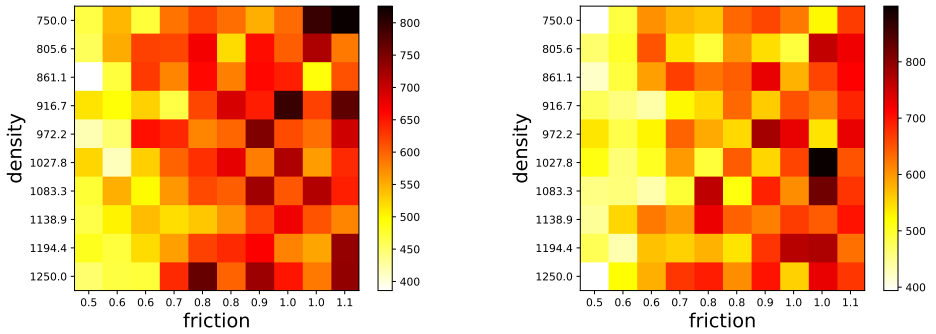
Therefore, combining (A.38)-(A.42), we have:

$$\eta(\pi_{k+1}|p_w^{k+1}) \approx \eta(\pi_{k+1}|p_w^k) \geq \mathbb{E}_{\mathrm{p}\sim P^{k+1}}[\eta(\pi_k|p_w^k)] = \eta(\pi_k|p_w^k). \tag{A.43}$$

where the approximation is made under the assumption that the expected returns of worst-case environment between two iterations are similar, which stems from the trust region constraint we impose on the update step between current and new policies, and can also be validated from experiments in Appendix A.5. □



(a) iteration $k = 300$ (b) iteration $k = 301$



(c) iteration $k = 350$ (d) iteration $k = 351$

Figure 1. Heatmaps of return between policy update at iterations $k = 300$ and $k = 350$, using MRPO on Hopper.

A.5. Empirical Verification of Assumption in Theorem 2

To verify the assumption made in Theorem 2, in Fig. 1, we study how the parameters of environments with poor performance scatter in the parameter space with different dimensions. Specifically, we plot the heatmap of return for the range of Hopper environments used for training, achieved by using MRPO to update the policy between two iterations. It can be validated that at the iteration $k = 300$, the poorly performing environments of the two policies before and after the MRPO update concentrate in the same region, i.e., the area of small frictions. The same result can be observed for the iteration $k = 350$.

For example, as shown in Figs. 1(a) and 1(b), at iteration $k = 300$, $p_w^{300} = (750, 0.5)$, the MC estimation of $\eta(\pi_{300}|p_w^{300})$ is 487.6 and that of $\eta(\pi_{301}|p_w^{300})$ is 532.0. At iteration $k = 301$, $p_w^{301} = (1027.8, 0.5)$ and the MC estimation of $\eta(\pi_{301}|p_w^{301})$ is 517.6. As shown in Figs. 1(c) and 1(d), at iteration $k = 350$, $p_w^{350} = (861.1, 0.5)$, the MC estimation of $\eta(\pi_{350}|p_w^{350})$ is 385.9 and that of $\eta(\pi_{351}|p_w^{350})$ is 422.2. At iteration $k = 351$, $p_w^{351} = (750, 0.5)$ and the MC estimation of $\eta(\pi_{351}|p_w^{351})$ is 394.0. In both cases, the empirical results can support the assumption that we made in (A.43), i.e., the expected returns of worst-case environment between two iterations are similar.

A.6. Bounded Reward Function Condition in Robot Control Tasks

In Theorem 1, we state the condition that reward function is bounded. Referring to the source code of OpenAI gym (Brockman et al., 2016), the reward function for the different robot control tasks evaluated in this paper are listed below.

Walker2d and Hopper:

$$R = x_{t+1} - x_t + b - 0.001|a_t|^2;$$

Halfcheetah:

$$R = x_{t+1} - x_t - 0.001|a_t|^2;$$

InvertedPendulum:

$$R = 1, \quad \text{if the pendulum does not fall down or the number of maximum time steps is reached;}$$

InvertedDoublePendulum:

$$R = b - c_{dist} - c_{vel}.$$

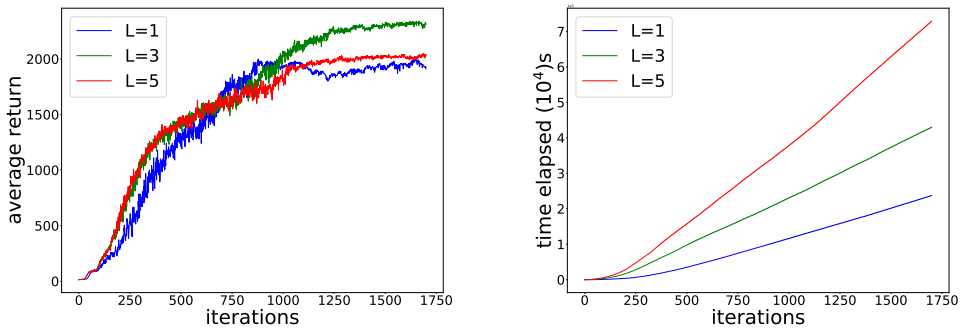
Cartpole:

$$R = 1, \quad \text{if the pole does not fall down or the number of maximum time steps is reached;}$$

In Walker2d, Hopper and Halfcheetah, x_{t+1} and x_t denote the positions of the robot at time step $t+1$ and t, respectively. For Walker2d and Hopper, $b \in \{0, 1\}$, and b equals 0 when the robot falls down or 1 otherwise. The squared norm of action represents the energy cost of the system. Since the maximum distance that the robot can move in one time step and the energy cost by taking an action at each timestep are bounded, these three tasks all have the bounded reward function. In InvertedPendulum and Cartpole, the reward is always 1. In InvertedDoublePendulum, b equals 0 when the pendulum falls down or 10 otherwise, c_{dist} is the distance between the robot and the centre, and c_{vel} is the weighted sum of the two pendulum's angular velocities. Since all the three parameters b, c_{dist} and c_{vel} are physically bounded, the reward function, as a linear combination of them, is also bounded.

A.7. Analysis of the Monte Carlo Estimation of $\eta(\pi|p)$

In Theorem 1, the worst-case environment parameter p_w needs to be selected according to the expected cumulative discounted reward $\eta(\pi|p)$ of environment p. However, $\eta(\pi|p)$ is infeasible to get in the practical implementation. Therefore, as a commonly used alternative approach as in (Rajeswaran et al., 2017), we use the mean of the cumulative discounted reward of L sampled trajectories $\sum_{j=0}^{L-1} G(\boldsymbol{\tau}_{i,j}|p_i)/L$ to approximate the expectation $\eta(\pi|p_i) = \mathbb{E}_{\boldsymbol{\tau}}\left[G(\boldsymbol{\tau}|p_i)\right]$ of any environment p_i, by using Monte Carlo method. We then determine the worst-case environment p_w based on $\sum_{j=0}^{L-1} G(\boldsymbol{\tau}_{i,j}|p_i)/L$ of a given set of environments $p_i{}_{i=0}^{M-1}$. In the following, we will analyze the impact of L on the MC estimation error.



(a) Performance (b) Time elapsed

Figure 2. (a) Training curves of average return of MRPO on Hopper with different L; (b) Elapsed time versus number of iterations curves during training.

Theoretical analysis of the impact of L: Referring to Chebyshev's inequality, for any environment p_i and any $\varepsilon \geq 0$, with probability of at least $1 - \frac{\sigma^2}{L\varepsilon^2}$, we have

$$\left| \frac{\sum_{j=0}^{L-1} G(\boldsymbol{\tau}_{i,j}|p_i)}{L} - \frac{\sum_{j=0}^{L-1} \mathbb{E}_{\boldsymbol{\tau}_{i,j}}[G(\boldsymbol{\tau}_{i,j}|p_i)]}{L} \right| = \left| \frac{\sum_{j=0}^{L-1} G(\boldsymbol{\tau}_{i,j}|p_i)}{L} - \eta(\pi|p_i) \right| \leq \varepsilon, \tag{A.44}$$

where $\sigma = Var(G(\boldsymbol{\tau})|p_i)$ is the variance of trajectory $\boldsymbol{\tau}$'s return. From the above equation, we find out that the variance of the return does affect the MC estimation of $\eta(\pi|p)$ and a larger L can guarantee a higher probability for the convergence of $\sum_{j=0}^{L-1} G(\boldsymbol{\tau}_{i,j}|p_i)/L$ to $\eta(\pi|p_i)$.

Empirical evaluation of the impact of L: In practice, we conduct experiment of MRPO on Hopper with different choices of L. We find out that a larger L would not greatly affect the performance in terms of average return as shown in Fig. 2(a), but will significantly increase the training time as shown in Fig. 2(b). In other words, for the same number of training iterations, a larger L would consume significantly longer running time than a smaller L, while the performance is similar. Therefore, we set $L = 1$ in our practical implementation of MRPO to strike a trade-off between the approximation accuracy and time complexity in training.

A.8. Analysis of the Lipschitz Assumption

In robot control tasks, classical optimal control methods commonly utilize the differential equation to formulate the dynamic model, which then indicates that the transition dynamics model is L_p-Lipschitz and this formulated dynamic function can be used to estimate the Lipschitz constant L_p.

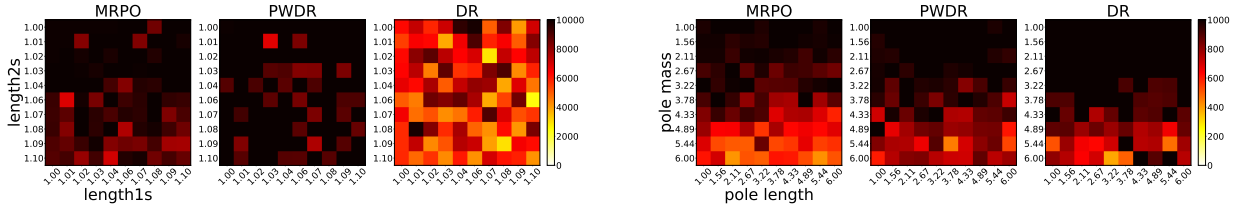
Here we illustrate the inverted pendulum system, which is one of our robot control tasks for validation. The single inverted pendulum has two state variables θ and $\dot{\theta}$, and one control input u, where θ and $\dot{\theta}$ represent the angular position from the inverted position and the angular velocity, respectively, and u is the torque. The system dynamics can therefore be described as

$$\ddot{\theta} = \frac{mgl \sin\theta + u - 0.1\dot{\theta}}{ml^2}, \tag{A.45}$$

where m is the mass, g is the Gravitational acceleration, and l is the length of pendulum. In our setting, we may choose m as the variable environment parameter p. Since the above system dynamics are differentiable w.r.t. m, it can be verified that the maximum value of the first derivative of the system dynamic model can be chosen as the Lipschitz constant L_p.

A.9. Generalization to Unseen Environments of Cartpole and InvertedDoublePendulum

In Fig. 3, we show the comparison results of MPRO, PR-DR and DR on unseen environments for the other two benchmarks, InvertedDoublePendulum and Cartpole, to provide empirical support for the generalization capability of MRPO.

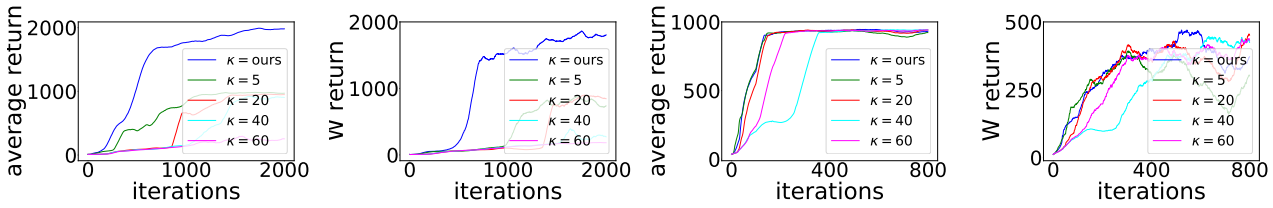


(a) InvertedDoublePendulum (b) Cartpole

Figure 3. Heatmap of return in unseen environments on InvertedDoublePendulum and Cartpole, with policies trained by MRPO, PW-DR and DR in the training environments.

A.10. Tuning of κ in MRPO on Walker2d and InvertedPendulum

As stated in section 4.3, κ was selected in a gradually increasing manner to strike tradeoff between average and worst-case return. 1) Theoretically, from Eqs. (3) and (9), a large κ increases the penalty to $||p - p_w||$, forcing environment distribution P in Eq. (4) more concentrated around p_w. Thus, we start from small κ to ensure policy trained with good average performance, and then increase κ gradually to focus on improving worst-case performance. 2) Empirically, we further test MRPO on Walker2d and InvertedPendulum with different choices of κ. Additional result of Walker2d and InvertedPendulum in Figs. 4(a)-4(d) validates a similar trend to that of Hopper in Fig. 3 in the main text.

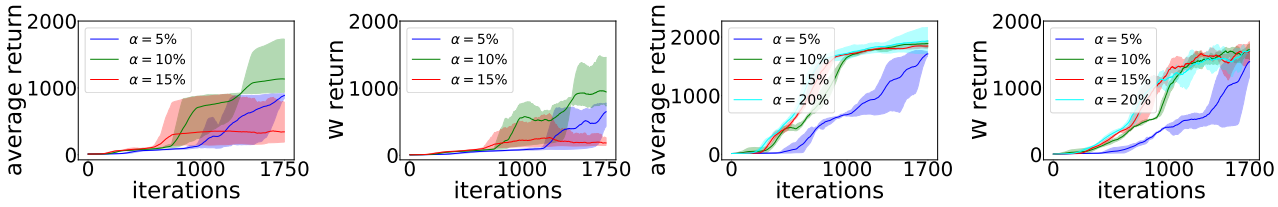


(a) W2D A (b) W2D W (c) IP A (d) IP W

Figure 4. Training curves of average return (A) and 10% worst-case return (W) of MRPO on W2D and IP with different κ.

A.11. Hyperparameter tuning for PW-DR

PW-DR needs to tune the hyperparameter α, standing for the α-percentage of worst performing trajectories. We perform the search of α on Walker2d and Hopper in Figs. 5(a)-5(d), which validates that $\alpha = 10\%$ achieves highest performance for PW-DR.



(a) W2D A (b) W2D W (c) HP A (d) HP W

Figure 5. Training curves of average return (A) and 10% worst-case return (W) of PW-DR on W2D and HP with different α.

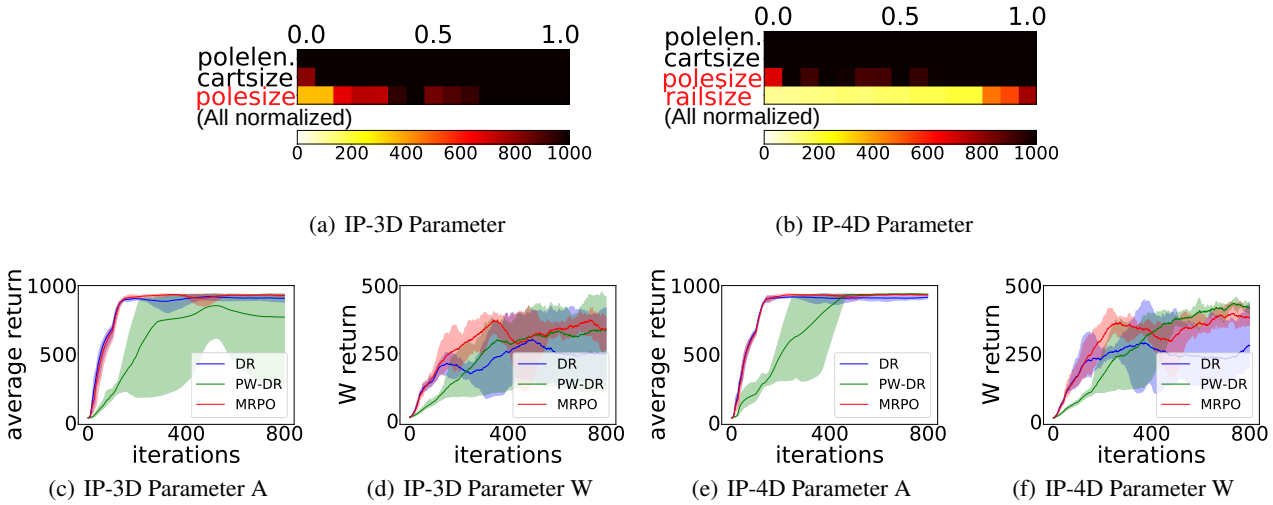
A.12. Generalization to higher dimensional randomization of environment parameters

Table 1. Range of environment parameter for InvertedPedulum with higher dimensional randomization

Task	Env. Param.	Training Range
IP-3D Parameter	Pole length	$[0.50, 2.00]$
	Cart size	$[0.05, 0.25]$
	Pole size	$[0.03, 0.068]$
IP-4D Parameter	Pole length	$[0.50, 2.00]$
	Cart size	$[0.05, 0.25]$
	Pole size	$[0.03, 0.068]$
	Rail size	$[0.01, 0.03]$

In the main text, we evaluate the performance of MRPO on environments under 2D parameters randomization. We now discuss the scalability of MRPO to higher dimensional randomization of environment parameters and test on InvertedPendulum with 3D and 4D parameters randomization as shown in Table 1. We propose here a possible extension of MRPO to address this issue: i) To compute $||p - p_w||$ in MRPO, we normaliz each dimension of p w.r.t. its training range to remove impact

of different units. ii) We then assign different weights to these dimensions w.r.t. their impact on performance as shown in Figs. 6(a)-6(b), i.e., the dimension that hurts policy's performance more (while others fixed as default in Roboschool) is given a higher weight. iii) We apply the same training strategy of MRPO on IP-3D Parameter and IP-4D Parameter, and the results in Figs. 6(c)-6(f) show that MRPO generalizes well to 3 and 4 dimensions on InvertedPendulum, and thus a potential to even higher dimensional cases.



(a) IP-3D Parameter

(b) IP-4D Parameter

(c) IP-3D Parameter A

(d) IP-3D Parameter W

(e) IP-4D Parameter A

(f) IP-4D Parameter W

Figure 6. Training curves of average return (A) and 10% worst-case return (W) of MRPO extending on IP-3D Parameter and IP-4D Parameter

References

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, pp. 12519–12530, 2019.

Rajeswaran, A., Ghotra, S., Ravindran, B., and Levine, S. Epopt: Learning robust neural network policies using model ensembles. 2017.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.