A Decentralized Policy Gradient Approach to Multi-Task Reinforcement Learning - Supplementary Material

Sihan Zeng[1] **Malik Aqeel Anwar**[1] **Thinh T. Doan**[2] **Arijit Raychowdhury**[1] **Justin Romberg**[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA
[2]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg Virginia, USA

A COMPUTATION DETAILS OF EXAMPLES IN SECTION 2.2

First, we look at the first example in Section 2.2 used to illustrate that deterministic optimal policy may not exist in MTRL in general. As we discussed, it is easy to see that the optimal policy in state S_2 and S_4 is to always take action L in order to reach the positive reward or to stay away from the negative reward, and all that is left to be figured out is the policy at state S_3.

There are 2 possible deterministic policies in state S_3, to always take action L or to always take action R. First, consider one policy $\pi_{d,l}$, which is to always take L.

We have $V_1^{\pi_{d,l}}(S_3) = \gamma$ as the agent reaches S_1 in 2 steps under $\pi_{d,l}$ and claims the $+1$ reward. However, this policy produces a zero value in environment 2, $V_2^{\pi_d}(S_3) = 0$, since an agent will move back and forth between S_3 and S_4 forever. Therefore, this deterministic policy achieves

$$V_1^{\pi_{d,l}}(S_3) + V_2^{\pi_{d,l}}(S_3) = \gamma + 0 = \gamma.$$

By symmetry, the value of the policy $\pi_{d,r}$, which is to always take action R in state S_3, is

$$V_1^{\pi_{d,r}}(S_3) + V_2^{\pi_{d,r}}(S_3) = 0 + \gamma = \gamma.$$

Now, let's consider a stochastic policy π_s, which we will show performs better than the two deterministic policies. This policy π_s takes the same deterministic actions as $\pi_{d,l}$ and $\pi_{d,r}$ in state S_2, S_4, and is defined as follows for state S_3.

$$\pi_s(a|S_3) = \begin{cases} p, & a = \text{left} \\ 1-p, & a = \text{right} \end{cases}$$

We compute cumulative rewards under π_s.

$$\begin{aligned} V_1^{\pi_s}(S_3) &= p\gamma + p(1-p)\gamma^3 + p(1-p)^2\gamma^5 + ... \\ &= p\gamma \sum_{k=0}^{\infty} \left((1-p)\gamma^2\right)^k \\ &= \frac{p\gamma}{1-(1-p)\gamma^2}. \end{aligned}$$

Similarly,

$$\begin{aligned} V_2^{\pi_s}(S_3) &= (1-p)\gamma + (1-p)p\gamma^3 + (1-p)p^2\gamma^5 + ... \\ &= (1-p)\gamma \sum_{k=0}^{\infty} \left(p\gamma^2\right)^k \end{aligned}$$

$$= \frac{(1-p)\gamma}{1-p\gamma^2}.$$

Then,

$$V_1^{\pi_s}(S_3) + V_2^{\pi_s}(S_3) = \frac{p\gamma}{1-(1-p)\gamma^2} + \frac{(1-p)\gamma}{1-p\gamma^2}.$$

Taking the derivative with respect to p and setting it to 0, we get

$$\frac{1}{(1-(1-p)\gamma^2)^2} = \frac{1}{(1-p\gamma^2)^2}, \tag{1}$$

which leads to $p = 0.5$.

The value of policy π_s at state S_3 is

$$\begin{aligned} V_1^{\pi_s}(S_3) + V_2^{\pi_s}(S_3) &= \frac{p\gamma}{1-(1-p)\gamma^2} + \frac{(1-p)\gamma}{1-p\gamma^2} \\ &= \frac{2\gamma}{2-\gamma^2}. \end{aligned}$$

Then, we explain how the three stationary points are computed in the second example in Section 2.2. Note that from Section B.4, we have that

$$\frac{\partial}{\partial \theta_{s,a}} V_i^{\pi_\theta}(\rho_i) = \frac{1}{1-\gamma_i} d_{i,\rho_i}^{\pi_\theta}(s)\pi_\theta(a|s)A_i^{\pi_\theta}(s,a) \tag{2}$$

We define $D_i^{\pi_\theta}$ to be the $|\mathcal{S}_i| \times |\mathcal{S}_i|$ matrix where the entry (i,j) is $d_i^{\pi_\theta}(s_i|s_j)$. It can be easily seen that

$$d_{i,\rho_i}^{\pi_\theta}(s) = D_i^{\pi_\theta}\rho_i. \tag{3}$$

Given $P_i^{\pi_\theta}$ the transition probability matrix of task i under policy π_θ (whose entry (j,k) denotes $P_i(j \mid k)$), the matrix $D_i^{\pi_\theta}$ can be computed as

$$D_i^{\pi} = (1-\gamma P_i^{\pi})^{-1}. \tag{4}$$

Given the small scale and the known dynamics of the problem, we can also compute the value function and the Q function of the policy π_θ in the two tasks by solving the Bellman equation, from which we get $A_i^{\pi_\theta}(s,a)$. Specifically, under a policy π, the value functions associated with the first and second tasks are

$$V_1^{\pi} = (I-\gamma(P_1^{\pi})^\top)^{-1}\begin{bmatrix} 0 \\ 1-p \\ 0 \\ -p \\ 0 \end{bmatrix}, \quad \text{and} \quad V_2^{\pi} = (I-\gamma(P_2^{\pi})^\top)^{-1}\begin{bmatrix} 0 \\ -p \\ 0 \\ 1-p \\ 0 \end{bmatrix}. \tag{5}$$

In addition, we can compute the Q functions

$$\begin{aligned} Q_1^{\pi}(\cdot,L) &= [0, \quad (1-p)+\gamma p V_1^{\pi}(S_3), \quad \gamma V_1^{\pi}(S_2), \quad \gamma(1-p)V_1^{\pi}(S_3)-p, \quad 0]^\top, \\ Q_1^{\pi}(\cdot,R) &= [0, \quad (1-p)+\gamma p V_1^{\pi}(S_3), \quad \gamma V_1^{\pi}(S_4), \quad \gamma(1-p)V_1^{\pi}(S_3)-p, \quad 0]^\top, \\ Q_2^{\pi}(\cdot,L) &= [0, \quad \gamma(1-p)V_2^{\pi}(S_3)-p, \quad \gamma V_2^{\pi}(S_2), \quad \gamma p V_2^{\pi}(S_3)+(1-p), \quad 0]^\top, \\ Q_2^{\pi}(\cdot,R) &= [0, \quad \gamma(1-p)V_2^{\pi}(S_3)-p, \quad \gamma V_2^{\pi}(S_4), \quad \gamma p V_2^{\pi}(S_3)+(1-p), \quad 0]^\top, \end{aligned} \tag{6}$$

from which the advantage function can be easily computed by taking the difference between the Q functions and the value functions. We also know $\pi_\theta(s,a)$ of the policy for which we would like to evaluate the gradient. Therefore, we can compute all the quantities in the gradient expression (2). Now we go through all three parameterizations and calculate the gradient and the cumulative return.

We first consider the policy π_1 under the parameterization $\theta_{S_3,L} = 1, \theta_{S_3,R} = \infty$, which implies $\pi_1(L \mid S_3) = 0$ and $\pi_1(R \mid S_3) = 1$. First, we can easily see that the transition probability matrices are

$$P_1^{\pi_1} = \begin{bmatrix} 1 & 1-p & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 1-p & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p & 1 \end{bmatrix}, \quad \text{and} \quad P_2^{\pi_1} = \begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1-p & 1 \end{bmatrix}.$$

Computing $D_i^{\pi_1}$ according to (4) using Gaussian elimination, we can derive

$$D_1^{\pi_1} = \begin{bmatrix} 1 & \gamma(1-p) & 0 & 0 & 0 \\ 0 & 1-\gamma & 0 & 0 & 0 \\ 0 & \frac{\gamma p(1-\gamma)}{(\gamma^2 p-\gamma^2+1)} & \frac{1-\gamma}{\gamma^2 p-\gamma^2+1} & \frac{\gamma(1-\gamma)(1-p)}{\gamma^2 p-\gamma^2+1} & 0 \\ 0 & \frac{\gamma^2 p(1-\gamma)}{(\gamma^2 p-\gamma^2+1)} & \frac{\gamma(1-\gamma)}{\gamma^2 p-\gamma^2+1} & \frac{1-\gamma}{\gamma^2 p-\gamma^2+1} & 0 \\ 0 & \frac{\gamma^3 p^2}{(\gamma^2 p-\gamma^2+1)} & \frac{\gamma^2 p}{\gamma^2 p-\gamma^2+1} & \frac{\gamma p}{\gamma^2 p-\gamma^2+1} & 1 \end{bmatrix}, \quad \text{and} \quad D_2^{\pi_1} = \begin{bmatrix} 1 & \gamma p & 0 & 0 & 0 \\ 0 & 1-\gamma & 0 & 0 & 0 \\ 0 & \frac{\gamma(1-\gamma)(1-p)}{1-\gamma^2 p} & \frac{1-\gamma}{1-\gamma^2 p} & \frac{\gamma(1-\gamma)p}{1-\gamma^2 p} & 0 \\ 0 & \frac{\gamma^2(1-\gamma)(1-p)}{1-\gamma^2 p} & \frac{\gamma(1-\gamma)}{1-\gamma^2 p} & \frac{1-\gamma}{1-\gamma^2 p} & 0 \\ 0 & \frac{\gamma^3(1-p)^2}{1-\gamma^2 p} & \frac{\gamma^2(1-p)}{1-\gamma^2 p} & \frac{\gamma(1-p)}{1-\gamma^2 p} & 1 \end{bmatrix}.$$

As we explained in (5) and (6), we can compute the advantage functions

$$\begin{aligned} A_1^{\pi_1}(\cdot, L) &= \left[0, 0, \quad \frac{\gamma(-\gamma^2 p^2 + (1-p)(\gamma^2 p - \gamma^2 + 1) + p)}{\gamma^2 p - \gamma^2 + 1}, 0, \quad 0\right]^\top, \\ A_1^{\pi_1}(\cdot, R) &= [0, \quad 0, \quad 0, \quad 0, \quad 0]^\top, \\ A_2^{\pi_1}(\cdot, L) &= \left[0, \quad 0, \quad \frac{\gamma(\gamma^2(1-p)^2 + p(\gamma^2 p - 1) - (1-p))}{1-\gamma^2 p}, \quad 0, \quad 0\right]^\top, \\ A_2^{\pi_1}(\cdot, R) &= [0, \quad 0, \quad 0, \quad 0, \quad 0]^\top. \end{aligned}$$

Recall (2). We have

$$\frac{\partial}{\partial \theta_{S_3,L}}(V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2)) = \frac{1}{1-\gamma} d_{1,\rho_1}^{\pi_1}(S_3)\pi_1(L|S_3)A_1^{\pi_1}(S_3, L) + \frac{1}{1-\gamma} d_{2,\rho_2}^{\pi_1}(S_3)\pi_1(L|S_3)A_2^{\pi_1}(S_3, L) = 0,$$

since $\pi_1(L \mid S_3) = 0$. In addition, we have

$$\frac{\partial}{\partial \theta_{S_3,R}}(V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2)) = \frac{1}{1-\gamma} d_{1,\rho_1}^{\pi_1}(S_3)\pi_1(R|S_3)A_1^{\pi_1}(S_3, R) + \frac{1}{1-\gamma} d_{2,\rho_2}^{\pi_1}(S_3)\pi_1(R|S_3)A_2^{\pi_1}(S_3, R) = 0,$$

since $A_1^{\pi_1}(S_3, R) = A_2^{\pi_1}(S_3, R) = 0$. The cumulative return under this policy is

$$V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2) = V_1^{\pi_1}(S_3) + V_2^{\pi_1}(S_3) = \frac{\gamma(-2\gamma^2 p + \gamma^2 + 2p - 1)}{\gamma^4 p^2 - \gamma^4 p + \gamma^2 - 1}.$$

By symmetry, the second policy π_2 under parameterization $\theta_{S_3,L} = \infty, \theta_{S_3,R} = 1$ is also a stationary point and has a cumulative return

$$V_1^{\pi_2}(\rho_1) + V_2^{\pi_2}(\rho_2) = \frac{\gamma(-2\gamma^2 p + \gamma^2 + 2p - 1)}{\gamma^4 p^2 - \gamma^4 p + \gamma^2 - 1}.$$

Finally, we look at the policy π_3 under parameterization $\theta_{S_3,L} = 1, \theta_{S_3,R} = 1$, which implies $\pi_3(L \mid S_3) = \pi_3(R \mid S_3) = 0.5$. We can see that the transition probability matrices are

$$P_1^{\pi_3} = \begin{bmatrix} 1 & 1-p & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & p & 0 & 1-p & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & p & 1 \end{bmatrix}, \quad \text{and} \quad P_2^{\pi_3} = \begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1-p & 1 \end{bmatrix}.$$

Computing $D_i^{\pi_3}$ according to (4) using Gaussian elimination, we can derive

$$D_1^{\pi_3} = \begin{bmatrix} 1 & \frac{\gamma(-\gamma^2p^2+2\gamma^2p-\gamma^2-2p+2)}{2-\gamma^2} & \frac{\gamma^2(1-p)}{2-\gamma^2} & \frac{\gamma^3(1-p)^2}{2-\gamma^2} & 0 \\ 0 & \frac{(1-\gamma)(\gamma^2p-\gamma^2+2)}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{\gamma^2(1-\gamma)(1-p)}{2-\gamma^2} & 0 \\ 0 & \frac{2\gamma(1-\gamma)(1-p)}{2-\gamma^2} & \frac{2(1-\gamma)}{2-\gamma^2} & \frac{2\gamma(1-\gamma)(1-p)}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^2(1-\gamma)p}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{(1-\gamma)(2-\gamma^2p)}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^3p^2}{2-\gamma^2} & \frac{\gamma^2p}{2-\gamma^2} & \frac{\gamma p(2-\gamma^2p)}{2-\gamma^2} & 1 \end{bmatrix},$$

$$\text{and} \quad D_2^{\pi_3} = \begin{bmatrix} 1 & \frac{\gamma p(2-\gamma^2p)}{2-\gamma^2} & \frac{\gamma^2p}{2-\gamma^2} & \frac{\gamma^2(1-\gamma)p}{2-\gamma^2} & 0 \\ 0 & \frac{(1-\gamma)(2-\gamma^2p)}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{\gamma^2(1-\gamma)p}{2-\gamma^2} & 0 \\ 0 & \frac{2\gamma(1-\gamma)(1-p)}{2-\gamma^2} & \frac{2(1-\gamma)}{2-\gamma^2} & \frac{2\gamma(1-\gamma)p}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^2(1-\gamma)(1-p)}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{(1-\gamma)(2-\gamma^2+\gamma^2p)}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^3(1-p)^2}{2-\gamma^2} & \frac{\gamma^2(1-p)}{2-\gamma^2} & \frac{\gamma(2-\gamma^2p^2+2\gamma^2p-\gamma^2-2p)}{2-\gamma^2} & 1 \end{bmatrix}.$$

As we explained in (5) and (6), we can compute the advantage functions

$$A_1^{\pi_3}(\cdot, L) = \left[0, 0, \quad \frac{\gamma(-2\gamma^2p^2+2\gamma^2p-\gamma^2+1)}{2-\gamma^2}, 0, \quad 0\right]^\top,$$
$$A_1^{\pi_3}(\cdot, R) = \left[0, \quad 0, \quad \frac{\gamma(2\gamma^2p^2-2\gamma^2p+\gamma^2-1)}{2-\gamma^2}, \quad 0, \quad 0\right]^\top,$$
$$A_2^{\pi_3}(\cdot, L) = \left[0, \quad 0, \quad \frac{\gamma(2\gamma^2p^2-2\gamma^2p+\gamma^2-1)}{2-\gamma^2}, \quad 0, \quad 0\right]^\top,$$
$$A_2^{\pi_3}(\cdot, R) = \left[0, \quad 0, \quad \frac{\gamma(-2\gamma^2p^2+2\gamma^2p-\gamma^2+1)}{2-\gamma^2}, \quad 0, \quad 0\right]^\top,$$

From (2), we have

$$\begin{aligned}\frac{\partial}{\partial\theta_{S_3,L}}(V_1^{\pi_3}(\rho_1)+V_2^{\pi_3}(\rho_2)) &= \frac{1}{1-\gamma}\pi_3(L|S_3)\left(d_{1,\rho_1}^{\pi_3}(S_3)A_1^{\pi_3}(S_3,L)+d_{2,\rho_2}^{\pi_3}(S_3)A_2^{\pi_3}(S_3,L)\right) \\ &= \frac{0.5}{1-\gamma}\left(\frac{2(1-\gamma)}{2-\gamma^2}\cdot\frac{\gamma(-2\gamma^2p^2+2\gamma^2p-\gamma^2+1)}{2-\gamma^2}+\frac{2(1-\gamma)}{2-\gamma^2}\cdot\frac{\gamma(2\gamma^2p^2-2\gamma^2p+\gamma^2-1)}{2-\gamma^2}\right) \\ &= 0.\end{aligned}$$

Similarly, one can show

$$\frac{\partial}{\partial\theta_{S_3,R}}(V_1^{\pi_3}(\rho_1)+V_2^{\pi_3}(\rho_2)) = 0.$$

The cumulative return under this policy is

$$V_1^{\pi_3}(\rho_1)+V_2^{\pi_3}(\rho_2) = V_1^{\pi_1}(S_3)+V_2^{\pi_1}(S_3) = \frac{\gamma(2-4p)}{2-\gamma^2}.$$

For computational simplicity, we choose $\gamma = \sqrt{0.5}$. Then,

$$V_1^{\pi_1}(\rho_1)+V_2^{\pi_1}(\rho_2) = \frac{\gamma(-2\gamma^2p+\gamma^2+2p-1)}{\gamma^4p^2-\gamma^4p+\gamma^2-1} = \frac{2p-1}{8\sqrt{2}(p-2)(p+1)},$$
$$\text{and} \quad V_1^{\pi_3}(\rho_1)+V_2^{\pi_3}(\rho_2) = V_1^{\pi_1}(S_3)+V_2^{\pi_1}(S_3) = \frac{\gamma(2-4p)}{2-\gamma^2} = \frac{4-8p}{3}.$$

If $p > 0.5$,

$$V_1^{\pi_1}(\rho_1)+V_2^{\pi_1}(\rho_2) = V_1^{\pi_2}(\rho_1)+V_2^{\pi_2}(\rho_2) = \frac{2p-1}{8\sqrt{2}(p-2)(p+1)} > \frac{4-8p}{3} = V_1^{\pi_3}(\rho_1)+V_2^{\pi_3}(\rho_2).$$

B PROOFS

In this appendix, we provide complete analysis for the results stated in the main paper. We first introduce the following notations used throughout this appendix.

$$\boldsymbol{\theta} \triangleq \left[\theta_1^T, \theta_2^T, ..., \theta_N^T\right]^T \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}|}, \qquad \boldsymbol{V}(\boldsymbol{\theta};\boldsymbol{\rho}) \triangleq \begin{pmatrix} V_1^{\pi_{\theta_1}}(\rho_1) \\ V_2^{\pi_{\theta_2}}(\rho_2) \\ \vdots \\ V_N^{\pi_{\theta_N}}(\rho_N) \end{pmatrix} \in \mathbb{R}^N,$$
$$\boldsymbol{\rho} = [\rho_1^T, \rho_2^T, ..., \rho_N^T]^T, \quad \boldsymbol{\mu} = [\mu_1^T, \mu_2^T, ..., \mu_N^T]^T, \quad \overline{\nabla \boldsymbol{V}}(\boldsymbol{\theta};\boldsymbol{\rho}) = \frac{1}{N}\sum_{i=1}^N \nabla_{\theta_i} V_i^{\pi_{\theta_i}}(\rho_i). \tag{7}$$

B.1 PROOF OF THEOREM 1

Define $D = 2N\lambda + \sum_{i=1}^N \frac{1}{(1-\gamma_i)^2}$. In the proof, we will need the following lemmas.

Lemma B.1. *For all k and $\boldsymbol{\mu}$, $||\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu})|| \le D$.*

Proof. By Eq. (37),

$$\begin{aligned} ||\nabla L_i^\lambda(\theta_i^k;\mu_i)|| &\le \sum_{s,a}\left|\frac{\partial L_i^\lambda(\theta_i^k;\mu_i)}{\partial \theta_{i\,s,a}^k}\right| \\ &\le \sum_{s,a}\left|\frac{1}{1-\gamma_i} d_{\mu_j}^{\pi_\theta}(s)\pi_\theta(a\,|\,s)A_i^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right)\right| \\ &\le \sum_{s,a}\frac{d_{\mu_j}^{\pi_\theta}(s)\pi_\theta(a\,|\,s)}{1-\gamma_i}\frac{1}{1-\gamma_i} + \sum_{s,a}\frac{\lambda}{|\mathcal{S}||\mathcal{A}|} + \sum_{s,a}\frac{\lambda}{|\mathcal{S}|}\pi_\theta(a\,|\,s) \\ &\le \frac{1}{(1-\gamma_i)^2} + 2\lambda, \end{aligned}$$

where the second last inequality uses (3). Using triangular inequality,

$$||\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu})|| \le \sum_{i=1}^N ||\nabla L_i^\lambda(\theta_i^k;\mu_i)|| \le 2N\lambda + \sum_{i=1}^N \frac{1}{(1-\gamma_i)^2}. \tag{8}$$

□

Lemma B.2. *Let $\bar\theta^k = \frac{1}{N}\sum_{i=1}^N \theta_i^k$. If each agent starts with the same initialization, i.e. $\theta_1^0 = \theta_2^0 = ... = \theta_N^0$, then*

$$||\theta_i^k - \bar\theta^k|| \le \frac{\alpha D}{1-\sigma_2}, \quad \forall i, k. \tag{9}$$

This is a standard result whose proof can be found in the existing literature, such as Yuan et al. [2016].

We made the assumption in Theorem 1 that the agents start with the same initialization. We denote $\theta^0 = \theta_i^0$, $\forall i$.

We define the Lyapunov function

$$\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) \triangleq -\mathbf{1}^T\boldsymbol{L}^\lambda(\boldsymbol{\theta};\boldsymbol{\mu}) + \frac{1}{2\alpha}||\boldsymbol{\theta}||_{I-W}^2, \tag{10}$$

where $||\boldsymbol{\theta}||_{I-W}^2 \triangleq \boldsymbol{\theta}^T((I-W)\otimes I)\boldsymbol{\theta}$.

Note that the sequence $\{\boldsymbol{\theta}^k\}$ generated by the distributed policy gradient algorithm is the same as the sequence generated by applying gradient descent on $\xi_{\alpha,\lambda}(\boldsymbol{\theta})$, if both algorithms use fixed step size α. This can be observed by re-writing the update equation (9).

$$\begin{aligned}\boldsymbol{\theta}^{k+1} &= (W\otimes I)\boldsymbol{\theta}^k + \alpha\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu}) \\ &= \boldsymbol{\theta}^k + \alpha\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu}) - ((I-W)\otimes I)\boldsymbol{\theta}^k \\ &= \boldsymbol{\theta}^k - \alpha(-\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \frac{1}{\alpha}((I-W)\otimes I)\boldsymbol{\theta}^k) \\ &= \boldsymbol{\theta}^k - \alpha\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) \end{aligned} \tag{11}$$

We have to establish the smoothness constant of $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu})$. Combining Lemma B.5 and Lemma B.6, $L_i^\lambda(\theta_i)$ is β_i^λ-smooth with

$$\beta_i^\lambda = \frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|}, \tag{12}$$

which implies $\sum_{i=1}^N L_i^\lambda(\theta_i)$ is β^λ-smooth, where

$$\beta^\lambda = \sum_{i=1}^N \left(\frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|}\right). \tag{13}$$

In addition, we know $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu})$ is $\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}$-smooth, with

$$\beta^{\boldsymbol{\xi}_{\alpha,\lambda}} = \beta^\lambda + \frac{1}{\alpha}\sigma_{\max}(I-W) = \beta^\lambda + \alpha^{-1}(1-\sigma_N). \tag{14}$$

By the $\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}$-smoothness of $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta})$, we have

$$\begin{aligned}\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^{k+1};\boldsymbol{\mu}) &\le \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \langle\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}), \boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k\rangle + \frac{\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}}{2}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \\ &= \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \langle -\frac{\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k}{\alpha}, \boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k\rangle + \frac{\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}}{2}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \\ &= \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + (\frac{\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}}{2} - \frac{1}{\alpha})||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \\ &= \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) - \frac{1}{2}(\alpha^{-1}(1+\sigma_N) - \beta^\lambda)||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \end{aligned} \tag{15}$$

Since $\alpha \le \frac{1+\sigma_N}{2\sum_{i=1}^N(\frac{8}{(1-\gamma_i)^3}+\frac{2\lambda}{|\mathcal{S}|})} = \frac{1+\sigma_N}{2\beta^\lambda}$, we know $\frac{1}{2}(\alpha^{-1}(1+\sigma_N)-\beta^\lambda) \ge 0$, $\forall k$. This implies $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})$ is a non-increasing sequence. Let $\tilde{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \xi_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu})$. We have

$$\begin{aligned}\sum_{k=0}^{K-1}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 &\le \sum_{k=0}^{K-1} 2(\alpha^{-1}(1+\sigma_N)-\beta^\lambda)^{-1}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^{k+1};\boldsymbol{\mu})) \\ &= c_1(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^{K-1};\boldsymbol{\mu})) \\ &\le c_1(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu})), \end{aligned} \tag{16}$$

where we define $c_1 = 2(\alpha^{-1}(1+\sigma_N)-\beta^\lambda)^{-1}$.

This implies

$$\min_{k<K}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \le \frac{c_1}{K}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu})). \tag{17}$$

From Eq. (11), $||\alpha\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 = ||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2$. Thus,

$$\min_{k<K}||\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 = \frac{1}{\alpha^2}\min_{k<K}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \le \frac{c_1}{K\alpha^2}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu})). \tag{18}$$

Taking derivative of Eq. (10),

$$\nabla \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) = -\nabla \boldsymbol{L}^{\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) + \frac{1}{\alpha}((I-W)\otimes I)\boldsymbol{\theta}, \tag{19}$$

Observe that $\mathbf{1}^T(I-W)=\mathbf{0}$ due to the double stochasticity of W, which leads to

$$\begin{aligned}\overline{\nabla \boldsymbol{\xi}}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) &= -\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) + \frac{1}{N\alpha}(\mathbf{1}^T(I-W)\otimes I)\boldsymbol{\theta}\\ &= -\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}).\end{aligned}$$

Now we can bound the gradient $\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})$.

$$\begin{aligned}\min_{k<K}||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 &= \min_{k<K}||\overline{\nabla \boldsymbol{\xi}}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2\\ &\le \min_{k<K}||\nabla \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2\\ &\le \frac{c_1}{K\alpha^2}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu}))\\ &= \frac{c_1}{K\alpha^2}(-\sum_{i=1}^N L_i^{\lambda}(\boldsymbol{\theta}^0;\mu_i) + \frac{1}{2\alpha}||\boldsymbol{\theta}^0||^2_{I-W} + \sum_{i=1}^N L_i^{\lambda}(\tilde{\boldsymbol{\theta}};\mu_i) - \frac{1}{2\alpha}||\tilde{\boldsymbol{\theta}}||^2_{I-W})\\ &\le \frac{c_1}{K\alpha^2}\sum_{i=1}^N (L_i^{\lambda}(\tilde{\boldsymbol{\theta}};\mu_i) - L_i^{\lambda}(\boldsymbol{\theta}^0;\mu_i))\\ &\le \frac{c_1}{K\alpha^2}\sum_{i=1}^N (V_i^{\pi_{\tilde{\theta}_i}}(\mu_i) - V_i^{\pi_{\theta_i^0}}(\mu_i) + \lambda \mathrm{RE}(\pi_{\theta_i^0}))\\ &\le \frac{c_1}{K\alpha^2}\sum_{i=1}^N (\frac{1}{1-\gamma_i} + \lambda \mathrm{RE}(\pi_{\theta^0})).\end{aligned} \tag{20}$$

The third line comes from (18). The fifth line uses our assumption that all agents start with the same parameter initialization, making $||\boldsymbol{\theta}^0||^2_{I-W}=0$.The second last inequality is from the fact that relative entropy is non-negative. The last inequality comes from the bounded value function (3).

Using the definition of $\boldsymbol{L}^{\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})$ in (7), we have

$$\begin{aligned}\min_{k<K}||\overline{\nabla V}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 &= \min_{k<K}||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \frac{\lambda}{N}\sum_{i=1}^N \nabla \mathrm{RE}(\pi_{\theta_i^k})||^2\\ &\le 2\min_{k<K}||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 + \frac{2}{N}\sum_{i=1}^N ||\nabla \lambda \mathrm{RE}(\pi_{\theta_i^k})||^2.\end{aligned} \tag{21}$$

The second term uses the smoothness of the regularizer, which we establish in Lemma B.6. The first term is bounded in (20). Therefore,

$$\begin{aligned}\min_{k<K}||\overline{\nabla V}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 &\le 2\min_{k<K}||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 + \frac{2}{N}\sum_{i=1}^N ||\nabla \lambda \mathrm{RE}(\pi_{\theta_i^k})||^2\\ &\le \frac{2c_1}{K\alpha^2}\sum_{i=1}^N (\frac{1}{1-\gamma_i} + \lambda \mathrm{RE}(\pi_{\theta^0})) + \frac{2}{N}\left(\frac{\lambda}{\sqrt{|\mathcal{A}|}} + \lambda\right)^2\\ &\le \frac{2c_1}{K\alpha^2}\sum_{i=1}^N (\frac{1}{1-\gamma_i} + \lambda \mathrm{RE}(\pi_{\theta^0})) + \frac{8\lambda^2}{N}\end{aligned} \tag{22}$$

Using the smoothness of V_i, which we show in Lemma B.5, we have

$$\begin{aligned}\min_{k<K} ||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_i^k;\mu_j)||^2 &= \min_{k<K} ||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_j^k;\mu_j) - \big(\nabla V_j(\theta_j^k;\mu_j) - \nabla V_j(\theta_i^k;\mu_j)\big)||^2 \\ &\leq \min_{k<K} 2||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_j^k;\mu_j)||^2 \\ &\qquad + \frac{2}{N}\sum_{j=1}^{N}||\nabla V_j(\theta_j^k;\mu_j) - \nabla V_j(\theta_i^k;\mu_j)||^2 \\ &\leq 2\min_{k<K}||\overline{\nabla V}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 + \frac{2}{N}\sum_{j=1}^{N}\frac{64}{(1-\gamma_j)^6}||\theta_i^k - \theta_j^k||^2. \end{aligned} \tag{23}$$

From Lemma B.2, we have

$$\begin{aligned}||\theta_i^k - \theta_j^k|| &= ||(\theta_i^k - \bar{\theta}^k) - (\bar{\theta}^k - \theta_j^k)|| \\ &\leq ||\theta_i^k - \bar{\theta}^k|| + ||\theta_j^k - \bar{\theta}^k|| \\ &\leq \frac{2\alpha D}{1-\sigma_2}. \end{aligned} \tag{24}$$

Plugging this inequality and (22) into (23), we get

$$\begin{aligned}&\min_{k<K} ||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_i^k;\mu_j)||^2 \\ &\leq \frac{4c_1}{K\alpha^2}\sum_{j=1}^{N}(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})) + \frac{16\lambda^2}{N} + \frac{2}{N}\sum_{j=1}^{N}\frac{64}{(1-\gamma_j)^6}\frac{4\alpha^2 D^2}{(1-\sigma_2)^2} \\ &\leq \frac{16}{K\alpha}\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})\right) + \frac{16\lambda^2}{N} + \sum_{j=1}^{N}\frac{512 D^2\alpha^2}{N(1-\sigma_2)^2(1-\gamma_j)^6}. \end{aligned}$$

The proof is completed by recognizing $\rho_i = \mu_i$, $\forall i$.

B.2 PROOF OF THEOREM 2

When condition (12) is observed, we can establish the global optimality condition under the tabular policy.

Proposition 1. *Let $\theta^* = \max_\theta V(\theta;\boldsymbol{\rho})$. For policy parameter θ, if $||\sum_{i=1}^{N}\nabla L_i^\lambda(\theta;\mu_i)|| \leq \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$, we have*

$$V(\theta^*;\boldsymbol{\rho}) - V(\theta;\boldsymbol{\rho}) \leq 2\lambda N \max_{s\in\mathcal{S}, i: s\in\mathcal{S}_i}\{\frac{d_{\rho_i}^{\pi_{\theta^\star}}(s)}{(1-\gamma_i)\mu_i(s)}\}$$

if the environment and the initial state distributions $\boldsymbol{\rho}$ and $\boldsymbol{\mu}$ jointly satisfies the discounted visitation match assumption.

The proof this proposition is in Section B.3. Using the proposition, we can guarantee that θ_i^k is an ϵ-optimal solution in the objective function by setting $\epsilon = 2N\lambda\max_{j,s}\{\frac{d_{\rho_j}^{\pi_{\theta^\star}}(s)}{(1-\gamma_j)\mu_j(s)}\}$ and ensuring $||\sum_{j=1}^{N}\nabla L_j^\lambda(\theta_i^k;\mu_j)|| \leq \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$. Solving for λ in terms of ϵ, we get

$$\lambda = \frac{\epsilon}{2N\max_{j,s}\{\frac{d_{\rho_j}^{\pi_{\theta^\star}}(s)}{(1-\gamma_j)\mu_j(s)}\}}. \tag{25}$$

Now we bound the norm of the gradient.

$$\begin{aligned}
\min_{k<K} ||\sum_{j=1}^{N} \nabla L_j^\lambda(\theta_i^k;\mu_j)|| &= \min_{k<K} ||\sum_{j=1}^{N} \nabla L_j^\lambda(\theta_j^k;\mu_j) + \sum_{j=1}^{N}\left(\nabla L_j^\lambda(\theta_i^k;\mu_j) - \nabla L_j^\lambda(\theta_j^k)\right)|| \\
&\le \min_{k<K} ||N\overline{\nabla \boldsymbol{L}}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu})|| + \sum_{j=1}^{N} ||\nabla L_j^\lambda(\theta_i^k;\mu_j) - \nabla L_j^\lambda(\theta_j^k;\mu_j)|| \\
&\le N\min_{k<K} ||\overline{\nabla \boldsymbol{L}}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu})|| + \sum_{j=1}^{N} \beta_i^\lambda ||\theta_i^k - \theta_j^k||, \qquad (26)
\end{aligned}$$

where the last inequality uses the smoothness property of L_i^λ.Combining Lemma B.5 and Lemma B.6, $\beta_i^\lambda = \frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|}$. We have a bound on the first term in (20), and now we bound the second term using Lemma B.2.

$$\begin{aligned}
||\theta_i^k - \theta_j^k|| &= ||(\theta_i^k - \bar{\theta}^k) - (\bar{\theta}^k - \theta_j^k)|| \\
&\le ||\theta_i^k - \bar{\theta}^k|| + ||\theta_j^k - \bar{\theta}^k|| \\
&\le \frac{2\alpha D}{1-\sigma_2} \qquad (27)
\end{aligned}$$

Plug this into (26),

$$\begin{aligned}
\min_{k<K} ||\sum_{j=1}^{N} \nabla L_j^\lambda(\theta_i^k;\mu_j)|| &\le N\min_{k<K} ||\overline{\nabla \boldsymbol{L}}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu})|| + \sum_{j=1}^{N} \beta_i^\lambda ||\theta_i^k - \theta_j^k|| \\
&\le N\sqrt{\frac{c_1}{K\alpha^2}\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})\right)} + \sum_{j=1}^{N}\beta_i^\lambda \frac{2\alpha D}{1-\sigma_2} \qquad (28) \\
&\le N\sqrt{\frac{c_1}{K\alpha^2}\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})\right)} + \frac{2\alpha\beta^\lambda D}{1-\sigma_2} \qquad (29)
\end{aligned}$$

To ensure $\min_{k<K} ||\sum_{j=1}^{N} \nabla L_j^\lambda(\theta_i^k;\mu_j)|| \le \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$, we make

$$N\sqrt{\frac{c_1}{K\alpha^2}\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})\right)} + \frac{2\alpha\beta^\lambda D}{1-\sigma_2} \le \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} \qquad (30)$$

Solving for K, we get

$$\begin{aligned}
K &\ge \frac{c_1 N^2\left(\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})\right)\right)}{\alpha^2\left(\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} - \frac{2\alpha\beta^\lambda D}{1-\sigma_2}\right)^2} \qquad (31) \\
&= \frac{c_1 N^2\left(\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j} + \lambda \mathrm{RE}(\pi_{\theta^0})\right)\right)}{\alpha^2\left(\frac{\epsilon c_2}{4|\mathcal{S}||\mathcal{A}|} - \frac{2\alpha D}{1-\sigma_2}\sum_{j=1}^{N}\left(\frac{8}{(1-\gamma_j)^3} + \frac{\epsilon c_2}{N|\mathcal{S}|}\right)\right)^2}, \qquad (32)
\end{aligned}$$

where we used the fact that $\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} - \frac{2\alpha\beta^\lambda D}{1-\sigma_2} > 0$, if $\alpha < \frac{\lambda N(1-\sigma_2)}{4\beta^\lambda D|\mathcal{S}||\mathcal{A}|}$.

B.3 PROOF OF PROPOSITION 1

From the assumption (12), we define

$$\frac{d_{i,\rho_i}^{\pi_{\theta^*}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)} = \frac{d_{j,\rho_j}^{\pi_{\theta^*}}(s)}{d_{j,\mu_j}^{\pi_\theta}(s)} \triangleq \tilde{d}(s), \qquad \forall s: s\in\mathcal{S}_i\cap\mathcal{S}_j,\ \forall i,j. \qquad (33)$$

Kakade and Langford [2002] introduced the performance difference lemma that relates the value function of two policies. We use this lemma in our analysis.

Lemma B.3. *For any policy π and $\tilde{\pi}$ operating in environment i under the initial state distribution ρ_i,*

$$V_i^{\pi}(\rho_i) - V_i^{\tilde{\pi}}(\rho_i) = \frac{1}{1-\gamma_i}\mathbb{E}_{s\sim d_{i,\rho_i}^{\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right]. \tag{34}$$

By Lemma B.3,

$$\begin{aligned}
V(\theta^*;\boldsymbol{\rho}) - V(\theta;\boldsymbol{\rho}) &= \sum_{i=1}^{N}\frac{1}{1-\gamma_i}\sum_{s\in\mathcal{S}_i}\sum_{a\in\mathcal{A}} d_{i,\rho_i}^{\pi_{\theta^\star}}(s)\pi_{\theta^\star}(a\,|\,s)A_i^{\pi_\theta}(s,a)\\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_{\theta^\star}(a\,|\,s)\sum_{i:s\in\mathcal{S}_i}\frac{1}{1-\gamma_i}d_{i,\rho_i}^{\pi_{\theta^\star}}(s)A_i^{\pi_\theta}(s,a)\\
&\le \sum_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}\sum_{i:s\in\mathcal{S}_i}\frac{1}{1-\gamma_i}d_{i,\rho_i}^{\pi_{\theta^\star}}(s)A_i^{\pi_\theta}(s,a)\\
&= \sum_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\rho_i}^{\pi_{\theta^\star}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\\
&= \sum_{s\in\mathcal{S}}\tilde{d}(s)\max_{a\in\mathcal{A}}\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\\
&\le \max_{s\in\mathcal{S},i:s\in\mathcal{S}_i}\{\frac{d_{i,\rho_i}^{\pi_{\theta^*}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\}\sum_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\\
&\le \max_{s\in\mathcal{S},i:s\in\mathcal{S}_i}\{\frac{d_{i,\rho_i}^{\pi_{\theta^*}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\}|\mathcal{S}|\frac{2\lambda N}{|\mathcal{S}|}\\
&= 2\lambda N\max_{s\in\mathcal{S},i:s\in\mathcal{S}_i}\{\frac{d_{i,\rho_i}^{\pi_{\theta^*}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\}\\
&= 2\lambda N\max_{s\in\mathcal{S},i:s\in\mathcal{S}_i}\{\frac{d_{i,\rho_i}^{\pi_{\theta^*}}(s)}{(1-\gamma_i)\mu_i(s)}\}
\end{aligned} \tag{35}$$

The sixth line follows since $\max_{a\in\mathcal{A}}\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\ge 0,\ \forall s$. The last inequality uses the fact that $d_{i,\mu_i}^{\pi}(s)\ge(1-\gamma_i)\mu_i(s)$, element-wise, $\forall\pi$, which simply follows from the definition of $d_{i,\mu_i}^{\pi}(s)$. The seventh line uses

$$\max_{a\in\mathcal{A}}\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\le\frac{2\lambda N}{|\mathcal{S}|}, \tag{36}$$

which we now prove. To show this, we only have to prove this is true for those (s,a) where $\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\ge 0$. The gradient of θ under the softmax parameterization in environment i is

$$\frac{\partial L_i^{\lambda}(\theta;\mu_i)}{\partial\theta_{s,a}} = \frac{1}{1-\gamma_i}d_{i,\mu_i}^{\pi_\theta}(s)\pi_\theta(a\,|\,s)A_i^{\pi_\theta}(s,a)+\frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|}-\pi_\theta(a|s)\right). \tag{37}$$

From our assumption $||\sum_{i=1}^{N}\nabla L_i^{\lambda}(\theta;\mu_i)||\le\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$, we know that for all (s,a) such that $\sum_{i:s\in\mathcal{S}_i}\frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i}A_i^{\pi_\theta}(s,a)\ge 0$,

$$\begin{aligned}
\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} &\ge \sum_{i=1}^{N}\frac{\partial L_i^{\lambda}(\theta;\mu_i)}{\partial\theta_{s,a}}\\
&= \sum_{i:s\in\mathcal{S}_i}\frac{1}{1-\gamma_i}d_{i,\mu_i}^{\pi_\theta}(s)\pi_\theta(a|s)A_i^{\pi_\theta}(s,a)+\sum_{i=1}^{N}\frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|}-\pi_\theta(a|s)\right)
\end{aligned}$$

$$\geq 0 + \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|} \left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s) \right)$$
$$\geq \frac{\lambda N}{|\mathcal{S}|} \left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s) \right). \tag{38}$$

Rearranging the terms,

$$\pi_\theta(a \mid s) \geq \frac{1}{|\mathcal{A}|} - \frac{|\mathcal{S}|}{\lambda N} \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} \geq \frac{1}{2|\mathcal{A}|}. \tag{39}$$

Re-writing Eq. (37) and summing over environments,

$$\begin{aligned}
\sum_{i=1}^{N} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) &= \sum_{i:s\in\mathcal{S}_i} \frac{1}{\pi_\theta(a|s)} \frac{\partial L_i^\lambda(\theta;\mu_i)}{\partial \theta_{s,a}} - \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|} \left(\frac{1}{\pi_\theta(a \mid s)|\mathcal{A}|} - 1 \right) \\
&\leq \frac{1}{\pi_\theta(a|s)} \sum_{i:s\in\mathcal{S}_i} \frac{\partial L_i^\lambda(\theta;\mu_i)}{\partial \theta_{s,a}} + \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|} \\
&\leq 2|\mathcal{A}| \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} + \frac{\lambda N}{|\mathcal{S}|} \\
&\leq \frac{2\lambda N}{|\mathcal{S}|},
\end{aligned} \tag{40}$$

where the second last line uses inequality (39).

B.4 DERIVATION OF THE GRADIENT (8)

Here we just derive the gradient for $V_i^{\pi_\theta}$. The gradient of L_i^λ can be easily computed from the gradient of $\nabla V_i^{\pi_\theta}$ by adding the gradient of the entropy regularizer.

By definition,

$$\begin{aligned}
V_i^{\pi_\theta}(s_i) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma_i^k \mathcal{R}_i(s_i^k, a_i^k) \mid s_i^0 = s_i \right], \qquad a_i^k \sim \pi_\theta(s_i^k) \\
&= \sum_{a_i\in\mathcal{A}} \pi_\theta(a_i \mid s_i) Q_i^{\pi_\theta}(s_i, a_i) \\
&= \sum_{a_i\in\mathcal{A}} \pi_\theta(a_i \mid s_i) \mathbb{E}_{s_i'\in\mathcal{S}_i} \left[\mathcal{R}(s_i, a_i) + \gamma_i V_i^{\pi_\theta}(s_i') \right],
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{\partial V_i^{\pi_\theta}(s_i)}{\partial \theta} &= \sum_{a_i\in\mathcal{A}} \left[Q_i^{\pi_\theta}(s_i,a_i) \frac{\partial \pi_\theta(a_i \mid s_i)}{\partial \theta} + \pi_\theta(a_i \mid s_i) \frac{\partial Q_i^{\pi_\theta}(s_i,a_i)}{\partial \theta} \right] \\
&= \sum_{a_i\in\mathcal{A}} \pi_\theta(a_i \mid s_i) Q_i^{\pi_\theta}(s_i,a_i) \frac{\nabla_\theta \pi_\theta(a_i \mid s_i)}{\pi_\theta(a_i \mid s_i)} \\
&\qquad + \sum_{a_i\in\mathcal{A}} \pi_\theta(a_i \mid s_i) \frac{\partial}{\partial \theta} \mathbb{E}_{s_i'\in\mathcal{S}_i} \left[\mathcal{R}(s_i,a_i) + \gamma_i V_i^{\pi_\theta}(s_i') \right] \\
&= \sum_{a_i\in\mathcal{A}} \pi_\theta(a_i \mid s_i) Q_i^{\pi_\theta}(s_i,a_i) \nabla_\theta \ln \pi_\theta(a_i \mid s_i) \\
&\qquad + \gamma_i \sum_{a_i\in\mathcal{A}} \pi_\theta(a_i \mid s_i) \sum_{s_i'\in\mathcal{S}_i} p_i(s_i' \mid s_i, a_i) \frac{\partial}{\partial \theta} V_i^{\pi_\theta}(s_i')
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{a_i \in \mathcal{A}} \pi_\theta(a_i \mid s_i) Q_i^{\pi_\theta}(s_i, a_i) \nabla_\theta \ln \pi_\theta(a_i \mid s_i) \\
&\qquad + \gamma_i \sum_{a_i \in \mathcal{A}} \pi_\theta(a_i \mid s_i) \sum_{s_i' \in \mathcal{S}_i} p_i(s_i' \mid s_i, a_i) \\
&\qquad\qquad \times \sum_{a_i' \in \mathcal{A}} \pi_\theta(a_i' \mid s_i') Q_i^{\pi_\theta}(s_i', a_i') \nabla_\theta \ln \pi_\theta(a_i' \mid s_i') \\
&\qquad + \gamma_i^2 \sum_{a_i' \in \mathcal{A}} \pi_\theta(a_i \mid s_i) \sum_{s_i' \in \mathcal{S}_i} p_i(s_i' \mid s_i, a_i) \\
&\qquad\qquad \times \sum_{a' \in \mathcal{A}} \pi_\theta(a_i' \mid s_i') \sum_{s_i'' \in \mathcal{S}_i} p_i(s_i'' \mid s_i', a_i') \frac{\partial}{\partial \theta} V_i^{\pi_\theta}(s_i'') \\
&= \sum_{k=0}^{\infty} \sum_{s_i' \in \mathcal{S}_i} \gamma_i^k P_{\pi_\theta}(s_i^k = s_i' \mid s_i) \sum_{a_i' \in \mathcal{A}} \pi_\theta(a_i' \mid s_i') Q_i^{\pi_\theta}(s_i', a_i') \nabla_\theta \ln \pi_\theta(a_i' \mid s_i') \\
&= \frac{1}{1-\gamma_i} \mathbb{E}_{s_i' \sim d_{s_i}^{\pi_\theta}(\cdot)} \mathbb{E}_{a_i' \sim \pi_\theta(\cdot \mid s_i')} Q_i^{\pi_\theta}(s_i', a_i') \nabla_\theta \ln \pi_\theta(a_i' \mid s_i') \\
&= \frac{1}{1-\gamma_i} \mathbb{E}_{s_i' \sim d_{s_i}^{\pi_\theta}(\cdot)} \mathbb{E}_{a_i' \sim \pi_\theta(\cdot \mid s_i')} \left(Q_i^{\pi_\theta}(s_i', a_i') - V_i^{\pi_\theta}(s_i') + V_i^{\pi_\theta}(s_i') \right) \nabla_\theta \ln \pi_\theta(a_i' \mid s_i') \\
&= \frac{1}{1-\gamma_i} \mathbb{E}_{s_i' \sim d_{s_i}^{\pi_\theta}(\cdot)} \mathbb{E}_{a_i' \sim \pi_\theta(\cdot \mid s_i')} A_i^{\pi_\theta}(s_i', a_i') \nabla_\theta \ln \pi_\theta(a_i' \mid s_i'),
\end{aligned}
$$

where the last equation follows since

$$
\sum_{a'} \pi_\theta(a'|s') V_i^{\pi_\theta}(s_i') \nabla_\theta \ln \pi_\theta(a_i' \mid s_i') = \sum_{a'} V_i^{\pi_\theta}(s_i') \nabla_\theta \pi_\theta(a'|s') = V_i^{\pi_\theta}(s_i') \nabla_\theta \sum_{a'} \pi_\theta(a'|s') = 0.
$$

Let $\mathbb{1}[\cdot]$ denote the indicator function of argument condition. We observe that under the softmax parameterization

$$
\begin{aligned}
\frac{\partial \ln \pi_\theta(a'|s')}{\partial \theta_{s,a}} &= \frac{\partial}{\partial \theta_{s,a}} \left(\theta_{s',a'} - \ln \sum_{a''} \exp(\theta_{s',a''}) \right) \\
&= \mathbb{1}(s' = s, a' = a) - \mathbb{1}(s' = s) \frac{\exp(\theta_{s',a})}{\sum_{a''} \exp(\theta_{s',a''})} \\
&= \mathbb{1}\left[s' = s\right] \left(\mathbb{1}\left[a' = a\right] - \pi_\theta\left(a|s'\right) \right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{s,a}} V_i^{\pi_\theta}(\rho_i) &= \frac{\partial}{\partial \theta_{s,a}} \mathbb{E}_{s_i \sim \rho_i} \left[V_i^{\pi_\theta}(s_i) \right] \\
&= \mathbb{E}_{s_i \sim \rho_i} \left[\frac{\partial V_i^{\pi_\theta}(s_i)}{\partial \theta_{s,a}} \right] \\
&= \frac{1}{1-\gamma_i} \mathbb{E}_{s_i \sim \rho_i} \mathbb{E}_{s_i' \sim d_{s_i}^{\pi_\theta}(\cdot)} \mathbb{E}_{a_i' \sim \pi_\theta(\cdot \mid s_i')} A_i^{\pi_\theta}(s_i', a_i') \frac{\partial \ln \pi_\theta(a_i' \mid s_i')}{\partial \theta_{s,a}} \\
&= \frac{1}{1-\gamma_i} \mathbb{E}_{s_i \sim \rho_i} \mathbb{E}_{s_i' \sim d_{s_i}^{\pi_\theta}(\cdot)} \mathbb{E}_{a_i' \sim \pi_\theta(\cdot \mid s_i')} A_i^{\pi_\theta}(s_i', a_i') \mathbb{1}\left[s_i' = s\right] \left(\mathbb{1}\left[a_i' = a\right] - \pi_\theta\left(a|s_i'\right) \right) \\
&= \frac{1}{1-\gamma_i} \mathbb{E}_{s_i \sim \rho_i} d_{s_i}^{\pi_\theta}(s) \pi_\theta(a|s) A_i^{\pi_\theta}(s, a) \\
&\qquad - \frac{\pi_\theta\left(a|s\right)}{1-\gamma_i} \mathbb{E}_{s_i \sim \rho_i} \mathbb{E}_{s_i' \sim d_{s_i}^{\pi_\theta}(\cdot)} \mathbb{1}\left[s_i' = s\right] \sum_{a_i'} \pi_\theta(a_i'|s_i') A_i^{\pi_\theta}(s_i', a_i') \\
&= \frac{1}{1-\gamma_i} d_{\rho_i}^{\pi_\theta}(s) \pi_\theta(a|s) A_i^{\pi_\theta}(s, a)
\end{aligned}
$$

B.5 LIPSCHITZ, SMOOTHNESS, AND HESSIAN LIPSCHITZ CONSTANTS

Lemma B.4. *Let $\pi_\alpha \triangleq \pi_{\boldsymbol{\theta}+\alpha\boldsymbol{u}}$, where $\boldsymbol{u}$ is a unit vector and $\tilde{V}_i(\alpha) \triangleq V_i^{\pi_\alpha}(s_i)$. If*

$$\sum_{a\in\mathcal{A}} \left| \frac{d\pi_\alpha(a|s_0)}{d\alpha}\Big|_{\alpha=0} \right| \leq C', \quad \sum_{a\in\mathcal{A}} \left| \frac{d^2\pi_\alpha(a|s_0)}{d\alpha^2}\Big|_{\alpha=0} \right| \leq C'', \quad \textit{and} \sum_{a\in\mathcal{A}} \left| \frac{d^3\pi_\alpha(a|s_0)}{d\alpha^3}\Big|_{\alpha=0} \right| \leq C''', \tag{41}$$

then we have

$$\begin{aligned} \max_{||\boldsymbol{u}||_2=1} \left| \frac{d\tilde{V}_i(\alpha)}{d\alpha}\Big|_{\alpha=0} \right| &\leq \frac{C'}{(1-\gamma_i)^2}, \\ \max_{||\boldsymbol{u}||_2=1} \left| \frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2}\Big|_{\alpha=0} \right| &\leq \frac{C''}{(1-\gamma_i)^2} + \frac{2\gamma_i C'^2}{(1-\gamma_i)^3}, \textit{ and} \\ \max_{||\boldsymbol{u}||_2=1} \left| \frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3}\Big|_{\alpha=0} \right| &\leq \frac{C'''}{(1-\gamma_i)^2} + \frac{6\gamma_i C' C''}{(1-\gamma_i)^3} + \frac{6\gamma_i^2 C'^3}{(1-\gamma_i)^4} \end{aligned} \tag{42}$$

Proof. The proof uses a similar technique to Lemma E.2 of Agarwal et al. [2020], which proves the second derivative is bounded. Here we also show the first and the third derivative is bounded. We use $\tilde{P}_i(\alpha)$ to denote the state-action transition matrix in environment i.

$$[\tilde{P}_i(\alpha)]_{(s,a)\to(s',a')} = \pi_\alpha(a'|s')\, P_i(s'|s,a) \tag{43}$$

Differentiating with respect to α, we get

$$\left[\frac{d\tilde{P}_i(\alpha)}{d\alpha}\Big|_{\alpha=0} \right]_{(s,a)\to(s',a')} = \frac{d\pi_\alpha(a'|s')}{d\alpha}\Big|_{\alpha=0} P_i(s'|s,a), \tag{44}$$

which implies that for any $\boldsymbol{x}$,

$$\left[\frac{d\tilde{P}_i(\alpha)}{d\alpha}\Big|_{\alpha=0} x \right]_{s,a} = \sum_{a',s'} \frac{d\pi_\alpha(a'|s')}{d\alpha}\Big|_{\alpha=0} P_i(s'|s,a)\, x_{a',s'} \tag{45}$$

We can bound the ℓ_∞ norm of this as

$$\begin{aligned} \max_{||\boldsymbol{u}||_2=1} \left\| \frac{d\tilde{P}_i(\alpha)}{d\alpha}\boldsymbol{x} \right\|_\infty &= \max_{s,a} \max_{||\boldsymbol{u}||_2=1} \left| \left[\frac{d\tilde{P}_i(\alpha)}{d\alpha}\Big|_{\alpha=0} \boldsymbol{x} \right]_{s,a} \right| \\ &= \max_{s,a} \max_{||\boldsymbol{u}||_2=1} \left| \sum_{a',s'} \frac{d\pi_\alpha(a'|s')}{d\alpha}\Big|_{\alpha=0} P_i(s'|s,a)\, \boldsymbol{x}_{a',s'} \right| \\ &\leq \max_{s,a} \sum_{a',s'} \left| \frac{d\pi_\alpha(a'|s')}{d\alpha}\Big|_{\alpha=0} \right| P_i(s'|s,a)\, |\boldsymbol{x}_{a',s'}| \\ &\leq \max_{s,a} \sum_{s'} P_i(s'|s,a)\, ||\boldsymbol{x}||_\infty \sum_{a'} \left| \frac{d\pi_\alpha(a'|s')}{d\alpha}\Big|_{\alpha=0} \right| \\ &\leq C' ||\boldsymbol{x}||_\infty \end{aligned} \tag{46}$$

Using the same approach, we can bound

$$\max_{||\boldsymbol{u}||_2=1} \left\| \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}\boldsymbol{x} \right\|_\infty \leq C''||\boldsymbol{x}||_\infty, \text{ and } \max_{||\boldsymbol{u}||_2=1} \left\| \frac{d^3\tilde{P}_i(\alpha)}{d\alpha^3}\boldsymbol{x} \right\|_\infty \leq C'''||\boldsymbol{x}||_\infty. \tag{47}$$

With $M(\alpha) := (\boldsymbol{I} - \gamma_i \tilde{P}_i(\alpha))^{-1}$, we re-writing the Bellman equation in the matrix form,

$$Q^\alpha(s_0, a_0) = e^T_{(s_0,a_0)}(\boldsymbol{I} - \gamma_i \tilde{P}_i(\alpha))^{-1} r = e^T_{(s_0,a_0)} M(\alpha) r. \tag{48}$$

Taking the first, second, and third derivative of $Q^\alpha(s_0, a_0)$ with respect to α,

$$\frac{dQ^\alpha(s_0, a)}{d\alpha} = \gamma_i e^T_{(s_0,a)} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r, \tag{49}$$

$$\begin{aligned}\frac{d^2 Q^\alpha(s_0, a_0)}{(d\alpha)^2} &= 2\gamma_i^2 e^T_{(s_0,a_0)} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \\ &\quad + \gamma_i e^T_{(s_0,a_0)} M(\alpha) \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2} M(\alpha) r,\end{aligned} \tag{50}$$

$$\begin{aligned}\frac{d^3 Q^\alpha(s_0, a_0)}{(d\alpha)^3} &= 6\gamma_i^3 e^T_{(s_0,a_0)} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \\ &\quad + 3\gamma_i^2 e^T_{(s_0,a_0)} M(\alpha) \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \\ &\quad + 3\gamma_i^2 e^T_{(s_0,a_0)} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2} M(\alpha) r \\ &\quad + \gamma_i e^T_{(s_0,a_0)} M(\alpha) \frac{d^3\tilde{P}_i(\alpha)}{d\alpha^3} M(\alpha) r\end{aligned} \tag{51}$$

Using $M(\alpha)\mathbf{1} = (\mathrm{I} - \gamma_i \tilde{P}_i(\alpha))^{-1}\mathbf{1} = \sum_{n=0}^{\infty} \gamma_i^n \tilde{P}_i(\alpha)^n \mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$ and inequalities (46) and (47), we have

$$\begin{aligned}\max_{||\boldsymbol{u}||_2=1} \left| \frac{dQ^\alpha(s_0, a)}{d\alpha} \Big|_{\alpha=0} \right| &\leq \left\| \gamma_i M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \right\|_\infty \\ &\leq \frac{\gamma_i C'}{(1-\gamma_i)^2},\end{aligned} \tag{52}$$

$$\begin{aligned}\max_{||\boldsymbol{u}||_2=1} \left| \frac{d^2 Q^\alpha(s_0, a_0)}{d\alpha^2} \Big|_{\alpha=0} \right| &\leq 2\gamma_i^2 \left\| M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \right\|_\infty \\ &\quad + \gamma_i \left\| M(\alpha) \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2} M(\alpha) r \right\|_\infty\end{aligned} \tag{53}$$

$$\leq \frac{2\gamma_i^2 C'^2}{(1-\gamma_i)^3} + \frac{\gamma_i C''}{(1-\gamma_i)^2} \tag{54}$$

$$\begin{aligned}\max_{||\boldsymbol{u}||_2=1} \left| \frac{d^3 Q^\alpha(s_0, a_0)}{d\alpha^3} \Big|_{\alpha=0} \right| &\leq 6\gamma_i^3 \left\| M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \right\|_\infty \\ &\quad + 3\gamma_i^2 \left\| M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2} M(\alpha) r \right\|_\infty \\ &\quad + 3\gamma_i^2 \left\| M(\alpha) \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2} M(\alpha) \frac{d\tilde{P}_i(\alpha)}{d\alpha} M(\alpha) r \right\|_\infty \\ &\quad + \gamma_i \left\| M(\alpha) \frac{d^3\tilde{P}_i(\alpha)}{d\alpha^3} M(\alpha) r \right\|_\infty \\ &\leq \frac{6\gamma_i^3 C'^3}{(1-\gamma_i)^4} + \frac{3\gamma_i^2 C' C''}{(1-\gamma_i)^3} + \frac{3\gamma_i^2 C' C''}{(1-\gamma_i)^3} + \frac{\gamma_i C'''}{(1-\gamma_i)^2}\end{aligned}$$

$$= \frac{6\gamma_i^3 C'^3}{(1-\gamma_i)^4} + \frac{6\gamma_i^2 C' C''}{(1-\gamma_i)^3} + \frac{\gamma_i C'''}{(1-\gamma_i)^2} \tag{55}$$

By the definition of $\tilde{V}_i(\alpha)$,

$$\tilde{V}_i(\alpha) = \sum_a \pi_\alpha \left(a|s_0\right) Q^\alpha \left(s_0, a\right). \tag{56}$$

Taking the first derivative of $\tilde{V}_i(\alpha)$ with respect to α,

$$\frac{d\tilde{V}_i(\alpha)}{d\alpha} = \sum_a \frac{d\pi_\alpha \left(a|s_0\right)}{d\alpha} Q_i^\alpha \left(s_0, a\right) + \sum_a \pi_\alpha \left(a|s_0\right) \frac{dQ_i^\alpha \left(s_0, a\right)}{d\alpha}. \tag{57}$$

Taking the second derivative of $\tilde{V}_i(\alpha)$ with respect to α,

$$\begin{aligned} \frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2} &= \sum_a \frac{d^2\pi_\alpha \left(a|s_0\right)}{d\alpha^2} Q_i^\alpha \left(s_0, a\right) + 2\sum_a \frac{d\pi_\alpha \left(a|s_0\right)}{d\alpha} \frac{dQ_i^\alpha \left(s_0, a\right)}{d\alpha} \\ &+ \sum_a \pi_\alpha \left(a|s_0\right) \frac{d^2 Q_i^\alpha \left(s_0, a\right)}{d\alpha^2}. \end{aligned} \tag{58}$$

Taking the third derivative of $\tilde{V}_i(\alpha)$ with respect to α,

$$\begin{aligned} \frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3} &= \sum_a \frac{d^3\pi_\alpha \left(a|s_0\right)}{d\alpha^3} Q^\alpha \left(s_0, a\right) + 3\sum_a \frac{d^2\pi_\alpha \left(a|s_0\right)}{d\alpha^2} \frac{dQ^\alpha \left(s_0, a\right)}{d\alpha} \\ &+ 3\sum_a \frac{d\pi_\alpha \left(a|s_0\right)}{d\alpha} \frac{d^2 Q^\alpha \left(s_0, a\right)}{d\alpha^2} + \sum_a \pi_\alpha \left(a|s_0\right) \frac{d^3 Q^\alpha \left(s_0, a\right)}{d\alpha^3}. \end{aligned} \tag{59}$$

Finally, we have

$$\max_{||\boldsymbol{u}||_2=1} \left| \frac{d\tilde{V}_i(\alpha)}{d\alpha} \bigg|_{\alpha=0} \right| \leq \frac{C'}{1-\gamma_i} + \frac{\gamma_i C'}{(1-\gamma_i)^2} = \frac{C'}{(1-\gamma_i)^2} \tag{60}$$

,

$$\begin{aligned} \max_{||\boldsymbol{u}||_2=1} \left| \frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2} \bigg|_{\alpha=0} \right| &\leq \frac{C''}{1-\gamma_i} + \frac{2C'^2}{(1-\gamma_i)^2} + \left(\frac{2\gamma_i C'^2}{(1-\gamma_i)^3} + \frac{\gamma_i C''}{(1-\gamma_i)^2} \right) \\ &= \frac{C''}{(1-\gamma_i)^2} + \frac{2\gamma_i C'^2}{(1-\gamma_i)^3} \end{aligned} \tag{61}$$

, and

$$\begin{aligned} \max_{||\boldsymbol{u}||_2=1} \left| \frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3} \bigg|_{\alpha=0} \right| &\leq \frac{C'''}{1-\gamma_i} + \frac{3\gamma_i C' C''}{(1-\gamma_i)^2} + 3C' \left(\frac{2\gamma_i^2 C'^2}{(1-\gamma_i)^3} + \frac{\gamma_i C''}{(1-\gamma_i)^2} \right) \\ &+ \frac{6\gamma_i^3 C'^3}{(1-\gamma_i)^4} + \frac{6\gamma_i^2 C' C''}{(1-\gamma_i)^3} + \frac{\gamma_i C'''}{(1-\gamma_i)^2} \\ &= \frac{C'''}{1-\gamma_i} + \frac{\gamma_i (6C' C'' + C''')}{(1-\gamma_i)^2} + \frac{6\gamma_i^2 (C'^3 + C' C'')}{(1-\gamma_i)^3} + \frac{6\gamma_i^3 C'^3}{(1-\gamma_i)^4} \\ &= \frac{C'''}{(1-\gamma_i)^2} + \frac{6\gamma_i C' C''}{(1-\gamma_i)^3} + \frac{6\gamma_i^2 C'^3}{(1-\gamma_i)^4} \end{aligned} \tag{62}$$

□

Lemma B.5. *Under the tabular softmax policy, $V_i^{\pi_\theta}(\mu)$ is Lipschitz, has a Lipschitz gradient and a Lipschtz Hessian for all i and μ, i.e.*

$$
\begin{aligned}
&||V_i^{\pi_{\theta'}}(\mu) - V_i^{\pi_{\theta''}}(\mu)|| \leq \frac{2}{(1-\gamma_i)^2}||\theta' - \theta''||, \\
&||\nabla_{\theta'} V_i^{\pi_{\theta'}}(\mu) - \nabla_{\theta''} V_i^{\pi_{\theta''}}(\mu)|| \leq \frac{8}{(1-\gamma_i)^3}||\theta' - \theta''||, \text{ and} \\
&||\nabla^2_{\theta'} V_i^{\pi_{\theta'}}(\mu) - \nabla^2_{\theta''} V_i^{\pi_{\theta''}}(\mu)|| \leq \frac{48}{(1-\gamma_i)^4}||\theta' - \theta''||.
\end{aligned} \tag{63}
$$

Proof. To show a function is Lipschitz, we show the derivative of the Hessian with respect to θ is bounded. Under the softmax parameterization, we have

$$
\nabla_{\theta_s} \pi_\theta(a|s) = \pi_\theta(a|s)\left(e_a - \pi(\cdot|s)\right), \tag{64}
$$

$$
\nabla^2_{\theta_s} \pi_\theta(a|s) = \pi_\theta(a|s)\left(e_a e_a^\top - e_a \pi(\cdot|s)^\top - \pi(\cdot|s) e_a^\top + 2\pi(\cdot|s)\pi(\cdot|s)^\top - \operatorname{diag}(\pi(\cdot|s))\right), \tag{65}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{s,a'}} \nabla^2_{\theta_s} \pi_\theta(a|s) = {} & \pi_\theta(a|s)(\mathbf{1}(a=a') - \pi_\theta(a'|s))\big(e_a e_a^\top - e_a \pi(\cdot|s)^\top - \pi(\cdot|s) e_a^\top \\
& +2\pi(\cdot|s)\pi(\cdot|s)^\top - \operatorname{diag}(\pi(\cdot|s))\big) \\
& + \pi_\theta(a|s)(-e_a \pi_\theta(a'|s) e_{a'}^T + e_a \pi_\theta(a'|s)\pi_\theta(\cdot|s)^T - e_{a'}\pi_\theta(a'|s) e_a^T \\
& + \pi_\theta(\cdot|s))\pi_\theta(a'|s) e_a^T + 4\pi_\theta(\cdot|s)\pi_\theta(a'|s) e_{a'}^T - 4\pi_\theta(\cdot|s)\pi_\theta\pi_\theta(\cdot|s)^T \\
& + \operatorname{diag}(\pi_\theta(a'|s) e_a) - \operatorname{diag}(\pi_\theta(a'|s)\pi_\theta(\cdot|s)^T))
\end{aligned} \tag{66}
$$

where e_a is a vector with all 0 and 1 at action a. Then, for any s,

$$
\begin{aligned}
\sum_{a\in\mathcal{A}} \left| \frac{d\pi_\alpha(a|s)}{d\alpha}\Big|_{\alpha=0} \right| &\leq \sum_{a\in\mathcal{A}} \left| \boldsymbol{u}^T \nabla_{\theta+\alpha\boldsymbol{u}} \pi_\alpha(a|s)\big|_{\alpha=0} \right| \\
&\leq \sum_{a\in\mathcal{A}} \pi_\theta(a|s) \left| \boldsymbol{u}_s^T e_a - \boldsymbol{u}_s^T \pi(\cdot|s) \right| \\
&\leq \max_{a\in\mathcal{A}} \left(\left|\boldsymbol{u}_s^T e_a\right| + \left|\boldsymbol{u}_s^T \pi(\cdot|s)\right| \right) \leq 2,
\end{aligned} \tag{67}
$$

$$
\begin{aligned}
\sum_{a\in\mathcal{A}} \left| \frac{d^2\pi_\alpha(a|s)}{d\alpha^2}\Big|_{\alpha=0} \right| &\leq \sum_{a\in\mathcal{A}} \left| \boldsymbol{u}^T \nabla^2_{\theta+\alpha\boldsymbol{u}} \pi_\alpha(a|s)\big|_{\alpha=0} \boldsymbol{u} \right| \\
&\leq \max_{a\in\mathcal{A}} \big(\left|\boldsymbol{u}_s^T e_a e_a^T \boldsymbol{u}_s\right| + \left|\boldsymbol{u}_s^T e_a \pi(\cdot|s)^T \boldsymbol{u}_s\right| + \left|\boldsymbol{u}_s^T \pi(\cdot|s) e_a^T \boldsymbol{u}_s\right| \\
&\quad +2\left|u_s^\top \pi(\cdot|s)\pi(\cdot|s)^\top u_s\right| + \left|u_s^\top \operatorname{diag}(\pi(\cdot|s)) u_s\right| \big) \\
&\leq 6.
\end{aligned} \tag{68}
$$

Similarly,

$$
\begin{aligned}
\sum_{a\in\mathcal{A}} \left| \frac{d^3\pi_\alpha(a|s)}{d\alpha^3}\Big|_{\alpha=0} \right| &\leq \sum_{a\in\mathcal{A}} \sum_{a'\in\mathcal{A}} \left| \boldsymbol{u}_{a'} \boldsymbol{u}^T \nabla^3_{\theta+\alpha\boldsymbol{u}} \pi_\alpha(a|s)\big|_{\alpha=0} \boldsymbol{u} \right| \\
&\leq 26
\end{aligned} \tag{69}
$$

Then we can use Lemma B.4 with $C' = 2, C'' = 6, C''' = 26$, and get

$$
\max_{||\boldsymbol{u}||_2=1} \left| \frac{d\tilde{V}_i(\alpha)}{d\alpha}\Big|_{\alpha=0} \right| \leq \frac{2}{(1-\gamma_i)^2},
$$

$$\max_{||\boldsymbol{u}||_2=1}\left|\frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2}\Big|_{\alpha=0}\right| \leq \frac{6}{(1-\gamma_i)^2}+\frac{8\gamma_i}{(1-\gamma_i)^3} \leq \frac{8}{(1-\gamma_i)^3},$$
$$\max_{||\boldsymbol{u}||_2=1}\left|\frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3}\Big|_{\alpha=0}.\right| \leq \frac{26}{(1-\gamma_i)^2}+\frac{72\gamma_i}{(1-\gamma_i)^3}+\frac{48\gamma_i^2}{(1-\gamma_i)^4} \leq \frac{48}{(1-\gamma_i)^4} \tag{70}$$

This is equivalent to

$$\begin{aligned}
&||V_i^{\pi_{\theta'}}(\mu)-V_i^{\pi_{\theta''}}(\mu)|| \leq \frac{2}{(1-\gamma_i)^2}||\theta'-\theta''||,\\
&||\nabla V_i^{\pi_{\theta'}}(\mu)-\nabla V_i^{\pi_{\theta''}}(\mu)|| \leq \frac{8}{(1-\gamma_i)^3}||\theta'-\theta''||, \text{ and}\\
&||\nabla^2 V_i^{\pi_{\theta'}}(\mu)-\nabla^2 V_i^{\pi_{\theta''}}(\mu)|| \leq \frac{48}{(1-\gamma_i)^4}||\theta'-\theta''||.
\end{aligned} \tag{71}$$

□

Lemma B.6. *The cross entropy regularizer is Lipschitz, has a Lipschitz gradient and a Lipschtz Hessian, i.e.*

$$\begin{aligned}
&||\lambda RE(\pi'_\theta)-\lambda RE(\pi''_\theta)|| \leq \lambda(\frac{1}{\sqrt{|\mathcal{A}|}}+1)||\theta'-\theta''||,\\
&||\nabla_{\theta'}\lambda RE(\pi'_\theta)-\nabla_{\theta''}\lambda RE(\pi''_\theta)|| \leq \frac{2\lambda}{|\mathcal{S}|}||\theta'-\theta''||, \textit{ and}\\
&||\nabla^2_{\theta'}\lambda RE(\pi'_\theta)-\nabla^2_{\theta''}\lambda RE(\pi''_\theta)|| \leq \frac{6\lambda}{|\mathcal{S}|}||\theta'-\theta''||.
\end{aligned} \tag{72}$$

Proof. Define

$$\zeta(\theta) \triangleq -\lambda \text{RE}(\pi_\theta) = \frac{\lambda}{|\mathcal{S}||\mathcal{A}|}\sum_{s,a}\log\pi_\theta(a|s). \tag{73}$$

We have

$$\begin{aligned}
\nabla_{\theta_s}\zeta(\theta) &= \frac{\lambda}{|\mathcal{S}|}(\frac{1}{|\mathcal{A}|}\mathbf{1}-\pi_\theta(\cdot|s)),\\
\nabla^2_{\theta_s}\zeta(\theta) &= \frac{\lambda}{|\mathcal{S}|}(-\text{diag}(\pi_\theta(\cdot|s))+\pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T),\\
\frac{\partial}{\partial\theta_{s,a'}}\nabla^2_{\theta_s}\zeta(\theta) &= \frac{\lambda}{|\mathcal{S}|}(-\pi_\theta(a'|s)e_{a'}e_{a'}^T+\pi_\theta(a'|s)\text{diag}(\pi_\theta(\cdot|s))\\
&\quad +2\pi_\theta(a'|s)\pi_\theta(\cdot|s)e_{a'}^T-2\pi_\theta(a'|s)\pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T).
\end{aligned} \tag{74}$$

Now we can bound the norm of the gradient, the norm of the Hessian, and the norm of the third level gradient.

$$\begin{aligned}
||\nabla_\theta\zeta(\theta)|| &= \sum_s||\nabla_{\theta_s}\zeta(\theta)||\\
&\leq \frac{\lambda}{|\mathcal{S}|}\sum_s||\frac{1}{|\mathcal{A}|}\mathbf{1}-\pi_\theta(\cdot|s)||\\
&\leq \frac{\lambda}{|\mathcal{S}|}\sum_s\left(||\frac{1}{|\mathcal{A}|}\mathbf{1}||+||\pi_\theta(\cdot|s)||\right)\\
&\leq \frac{\lambda}{|\mathcal{S}|}\sum_s\left(\frac{1}{\sqrt{|\mathcal{A}|}}+1\right)
\end{aligned}$$

$$\leq \lambda(\frac{1}{\sqrt{|\mathcal{A}|}}+1). \tag{75}$$

For any vector $u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with $||u||_2 = 1$,

$$\begin{aligned}
\left|u^T \nabla_\theta^2 \zeta(\theta) u\right| &= \left|\sum_s u_s^T \nabla_{\theta_s}^2 \zeta(\theta) u_s\right| \\
&\leq \frac{\lambda}{|\mathcal{S}|} \sum_s \left|u_s^T \mathrm{diag}(\pi_\theta(\cdot|s)) u_s - u_s^T \pi_\theta(\cdot|s) \pi_\theta(\cdot|s)^T u_s\right| \\
&\leq \frac{2\lambda}{|\mathcal{S}|} \sum_s ||u_s||_\infty^2 \\
&\leq \frac{2\lambda}{|\mathcal{S}|} ||u||_2^2 \\
&\leq \frac{2\lambda}{|\mathcal{S}|},
\end{aligned} \tag{76}$$

where the first equality follows since $\nabla_{\theta_{s'}} \nabla_{\theta_{s''}} \zeta(\theta) = 0$, $\forall s' \neq s''$. Using this method, we can further get

$$\begin{aligned}
\left|\sum_{s',a'} u_{s',a'} u^T \nabla_\theta^2 \zeta(\theta) u\right| &= \left|\sum_s \sum_{a'} u_{s,a'} u_s^T \nabla_{\theta_s}^2 \zeta(\theta) u_s\right| \\
&\leq \frac{\lambda}{|\mathcal{S}|} \sum_s \Bigg| - \sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s) e_{a'} e_{a'}^T u_s \\
&\qquad + \sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s) \mathrm{diag}(\pi_\theta(\cdot|s)) u_s \\
&\qquad + 2 \sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s) \pi_\theta(\cdot|s) e_{a'}^T u_s \\
&\qquad - 2 \sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s) \pi_\theta(\cdot|s) \pi_\theta(\cdot|s)^T u_s \Bigg| \\
&\leq \frac{6\lambda}{|\mathcal{S}|} \sum_s ||u_s||_\infty^3 \\
&\leq \frac{6\lambda}{|\mathcal{S}|} ||u||_3^3 \\
&\leq \frac{6\lambda}{|\mathcal{S}|},
\end{aligned} \tag{77}$$

where the last inequality uses $||u||_3 \leq ||u||_2$. This implies that $\zeta(\theta)$ is $\lambda(\frac{1}{\sqrt{|\mathcal{A}|}}+1)$-Lipschitz, $\frac{2\lambda}{|\mathcal{S}|}$-smooth, and has $\frac{6\lambda}{|\mathcal{S}|}$-Lipschitz Hessian. □

C EXPERIMENTS DETAILS

C.1 DRONE EXPERIMENTS

The framework used for the drone experiment is PEDRA [PED], a 3D realistically stimulated drone navigation platform powered by Unreal Engine. In the simulated environment, a drone agent is equipped with a front-facing camera, and can implement actions to control its flight. To model the problem as an MDP, the state is represented by the monocular RGB images captured by the camera of the drone, which has dimension $103(height) \times 103(width) \times 3(color)$. There are a total number of 25 actions, corresponding to the drone controlling the yaw and pitch by various angles. Reward is calculated based on dynamic windowing of the simulated depth map, and is designed to encourage the drone to stay away from obstacles, as used in Anwar and Raychowdhury [2018].

We select 4 indoor environments on the PEDRA platform: indoor long, indoor cloud, indoor frogeyes, and indoor pyramid. They contain widely different lighting conditions, wall colors, furniture objects, and hallway structures (Fig. 1).

Figure 1: Environments used in drone navigation.

Every agent uses a 5-layer neural network as the function approximation. The exact architecture is shown in Figure 2. The agents use the ADAM optimizer with a constant step size of 1e-4, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Communication happens every episode, and follows a cyclic communication graph (ring graph). The same discount factor $\gamma = 0.99$ is used by all agents. The weight of the cross entropy regularizer is chosen to be 0.03. We conducted three sets of experiments, where the local gradient g_i^k is estimated using REINFORCE, advantage actor-critic (A2C), and proximal policy optimization (PPO), respectively. The discounted cumulative reward is estimated by the every visit Monte-Carlo method in all experiments. For PPO, we choose the clipping parameter ϵ to be 0.2. We train the agents for 4000 episodes in all experiments. Using two RTX2080 GPUs, each set of experiments takes about 25 hours to complete.

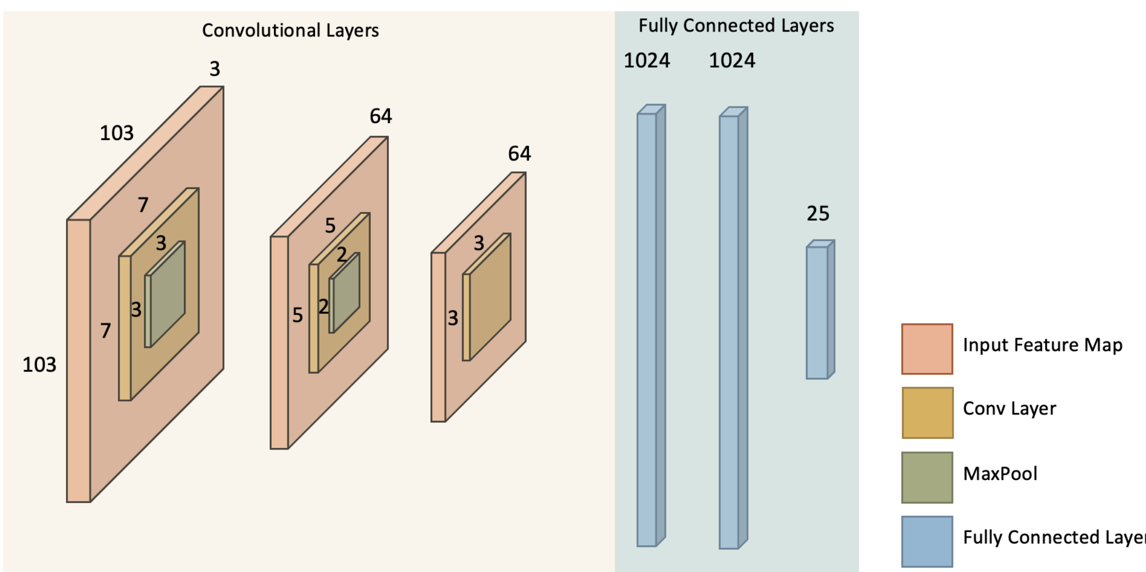

Figure 2: Network architecture for drone experiments

References

https://icsrl.ece.gatech.edu/pedra.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. volume 125 of *Proceedings of Machine Learning Research*, pages 64–66, 2020.

Malik Aqeel Anwar and Arijit Raychowdhury. Navren-rl: Learning to fly in real environment via end-to-end deep reinforcement learning using monocular images. In *2018 25th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pages 1–6. IEEE, 2018.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.