A. Proof of Lemma 3

Lemma. *For Reparameterizable RL, given assumptions 1, 2, and 3, the empirical reward R defined in (10), as a function of the parameter θ, has a Lipschitz constant of*

$$\beta = \sum_{t=0}^{T} \gamma^t L_r L_{t_2} L_{\pi 2} \frac{\nu^t - 1}{\nu - 1}$$

where $\nu = L_{t1} + L_{t2} L_{\pi 1}$.

Proof. Let's denote $s'_t = s_t(\theta')$, and $s_t = s_t(\theta)$. We start by investigating the policy function across different time steps:

$$\begin{aligned} &\|\pi(s'_t;\theta') - \pi(s_t;\theta)\| \\ &= \|\pi(s'_t;\theta') - \pi(s_t;\theta') + \pi(s_t;\theta') - \pi(s_t;\theta)\| \\ &\le \|\pi(s'_t;\theta') - \pi(s_t;\theta')\| + \|\pi(s_t;\theta') - \pi(s_t;\theta)\| \\ &\le L_{\pi 1}\|s'_t - s_t\| + L_{\pi 2}\|\theta' - \theta\| \end{aligned} \tag{17}$$

The first inequality is the triangle inequality, and the second is from our Lipschitz assumption 2.

If we look at the change of states as the episode proceeds:

$$\begin{aligned} &\|s'_t - s_t\| \\ &= \|\mathcal{T}(s'_{t-1}, \pi(s'_{t-1};\theta'), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s_{t-1};\theta), \xi_{t-1})\| \\ &\le \|\mathcal{T}(s'_{t-1}, \pi(s'_{t-1};\theta'), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s'_{t-1};\theta'), \xi_{t-1})\| \\ &+ \|\mathcal{T}(s_{t-1}, \pi(s'_{t-1};\theta'), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s_{t-1};\theta), \xi_{t-1})\| \\ &\le L_{t1}\|s'_{t-1} - s_{t-1}\| + L_{t2}\|\pi(s'_{t-1};\theta') - \pi(s_{t-1};\theta)\| \end{aligned} \tag{18}$$

Now combine both (17) and (18),

$$\begin{aligned} &\|s'_t - s_t\| \\ &\le L_{t1}\|s'_{t-1} - s_{t-1}\| \\ &+ L_{t2}(L_{\pi 1}\|s'_{t-1} - s_{t-1}\| + L_{\pi 2}\|\theta' - \theta\|) \\ &\le (L_{t1} + L_{t2}L_{\pi 1})\|s'_{t-1} - s_{t-1}\| + L_{t2}L_{\pi 2}\|\theta' - \theta\| \end{aligned}$$

In the initialization, we know $s'_0 = s_0$ since the initialization process does not involve any computation using the parameter θ in the policy π.

By recursion, we get

$$\begin{aligned} \|s'_t - s_t\| &\le L_{t_2}L_{\pi 2}\|\theta' - \theta\| \sum_{t=0}^{t-1} (L_{t1} + L_{t2}L_{\pi 1})^t \\ &= L_{t_2}L_{\pi 2}\frac{\nu^t - 1}{\nu - 1}\|\theta' - \theta\| \end{aligned}$$

where $\nu = L_{t1} + L_{t2}L_{\pi 1}$.

By assumption 3, $r(s)$ is L_r-Lipschitz, so

$$\begin{aligned} \|r(s'_t) - r(s_t)\| &\le L_r\|s'_t - s_t\| \\ &\le L_r L_{t_2} L_{\pi 2}\frac{\nu^t - 1}{\nu - 1}\|\theta' - \theta\| \end{aligned}$$

So the reward

$$\begin{aligned} &|R(s') - R(s)| = |\sum_{t=0}^{T}\gamma^t r(s'_t) - \sum_{t=0}^{T}\gamma^t r(s_t)| \\ &\le |\sum_{t=0}^{T}\gamma^t (r(s'_t) - r(s_t))| \le \sum_{t=0}^{T}\gamma^t |r(s'_t) - r(s_t))| \\ &\le \sum_{t=0}^{T}\gamma^t L_r L_{t_2} L_{\pi 2}\frac{\nu^t - 1}{\nu - 1}\|\theta' - \theta\| = \beta\|\theta' - \theta\| \end{aligned}$$

□

B. Proof of Lemma 6

Lemma. *In reparameterizable RL, suppose the initialization function $\mathcal{I}'$ in the test environment satisfies $\|(\mathcal{I}' - \mathcal{I})(\xi)\| \le \delta$, and the transition function is the same for both training and testing environment. If assumptions (1), (2), and (3) hold then*

$$|\mathbb{E}_\xi[R(s(\xi;\mathcal{I}'))] - \mathbb{E}_\xi[R(s(\xi;\mathcal{I}))]| \le \sum_{t=0}^{T}\gamma^t L_r (L_{t1} + L_{t2}L_{\pi 1})^t \delta$$

Proof. Denote the states at time t with $\mathcal{I}'$ as the initialization function as s'_t. Again we look at the difference between s'_t and s_t. By triangle inequality and assumptions 1 and 2,

$$\begin{aligned} &\|s'_t - s_t\| \\ &= \|\mathcal{T}(s'_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s_{t-1}), \xi_{t-1})\| \\ &\le \|\mathcal{T}(s'_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s'_{t-1}), \xi_{t-1})\| \\ &+ \|\mathcal{T}(s_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s_{t-1}), \xi_{t-1})\| \\ &\le L_{t1}\|s'_{t-1} - s_{t-1}\| + L_{t2}\|\pi(s'_{t-1}) - \pi(s_{t-1})\| \\ &\le L_{t1}\|s'_{t-1} - s_{t-1}\| + L_{t2}L_{\pi 1}\|s'_{t-1} - s_{t-1}\| \\ &= (L_{t1} + L_{t2}L_{\pi 1})\|s'_{t-1} - s_{t-1}\| \\ &\le (L_{t1} + L_{t2}L_{\pi 1})^t\|s'_0 - s_0\| \\ &\le (L_{t1} + L_{t2}L_{\pi 1})^t\delta \end{aligned}$$

where the last inequality is due to the assumption that

$$\|s'_0 - s_0\| = \|\mathcal{I}'(\xi) - \mathcal{I}(\xi)\| \le \delta$$

Also since $r(s)$ is also Lipschitz,

$$|R(s') - R(s)| = |\sum_{t=0}^{T} \gamma^t r(s'_t) - \sum_{t=0}^{T} \gamma^t r(s_t)|$$
$$\leq \sum_{t=0}^{T} \gamma^t |r(s'_t) - r(s_t)| \leq \sum_{t=0}^{T} \gamma^t L_r \|s'_t - s_t\|$$
$$\leq L_r \delta \sum_{t=0}^{T} \gamma^t (L_{t1} + L_{t2} L_{\pi 1})^t$$

The argument above holds for any given random input ξ, so

$$|\mathbb{E}_\xi[R(s'(\xi)] - \mathbb{E}_\xi[R(s(\xi)]|$$
$$\leq \left| \int_\xi (R(s'(\xi)) - R(s(\xi))) \right|$$
$$\leq \int_\xi |R(s'(\xi)) - R(s(\xi))|$$
$$\leq L_r \delta \sum_{t=0}^{T} \gamma^t (L_{t1} + L_{t2} L_{\pi 1})^t$$

□

C. Proof of Lemma 7

Lemma. *In reparameterizable RL, suppose the transition $\mathcal{T}'$ in the test environment satisfies $\forall x, y, z, \|(\mathcal{T}' - \mathcal{T})(x, y, z)\| \leq \delta$, and the initialization is the same for both the training and testing environment. If assumptions (1), (2) and (3) hold then*

$$|\mathbb{E}_\xi[R(s(\xi; \mathcal{T}'))] - \mathbb{E}_\xi[R(s(\xi; \mathcal{T}))]| \leq \sum_{t=0}^{T} \gamma^t L_r \frac{1 - \nu^t}{1 - \nu} \delta \tag{19}$$

where $\nu = L_{t1} + L_{t2} L_{\pi 1}$

Proof. Again let's denote the state at time t with the new transition function $\mathcal{T}'$ as s'_t, and the state at time t with the original transition function $\mathcal{T}$ as s_t, then

$$\|s'_t - s_t\|$$
$$= \|\mathcal{T}'(s'_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s_{t-1}), \xi_{t-1})\|$$
$$\leq \|\mathcal{T}'(s'_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}'(s_{t-1}, \pi(s_{t-1}), \xi_{t-1})\| + \|\mathcal{T}'(s_{t-1}, \pi(s_{t-1}), \xi_{t-1}) - \mathcal{T}(s_{t-1}, \pi(s_{t-1}), \xi_{t-1})\|$$
$$\leq \|\mathcal{T}'(s'_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}'(s_{t-1}, \pi(s'_{t-1}), \xi_{t-1})\|$$
$$+ \|\mathcal{T}'(s_{t-1}, \pi(s'_{t-1}), \xi_{t-1}) - \mathcal{T}'(s_{t-1}, \pi(s_{t-1}), \xi_{t-1})\| + \delta$$
$$\leq L_{t1} \|s'_{t-1} - s_{t-1}\| + L_{t2} \|\pi(s'_{t-1}) - \pi(s_{t-1})\| + \delta$$
$$\leq L_{t1} \|s'_{t-1} - s_{t-1}\| + L_{t2} L_{\pi 1} \|s'_{t-1} - s_{t-1}\| + \delta$$
$$= (L_{t1} + L_{t2} L_{\pi 1}) \|s'_{t-1} - s_{t-1}\| + \delta$$

Again we have the initialization condition

$$s'_0 = s_0$$

since the initialization procedure $\mathcal{I}$ stays the same. By recursion we have

$$\|s'_t - s_t\| \leq \delta \sum_{t=0}^{t-1} (L_{t1} + L_{t2} L_{\pi 1})^t \tag{20}$$

By assumption 3,

$$|R(s') - R(s)| = |\sum_{t=0}^{T} \gamma^t r(s'_t) - \sum_{t=0}^{T} \gamma^t r(s_t)|$$
$$\leq \sum_{t=0}^{T} \gamma^t |r(s'_t) - r(s_t)| \leq \sum_{t=0}^{T} \gamma^t L_r \|s'_t - s_t\|$$
$$\leq L_r \delta \sum_{t=0}^{T} \gamma^t \left(\sum_{k=0}^{t-1} (L_{t1} + L_{t2} L_{\pi 1})^k \right)$$
$$\leq L_r \delta \sum_{t=0}^{T} \gamma^t \frac{\nu^t - 1}{\nu - 1}$$

where $\nu = L_{t1} + L_{t2} L_{\pi 1}$. Again the argument holds for any given random input ξ, so

$$|\mathbb{E}_\xi[R(s'(\xi)] - \mathbb{E}_\xi[R(s(\xi)]|$$
$$\leq \left| \int_\xi (R(s'(\xi)) - R(s(\xi))) \right|$$
$$\leq \int_\xi |R(s'(\xi)) - R(s(\xi))|$$
$$\leq L_r \delta \sum_{t=0}^{T} \gamma^t \frac{\nu^t - 1}{\nu - 1}$$

□

D. Proof of Theorem 1

Theorem. *In **reparameterizable RL**, suppose the transition $\mathcal{T}'$ in the test environment satisfies $\forall x, y, z, \|(\mathcal{T}' - \mathcal{T})(x, y, z)\| \leq \zeta$, and suppose the initialization function $\mathcal{I}'$ in the test environment satisfies $\forall \xi, \|(\mathcal{I}' - \mathcal{I})(\xi)\| \leq \epsilon$. If assumptions (1), (2) and (3) hold, the peripheral random variables ξ^i for each episode are i.i.d., and the reward is bounded $|R(\boldsymbol{s})| \leq c/2$, then with probability at least $1 - \delta$, for all policy $\pi \in \Pi$,*

$$|\mathbb{E}_\xi[R(\boldsymbol{s}(\xi; \pi, \mathcal{T}', \mathcal{I}'))] - \frac{1}{n} \sum_i R(\boldsymbol{s}(\xi^i; \pi, \mathcal{T}, \mathcal{I}))|$$
$$\leq Rad(R_{\pi, \mathcal{T}, \mathcal{I}}) + L_r \zeta \sum_{t=0}^{T} \gamma^t \frac{\nu^t - 1}{\nu - 1} + L_r \epsilon \sum_{t=0}^{T} \gamma^t \nu^t$$
$$+ O\left(c \sqrt{\frac{\log(1/\delta)}{n}} \right)$$

where $\nu = L_{t1} + L_{t2}L_{\pi 1}$, *and*

$$Rad(R_{\pi,\mathcal{T},\mathcal{I}}) = \mathbb{E}_\xi \mathbb{E}_\sigma \left[\sup_\pi \frac{1}{n} \sum_{i=1}^n \sigma_i R(\boldsymbol{s}^i(\xi^i; \pi, \mathcal{T}, \mathcal{I})) \right]$$

is the Rademacher complexity of $R(\boldsymbol{s}(\xi; \pi, \mathcal{T}, \mathcal{I}))$ *under the training transition* $\mathcal{T}$*, the training initialization* $\mathcal{I}$*, and* n *is the number if training episodes.*

Proof. Note

$$
\begin{aligned}
&|\frac{1}{n}\sum_i R(s(\xi^i;\pi,\mathcal{T},\mathcal{I})) - \mathbb{E}_\xi[R(s(\xi;\pi,\mathcal{T}',\mathcal{I}'))]| \\
&\le |\frac{1}{n}\sum_i R(s(\xi^i;\pi,\mathcal{T},\mathcal{I})) - \mathbb{E}_\xi[R(s(\xi;\pi,\mathcal{T},\mathcal{I}))]| \\
&+ |\mathbb{E}_\xi[R(s(\xi;\pi,\mathcal{T},\mathcal{I}))] - \mathbb{E}_\xi[R(s(\xi;\pi,\mathcal{T}',\mathcal{I}))]| \\
&+ |\mathbb{E}_\xi[R(s(\xi;\pi,\mathcal{T}',\mathcal{I}))] - \mathbb{E}_\xi[R(s(\xi;\pi,\mathcal{T}',\mathcal{I}'))]|
\end{aligned}
$$

Then theorem 1 is a direct consequence of Lemma 2, Lemma 6, and Lemma 7. □