A New Representation of Successor Features for Transfer across Dissimilar Environments (Supplementary Materials)

1. Theoretical Proofs

We show our theoretical analysis for both cases of finite $\mathbb{X}$ and infinite $\mathbb{X}$.

1.1. Finite $\mathbb{X}$

Theorem 1

Let $\mathbf{S}_i$ and $\mathbf{S}_j$ be two different source environments with dissimilar transition dynamics p_i and p_j respectively. Let $\delta_{ij} \triangleq \max_{s,a} |r_i(s,a) - r_j(s,a)|$, where $r_i(.,.)$ and $r_j(.,.)$ are the reward functions of environment $\mathbf{S}_i$ and $\mathbf{S}_j$ respectively. We denote π_i^* and π_j^* as optimal policies in $\mathbf{S}_i$ and $\mathbf{S}_j$. It can be shown that the difference of their action-value functions is upper bounded as:

$$\begin{aligned} Q_i^{\pi_i^*}(s,a) - Q_i^{\pi_j^*}(s,a) \leq & \frac{2\delta_{ij}}{1-\gamma} \\ & + \frac{\gamma \Big|\Big| \mathbf{P}_i(s,a) - \mathbf{P}_j(s,a) \Big|\Big|}{(1-\gamma)} \\ & \times \frac{\Big(\Big|\Big| \mathbf{Q}_i^i - \mathbf{Q}_j^j \Big|\Big| + \Big|\Big| \mathbf{Q}_j^j - \mathbf{Q}_i^j \Big|\Big| \Big)}{(1-\gamma)}, \quad (1) \end{aligned}$$

where $Q_i^{\pi_k^*}$ shows the action-value function in environment $\mathbf{S}_i$ by following an optimal policy that is learned in the environment $\mathbf{S}_k \in \{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$. We also define $\mathbf{P}_i(s,a) = [p_i(s'|s,a), ...]_{\forall s' \in \mathcal{S}}$, $\mathbf{P}_j(s,a) = [p_j(s'|s,a), ...]_{\forall s' \in \mathcal{S}}$, $\mathbf{Q}_i^i = [\max\limits_{b \in \mathcal{A}} Q_i^{\pi_i^*}(s',b), ...]_{\forall s' \in \mathcal{S}}$, $\mathbf{Q}_j^j = [\max\limits_{b \in \mathcal{A}} Q_j^{\pi_j^*}(s',b), ...]_{\forall s' \in \mathcal{S}}$, $\mathbf{Q}_i^j = [\max\limits_{b \in \mathcal{A}} Q_i^{\pi_j^*}(s',b), ...]_{\forall s' \in \mathcal{S}}$, γ as the discount factor, and $||.||$ to be $2-$norm (Euclidean norm). **Proof:** We start by following the steps from (Barreto et al., 2018). The left side of the inequality (1) can be rewritten as:

$$\begin{aligned} Q_i^{\pi_i^*}(s,a) - Q_i^{\pi_j^*}(s,a) = & Q_i^{\pi_i^*}(s,a) - Q_j^{\pi_j^*}(s,a) \\ & + Q_j^{\pi_j^*}(s,a) - Q_i^{\pi_j^*}(s,a) \\ \leq & \underbrace{\Big| Q_i^{\pi_i^*}(s,a) - Q_j^{\pi_j^*}(s,a) \Big|}_{\text{(I)}} \\ & + \underbrace{\Big| Q_j^{\pi_j^*}(s,a) - Q_i^{\pi_j^*}(s,a) \Big|}_{\text{(II)}}. \end{aligned}$$

For **(I)**, it can be shown that:

$$\begin{gathered} \Big| Q_i^{\pi_i^*}(s,a) - Q_j^{\pi_j^*}(s,a) \Big| \leq \Big| r_i(s,a) + \gamma \sum_{s'} p_i(s'|s,a) \\ \max_{\mathrm{s'}} Q_i^{\pi_i^*}(s',b) - r_j(s,a) - \gamma \sum_{s'} p_j(s'|s,a) \max_{\mathrm{b}} Q_j^{\pi_j^*}(s',b) \Big| \leq \\ \Big| r_i(s,a) - r_j(s,a) \Big| + \Big| \gamma \sum_{s'} p_i(s'|s,a) \max_{\mathrm{b}} Q_i^{\pi_i^*}(s',b) - \\ \gamma \sum_{s'} p_j(s'|s,a) \max_{\mathrm{b}} Q_j^{\pi_j^*}(s',b) \Big| \leq \\ \Big| r_i(s,a) - r_j(s,a) \Big| + \gamma \Big(\Big|\Big| (\mathbf{P}_i - \mathbf{P}_j).(\mathbf{Q}_i^i - \mathbf{Q}_j^j) \Big|\Big| \Big) \leq \\ \Big| r_i(s,a) - r_j(s,a) \Big| + \gamma \Big|\Big| \mathbf{P}_i - \mathbf{P}_j \Big|\Big| \times \Big|\Big| \mathbf{Q}_i^i - \mathbf{Q}_j^j \Big|\Big| = \\ \delta_{ij} + \gamma \Big|\Big| \mathbf{P}_i - \mathbf{P}_j \Big|\Big| \times \Big|\Big| \mathbf{Q}_i^i - \mathbf{Q}_j^j \Big|\Big| \leq \\ \frac{\delta_{ij}}{1-\gamma} + \frac{\gamma}{1-\gamma} \Big|\Big| \mathbf{P}_i - \mathbf{P}_j \Big|\Big| \times \Big|\Big| \mathbf{Q}_i^i - \mathbf{Q}_j^j \Big|\Big|, \ \forall s' \in \mathcal{S} \end{gathered}$$

For **(II)**, it can be shown that:

$$\begin{gathered} \Big| Q_j^{\pi_j^*}(s,a) - Q_i^{\pi_j^*}(s,a) \Big| \leq \Big| r_j(s,a) + \gamma \sum_{s'} p_j(s'|s,a) \\ \max_{\mathrm{b}} Q_j^{\pi_j^*}(s',b) - r_i(s,a) - \gamma \sum_{s'} p_i(s'|s,a) \max_{\mathrm{b}} Q_i^{\pi_j^*}(s',b) \Big| \leq \\ \Big| r_j(s,a) - r_i(s,a) \Big| + \Big| \gamma \sum_{s'} p_j(s'|s,a) \max_{\mathrm{b}} Q_j^{\pi_j^*}(s',b) - \\ \gamma \sum_{s'} p_i(s'|s,a) \max_{\mathrm{b}} Q_i^{\pi_j^*}(s',b) \Big| \leq \\ \Big| r_j(s,a) - r_i(s,a) \Big| + \gamma \Big(\Big|\Big| (\mathbf{P}_j - \mathbf{P}_i).(\mathbf{Q}_j^j - \mathbf{Q}_i^j) \Big|\Big| \Big) \leq \\ \Big| r_j(s,a) - r_i(s,a) \Big| + \gamma \Big|\Big| \mathbf{P}_j - \mathbf{P}_i \Big|\Big| \times \Big|\Big| \mathbf{Q}_j^j - \mathbf{Q}_i^j \Big|\Big| = \\ \delta_{ij} + \gamma \Big|\Big| \mathbf{P}_j - \mathbf{P}_i \Big|\Big| \times \Big|\Big| \mathbf{Q}_j^j - \mathbf{Q}_i^j \Big|\Big| \leq \\ \frac{\delta_{ij}}{1-\gamma} + \frac{\gamma}{1-\gamma} \Big|\Big| \mathbf{P}_j - \mathbf{P}_i \Big|\Big| \times \Big|\Big| \mathbf{Q}_j^j - \mathbf{Q}_i^j \Big|\Big| = \\ \frac{\delta_{ij}}{1-\gamma} + \frac{\gamma}{1-\gamma} \Big|\Big| \mathbf{P}_i - \mathbf{P}_j \Big|\Big| \times \Big|\Big| \mathbf{Q}_j^j - \mathbf{Q}_i^j \Big|\Big|, \ \forall s' \in \mathcal{S} \end{gathered}$$

Considering (I) and (II), it leads to the upper bound:

$$Q_i^{\pi_i^*}(s,a) - Q_i^{\pi_j^*}(s,a) \leq \frac{2\delta_{ij}}{1-\gamma} + \frac{\gamma\Big|\Big|\mathbf{P}_i - \mathbf{P}_j\Big|\Big|}{(1-\gamma)} \times \frac{\Big(\Big|\Big|\mathbf{Q}_i^i - \mathbf{Q}_j^j\Big|\Big| + \Big|\Big|\mathbf{Q}_j^j - \mathbf{Q}_i^j\Big|\Big|\Big)}{(1-\gamma)}.$$

♠

Lemma 1

Let $\pi_1^*, ..., \pi_N^*$ be N optimal policies for $\mathbf{S}_1, \dots, \mathbf{S}_N$ respectively and $\tilde{Q}_{\mathcal{T}}^{\pi_j^*} = \big(\tilde{\boldsymbol{\psi}}^{\pi_j^*}\big)^{\mathrm{T}}\tilde{\mathbf{w}}_{\mathcal{T}}$ denote the action-value function of an optimal policy learned in $\mathbf{S}_j$ and executed in the target environment $\mathcal{T}$. Let $\tilde{\boldsymbol{\psi}}^{\pi_j^*}$ denote the estimated successor feature function from the combined source and target observations from $\mathbf{S}_j$ and $\mathcal{T}$ as defined in Eq. (9)(in paper), and $\tilde{\mathbf{w}}_{\mathcal{T}}$ is the estimated reward mapper for environment $\mathcal{T}$ by using loss function in Eq. (6)(in paper). It can be shown that the difference of the true action-value function and the estimated one through successor feature functions and reward mapper, is bounded as:

$$\Pr\Big(\Big|Q_{\mathcal{T}}{}^{\pi_j^*}(s,a) - \tilde{Q}_{\mathcal{T}}^{\pi_j^*}(s,a)\Big| \leq \varepsilon(m)\; \forall s,a\Big) \geq 1-\delta,$$

where $\varepsilon(m) = \sqrt{2\mathrm{log}(|\mathbb{X}|u_m/\delta)}\sigma_{m,d}(\mathbf{x}),\;\; \mathbf{x} \in \mathbb{X}\;\; \delta \in (0,1)$, $u_m = \frac{\pi^2 m^2}{6}$, m being the number of observations in environment $\mathcal{T}$, and $\mathbf{x} = (s,a)$. $\sigma_{m,d}(\mathbf{x})$ is the square root of posterior variance as defined in Eq. (10)(in paper).

Proof: For proving this Lemma, we first follow the properties of Normal distribution. Let us assume that $l \sim \mathcal{N}(0,1)$, and $\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) \sim \mathcal{N}\big(\mu_{m,d}(\mathbf{x}), \sigma_{m,d}^2(\mathbf{x})\big)$, $\mathbf{x} \in \mathbb{X}$, $\mathbf{x} = (s,a)$, and $d = \{1, \dots, D\}$ as defined in Eq. (9) and Eq. (10)(in paper). m target observations are assumed to be available. We follow (Srinivas et al., 2009), based on the properties of Normal distribution, $\Pr(l > c),\; c > 0$ is calculated as:

$$\Pr(l > c) = e^{-c^2/2}(2\pi)^{-1/2}\int e^{-(l-c)^2 - c(l-c)}dl \leq e^{-c^2/2}\Pr(l>0) = \frac{1}{2}(e^{-c^2/2}).$$

As $c > 0$, we know $e^{-c(l-c)} \leq 1$ for $l \geq c$. Accordingly, $\Pr\Big[|\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \tilde{\psi}_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x})| > \beta_m^{1/2}\sigma_{m,d}(\mathbf{x})\Big] \leq e^{-\beta_m/2},\;\; \beta_m = 2\mathrm{log}(|\mathbb{X}|u_m/\delta)$. Now assuming $l = \big(\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \tilde{\psi}_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x})\big)/\sigma_{m,d}(\mathbf{x})$ and $c = \beta_m^{1/2}$, the error of modelling successor feature function can be written as:

$$\Pr\Big(\Big|\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \tilde{\psi}_{\mathcal{T},d}^{\pi_j^*}(s,a)\Big| \leq \beta_m^{1/2}\sigma_{m,d}(\mathbf{x})\Big) \geq 1 - |\mathbb{X}|e^{-\beta_m/2}.\;\; \forall \mathbf{x} \in \mathbb{X}. \tag{2}$$

If $|\mathbb{X}|e^{-\beta_m/2} = \frac{\delta}{u_m}$, the inequality 2 holds for $u_m = \pi^2 m^2/6$. We follow the assumption in (Barreto et al., 2017; 2019; 2020), $\exists\tilde{\mathbf{w}}_{\mathcal{T}},\;\; s.t.\; \tilde{Q}_{\mathcal{T}}^{\pi_j^*} = \big(\tilde{\boldsymbol{\psi}}^{\pi_j^*}\big)^{\mathrm{T}}\tilde{\mathbf{w}}_{\mathcal{T}}$ given all dimensions of successor feature function, hence Lemma 1 holds for $\forall\mathbf{x} \in \mathbb{X}$:

$$\Pr\Big(\Big|Q_{\mathcal{T}}{}^{\pi_j^*}(\mathbf{x}) - \tilde{Q}_{\mathcal{T}}^{\pi_j^*}(\mathbf{x})\Big| \leq \varepsilon(m)\; \forall\mathbf{x} \in \mathbb{X}\Big) \geq 1-\delta.$$

♠

We note $\varepsilon(m)$ decreases $\varepsilon(m) \in \big(\mathcal{O}\mathrm{log}(m)^{-1}\big)$ as $\sigma_{m,d}(\mathbf{x}) \in \mathcal{O}\big(\mathrm{log}(m)^{-2}\big)$ (Lederer et al., 2019) and $\sqrt{2\mathrm{log}(|\mathbb{X}|u_m/\delta)} \in \mathcal{O}\big(\mathrm{log}(m)\big)$. This guarantees that the error of modelling convergence to zero as $m \to \infty$. Before starting the proof of Theorem 2, we present Remark 1 based on the concept of GPI (Barreto et al., 2018) as follows:

Remark 1 Let $\pi_1, ..., \pi_N$ be N decision policies and correspondingly $\tilde{Q}^{\pi_1}, \tilde{Q}^{\pi_2}, ..., \tilde{Q}^{\pi_N}$ are the respective estimated action-value functions (Lemma 1) such that:

$$\Big|Q^{\pi_i}(\mathbf{x}) - \tilde{Q}^{\pi_i}(\mathbf{x})\Big| \leq \varepsilon(m)\;\; \forall\mathbf{x} = (s,a),$$

where m is the number of target observations. Defining: $\pi(s) \in \mathrm{argmax_a}\, \mathrm{max_i}\, \tilde{Q}^{\pi_i}(s,a)$, Then:

$$Q^{\pi}(\mathbf{x}) \geq \max_i Q^{\pi_i}(\mathbf{x}) - \frac{2}{1-\gamma}\varepsilon(m)\;\; \forall\mathbf{x} = (s,a).$$

Proof:

We start the proof by extending $P^{\pi}\max_{\mathrm{i}} \tilde{Q}^{\pi_i}(s,a)$, where P^{π} is the Bellman operator (Barreto et al., 2018):

$$\begin{aligned}
P^{\pi}\max_{\mathrm{i}} \tilde{Q}^{\pi_i}(s,a) &= r(s,a) + \gamma\sum_{s'} p_i(s'|s,a)\max_{\mathrm{i}} \tilde{Q}^{\pi_i}(s',\pi(s')) \\
&\geq r(s,a) + \gamma\sum_{s'} p_i(s'|s,a)\max_{\mathrm{b}} Q^{\pi_i}(s',b) - \gamma\varepsilon(m) \\
&\geq r(s,a) + \gamma\sum_{s'} p_i(s'|s,a)\max_{\mathrm{i}} Q^{\pi_i}(s',\pi_i(s')) - \gamma\varepsilon(m) \\
&\geq r(s,a) + \gamma\sum_{s'} p_i(s'|s,a)\, Q^{\pi_i}(s',\pi_i(s')) - \gamma\varepsilon(m) \\
&= Q^{\pi_i}(s,a) - \gamma\varepsilon(m)
\end{aligned}$$

as $P^{\pi}\max_{\mathrm{i}} \tilde{Q}^{\pi_i}(s,a) \geq Q^{\pi_i}(s,a) - \gamma\varepsilon(m),\; \forall i = \{1, \dots, N\}$:

$$\begin{aligned}
P^{\pi}\max_{\mathrm{i}} \tilde{Q}^{\pi_i}(s,a) &\geq \max_{\mathrm{i}} Q^{\pi_i}(s,a) - \gamma\varepsilon(m) \\
&\geq \tilde{Q}^{\pi_i}(s,a) - \gamma\varepsilon(m) - \varepsilon(m) \\
&\geq \tilde{Q}^{\pi_i}(s,a) - \frac{1+\gamma}{1-\gamma}\varepsilon(m) \\
&\geq Q^{\pi_i}(s,a) - \varepsilon(m) - \frac{1+\gamma}{1-\gamma}\varepsilon(m)
\end{aligned}$$

We also know that $Q^{\pi}(\mathbf{x}) = \lim_{k\to\infty}(P^{\pi})^{k}\max_{\mathrm{i}} \tilde{Q}^{\pi_i}(\mathbf{x})$. Then, it follows:

$$Q^{\pi}(\mathbf{x}) \geq \max_{\mathrm{i}} Q^{\pi_i}(\mathbf{x}) - \frac{2}{1-\gamma}\varepsilon(m)\ \forall \mathbf{x} = (s,a), \mathbf{x} \in \mathbb{X}.$$

♠

Theorem 2

Let $\mathbf{S}_{i=1\dots N}$ be N different source environments with dissimilar transition functions $p_{i=1\dots N}$. Let us denote the optimal policy π that is defined based on the GPI as:

$$\pi(s) \in \underset{\mathrm{a}\in\mathcal{A}}{\operatorname{argmax}} \max_{\mathrm{j}\in\{1\dots\mathrm{N}\}} \tilde{Q}_{\mathcal{T}}^{\pi_j^*}(s,a), \tag{3}$$

where $\tilde{Q}_{\mathcal{T}}^{\pi_j^*} = \left(\tilde{\boldsymbol{\psi}}^{\pi_j^*}\right)^{\mathrm{T}}\tilde{\mathbf{w}}_{\mathcal{T}}$ being the action-value function of an optimal policy learned in $\mathbf{S}_j$ and executed in target environment $\mathcal{T}$, $\tilde{\boldsymbol{\psi}}^{\pi_j^*}$ is the estimated successor feature from the combined source and target observations from $\mathbf{S}_j$ and $\mathcal{T}$ as defined in Eq. (9)(in paper), and $\tilde{\boldsymbol{w}}_{\mathcal{T}}$ is the estimated reward mapper for target environment from Eq. (6)(in paper). Considering Lemma 1 and Eq. (13)(in paper), the difference of optimal action-value function in the target environment and our GPI-derived action value function is upper bounded as:

$$\begin{aligned}Q_{\mathcal{T}}^{*}(s,a) - \tilde{Q}_{\mathcal{T}}^{\pi\in\pi_j^*}(s,a) \leq{}& \frac{2\phi_{\max}}{1-\gamma}\Big|\Big|\tilde{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_j\Big|\Big| \\ &+ \frac{\gamma\Big|\Big|\mathbf{P}_{\mathcal{T}}(s,a) - \mathbf{P}_j(s,a)\Big|\Big|}{(1-\gamma)} \\ &\times \frac{\left(\Big|\Big|\mathbf{Q}_{\mathcal{T}}^{*} - \mathbf{Q}_j^{j}\Big|\Big| + \Big|\Big|\mathbf{Q}_j^{j} - \mathbf{Q}_{\mathcal{T}}^{j}\Big|\Big|\right)}{(1-\gamma)} \\ &+ \frac{2\varepsilon(m)}{(1-\gamma)}.\end{aligned} \tag{4}$$

where $\phi_{\max} = \max_{s,a}||\phi(s,a)||$. We also define $\mathbf{P}_{\mathcal{T}} = [p_{\mathcal{T}}(s'|s,a),\dots]_{\forall s'\in\mathcal{S}}$, $\mathbf{P}_j = [p_j(s'|s,a),\dots]_{\forall s'\in\mathcal{S}}$, $\mathbf{Q}_{\mathcal{T}}^{*} = [\max_{b\in\mathcal{A}} Q_{\mathcal{T}}^{*}(s',b),\dots]_{\forall s'\in\mathcal{S}}$, $\mathbf{Q}_j^{j} = [\max_{b\in\mathcal{A}} \tilde{Q}_j^{\pi_j^*}(s',b),\dots]_{\forall s'\in\mathcal{S}}$, $\mathbf{Q}_{\mathcal{T}}^{j} = [\max_{b\in\mathcal{A}} \tilde{Q}_{\mathcal{T}}^{\pi_j^*}(s',b),\dots]_{\forall s'\in\mathcal{S}}$, and γ as the discount factor.

Proof: $Q_{\mathcal{T}}^{*}(s,a) - \tilde{Q}_{\mathcal{T}}^{\pi\in\pi_j^*}(s,a)$ is defined as the difference of the optimal action-value function, and the action-value function derived from GPI. It can be shown that:

$$\begin{aligned}&Q_{\mathcal{T}}^{*}(s,a) - \tilde{Q}_{\mathcal{T}}^{\pi}(s,a) \leq Q_{\mathcal{T}}^{*}(s,a) - Q_{\mathcal{T}}^{\pi_j^*}(s,a) + \frac{2}{1-\gamma}\varepsilon(m) \\ &\leq \frac{2\delta_{ij}}{1-\gamma} + \frac{\gamma\Big|\Big|\mathbf{P}_i - \mathbf{P}_j\Big|\Big| \times \Big|\Big|\mathbf{Q}_{\mathcal{T}}^{*} - \mathbf{Q}_j^{j}\Big|\Big| + \Big|\Big|\mathbf{Q}_j^{j} - \mathbf{Q}_{\mathcal{T}}^{j}\Big|\Big|}{(1-\gamma)} \\ &\qquad + \frac{2}{1-\gamma}\varepsilon(m) \quad //\mathrm{Theorem1} \\ &\qquad \leq \frac{2}{1-\gamma}\max_{s,a}||\phi(s,a)||\ ||\mathrm{w}_{\mathcal{T}} - \mathbf{w}_j|| \\ &+ \frac{\gamma\Big|\Big|\mathbf{P}_i - \mathbf{P}_j\Big|\Big| \times \Big|\Big|\mathbf{Q}_{\mathcal{T}}^{*} - \mathbf{Q}_j^{j}\Big|\Big| + \Big|\Big|\mathbf{Q}_j^{j} - \mathbf{Q}_{\mathcal{T}}^{j}\Big|\Big|}{(1-\gamma)} + \frac{2}{1-\gamma}\varepsilon(m).\end{aligned}$$

♠

As explained in Lemma 1, $\varepsilon(m) \to 0$ with $\varepsilon(m) \in \left(\mathcal{O}\log(m)^{-1}\right)$, as $m \to \infty$ - that is, the number of target observations tend to infinity. Note that the remaining terms of the upper bound depends on the amount of dissimilarity of source and target environments as explained in Section 3.1 of the paper.

1.2. Infinite $\mathbb{X}$

We now continue our analysis on the cases that the action-state space ($\mathbb{X}$) is infinite - i.e. there may be infinite observations coming from the target environment. In that case, Lemma 1 will not hold and further steps need to be taken.

Let us assume $\mathbb{X}_m \subset \mathbb{X}$ represents a subset of infinite $\mathbb{X}$ at time step m, where m target observations are seen. Clearly, Lemma 1 will hold with this assumption if $\beta_m = 2\log(|\mathbb{X}_m|u_m/\delta)$. The main question in here is if we can extend this to the whole search space $\mathbb{X}$.

Following Boole's inequality - known as union bound, it can be shown that for some constants $a, b, L > 0$ (Srinivas et al., 2009):

$$\Pr\Big(\forall i = \{1,2\}, \forall \mathbf{x} \in \mathbb{X}, |\frac{\partial \psi_{\mathcal{T},d}^{\pi_j^*}}{\partial \mathbf{x}_i}| < L\Big) \geq 1 - 2ab^{\frac{L^2}{b^2}},$$

that implies:

$$\Big(\forall \mathbf{x} \in \mathbb{X}, |\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}')|\Big) \leq L|\mathbf{x} - \mathbf{x}'|,\ \ \mathbf{x}, \mathbf{x}' \in \mathbb{X}. \tag{5}$$

Eq. (5) enables us to perform a discretisation on the search space $\mathbb{X}_m$ with size of τ_m^2 so that:

$$|\mathbf{x} - [\mathbf{x}]_{\mathrm{m}}| \leq 2\mathrm{r}/\tau_m,$$

where $[\mathbf{x}]_m$ denotes the closest point from $\mathbb{X}_m$ to $\mathbf{x} \in \mathbb{X}$ and τ_m implies the number uniformly spaced points on both coordinates of $\mathbb{X}_m$ that is a discretisation factor. We now proceed to Lemma 2 as an extension of successor feature function modelling error in infinite $\mathbb{X}$ space.

Lemma 2

Let $\mathbf{x} = (s, a) \in \mathbb{X}$, and $\mathbb{X}$ is infinite state-action space. $\pi_1^*, ..., \pi_N^*$ is N optimal policies for $\mathbf{S}_1, \ldots, \mathbf{S}_N$ respectively and $\tilde{Q}_{\mathcal{T}}^{\pi_j^*} = \big(\tilde{\boldsymbol{\psi}}^{\pi_j^*}\big)^{\mathrm{T}} \tilde{\mathbf{w}}_{\mathcal{T}}$ denote the action-value function of an optimal policy learned in $\mathbf{S}_j$ and executed in the target environment $\mathcal{T}$. Let $\tilde{\boldsymbol{\psi}}^{\pi_j^*}$ denote the estimated successor feature function from the combined source and target observations from $\mathbf{S}_j$ and $\mathcal{T}$ as defined in Eq. (9)(in paper), and $\tilde{\mathbf{w}}_{\mathcal{T}}$ is the estimated reward mapper for environment $\mathcal{T}$ by using loss function in Eq. (6)(in paper). It can be shown that the difference of the true action-value function and the estimated one through successor feature functions and reward mapper, is bounded as:

$$\Pr\Big(\Big|Q_{\mathcal{T}}{}^{\pi_j^*}(s,a) - \tilde{Q}_{\mathcal{T}}^{\pi_j^*}(s,a)\Big| \leq \varepsilon(m)\, \forall s, a\Big) \geq 1 - \delta,$$

where
$\varepsilon(m) = \sqrt{2\text{log}(2u_m/\delta) + 8\text{log}(2mbr\sqrt{\text{log}(4a/\delta)})}$
$\sigma_{m,d}([\mathbf{x}]_m) + \frac{1}{m^2}$, $\mathbf{x} \in \mathbb{X}\ \delta \in (0,1)$, $u_m = \frac{\pi^2 m^2}{6}$, $m > 1$ being the number of observations in environment $\mathcal{T}$, $[\mathbf{x}]_m$ is the closest points in $\mathbb{X}_m$ to $\mathbf{x} \in \mathbb{X}$. $\sigma_{m,d}(.)$ is the square root of posterior variance as defined in Eq. (10)(in paper). $a, b > 0$ are constants.

Proof: As explained in Section 1.2, we know:

$$\Pr\Big(\forall \mathrm{x} \in \mathbb{X}, |\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}')| < b\sqrt{\log(4a/\delta)}\Big)|\mathbf{x} - \mathbf{x}'|. \tag{6}$$

Accordingly, by replacing $\mathbf{x}'$:

$$\forall \mathbf{x} \in \mathbb{X}_m, |\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \psi_{\mathcal{T},d}^{\pi_j^*}([\mathbf{x}]_m)| \leq 2rb\sqrt{\log(4a/\delta)}/\tau_m.$$

By selecting the discretisation factor as $\tau_m = 4m^2br\sqrt{\log(4a/\delta)}$:

$$\forall \mathbf{x} \in \mathbb{X}_m, |\psi_{\mathcal{T},d}^{\pi_j^*}(\mathbf{x}) - \psi_{\mathcal{T},d}^{\pi_j^*}([\mathbf{x}]_m)| \leq \frac{1}{m^2}.$$

This implies $|\mathbb{X}_m| = (4m^2br\sqrt{\log(4a/\delta)})^2$. By replacing $|\mathbb{X}_m|$ in β_m defined in Section 1.2, the proof is completed. ♠

Hence, if $\mathbb{X}$ is infinite set, Theorem 2 holds with $\varepsilon(m) = \sqrt{2\text{log}(2u_m/\delta) + 8\text{log}(2mbr\sqrt{\text{log}(4a/\delta)})}\sigma_{m,d}([\mathbf{x}]_m) + \frac{1}{m^2}$.

2. Experimental Details

Maze (navigation problem): For the task of navigation, we designed a maze environment with following properties: (1) We set the ε-greedy exploration rate to $\varepsilon_e = 0.5$ for the adaptation phase with decay rate of 0.9999, this value is set to zero in the testing phase. (2) Discount factor value is set to $\gamma = 0.9$. (3) $\alpha = 0.05$ is the learning rate. Agent is allowed to reach to the goal in maximum of 100 steps, otherwise it terminates.

For 12 source environments, 25 obstacles are randomly generated, the agent always start from top left, and the goal is also randomly placed in these environments. We used generic Q-learning with replay buffer size 10^4 and Adam optimizer with batch size 64 to find the optimal policies in all these 12 environments. Algorithm 1 in the paper is then used to estimate the successor feature functions in the environment. Figure 2 demonstrates our toy environment with agent at the top left, red obstacles, and the green goal. Our proposed feature function for this problem is a MLP with 4 hidden layers with a linear activation function in the last hidden layer to represent the reward mapper of the task. The remaining hidden layers have ReLU activation function. The output of this network is the predicted value of the reward for a state and action. Note that this network minimises the loss function introduced in Eq. (6)(in paper). We used SE kernel and maximising the log marginal likelihood for finding the best set of hyperparameters for GP.

CartPole: As mentioned, to translate the image data into states, we used a CNN with: (1) First hidden layer with 64 filters of 5×5 with stride 3 with a ReLU activation, second hidden layer with 64 filters of 4×4 with stride 2 and a ReLU activation, third hidden layer with 64 filters of 3×3 with stride 1 and a ReLU activation. The final hidden layer is "features" we used in an image that is fully connected Flatten units.

Maximum number of steps for the CartPole problem is set to 200. We set the ε-greedy exploration rate to $\varepsilon_e = 0.5$ for the adaptation phase with decay rate of 0.9999, this value is set to zero in testing phase. The source learned policy is with pole's length of $0.5m$ and accordingly, using Algorithm 1 in the paper the corresponding successor features are extracted. We used generic Q-learning with replay buffer size 10^5 and Adam optimizer with batch size 64 to find this optimal policy. The target environment is then modified to incorporate the change of environment. Figure 1 demonstrates the change of dynamics. We used the same structure of feature function in Maze problem for this experiment.

FSF: This environment (Barreto et al., 2020) is a 10×10 grid with 10 objects and an agent occupying one cell at each time step. There are 2 types of objects each with a reward associated with it. We randomly initialise those 10 objects by sampling both their type and position from a uniform distributions over the corresponding sets. Likewise, the initial position of the agent is a uniform sample of all possible

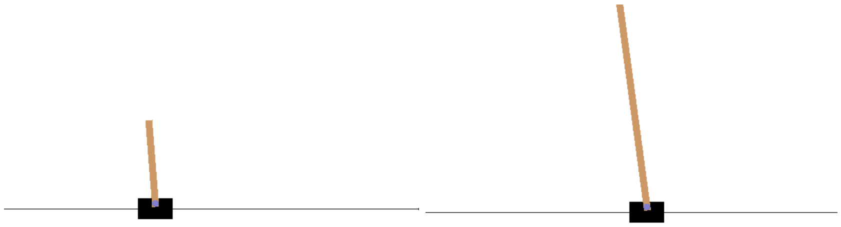

Figure 1. Illustration of the change in the dynamics for the CartPole problem. (Left) Pole's length is $0.5m$ in the source environment and (Right) Pole's length changed to $3m$ in the target environment.

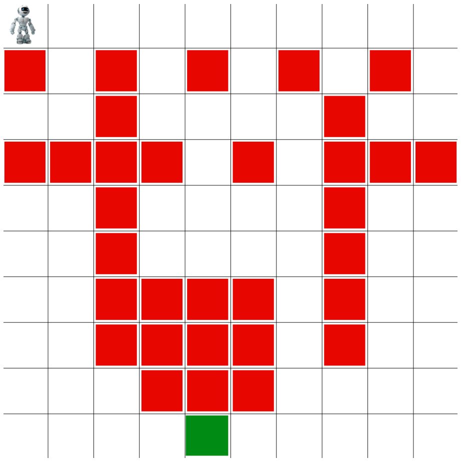

Figure 2. Illustration of the maze environment.

positions in the grid. The reward function is defined by the object type that has been picked up by the agent. e.g. Picking up red object is $+1$ reward and picking up blue is -1. Agent picks up an object if it occupies that particular cell in which the object exists. If agent picks up an object, another one will be generated randomly (in terms of location and type) in the grid. At each step the agent receives an observation representing the configuration of the environment (Barreto et al., 2020). These are $11 \times 11 \times (\mathbb{D} + 1)$ tensors that can be seen as 11×11 images with $(\mathbb{D} + 1)$ channels that are used to identify objects and walls (Barreto et al., 2020). The observations are shifted so that the the agent is always at the top-left cell of the grid. Figure 3 shows an example of this environment. The two source policies used in this experiment have $\mathrm{w}_1 = [1, 0], \mathrm{w}_2 = [1, 0]$, respectively. Intuitively, the reward mappers indicate picking up an object of particular type and ignoring the other type. However, in the target environment, the change of reward function is to pick up the first object type and "avoid" the second one with negative reward - i.e. $\boldsymbol{w}_{\mathcal{T}} = [1, -1]$. For the dissimilarity of dynamics, we added 5% noise to the transitions of the agent and also randomly placed a terminal state in the target environment with -1 reward.

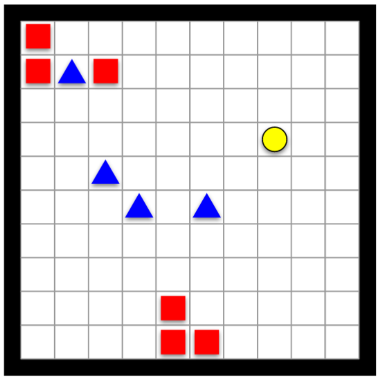

Figure 3. An example of the environment described in FSF (Barreto et al., 2020).

2.1. Additional Experiments

In this section, we compare our results in navigation problem with generic Q-Learning. We relaxed the assumption adaptation and testing phase and Q-Learning is allowed to use all the observations from the target environment. Figure 4 shows Q-Learning also converges to the same amount of avg. reward, however, since no transfer is involved, it is significantly slower than other baselines. For Q-Learning, we set $\varepsilon_e = 0.9$ with a decay rate of 0.9999 in 10^4 iterations.

References

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pp. 4055–4065, 2017.

Barreto, A., Borsa, D., Quan, J., Schaul, T., Silver, D., Hessel, M., Mankowitz, D., Zidek, A., and Munos, R. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning*, pp. 501–510. PMLR, 2018.

Barreto, A., Borsa, D., Hou, S., Comanici, G., Aygün, E.,

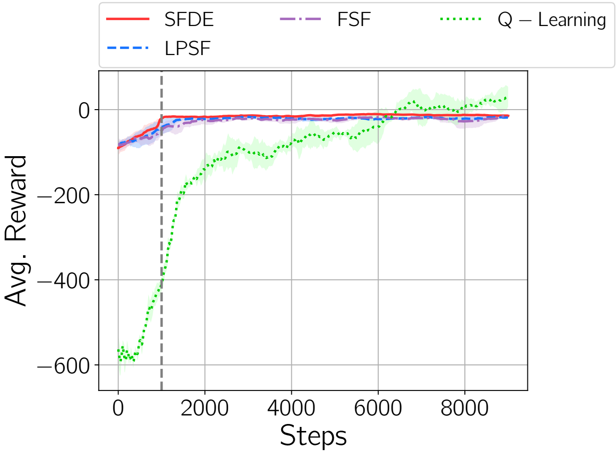

Figure 4. Performance of Q-Learning in navigation problem.

Hamel, P., Toyama, D., Mourad, S., Silver, D., Precup, D., et al. The option keyboard: Combining skills in reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 13052–13062, 2019.

Barreto, A., Hou, S., Borsa, D., Silver, D., and Precup, D. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117(48):30079–30087, 2020.

Lederer, A., Umlauft, J., and Hirche, S. Uniform error bounds for gaussian process regression with application to safe control. *arXiv preprint arXiv:1906.01376*, 2019.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.