Supplementary Material

Theorem 1. Given any constants $\delta > 0$ and $M < \infty$, there exist a time instant $t_0 < \infty$ and a resolution parameter $n_0 < \infty$ such that $\forall t > t_0$ and $\forall n > n_0$:

$$\Pr\left\{G_i\left(t\right) > M\right\} > 1 - \delta, i \in N_r.$$

Proof: Theorem 1 is equivalent to

$$\Pr\left\{G_i\left(t\right) \leq M\right\} \leq \delta, i \in N_r. \tag{1}$$

The events $\{G_i\left(t\right) = k\}$ and $\{G_i\left(t\right) = j\}$ are mutually exclusive when $j \neq k$. So (1) is equivalent to

$$\sum_{k=1}^{M} \Pr\left\{G_i\left(t\right) = k\right\} \leq \delta \tag{2}$$

where

$$\begin{aligned}\Pr\left\{G_i\left(t\right) = k\right\} = C\left(t, k\right) \left(\Pr\left\{a_i \text{ is chosen}\right\}\right)^k \\ \left(\Pr\left\{a_i \text{ is not chosen}\right\}\right)^{t-k}.\end{aligned} \tag{3}$$

A threshold of resolution parameter $n(i)$ is given for each a_i, by which $p_i(t) > 0$ if $t \leq t(i)$. $t(i)$ is a threshold of time. $r\left(t\right) = \left|\left\{a_i | p_i\left(t\right) \neq 0\right\}\right|$, where $|\cdot|$ denotes the cardinality of a set. When LA has not converged, $r(t) \geq 2$.

First, for any iteration of DEP$_{RI}$,

$$\Pr\left\{a_i \text{ is chosen}\right\} \leq 1. \tag{4}$$

Then, for action a_i, $t \leq t(i)$. If $n = n(i)$ and a_i is punished until $t(i)$, $p_i\left(t(i)\right) = \left(p_i\left(0\right) - \sum_{j=1}^{t(i)} \frac{1}{r(t)n(i)}\right)$. Therefore, for all $n > n(i)$ and $t < t(i)$,

$$\begin{aligned} p_i\left(t\right) &\geq \left(p_i\left(0\right) - \sum_{j=1}^{t(i)} \frac{1}{r\left(t\right) n\left(i\right)}\right) \\ &= \left(p_i\left(0\right) - \frac{t(i)}{r\left(t\right) n\left(i\right)}\right) \\ &\geq \left(p_i\left(0\right) - \frac{t\left(i\right)}{2n\left(i\right)}\right), i \in N_r. \end{aligned} \tag{5}$$

The second line of (5) is equal to the first line's right part because $\sum_{j=1}^{t(i)} \frac{1}{r(t)n(i)} = \frac{t(i)}{r(t)n(i)}$. It is greater than the item in the third line because $r(t) \geq 2$ when LA has not converged.

To make $p_i\left(0\right) - \frac{t(i)}{2n(i)} > 0$, $n(i)$ can be set as

$$n(i) = r \cdot t(i). \tag{6}$$

Then, at time $t \leq t(i)$, we have

$$\text{Pr}\{a_i \text{ is not chosen}\} \leq 1 - \frac{1}{2r} < 1. \tag{7}$$

Then, according to (3),(4) and (7), we have

$$\begin{aligned}
\text{Pr}\left\{G_i\left(t\right) < M\right\} &= \sum_{k=1}^{M} \text{Pr}\left\{G_i\left(t\right) = k\right\} \\
&\leq \sum_{k=1}^{M} C\left(t,k\right)\left(1\right)^k \left(1 - p_i\left(0\right) + \frac{t\left(i\right)}{2n\left(i\right)}\right)^{t-k} \\
&\leq \sum_{k=1}^{M} C\left(t,k\right)\left(1\right)^k \left(1 - \frac{1}{2r}\right)^{t-k}.
\end{aligned} \tag{8}$$

To prove the sum of M terms less than δ, we just need to make each element less than δ/M. It suffices to prove that M times the k'th term is less than δ. It can be observed that $C\left(t,k'\right) \leq t^{k'}$. Let $\psi = 1 - \frac{1}{2r}$. So we have to prove that:

$$\text{Pr}\left\{G_i\left(t\right) < M\right\} \leq M t^{k'} \psi^{t-k'} \leq \delta. \tag{9}$$

By using l'Hopital's rule k' times, we obtain:

$$\begin{aligned}
\lim_{t\to\infty} M t^{k'} \psi^{t-k'} &= M \lim_{t\to\infty} \frac{t^{k'}}{\left(1/\psi\right)^{t-k'}} \\
&= M \lim_{t\to\infty} \frac{k'!}{\left(\ln\left(1/\psi\right)\right)^{k'}\left(1/\psi\right)^{t-k'}} \\
&= 0.
\end{aligned} \tag{10}$$

Thus, for any given constants $\delta > 0$ and $M < \infty$, there exists a threshold $t(i)$ such that for all $t > t(i)$ and $n > n(i) = r \cdot t(i)$, (9) is satisfied. Since, we can repeat this argument for all the actions. t_0 and n_0 are defined as follows:

$$t_0 = \max_{i\in N_r}\{t(i)\}, \tag{11}$$

$$n_0 = \max_{i\in N_r}\{n(i)\} = \max_{i\in N_r}\{r \cdot t(i)\}. \tag{12}$$

Therefore, for all i, it is true that for all $t > t_0$ and $n > n_0$, $\text{Pr}\left\{G_i\left(t\right) > M\right\} > 1 - \delta$, which completes the proof. ■

Theorem 2. Given any $\delta \in (0,1)$, there exists $t_0 < \infty$, such that $\text{Pr}\left\{\overline{C}\left(t_0\right)\right\} = 1$, where $\overline{C}\left(t_0\right)$

is defined as follows:

$$q_j(t) = \Pr\left\{\left|\widetilde{d}_j(t) - d_j\right| < \frac{w}{2}\right\}, \tag{13}$$

$$q(t) = \Pr\left\{\left|\tilde{d}_j(t) - d_j\right| < \frac{w}{2},\ \forall j \in N_r\right\} = \prod_{j \in N_r} q_j(t), \tag{14}$$

$$C(t) = \{q(t) > 1 - \delta\},\ \delta \in (0,1), \tag{15}$$

$$\overline{C}(t_0) = \left\{\bigcap_{t > t_0} \{q(t) > 1 - \delta\}\right\},\ \delta \in (0,1) \tag{16}$$

where w is the minimum difference between the reward probabilities of any two actions.

Proof: Theorem 2 is equivalent to

$$\Pr\left\{\left|\widetilde{d}_i(t) - d_i\right| < \frac{w}{2}, \forall i \in N_r, \forall t > t_0\right\} > 1 - \delta. \tag{17}$$

By the weak law of large numbers, for a given $\delta > 0$, $\exists M_i < \infty$, such that if a_i is chosen at least M_i times:

$$\Pr\left\{\left|\widetilde{d}_i(t) - d_i\right| < \frac{w}{2}, \forall t > t_0\right\} > 1 - \delta. \tag{18}$$

Let $M = \max_{1 \le i \le r}\{M_i\}$. According to Theorem 1, there exist a time instant $t_0 < \infty$ and a resolution parameter $n_0 < \infty$ such that $\forall t > t_0$ and $\forall n > n_0$:

$$\Pr\{G_i(t) > M\} > 1 - \delta, i \in N_r.$$

Thus, if all actions are chosen at least M times, then each of the $\widetilde{d}_i(t)$ will be in an $w/2$ neighborhood of d_i with a probability greater than $1 - \delta$, which completes the proof. ■

Theorem 3. $p_m(t)_{t>t_0} = \sum_{i \in \widetilde{X}(t)} p_i(t)$ is submartingale under DEP$_{RI}$.

Proof: First,

$$E[p_m(t)] \le 1 < \infty. \tag{19}$$

Then, according to the updating rule of DEP$_{RI}$, we have:

$$\begin{aligned}
&E\left[p_m\left(t+1\right)|Q\left(t\right)\right]\\
&\geq \sum_{j\in N_r} p_j(t)\left(d_j\left(s\left(t\right)\left(p_m\left(t\right)+c_t\Delta\right)\right.\right.\\
&\quad +o\left(t\right)\left(p_m\left(t\right)+f_t\Delta\right)\\
&\quad +\left(1-s\left(t\right)-o\left(t\right)\right)\left(p_m\left(t\right)+\frac{c_t}{\widehat{r}}\left(\widehat{r}-e_t\right)\Delta-e_t\Delta\right)\Big)\\
&\quad \left.+\left(1-d_j\right)p_m\left(t\right)\right)\\
&\geq \sum_{j\in N_r} p_j(t)\left(d_j\left(s\left(t\right)\left(p_m\left(t\right)+c_t\Delta\right)\right.\right.\\
&\quad +\left(1-s\left(t\right)\right)\left(p_m\left(t\right)+\min\{f_t,\frac{c_t\widehat{r}-c_te_t-\widehat{r}e_t}{\widehat{r}}\}\Delta\right)\\
&\quad \left.+\left(1-d_j\right)p_m\left(t\right)\right)\\
&= p_m\left(t\right)+\sum_{j\in N_r} p_j(t)d_j\\
&\left(s\left(t\right)\left(c_t-\min\{f_t,\frac{c_t\widehat{r}-c_te_t-\widehat{r}e_t}{\widehat{r}}\}\right)\Delta\right.\\
&\quad \left.+\min\{f_t,\frac{c_t\widehat{r}-c_te_t-\widehat{r}e_t}{\widehat{r}}\}\Delta\right)
\end{aligned}\tag{20}$$

where $c_t=1,2,...,r-\widehat{r}$, $f_t\in[-\frac{\widehat{r}-1}{\widehat{r}^2\Delta},\frac{\widehat{r}-1}{\widehat{r}^2\Delta}]$ and $e_t=1,2,...,\widehat{r}$. $s(t)$ is the probability that all actions in $\widehat{A}$ are in the currently estimated arbitrary action subset $\widetilde{A}(t)$ at time t and we can see that:

$$\begin{aligned}
s(t)&=\Pr\{a_i\in\widetilde{A}(t),\forall a_i\in\widehat{A}\}\\
&\geq \prod_{j\in\widetilde{X}(t)} q_j\left(t\right)
\end{aligned}\tag{21}$$

and $o(t)$ is the probability that actions in $\widetilde{A}(t)$ have changed after being updated at time t and $\exists a_i\in\widehat{A}$ is not in $\widetilde{A}(t)$ after the change.

We denote $g_t=\frac{c_t\widehat{r}-c_te_t-\widehat{r}e_t}{\widehat{r}}$. Then, we have

$$\begin{aligned}
&E\left[p_m\left(t+1\right)|Q\left(t\right)\right]-p_m(t)\\
&\geq \sum_{j\in N_r} p_j(t)d_j\left(s\left(t\right)\left(c_t-\min\{f_t,g_t\}\right)\Delta+\min\{f_t,g_t\}\Delta\right).
\end{aligned}\tag{22}$$

Given that $p_j(t)>0$ and $d_j>0$, we denote

$$Z_t=\frac{\min\{f_t,g_t\}}{\min\{f_t,g_t\}-c_t}$$

and

$$\max\left\{Z_t\right\} = \left\{ \begin{array}{l} \frac{\widehat{r}}{\widehat{r}+1}, \varDelta \le \frac{\widehat{r}-1}{\widehat{r}^3} \\ \frac{\widehat{r}-1}{\widehat{r}^2\varDelta+\widehat{r}-1}, \varDelta > \frac{\widehat{r}-1}{\widehat{r}^3} \end{array} \right. .$$

Let $1-\delta = \max\{Z_t\}$. According to Theorem 2, there exists t_0 such that $\forall t > t_0$, $\prod_{j\in\widetilde{X}(t)} q_j\left(t\right) \ge \max\{Z_t\}$. Thus,

$$E\left[p_m\left(t+1\right)\left|Q\left(t\right)\right] - p_m(t) \ge 0. \tag{23}$$

$p_m(t)_{t>t_0}$ is a submartingale, which completes the proof. ■

Based on the Martingale convergence theory and Theorem 3, we have Corollary 1.

Corollary 1. Under DEP$_{RI}$,

$$p_m(\infty) = 0 \ or \ 1.$$

Theorem 4. In all stationary environments, DEP$_{RI}$ is ϵ-optimal, i.e., given any $1-\delta \ge \max\{Z_t\}$, there exists $t_0 < \infty$ and $n_0 < \infty$, such that $\forall t > t_0$ and $\forall n > n_0$, Pr$\{p_m(\infty) = 1\} \to 1$.

Proof: We need to prove that

$$\varGamma_m(P) = \Pr\{p_m(\infty) = 1|P(0) = P\} \to 1. \tag{24}$$

Define $\varPhi_m(P) = e^{-x_m p_m}$, where x_m is a positive constant. An operator U is defined as

$$U(\varPhi_m(P)) = E[\varPhi_m(P(t+1))|P(t) = P]. \tag{25}$$

Therefore,

$$\begin{aligned}
&U(\varPhi_m(P)) - \varPhi_m(P) \\
&= E[\varPhi_m(P(t+1))|P(t) = P] - \varPhi_m(P) \\
&\le \sum_{j\in N_r} p_j(t)\left(d_j\left(s\left(t\right)e^{-x_m(p_m(t)+c_t\varDelta)}\right.\right. \\
&\quad +o\left(t\right)e^{-x_m(p_m(t)+f_t\varDelta)} \\
&\quad \left.+\left(1-s\left(t\right)-o\left(t\right)\right)e^{-x_m(p_m(t)+g_t\varDelta)}\right) \\
&\quad \left.+\left(1-d_j\right)e^{-x_m p_m(t)}\right) - \sum_{j\in N_r} p_j\left(t\right)e^{-x_m p_m(t)} \\
&= \sum_{j\in N_r} p_j\left(t\right)d_j e^{-x_m p_m(t)}\left(s\left(t\right)\left(e^{-x_m c_t\varDelta} - e^{-x_m g_t\varDelta}\right)\right. \\
&\quad \left.+o\left(t\right)\left(e^{-x_m f_t\varDelta} - e^{-x_m g_t\varDelta}\right) + \left(e^{-x_m g_t\varDelta} - 1\right)\right).
\end{aligned} \tag{26}$$

$U(\varPhi_m(P)) - \varPhi_m(P) \le 0$ is equivalent to the following formula:

$$\begin{aligned}
&s\left(t\right)\left(e^{-x_m c_t\varDelta} - e^{-x_m g_t\varDelta}\right) + o\left(t\right)\left(e^{-x_m f_t\varDelta} - e^{-x_m g_t\varDelta}\right) \\
&\quad + \left(e^{-x_m g_t\varDelta} - 1\right) \le 0.
\end{aligned} \tag{27}$$

If $b > 0$ and $x \to 0$, $b^x \doteq 1 + (\ln b)\, x + \frac{(\ln b)^2}{2} x^2$. So let $b = e^{-x_m}$, as $\Delta \to 0$, we have

$$\begin{aligned} & s\,(t) \left((\ln b)\,(c_t - g_t)\,\Delta + \frac{(\ln b)^2}{2} \left(c_t^2 - g_t^2 \right) \Delta^2 \right) \\ & + o\,(t) \left((\ln b)\,(f_t - g_t)\,\Delta + \frac{(\ln b)^2}{2} \left(f_t^2 - g_t^2 \right) \Delta^2 \right) \\ & + (\ln b)\, g_t \Delta + \frac{(\ln b)^2}{2} g_t^2 \Delta^2 \leq 0. \end{aligned} \tag{28}$$

Replacing b with e^{-x_m}, then

$$x_m(x_m - \frac{2(s(t)(c_t - g_t) + o(t)(f_t - g_t) + g(t))}{\Delta(s(t)(c_t^2 - g_t^2) + o(t)(f_t^2 - g_t^2) + g_t^2)}) \leq 0. \tag{29}$$

Thus,

$$0 < x_m \leq \frac{2(s(t)(c_t - g_t) + o(t)(f_t - g_t) + g(t))}{\Delta(s(t)(c_t^2 - g_t^2) + o(t)(f_t^2 - g_t^2) + g_t^2)}. \tag{30}$$

Denote $x_{m_0} = \frac{2(s(t)(c_t - g_t) + o(t)(f_t - g_t) + g(t))}{\Delta(s(t)(c_t^2 - g_t^2) + o(t)(f_t^2 - g_t^2) + g_t^2)}$, when $\Delta \to 0$, $x_{m_0} \to \infty$ as $s(t) > \max\{Z_t\}$. Thus, $\Phi_m(P)$ is superregular.

Denote $\phi_m(P) = \frac{1 - e^{-x_m p_m(t)}}{1 - e^{-x_m}}$. Obviously, $0 \leq \phi_m(P) \leq 1$. According to (30), $\phi_m(P)$ is a subregular.

Therefore,

$$\Gamma_m(P) \geq \phi_m(P) = \frac{1 - e^{-x_m p_m(t)}}{1 - e^{-x_m}}. \tag{31}$$

As $x_{m_0} \to \infty$, $\Gamma_m(P) \to 1$, implying that DEP$_{RI}$ is ϵ-optimality. ■