Dynamic Policy Programming with Function Approximation: Supplementary Material

Mohammad Gheshlaghi Azar
Radboud University Nijmegen
Geert Grooteplein Noord 21
6525 EZ Nijmegen Netherlands
m.azar@science.ru.nl

Vicenç Gómez
Radboud University Nijmegen
Geert Grooteplein Noord 21
6525 EZ Nijmegen Netherlands
v.gomez@science.ru.nl

Hilbert J. Kappen
Radboud University Nijmegen
Geert Grooteplein Noord 21
6525 EZ Nijmegen Netherlands
m.azar@science.ru.nl

Appendices

A Definitions

A.1 Markov Decision Processes and Bellman Equation

A stationary MDP is a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, T, \gamma)$, where $\mathcal{S}, \mathcal{A}, \mathcal{R}$ are, respectively, the set of all system states, the set of actions that can be taken and the set of rewards that may be issued, such that $r^a_{ss'}$ denotes the reward of the next state s' given that the current state is s and the action is a. T is a set of matrices of dimension $|\mathcal{S} \times \mathcal{S}|$, one for each $a \in \mathcal{A}$ such that $T^a_{ss'}$ denotes the probability of the next state s' given that the current state is s and the action is a. $\gamma \in (0, 1)$ denotes the discount factor.

Assumption 1. *We assume that for every 3-tuple $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the magnitude of the immediate reward, $|r^a_{ss'}|$ is bounded from above by R_{max}.*

A stationary policy is a mapping π that assigns to each state s a probability distribution over the action space $\mathcal{A}$, one for each $(s, a) \in \mathcal{S} \times \mathcal{A}$ such that $\pi_s(a)$ denotes the probability of the action a given the current state is s. Given the policy π, its corresponding value function V^π denotes the expected value of the long-term discounted sum of rewards in each state s, when the action is chosen by policy π. The goal is to find a policy π^* that attains the optimal value function $V^*(s)$, such that $V^*(s)$ satisfies a Bellman equation:

$$V^*(s) = \max_{\pi_s} \sum_{a \in \mathcal{A}} \pi_s(a) \sum_{s' \in \mathcal{S}} T^a_{ss'} \left(r^a_{ss'} + \gamma V^*(s') \right), \qquad \forall s \in \mathcal{S}. \tag{1}$$

A.2 Matrix notation

We find it convenient to use matrix and vector notation for analysis of theorem 1, 2 and 3. We begin by re-defining the variables of the MDP. $\mathbf{T}$ and $\mathbf{r}$ are, respectively, a $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ transition matrix with entries $\mathbf{T}(sa, s') = T^a_{ss'}$ and a $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ reward matrix with entries $\mathbf{r}(sa, s') = r^a_{ss'}$. $\bar{\mathbf{r}}$ is a $|\mathcal{S}||\mathcal{A}| \times 1$ column vector of expected rewards with entries $\bar{\mathbf{r}}(sa) = \sum_{s' \in \mathcal{S}} T^a_{ss'} r^a_{ss'}$.

The policy can also be re-expressed as a $|\mathcal{S}||\mathcal{A}| \times 1$ vector $\boldsymbol{\pi}$ with entries $\boldsymbol{\pi}(sa) = \pi_s(a)$. In addition, it will be convenient to introduce a $|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|$ matrix $\mathbf{\Pi}$, given by:

$$\mathbf{\Pi} = \begin{pmatrix} \pi_{s_1}(a_1) \cdots \pi_{s_1}(a_{|\mathcal{A}|}) & & & \\ & \pi_{s_2}(a_1) \cdots \pi_{s_2}(a_{|\mathcal{A}|}) & & \\ & & \ddots & \\ & & & \pi_{s_{|\mathcal{S}|}}(a_1) \cdots \pi_{s_{|\mathcal{S}|}}(a_{|\mathcal{A}|}) \end{pmatrix}, \tag{2}$$

where $\mathbf{\Pi}$ consists of $|\mathcal{S}|$ row blocks, each of length $|\mathcal{A}|$, which are arranged diagonally. The policy matrix $\mathbf{\Pi}$ is related to the policy vector $\boldsymbol{\pi}$ by $\boldsymbol{\pi} = \mathbf{\Pi}^{\mathsf{T}}\mathbf{1}$ with $\mathbf{1}$ denotes a $|\mathcal{S}| \times 1$ vector of all 1s. Further, one can easily verify that the matrix-product $\mathbf{\Pi T}$ gives the state to state transition matrix for the policy $\boldsymbol{\pi}$.

One can also present the value function in vector space. $\mathbf{v}^{\boldsymbol{\pi}}$ is a $|\mathcal{S}| \times 1$ vector with entries $\mathbf{v}^{\boldsymbol{\pi}}(s) = V^{\pi}(s)$. The Bellman equation (1) can now be re-expressed in matrix notation as:

$$\mathbf{v}^* = \boldsymbol{\mathcal{M}}_\infty\big(\bar{\mathbf{r}} + \gamma \mathbf{T}\mathbf{v}^*\big), \tag{3}$$

where $\boldsymbol{\mathcal{M}}_\infty$ is the max operator on the $|\mathcal{S}||\mathcal{A}| \times 1$ vector $\bar{\mathbf{r}} + \gamma\mathbf{T}\mathbf{v}^*$, such that $\boldsymbol{\mathcal{M}}_\infty\big(\bar{\mathbf{r}} + \gamma\mathbf{T}\mathbf{v}^*\big)(s) = \max_{a\in\mathcal{A}}\big(\bar{\mathbf{r}}(sa) + \gamma\mathbf{T}\mathbf{v}^*(sa)\big)$.

Often it is convenient to associate value functions not with states but with state-action pairs. Therefore, we introduce a $|\mathcal{S}||\mathcal{A}| \times 1$ vector of the action-values $\mathbf{q}^{\boldsymbol{\pi}}$ whose entries $\mathbf{q}^{\boldsymbol{\pi}}(sa)$ denotes the expected value of the sum of future rewards for all state-action $(s, a) \in \mathcal{S} \times \mathcal{A}$, provided the future actions are chosen by the policy $\boldsymbol{\pi}$. $\mathbf{q}^{\boldsymbol{\pi}}$ satisfies the following Bellman equation:

$$\mathbf{q}^{\boldsymbol{\pi}} = \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}}\mathbf{q}^{\boldsymbol{\pi}} = \bar{\mathbf{r}} + \gamma\mathbf{T}\mathbf{\Pi}\mathbf{q}^{\boldsymbol{\pi}}, \tag{4}$$

where we introduce the Bellman operator $\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}}$ and $\mathbf{q}^{\boldsymbol{\pi}}$ denotes the fixed point of $\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}}$. The optimal $\mathbf{q}$, $\mathbf{q}^*$, also satisfies a Bellman equation:

$$\mathbf{q}^* = \boldsymbol{\mathcal{T}}\mathbf{q}^* = \bar{\mathbf{r}} + \gamma\mathbf{T}\boldsymbol{\mathcal{M}}_\infty\mathbf{q}^*, \tag{5}$$

where we introduce the Bellman operator $\boldsymbol{\mathcal{T}}$. $\mathbf{q}^*$ denotes the fixed point of the operator $\boldsymbol{\mathcal{T}}$.

Both $\boldsymbol{\mathcal{T}}$ and $\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}}$ are contraction mappings with the factor γ (?, chap. 1). In other words, for any two vectors $\mathbf{q}$ and $\mathbf{q}'$, we have:

$$\|\boldsymbol{\mathcal{T}}\mathbf{q} - \boldsymbol{\mathcal{T}}\mathbf{q}'\|_\infty \le \gamma\,\|\mathbf{q} - \mathbf{q}'\|_\infty\,, \qquad \|\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}}\mathbf{q} - \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}}\mathbf{q}'\|_\infty \le \gamma\,\|\mathbf{q} - \mathbf{q}'\|_\infty\,, \tag{6}$$

where $\|\cdot\|_\infty$ denotes an L_∞-norm on $\Re^{|\mathcal{S}||\mathcal{A}|}$.

The operator $\mathcal{O}$ defined can be written in matrix notation as:

$$\boldsymbol{\mathcal{O}}\mathbf{p} = \mathbf{p} + \bar{\mathbf{r}} + \gamma\mathbf{T}\boldsymbol{\mathcal{M}}_\eta\mathbf{p} - \mathbf{\Xi}\boldsymbol{\mathcal{M}}_\eta\mathbf{p}. \tag{7}$$

Here, $\mathbf{p}$ denotes a $|\mathcal{S}||\mathcal{A}| \times 1$ vector of the action preferences. $\boldsymbol{\mathcal{M}}_\eta$ is the soft-max operator on the vector $\mathbf{p}$ with $\boldsymbol{\mathcal{M}}_\eta\mathbf{p}(s) = \mathcal{M}_\eta P(s)$. $\mathbf{\Xi}$ is a $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ matrix given by:

$$\mathbf{\Xi} = \begin{pmatrix} 1\cdots1 & & & \\ & 1\cdots1 & & \\ & & \ddots & \\ & & & 1\cdots1 \end{pmatrix}^{\mathsf{T}}, \tag{8}$$

where $\mathbf{\Xi}$ consists of $|\mathcal{S}|$ column blocks of 1s, each of length $|\mathcal{A}|$, which are arranged diagonally. $\mathbf{\Xi}$ transforms the $|\mathcal{S}| \times 1$ vector $\boldsymbol{\mathcal{M}}_\eta \mathbf{p}$ to the $|\mathcal{S}||\mathcal{A}| \times 1$ vector $\mathbf{\Xi}\boldsymbol{\mathcal{M}}_\eta \mathbf{p}$ with $\mathbf{\Xi}\boldsymbol{\mathcal{M}}_\eta \mathbf{p}(sa) = \boldsymbol{\mathcal{M}}_\eta \mathbf{p}(s)$. Further, one can easily verify that for any policy matrix defined by (2):

$$\mathbf{\Pi\Xi} = \begin{pmatrix} \sum\limits_{a\in\mathcal{A}} \boldsymbol{\pi}(s_1 a) & & & \\ & \sum\limits_{a\in\mathcal{A}} \boldsymbol{\pi}(s_2 a) & & \\ & & \ddots & \\ & & & \sum\limits_{a\in\mathcal{A}} \boldsymbol{\pi}(s_{|\mathcal{S}|} a) \end{pmatrix} = \mathbf{I}, \tag{9}$$

where $\mathbf{I}$ is a $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix.

We know that $\lim_{\eta\to\infty} \boldsymbol{\mathcal{M}}_\eta \mathbf{p} = \boldsymbol{\mathcal{M}}_\infty \mathbf{p}$. Further, It is not difficult to show that the following inequality holds for $\|\boldsymbol{\mathcal{M}}_\infty \mathbf{p} - \boldsymbol{\mathcal{M}}_\eta \mathbf{p}\|_\infty$:

Lemma 1. *Let* $\mathbf{p}$ *be a* $|\mathcal{S}||\mathcal{A}| \times 1$ *vector of action preferences and* η *be a positive constant, then an upper bound for* $\|\boldsymbol{\mathcal{M}}_\eta \mathbf{p} - \boldsymbol{\mathcal{M}}_\infty \mathbf{p}\|_\infty$ *can be obtained as follows:*

$$\|\boldsymbol{\mathcal{M}}_\eta \mathbf{p} - \boldsymbol{\mathcal{M}}_\infty \mathbf{p}\|_\infty \leq \frac{1}{\eta}\log(|\mathcal{A}|).$$

Proof. (sketch) First, we note that $\boldsymbol{\mathcal{M}}_\eta \mathbf{p} - \boldsymbol{\mathcal{M}}_\infty \mathbf{p} = \boldsymbol{\mathcal{M}}_\eta(\mathbf{p} - \boldsymbol{\mathcal{M}}_\infty \mathbf{p}) \leq 0$. Then we apply equation (12) of the main article to each entries of $\mathbf{p} - \boldsymbol{\mathcal{M}}_\infty \mathbf{p}$. Now, one can easily show that $\mathcal{L}_\eta(\mathbf{p} - \boldsymbol{\mathcal{M}}_\infty \mathbf{p})(s) \leq 0$ for all $s \in \mathcal{S}$. This combined with the fact that for the probability distribution π: $H_\pi(s) \leq \log(|\mathcal{A}|)$ derive $0 \leq \boldsymbol{\mathcal{M}}_\infty \mathbf{p} - \boldsymbol{\mathcal{M}}_\eta \mathbf{p} \leq \frac{1}{\eta}\log(|\mathcal{A}|)$. The result then follows by taking the sup-norm. □

Defining $\mathbf{p}_n$ as the action preference resulted by the n^{th} iteration of (7), we have:

$$\begin{aligned} \mathbf{p}_n &= \mathbf{p}_{n-1} + \bar{\mathbf{r}} + \gamma \mathbf{T}\boldsymbol{\mathcal{M}}_\eta \mathbf{p}_{n-1} - \mathbf{\Xi}\boldsymbol{\mathcal{M}}_\eta \mathbf{p}_{n-1} \\ &= \mathbf{p}_{n-1} + \bar{\mathbf{r}} + \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-1} - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-1} \end{aligned}, \qquad n = 1, 2, 3, \cdots, \tag{10}$$

where $\mathbf{p}_0 = \mathbf{p}$ and $\mathbf{\Pi}_n$ is a policy distribution matrix associated with the policy distribution vector $\boldsymbol{\pi}_n$ given by:

$$\boldsymbol{\pi}_n(sa) = \frac{\exp(\eta \mathbf{p}_n(sa))}{Z(s)}, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{11}$$

and $Z(s) = \sum_{a\in\mathcal{A}} \exp(\eta \mathbf{p}_n(sa))$ is the normalization factor.

B Proof of Lemma 1 of The Main Article

The optimal value value function is defined as follows:

$$V_{\bar{\pi}}^*(s) = \max_{\pi_s \in \Pi_s} \left[\sum_{a\in\mathcal{A}} \pi_s(a) \sum_{s'\in\mathcal{S}} T_{ss'}^a \left(r_{ss'}^a + \gamma V_{\bar{\pi}}^*(s')\right) - \frac{1}{\eta}\mathrm{KL}\left(\pi_s \| \bar{\pi}_s\right) \right] \tag{12}$$

The maximization in (12) can be performed in closed form using Lagrange multipliers:

$$\mathcal{L}(s, \lambda_s) = \sum_{a\in\mathcal{A}} \pi_s(a) \sum_{s'\in\mathcal{S}} T_{ss'}^a \left(r_{ss'}^a + \gamma V_{\bar{\pi}}^*(s')\right) - \frac{1}{\eta}\mathrm{KL}\left(\pi_s \| \bar{\pi}_s\right) - \lambda_s \left[\sum_{a\in\mathcal{A}} \pi_s(a) - 1 \right].$$

The necessary condition for the extremum with respect to π_s is:

$$0 = \frac{\partial \mathcal{L}(s, \lambda_s)}{\partial \pi_s(a)} = \sum_{s\in\mathcal{S}} T_{ss'}^a \left(r_{ss'}^a + \gamma V_{\bar{\pi}}^*(s)\right) - \frac{1}{\eta} - \frac{1}{\eta}\log\left(\frac{\pi_s(a)}{\bar{\pi}_s(a)}\right) - \lambda_s.$$

So, the solution is:

$$\pi_s(a) = \bar{\pi}_s(a) \exp\left(-\eta\lambda_s - 1\right) \exp\left[\eta \sum_{s' \in \mathcal{S}} T^a_{ss'}\left(r^a_{ss'} + \gamma V^*_{\bar{\pi}}(s')\right)\right], \qquad \forall s \in \mathcal{S}. \tag{13}$$

The Lagrange multipliers can be solved from the constraints:

$$\begin{aligned} 1 = \sum_{a \in \mathcal{A}} \pi_s(a) &= \exp\left(-\eta\lambda_s - 1\right) \sum_{a \in \mathcal{A}} \bar{\pi}_s(a) \exp\left[\eta \sum_{s' \in \mathcal{S}} T^a_{ss'}\left(r^a_{ss'} + \gamma V^*_{\bar{\pi}}(s')\right)\right], \\ \lambda_s &= \frac{1}{\eta} \log \sum_{a \in \mathcal{A}} \bar{\pi}_s(a) \exp\left[\eta \sum_{s' \in \mathcal{S}} T^a_{ss'}\left(r^a_{ss'} + \gamma V^*_{\bar{\pi}}(s')\right)\right] - \frac{1}{\eta}. \end{aligned} \tag{14}$$

By plugging (14) in to (13) we have:

$$\pi_s(a) = \frac{\bar{\pi}_s(a) \exp\left[\eta \sum_{s' \in \mathcal{S}} T^a_{ss'}\left(r^a_{ss'} + \gamma V^*_{\bar{\pi}}(s')\right)\right]}{\sum_{a \in \mathcal{A}} \bar{\pi}_s(a) \exp\left[\eta \sum_{s' \in \mathcal{S}} T^a_{ss'}\left(r^a_{ss'} + \gamma V^*_{\bar{\pi}}(s')\right)\right]}, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{15}$$

Substituting policy (15) in (12), we obtain:

$$V^*_{\bar{\pi}}(s) = \frac{1}{\eta} \log \sum_{a \in \mathcal{A}} \bar{\pi}_s(a) \exp\left[\eta \sum_{s' \in \mathcal{S}} T^a_{ss'}\left(r^a_{ss'} + \gamma V^*_{\bar{\pi}}(s')\right)\right].$$

C Proof of Theorem 1

This section provides a formal analysis of the convergence behavior of DPP. Our objective is to establish a rate of convergence for the value function of the policy induced by DPP.

Our main result is that at iteration n of DPP we have:

$$\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty < \delta_n, \qquad \text{where} \qquad \delta_n = O\left(\frac{1}{n}\right). \tag{16}$$

Here, $\mathbf{q}^{\boldsymbol{\pi}_n}$ is a vector of the action-values under the policy $\boldsymbol{\pi}_n$ and $\boldsymbol{\pi}_n$ is the policy induced by DPP at iteration n. Equation (16) implies that, asymptotically, the policy induced by DPP, $\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\pi}_n$, converges to the optimal policy $\boldsymbol{\pi}^*$.

To derive (16) one needs to relate $\mathbf{q}^{\boldsymbol{\pi}_n}$ to the optimal $\mathbf{q}^*$. Unfortunately, finding a direct relation between $\mathbf{q}^{\boldsymbol{\pi}_n}$ and $\mathbf{q}^*$ is not an easy task. Instead, we can relate $\mathbf{q}^{\boldsymbol{\pi}_n}$ to $\mathbf{q}^*$ via an auxiliary $\mathbf{q}^A_n$, which we define later in this section. In the remainder of this section, we first express $\boldsymbol{\pi}_n$ in terms of $\mathbf{q}^A_n$. Then, we obtain an upper bound on the normed error $\left\|\mathbf{q}^A_n - \mathbf{q}^*\right\|_\infty$ in lemma 2. Finally, we use these two results to derive a bound on the normed error $\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty$.

In order to express $\boldsymbol{\pi}_n$ in terms of $\mathbf{q}^A_n$, we expand the corresponding action preference $\mathbf{p}_n$ by recursive substitution of (10):

$$\begin{aligned} \mathbf{p}_n &= \mathbf{p}_{n-1} + \bar{\mathbf{r}} + \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-1} - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-1} \\ &= \mathbf{p}_{n-2} + 2\bar{\mathbf{r}} + \gamma \mathbf{T}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} - \mathbf{\Xi}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} + \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-2} + \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\bar{\mathbf{r}} \\ &\quad + \gamma^2 \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{T}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} - \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{\Xi}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-2} - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\bar{\mathbf{r}} \\ &\quad - \gamma \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{T}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} + \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{\Xi}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} \\ &= 2\bar{\mathbf{r}} + \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\bar{\mathbf{r}} + \mathbf{p}_{n-2} + \gamma \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-2} + \gamma^2 \mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{T}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\bar{\mathbf{r}} \\ &\quad - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{p}_{n-2} - \gamma \mathbf{\Xi}\mathbf{\Pi}_{n-1}\mathbf{T}\mathbf{\Pi}_{n-2}\mathbf{p}_{n-2} \qquad \text{by (9).} \end{aligned} \tag{17}$$

Equation (17) expresses $\mathbf{p}_n$ in terms of $\mathbf{p}_{n-2}$, where in the last step we have canceled some terms because of (9). To express $\mathbf{p}_n$ in terms of $\mathbf{p}_0$, we proceed with the expansion of (10):

$$\begin{aligned}
\mathbf{p}_n &= n\bar{\mathbf{r}} + \sum_{k=1}^{n-1} (n-k)\gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\Big)\bar{\mathbf{r}} + \mathbf{p}_0 + \sum_{k=1}^{n} \gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\Big)\mathbf{p}_0 \\
&\quad - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\Big[(n-1)\bar{\mathbf{r}} + \sum_{k=1}^{n-2}(n-k-1)\gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j-1}\Big)\bar{\mathbf{r}}\Big] \\
&\quad - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\Big(\mathbf{p}_0 + \sum_{k=1}^{n-1} \gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j-1}\Big)\mathbf{p}_0\Big) \\
&= n\Big(\bar{\mathbf{r}} + \sum_{k=1}^{n-1} \gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\Big)\bar{\mathbf{r}}\Big) - \gamma \sum_{k=1}^{n-1} k\gamma^{k-1} \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\Big)\bar{\mathbf{r}} + \mathbf{p}_0 \\
&\quad + \sum_{k=1}^{n} \gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\Big)\mathbf{p}_0 - (n-1)\mathbf{\Xi}\mathbf{\Pi}_{n-1}\Big(\bar{\mathbf{r}} + \sum_{k=1}^{n-2} \gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j-1}\Big)\bar{\mathbf{r}}\Big) \\
&\quad + \mathbf{\Xi}\mathbf{\Pi}_{n-1}\Big[\gamma \sum_{k=1}^{n-2} k\gamma^{k-1} \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j-1}\Big)\bar{\mathbf{r}} - \mathbf{p}_0 - \sum_{k=1}^{n-1} \gamma^k \Big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j-1}\Big)\mathbf{p}_0\Big].
\end{aligned} \tag{18}$$

We define the auxiliary action-values $\mathbf{q}_n^A = \bar{\mathbf{r}} + \sum_{k=1}^{n-1} \gamma^k \big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\big)\bar{\mathbf{r}}$ and $\mathbf{q}_n^{\mathbf{P}} = \mathbf{p}_0 + \sum_{k=1}^{n} \gamma^k \big(\prod_{j=1}^{k} \mathbf{T}\mathbf{\Pi}_{n-j}\big)\mathbf{p}_0$ and write (18) as:

$$\mathbf{p}_n = n\mathbf{q}_n^A - \gamma\frac{\partial \mathbf{q}_n^A}{\partial\gamma} + \mathbf{q}_n^{\mathbf{P}} - \mathbf{\Xi}\mathbf{\Pi}_{n-1}\Big((n-1)\mathbf{q}_{n-1}^A - \gamma\frac{\partial \mathbf{q}_{n-1}^A}{\partial\gamma} + \mathbf{q}_{n-1}^{\mathbf{P}}\Big) = n\mathbf{q}_n^A + \mathbf{c}_n - \mathbf{\Xi}\mathbf{d}_n. \tag{19}$$

Here, $\mathbf{c}_n = \mathbf{q}_n^{\mathbf{P}} - \gamma\partial\mathbf{q}_n^A/\partial\gamma$ is a $|\mathcal{S}||\mathcal{A}| \times 1$ vector bounded by:

$$\|\mathbf{c}_n\|_\infty \leq c_{\max} = \frac{\gamma}{(1-\gamma)^2}\bar{R}_{\max} + \frac{1}{(1-\gamma)}\|\mathbf{p}_0\|_\infty, \qquad n = 1, 2, 3, \cdots, \tag{20}$$

and $\mathbf{d}_n = \mathbf{\Pi}_{n-1}\big((n-1)\mathbf{q}_{n-1}^A - \gamma\partial\mathbf{q}_{n-1}^A/\partial\gamma + \mathbf{q}_{n-1}^{\mathbf{P}}\big)$ is a $|\mathcal{S}| \times 1$ vector. From its definition, it is easy to see that $\mathbf{q}_n^A$ satisfies the following Bellman equation:

$$\mathbf{q}_n^A = \bar{\mathbf{r}} + \gamma\mathbf{T}\mathbf{\Pi}_{n-1}\mathbf{q}_{n-1}^A. \tag{21}$$

Note the difference between (21) and (4): whereas $\mathbf{q}^\pi$ evolves according to a fixed policy π, $\mathbf{q}_n^A$ evolves according to a policy that evolves according to DPP.

Finally, we express $\boldsymbol{\pi}_n$ in terms of $\mathbf{q}_n^A$. By plugging (19) in (11) and taking in to account that $\mathbf{\Xi}\mathbf{d}_n(sa) = \mathbf{d}_n(s)$, we obtain:

$$\boldsymbol{\pi}_n(sa) = \frac{\exp\left(\eta\left(n\mathbf{q}_n^A(sa) + \mathbf{c}_n(sa) - \mathbf{d}_n(s)\right)\right)}{Z(s)} = \frac{\exp\left(\eta\left(n\mathbf{q}_n^A(sa) + \mathbf{c}_n(sa)\right)\right)}{Z'(s)}, \quad (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{22}$$

where $Z'(s) = Z(s)\exp(\eta\mathbf{d}_n(s))$ is the normalization factor. Equation (22) establishes the relation between $\boldsymbol{\pi}_n$ and $\mathbf{q}_n^A$. To relate $\mathbf{q}_n^A$ and $\mathbf{q}^*$ we state the following lemma, that establishes a bound on $\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty$:

Lemma 2 (A bound on $\|\mathbf{q}_n^A - \mathbf{q}^*\|_\infty$)**.** *Let assumption 1 hold, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$ and n be a positive integer, also, for keeping the representation succinct, assume that both $\|\bar{\mathbf{r}}\|_\infty$ and $\|\mathbf{p}_0\|_\infty$ are bounded from above by some constant $L > 0$, then the following inequality holds:*

$$\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty \leq \delta_n^1,$$

where δ_n^1 is given by:

$$\delta_n^1 = 2\gamma\frac{(1-\gamma)^2\log(|\mathcal{A}|)/\eta + 2L}{n(1-\gamma)^4} + 2\gamma^{n-1}\frac{L}{1-\gamma}. \tag{23}$$

Proof.

$$\begin{aligned}\left\|\mathbf{q}_n^A-\mathbf{q}^*\right\|_\infty &= \left\|\sum_{k=1}^{n-1}\left(\boldsymbol{\mathcal{T}}^{k-1}\mathbf{q}_{n-k+1}^A-\boldsymbol{\mathcal{T}}^{k}\mathbf{q}_{n-k}^A\right)+\boldsymbol{\mathcal{T}}^{n-1}\mathbf{q}_1^A-\mathbf{q}^*\right\|_\infty \\ &\leq \sum_{k=1}^{n-1}\left\|\boldsymbol{\mathcal{T}}^{k-1}\mathbf{q}_{n-k+1}^A-\boldsymbol{\mathcal{T}}^{k}\mathbf{q}_{n-k}^A\right\|_\infty+\left\|\boldsymbol{\mathcal{T}}^{n-1}\mathbf{q}_1^A-\mathbf{q}^*\right\|_\infty \\ &\leq \sum_{k=1}^{n-1}\gamma^{k-1}\left\|\mathbf{q}_{n-k+1}^A-\boldsymbol{\mathcal{T}}\mathbf{q}_{n-k}^A\right\|_\infty+\gamma^{n-1}\left\|\mathbf{q}_1^A-\mathbf{q}^*\right\|_\infty \qquad \text{by (6).}\end{aligned} \tag{24}$$

For $1\leq k\leq n-1$ we obtain:

$$\begin{aligned}\left\|\mathbf{q}_{n-k+1}^A-\boldsymbol{\mathcal{T}}\mathbf{q}_{n-k}^A\right\|_\infty &= \gamma\left\|\mathbf{T}\boldsymbol{\Pi}_{n-k}\mathbf{q}_{n-k}^A-\mathbf{T}\boldsymbol{\mathcal{M}}_\infty\mathbf{q}_{n-k}^A\right\|_\infty \quad \text{by (21) and (5)} \\ &\leq \gamma\left\|\boldsymbol{\Pi}_{n-k}\mathbf{q}_{n-k}^A-\boldsymbol{\mathcal{M}}_\infty\mathbf{q}_{n-k}^A\right\|_\infty \quad \text{by Hölder's inequality.}\end{aligned} \tag{25}$$

By monotonicity property for any action-value vector $\mathbf{q}$ and distribution matrix $\boldsymbol{\Pi}$ we have:

$$\boldsymbol{\Pi}\mathbf{q}(s)=\sum_{a\in\mathcal{A}}\boldsymbol{\pi}(sa)\mathbf{q}(sa)\leq\boldsymbol{\mathcal{M}}_\infty\mathbf{q}(s), \qquad \forall s\in\mathcal{S}. \tag{26}$$

By comparing (26) with (25), we obtain:

$$\begin{aligned}\left\|\mathbf{q}_{n-k+1}^A-\boldsymbol{\mathcal{T}}\mathbf{q}_{n-k}^A\right\|_\infty &\leq \gamma\max_{s\in\mathcal{S}}\left(\boldsymbol{\mathcal{M}}_\infty\mathbf{q}_{n-k}^A(s)-\boldsymbol{\Pi}_{n-k}\mathbf{q}_{n-k}^A(s)\right) \\ &\leq \gamma\max_{s\in\mathcal{S}}\left(\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_{n-k}^A+\frac{\mathbf{c}_{n-k}}{n-k}\right)(s)+\frac{c_{\max}}{n-k}-\boldsymbol{\Pi}_{n-k}\mathbf{q}_{n-k}^A(s)\right) \\ &= \gamma\left\|\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_{n-k}^A+\frac{\mathbf{c}_{n-k}}{n-k}\right)-\boldsymbol{\Pi}_{n-k}\mathbf{q}_{n-1}^A\right\|_\infty+\frac{\gamma c_{\max}}{n-k} \\ &\leq \gamma\left\|\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_{n-k}^A+\frac{\mathbf{c}_{n-k}}{n-k}\right)-\boldsymbol{\Pi}_{n-k}\left(\mathbf{q}_{n-k}^A+\frac{\mathbf{c}_{n-k}}{n-k}\right)\right\|_\infty+\frac{2\gamma c_{\max}}{n-k} \\ &= \gamma\left\|\boldsymbol{\mathcal{M}}_{\eta(n-k)}\left(\mathbf{q}_{n-k}^A+\frac{\mathbf{c}_{n-k}}{n-k}\right)-\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_{n-k}^A+\frac{\mathbf{c}_{n-k}}{n-k}\right)\right\|_\infty+\frac{2\gamma c_{\max}}{n-k} \\ &\leq \gamma\frac{\log(|\mathcal{A}|)/\eta+2c_{\max}}{n-k} \qquad \text{by lemma 1.}\end{aligned} \tag{27}$$

By substitution of (27) in (24), we obtain:

$$\left\|\mathbf{q}_n^A-\mathbf{q}^*\right\|_\infty\leq\left(\log(|\mathcal{A}|)/\eta+2c_{\max}\right)\sum_{k=1}^{n-1}\frac{\gamma^k}{n-k}+\frac{2\gamma^{n-1}L}{1-\gamma}. \tag{28}$$

It is not difficult to show that $\sum_{k=1}^{n-1}\frac{\gamma^k}{n-k}\leq 2\gamma/(n(1-\gamma)^2)$. Combining this with (28) yields:

$$\begin{aligned}\left\|\mathbf{q}_n^A-\mathbf{q}^*\right\|_\infty &\leq 2\gamma\frac{\log(|\mathcal{A}|)/\eta+2c_{\max}}{n(1-\gamma)^2}+2\gamma^{n-1}\frac{L}{1-\gamma} \\ &\leq 2\gamma\frac{(1-\gamma)^2\log(|\mathcal{A}|)/\eta+2L}{n(1-\gamma)^4}+2\gamma^{n-1}\frac{L}{1-\gamma} \qquad \text{by (37).}\end{aligned}$$

□

Equation (22) expresses the policy $\boldsymbol{\pi}_n$ in terms of the auxiliary $\mathbf{q}_n^A$. Lemma 2 provides an upper bound on the normed-error $\left\|\mathbf{q}_n^A-\mathbf{q}^*\right\|_\infty$. We use these two results to derive a bound on the normed-error $\left\|\mathbf{q}^{\boldsymbol{\pi}_n}-\mathbf{q}^*\right\|_\infty$:

We start by noting that $\mathbf{q}^{\boldsymbol{\pi}_n}$ is obtained by infinite application of the operator $\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n}$ to an arbitrary initial $\mathbf{q}$-vector, for which we take $\mathbf{q}^*$, then:

$$
\begin{aligned}
\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty &= \lim_{m\to\infty} \left\| \sum_{k=1}^{m} \Big(\boldsymbol{\mathcal{T}}^k_{\boldsymbol{\pi}_n} \mathbf{q}^* - \boldsymbol{\mathcal{T}}^{k-1}_{\boldsymbol{\pi}_n} \boldsymbol{\mathcal{T}} \mathbf{q}^* \Big) \right\|_\infty \\
&\leq \lim_{m\to\infty} \sum_{k=1}^{m} \left\| \boldsymbol{\mathcal{T}}^k_{\boldsymbol{\pi}_n} \mathbf{q}^* - \boldsymbol{\mathcal{T}}^{k-1}_{\boldsymbol{\pi}_n} \boldsymbol{\mathcal{T}} \mathbf{q}^* \right\|_\infty \\
&\leq \lim_{m\to\infty} \sum_{k=1}^{m} \gamma^{k-1} \left\| \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n} \mathbf{q}^* - \boldsymbol{\mathcal{T}} \mathbf{q}^* \right\|_\infty \qquad \text{by (6)} \\
&\leq \frac{1}{1-\gamma} \left\| \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n} \mathbf{q}^* - \boldsymbol{\mathcal{T}} \mathbf{q}^* \right\|_\infty \\
&= \frac{\gamma}{1-\gamma} \left\| \mathbf{T}\boldsymbol{\Pi}_n \mathbf{q}^* - \mathbf{T}\boldsymbol{\mathcal{M}}_\infty \mathbf{q}^* \right\|_\infty \\
&\leq \frac{\gamma}{1-\gamma} \left\| \boldsymbol{\Pi}_n \mathbf{q}^* - \boldsymbol{\mathcal{M}}_\infty \mathbf{q}^* \right\|_\infty \qquad \text{by Hölder's inequality.}
\end{aligned}
\tag{29}
$$

Along similar lines with the proof of lemma 2, we obtain:

$$
\begin{aligned}
\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty &\leq \frac{\gamma}{1-\gamma} \left\| \boldsymbol{\Pi}_n \mathbf{q}^* - \boldsymbol{\mathcal{M}}_\infty \mathbf{q}^* \right\|_\infty \\
&= \frac{\gamma}{1-\gamma} \max_{s\in\mathcal{S}} \left(\boldsymbol{\mathcal{M}}_\infty \mathbf{q}^*(s) - \boldsymbol{\Pi}_n \mathbf{q}^*(s) \right) \\
&\leq \frac{\gamma}{1-\gamma} \left[\max_{s\in\mathcal{S}} \Big(\boldsymbol{\mathcal{M}}_\infty \mathbf{q}^A_n(s) - \boldsymbol{\Pi}_n \mathbf{q}^*(s) \Big) + \delta^1_n \right] && \text{by lemma 2} \\
&\leq \frac{\gamma}{1-\gamma} \max_{s\in\mathcal{S}} \left(\boldsymbol{\mathcal{M}}_\infty \left(\mathbf{q}^A_n + \frac{\mathbf{c}_n}{n} \right)(s) - \boldsymbol{\Pi}_n \mathbf{q}^*(s) \right) \\
&\quad + \frac{\gamma}{1-\gamma} \left(\delta^1_n + \frac{c_{\max}}{n} \right) && \text{by (37)} \\
&\leq \frac{\gamma}{1-\gamma} \left\| \boldsymbol{\mathcal{M}}_\infty \left(\mathbf{q}^A_n + \frac{\mathbf{c}_n}{n} \right) - \boldsymbol{\Pi}_n \left(\mathbf{q}^A_n + \frac{\mathbf{c}_n}{n} \right) \right\|_\infty \\
&\quad + \frac{\gamma}{1-\gamma} \left(2\delta^1_n + \frac{2c_{\max}}{n} \right) \\
&= \frac{\gamma}{1-\gamma} \left\| \boldsymbol{\mathcal{M}}_\infty \left(\mathbf{q}^A_n + \frac{\mathbf{c}_n}{n} \right) - \boldsymbol{\mathcal{M}}_{\eta n} \left(\mathbf{q}^A_n + \frac{\mathbf{c}_n}{n} \right) \right\|_\infty \\
&\quad + \frac{\gamma}{1-\gamma} \left(2\delta^1_n + \frac{2c_{\max}}{n} \right) \\
&\leq \frac{\gamma}{1-\gamma} \left[\frac{\log(|\mathcal{A}|)}{n\eta} + 2\delta^1_n + \frac{2c_{\max}}{n} \right] && \text{by lemma 1} \\
&= \frac{\gamma}{1-\gamma} \left[\frac{(1-\gamma)^2 \log(|\mathcal{A}|)/\eta + 2L}{n(1-\gamma)^2} + 2\delta^1_n \right] && \text{by (37)} \\
&= \frac{\gamma}{1-\gamma} \left[\frac{(1-\gamma)^2 \log(|\mathcal{A}|)/\eta + 2L}{n(1-\gamma)^2} + 4\gamma \frac{(1-\gamma)^2 \log(|\mathcal{A}|)/\eta + 2L}{n(1-\gamma)^4} + 4\gamma^{n-1} \frac{L}{1-\gamma} \right] \\
&= \frac{\gamma}{1-\gamma} \left[4\frac{(1-\gamma)^2 \log(|\mathcal{A}|)/\eta + 2L}{n(1-\gamma)^4} + 4\gamma^{n-1} \frac{L}{1-\gamma} \right]
\end{aligned}
$$

This completes the proof.

D Proof of Theorem 2

First, we show that there exists a limit for $\mathbf{p}_n$ in infinity. Then, we compute this limit. Before we proceed with the proof we review the following assumption and corollary from the main article

Assumption 2. *We assume that MDP has a unique deterministic optimal policy* $\boldsymbol{\pi}^*$ *given by:*

$$\boldsymbol{\pi}^*(sa) = \begin{cases} 1 & a = a^*(s) \\ 0 & \text{otherwise} \end{cases}, \qquad \forall s \in \mathcal{S},$$

where $a^*(s) = \arg\max_{a\in\mathcal{A}} \mathbf{q}^*(sa)$.

Corollary 1. *The following relation holds in limit:*

$$\lim_{n\to+\infty} \mathbf{q}^{\boldsymbol{\pi}_n} = \mathbf{q}^*, \quad \forall(s,a) \in \mathcal{S}\times\mathcal{A}.$$

The convergence of $\mathbf{p}_n$ to a limit can be established by proving the convergence of all the terms in RHS of (19). We have already proven the convergence of $\mathbf{q}_n^A$ to $\mathbf{q}^*$ in lemma 2. Then, the convergence of $\partial\mathbf{q}_n^A/\partial\gamma$ to $\partial\mathbf{q}^*/\partial\gamma$ is immediate. Corollary 1 together with assumption 2 yields the convergence of $\boldsymbol{\pi}_n$ to $\boldsymbol{\pi}^*$. The convergence of $\mathbf{q}_n^{\mathbf{P}}$ to some limit $\mathbf{q}^{\mathbf{P}}$ then follows. Accordingly, there exists a limit for $\mathbf{p}_n$.

Now, we compute the limit of $\mathbf{p}_n$. Combining corollary 1 with (19) and taking in to account that $\mathbf{v}^* = \boldsymbol{\Pi}^*\mathbf{q}^*$ yields:

$$\begin{aligned}\lim_{n\to\infty}\mathbf{p}_n(sa) &= \lim_{n\to\infty}\left[n\mathbf{q}^*(sa) - \gamma\frac{\partial\mathbf{q}^*(sa)}{\partial\gamma} + \mathbf{q}^{\mathbf{P}}(sa) - (n-1)\mathbf{v}^*(s) + \gamma\frac{\partial\mathbf{v}^*(s)}{\partial\gamma} - \mathbf{q}^{\mathbf{P}}(sa^*)\right] \\ &= \lim_{n\to\infty} n(\mathbf{q}^*(sa) - \mathbf{v}^*(s)) \\ &\quad + \gamma\Big(\frac{\partial\mathbf{v}^*(s)}{\partial\gamma} - \frac{\partial\mathbf{q}^*(sa)}{\partial\gamma}\Big) + \mathbf{q}^{\mathbf{P}}(sa) - \mathbf{q}^{\mathbf{P}}(sa^*) + \mathbf{v}^*(s) \\ &= \begin{cases} \mathbf{v}^*(s) & a = a^*(s) \\ -\infty & \text{otherwise} \end{cases}.\end{aligned}$$

E Proof of Theorem 3

This section provides a formal theoretical analysis of the performance of dynamic policy programming in the presence of approximation error. Each iteration of approximate dynamic programming can be characterized by (30). Our objective is to establish a L_∞-norm performance loss bound of the policy induced by approximate DPP.

$$\begin{aligned}\mathbf{p}_n &= \mathbf{p}_{n-1} + \bar{\mathbf{r}} + \gamma\mathbf{T}\boldsymbol{\mathcal{M}}_\eta\mathbf{p}_{n-1} - \boldsymbol{\Xi}\boldsymbol{\mathcal{M}}_\eta\mathbf{p}_{n-1} + \boldsymbol{\epsilon}_{n-1} \\ &= \mathbf{p}_{n-1} + \bar{\mathbf{r}} + \gamma\mathbf{T}\boldsymbol{\Pi}_{n-1}\mathbf{p}_{n-1} - \boldsymbol{\Xi}\boldsymbol{\Pi}_{n-1}\mathbf{p}_{n-1} + \boldsymbol{\epsilon}_{n-1}\end{aligned}, \qquad n = 1,2,3,\cdots, \tag{30}$$

Our main result that at iteration n of approximate dynamic policy programming, we have:

$$\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty \le \delta_n, \tag{31}$$

where:

$$\delta_n = 4\gamma\frac{(1-\gamma)^2\log(|\mathcal{A}|)/\eta + 4L}{n(1-\gamma)^5} \quad + 4\gamma^n\frac{L}{(1-\gamma)^2} + \frac{2\gamma}{1-\gamma}\sum_{k=1}^{n}\gamma^{n-k}\bar{\varepsilon}_k.$$

with $\bar{\varepsilon}_k = \left\|1/k\sum_{j=0:k-1}\boldsymbol{\epsilon}_j\right\|_\infty$. Here, $\mathbf{q}^{\boldsymbol{\pi}_n}$ is a vector of the action-values under the policy $\boldsymbol{\pi}_n$ and $\boldsymbol{\pi}_n$ is the policy induced after n iteration of approximate DPP.

To relate $\mathbf{q}^*$ with $\mathbf{q}^{\boldsymbol{\pi}_n}$, we can relate $\mathbf{q}^{\boldsymbol{\pi}_n}$ to $\mathbf{q}^*$ via an auxiliary $\mathbf{q}_n^A$, which we define later in this section. In the remainder of this section, we first express $\boldsymbol{\pi}_n$ in terms of $\mathbf{q}_n^A$ in lemma 3. Then, we obtain an upper bound on the normed error $\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty$ in lemma 4. Finally, we use these two results to derive (31).

We begin our analysis with the following lemma:

Lemma 3. *Let n be a positive integer and $\mathbf{p}_0$ denotes the initial action preferences. Also, lets define the auxiliary action-value functions $\mathbf{q}_n^A$ and $\mathbf{q}_n^{\mathbf{P}}$ as:*

$$\mathbf{q}_n^A = \bar{\mathbf{r}} + \bar{\boldsymbol{\epsilon}}_n + \sum_{k=0}^{n-1} \gamma^k \big(\textstyle\prod_{j=1}^{k} \mathbf{T\Pi}_{n-j}\big)(\bar{\mathbf{r}} + \bar{\boldsymbol{\epsilon}}_{n-k}) \qquad \mathbf{q}_n^{\mathbf{P}} = \mathbf{p}_0 + \sum_{k=1}^{n} \gamma^k \big(\textstyle\prod_{j=1}^{k} \mathbf{T\Pi}_{n-j}\big)\mathbf{p}_0 \tag{32}$$

with $\bar{\boldsymbol{\epsilon}}_n = \sum_{l=0}^{n-1} \boldsymbol{\epsilon}_l / n$, then we have:

$$\mathbf{p}_n = n\mathbf{q}_n^A - \gamma \frac{\partial \mathbf{q}_n^A}{\partial \gamma} + \mathbf{q}_n^{\mathbf{P}} - \boldsymbol{\Xi\Pi}_{n-1}\Big((n-1)\mathbf{q}_{n-1}^A - \gamma \frac{\partial \mathbf{q}_{n-1}^A}{\partial \gamma} + \mathbf{q}_{n-1}^{\mathbf{P}}\Big) \tag{33}$$

Proof. In order to express $\boldsymbol{\pi}_n$ in terms of $\mathbf{q}_n^A$, we expand the corresponding action preference $\mathbf{p}_n$ by recursive substitution of (30):

$$\begin{aligned}
\mathbf{p}_n &= \mathbf{p}_{n-1} + \bar{\mathbf{r}} + \gamma\mathbf{T\Pi}_{n-1}\mathbf{p}_{n-1} - \boldsymbol{\Xi\Pi}_{n-1}\mathbf{p}_{n-1} + \boldsymbol{\epsilon}_{n-1} \\
&= \mathbf{p}_{n-2} + \boldsymbol{\epsilon}_{n-2} + 2\bar{\mathbf{r}} + \gamma\mathbf{T\Pi}_{n-2}\mathbf{p}_{n-2} + \gamma\mathbf{T\Pi}_{n-2}\boldsymbol{\epsilon}_{n-2} - \boldsymbol{\Xi\Pi}_{n-2}\mathbf{p}_{n-2} + \gamma\mathbf{T\Pi}_{n-1}\mathbf{p}_{n-2} \\
&\quad + \gamma^2\mathbf{T\Pi}_{n-1}\mathbf{T\Pi}_{n-2}\mathbf{p}_{n-2} + \gamma\mathbf{T\Pi}_{n-1}\bar{\mathbf{r}} - \gamma\mathbf{T\Pi}_{n-1}\boldsymbol{\Xi\Pi}_{n-2}\mathbf{p}_{n-2} - \boldsymbol{\Xi\Pi}_{n-1}\mathbf{p}_{n-2} \\
&\quad - \boldsymbol{\Xi\Pi}_{n-1}\boldsymbol{\epsilon}_{n-2} - \boldsymbol{\Xi\Pi}_{n-1}\bar{\mathbf{r}} - \gamma\boldsymbol{\Xi\Pi}_{n-1}\mathbf{T\Pi}_{n-2}\mathbf{p}_{n-2} + \boldsymbol{\Xi\Pi}_{n-1}\boldsymbol{\Xi\Pi}_{n-2}\mathbf{p}_{n-2} + \boldsymbol{\epsilon}_{n-1} \\
&= 2\bar{\mathbf{r}} + \gamma\mathbf{T\Pi}_{n-1}\bar{\mathbf{r}} + \mathbf{p}_{n-2} + +\gamma\mathbf{T\Pi}_{n-1}\mathbf{p}_{n-2} + \gamma^2\mathbf{T\Pi}_{n-1}\mathbf{T\Pi}_{n-2}\mathbf{p}_{n-2} - \boldsymbol{\Xi\Pi}_{n-1}\bar{\mathbf{r}} \\
&\quad - \boldsymbol{\Xi\Pi}_{n-1}\mathbf{p}_{n-2} - \gamma\boldsymbol{\Xi\Pi}_{n-1}\mathbf{T\Pi}_{n-2}\mathbf{p}_{n-2} + \boldsymbol{\epsilon}_{n-1} + \boldsymbol{\epsilon}_{n-2} + \gamma\mathbf{T\Pi}_{n-2}\boldsymbol{\epsilon}_{n-2} - \boldsymbol{\Xi\Pi}_{n-1}\boldsymbol{\epsilon}_{n-2}
\end{aligned} \tag{34}$$

Equation (34) expresses $\mathbf{p}_n$ in terms of $\mathbf{p}_{n-2}$, where in the last step we have canceled some terms because of (9). To express $\mathbf{p}_n$ in terms of $\mathbf{p}_0$, we proceed with the expansion of (30):

$$\begin{aligned}
\mathbf{p}_n &= \sum\nolimits_{k=0}^{n-1} (n-k)\gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\bar{\mathbf{r}} + \sum\nolimits_{k=0}^{n} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\mathbf{p}_0 \\
&\quad + \sum\nolimits_{k=0}^{n-1} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big) \sum\nolimits_{l=0}^{n-k-1} \boldsymbol{\epsilon}_l - \boldsymbol{\Xi\Pi}_{n-1} \sum\nolimits_{k=0}^{n-2} (n-k-1)\gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\bar{\mathbf{r}} \\
&\quad - \boldsymbol{\Xi\Pi}_{n-1}\Big(\sum\nolimits_{k=0}^{n-1} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\mathbf{p}_0 + \sum\nolimits_{k=0}^{n-2} \gamma^k \prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1} \sum\nolimits_{l=0}^{n-k-2} \boldsymbol{\epsilon}_l\Big) \\
&= n\sum\nolimits_{k=0}^{n-1} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\bar{\mathbf{r}} - \gamma \sum\nolimits_{k=1}^{n-1} k\gamma^{k-1} \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\bar{\mathbf{r}} \\
&\quad + n\sum\nolimits_{k=0}^{n-1} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\bar{\boldsymbol{\epsilon}}_{n-k} - \gamma \sum\nolimits_{k=1}^{n-1} k\gamma^{k-1} \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\bar{\boldsymbol{\epsilon}}_{n-k} \\
&\quad + \sum\nolimits_{k=0}^{n} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j}\Big)\mathbf{p}_0 - (n-1)\boldsymbol{\Xi\Pi}_{n-1} \sum\nolimits_{k=0}^{n-2} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\bar{\mathbf{r}} \\
&\quad - (n-1)\boldsymbol{\Xi\Pi}_{n-1} \sum\nolimits_{k=0}^{n-2} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\bar{\boldsymbol{\epsilon}}_{n-k-1} \\
&\quad + \gamma\boldsymbol{\Xi\Pi}_{n-1} \sum\nolimits_{k=1}^{n-2} k\gamma^{k-1} \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\bar{\boldsymbol{\epsilon}}_{n-k-1} \\
&\quad + \boldsymbol{\Xi\Pi}_{n-1}\Big[\gamma \sum\nolimits_{k=1}^{n-2} k\gamma^{k-1} \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\bar{\mathbf{r}} - \sum\nolimits_{k=0}^{n-1} \gamma^k \Big(\prod\nolimits_{j=1}^{k} \mathbf{T\Pi}_{n-j-1}\Big)\mathbf{p}_0\Big]
\end{aligned} \tag{35}$$

Equation (33) then follows by comparing (35) with (32). □

It is easy to see that $\mathbf{q}_n^A$, defined in lemma 3, satisfies the following Bellman equation:

$$\mathbf{q}_n^A = \bar{\mathbf{r}} + \gamma\mathbf{T\Pi}_{n-1}\mathbf{q}_{n-1}^A + \bar{\boldsymbol{\epsilon}}_n \tag{36}$$

Note the difference between (36) and (4): whereas $\mathbf{q}^\pi$ evolves according to a fixed policy π, $\mathbf{q}_n^A$ evolves according to a policy that evolves according to approximate DPP.

For brevity we introduce the new variables $\mathbf{c}_n = \mathbf{q}_n^{\mathbf{P}} - \gamma\partial\mathbf{q}_n^A/\partial\gamma$[1] and $\mathbf{d}_n = \mathbf{\Pi}_{n-1}\big((n-1)\mathbf{q}_{n-1}^A - \gamma\partial\mathbf{q}_{n-1}^A/\partial\gamma + \mathbf{q}_{n-1}^{\mathbf{P}}\big)$ and re-express (33) as:

$$\mathbf{p}_n = n\mathbf{q}_n^A + \mathbf{c}_n - \mathbf{\Xi}\mathbf{d}_n \tag{38}$$

Finally, we express $\boldsymbol{\pi}_n$ in terms of $\mathbf{q}_n^A$. By plugging (33) in (11) and taking in to account that $\mathbf{\Xi}\mathbf{d}_n(sa) = \mathbf{d}_n(s)$, we obtain:

$$\begin{aligned}\boldsymbol{\pi}_n(sa) &= \frac{\exp\left(\eta(n\mathbf{q}_n^A(sa) + \mathbf{c}_n(sa) - \mathbf{d}_n(s))\right)}{Z(s)} \\ &= \frac{\exp\left(\eta\left(n\mathbf{q}_n^A(sa) + \mathbf{c}_n(sa)\right)\right)}{Z'(s)},\end{aligned} \qquad (s,a)\in\mathcal{S}\times\mathcal{A}, \tag{39}$$

where $Z'(s) = Z(s)\exp(\mathbf{d}_n(s))$ is the normalization factor. Equation (39) establishes the relation between $\boldsymbol{\pi}_n$ and $\mathbf{q}_n^A$. To relate $\mathbf{q}_n^A$ and $\mathbf{q}^*$, we state the following lemma, that establishes a bound on $\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty$:

Lemma 4 (L_∞ bound on $\mathbf{q}_n^A - \mathbf{q}^*$). *Let assumption 1 hold and $\mathbf{q}_n^A$ defined according to* (32). *Let $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$ and n be a positive integer, also, for keeping the representation succinct, assume that both $\|\bar{\mathbf{r}}\|_\infty$ and $\|\mathbf{p}_0\|_\infty$ are bounded from above by some constant $L>0$, then the following inequality holds:*

$$\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty \le \delta_n^2,$$

where δ_n^2 is given by:

$$\delta_n^2 = 2\gamma\frac{(1-\gamma)^2\log(|\mathcal{A}|)/\eta + 4L}{n(1-\gamma)^4} + 2\gamma^{n-1}\frac{L}{1-\gamma} + \sum_{k=1}^{n}\gamma^{n-k}\bar{\varepsilon}_k. \tag{40}$$

with $\bar{\varepsilon}_k = \|\boldsymbol{\epsilon}_k\|_\infty$.

Proof.

$$\begin{aligned}\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty &= \left\|\sum_{k=1}^{n-1}\left(\boldsymbol{\mathcal{T}}^{k-1}\mathbf{q}_{n-k+1}^A - \boldsymbol{\mathcal{T}}^{k}\mathbf{q}_{n-k}^A\right) + \boldsymbol{\mathcal{T}}^{n-1}\mathbf{q}_1^A - \mathbf{q}^*\right\|_\infty \\ &\le \sum_{k=1}^{n-1}\left\|\boldsymbol{\mathcal{T}}^{k-1}\mathbf{q}_{n-k+1}^A - \boldsymbol{\mathcal{T}}^{k}\mathbf{q}_{n-k}^A\right\|_\infty + \left\|\boldsymbol{\mathcal{T}}^{n-1}\mathbf{q}_1^A - \mathbf{q}^*\right\|_\infty \\ &\le \sum_{k=1}^{n-1}\gamma^{k-1}\left\|\mathbf{q}_{n-k+1}^A - \boldsymbol{\mathcal{T}}\mathbf{q}_{n-k}^A\right\|_\infty + \gamma^{n-1}\left\|\mathbf{q}_1^A - \mathbf{q}^*\right\|_\infty \qquad \text{by (6).}\end{aligned} \tag{41}$$

For all $l\in\mathcal{N}: 2\le l\le n-1$ we obtain:

$$\begin{aligned}\left\|\mathbf{q}_{l+1}^A - \boldsymbol{\mathcal{T}}\mathbf{q}_l^A\right\|_\infty &= \left\|\gamma\left(\mathbf{T}\mathbf{\Pi}_l\mathbf{q}_l^A - \mathbf{T}\boldsymbol{\mathcal{M}}_\infty\mathbf{q}_l^A\right) + \bar{\boldsymbol{\epsilon}}_{l+1}\right\|_\infty && \text{by (36) and (5)} \\ &\le \gamma\left\|\mathbf{T}\mathbf{\Pi}_l\mathbf{q}_l^A - \mathbf{T}\boldsymbol{\mathcal{M}}_\infty\mathbf{q}_l^A\right\|_\infty + \bar{\varepsilon}_{l+1} \\ &\le \gamma\left\|\mathbf{\Pi}_l\mathbf{q}_l^A - \boldsymbol{\mathcal{M}}_\infty\mathbf{q}_l^A\right\|_\infty + \bar{\varepsilon}_{l+1} && \text{by Hölder's inequality.}\end{aligned} \tag{42}$$

[1] Note that:

$$\|\mathbf{c}_n\|_\infty \le c_{\max} = \frac{\gamma}{(1-\gamma)^2}(\bar{R}_{\max} + \epsilon_{\max}) + \frac{1}{(1-\gamma)}\|\mathbf{p}_0\|_\infty, \qquad n = 1,2,3,\cdots, \tag{37}$$

with $\epsilon_{\max} = \max_{k=1,2,3,\ldots}\|\boldsymbol{\epsilon}_k\|_\infty$

Taking in to account that $\boldsymbol{\mathcal{M}}_\infty \mathbf{q}_l^A \geq \mathbf{\Pi}_l \mathbf{q}_l^A(s)$ we obtain:

$$
\begin{aligned}
\left\|\mathbf{q}_{l+1}^A - \boldsymbol{\mathcal{T}}\mathbf{q}_l^A\right\|_\infty &\leq \gamma \max_{s\in\mathcal{S}} \left(\boldsymbol{\mathcal{M}}_\infty \mathbf{q}_l^A(s) - \mathbf{\Pi}_l \mathbf{q}_l^A(s)\right) + \bar{\varepsilon}_{l+1} \\
&\leq \gamma \max_{s\in\mathcal{S}} \Big(\boldsymbol{\mathcal{M}}_\infty \Big(\mathbf{q}_l^A + \frac{\mathbf{c}_l}{l}\Big)(s) + \frac{c_{\max}}{l} - \mathbf{\Pi}_l \mathbf{q}_l^A(s)\Big) + \bar{\varepsilon}_{l+1} \\
&= \gamma \left\|\boldsymbol{\mathcal{M}}_\infty \Big(\mathbf{q}_l^A + \frac{\mathbf{c}_l}{l}\Big) - \mathbf{\Pi}_l \mathbf{q}_l^A\right\|_\infty + \frac{\gamma c_{\max}}{l} + \bar{\varepsilon}_{l+1} \\
&\leq \gamma \left\|\boldsymbol{\mathcal{M}}_\infty \Big(\mathbf{q}_l^A + \frac{\mathbf{c}_l}{l}\Big) - \mathbf{\Pi}_l \Big(\mathbf{q}_l^A + \frac{\mathbf{c}_l}{l}\Big)\right\|_\infty + \frac{2\gamma c_{\max}}{l} + \bar{\varepsilon}_{l+1} \\
&= \gamma \left\|\boldsymbol{\mathcal{M}}_{\eta l} \Big(\mathbf{q}_l^A + \frac{\mathbf{c}_l}{l}\Big) - \boldsymbol{\mathcal{M}}_\infty \Big(\mathbf{q}_l^A + \frac{\mathbf{c}_l}{l}\Big)\right\|_\infty + \frac{2\gamma c_{\max}}{l} + \bar{\varepsilon}_{l+1} \\
&\leq \gamma \frac{\log(|\mathcal{A}|)/\eta + 2c_{\max}}{l} + \bar{\varepsilon}_{l+1} \qquad \text{by lemma 1.}
\end{aligned}
\tag{43}
$$

By substitution of (43) in (41), we obtain:

$$
\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty \leq \left(\log(|\mathcal{A}|)/\eta + 2c_{\max}\right) \sum_{k=1}^{n-1} \frac{\gamma^{n-k}}{k} + \frac{2\gamma^{n-1}L}{1-\gamma} + \sum_{k=1}^{n} \gamma^{n-k-1} \bar{\varepsilon}_k. \tag{44}
$$

It is not difficult to show that $\sum_{k=1}^{n-1} \frac{\gamma^k}{n-k} \leq 2\gamma/(n(1-\gamma)^2)$. Combining this with (44) yields:

$$
\begin{aligned}
\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty &\leq 2\gamma \frac{\log(|\mathcal{A}|)/\eta + 2c_{\max}}{n(1-\gamma)^2} + 2\gamma^{n-1} \frac{L}{1-\gamma} + \sum_{k=1}^{n} \gamma^{n-k} \bar{\varepsilon}_k \\
&\leq 2\gamma \frac{(1-\gamma)^2 \log(|\mathcal{A}|)/\eta + 4L}{n(1-\gamma)^4} + 2\gamma^{n-1} \frac{L}{1-\gamma} + \sum_{k=1}^{n} \gamma^{n-k} \bar{\varepsilon}_k.
\end{aligned}
$$

□

Equation (39) expresses the policy $\boldsymbol{\pi}_n$ in terms of the auxiliary $\mathbf{q}_n^A$. Lemma 4 provides a upper-bound on the normed-error $\left\|\mathbf{q}_n^A - \mathbf{q}^*\right\|_\infty$. We use these two results to derive a bound on the normed-error $\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty$.

We start by noting that $\mathbf{q}^{\boldsymbol{\pi}_n}$ is obtained by infinite application of the operator $\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n}$ to an arbitrary initial $\mathbf{q}$-vector, for which we take $\mathbf{q}^*$, then:

$$
\begin{aligned}
\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty &= \lim_{m\to\infty} \left\| \sum_{k=1}^{m} \Big(\boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n}^{k} \mathbf{q}^* - \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n}^{k-1} \boldsymbol{\mathcal{T}} \mathbf{q}^*\Big) \right\|_\infty \\
&\leq \lim_{m\to\infty} \sum_{k=1}^{m} \left\| \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n}^{k} \mathbf{q}^* - \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n}^{k-1} \boldsymbol{\mathcal{T}} \mathbf{q}^* \right\|_\infty \\
&\leq \lim_{m\to\infty} \sum_{k=1}^{m} \gamma^{k-1} \left\| \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n} \mathbf{q}^* - \boldsymbol{\mathcal{T}} \mathbf{q}^* \right\|_\infty \qquad \text{by (6)} \\
&\leq \frac{1}{1-\gamma} \left\| \boldsymbol{\mathcal{T}}_{\boldsymbol{\pi}_n} \mathbf{q}^* - \boldsymbol{\mathcal{T}} \mathbf{q}^* \right\|_\infty \\
&= \frac{\gamma}{1-\gamma} \left\| \mathbf{T}\mathbf{\Pi}_n \mathbf{q}^* - \mathbf{T}\boldsymbol{\mathcal{M}}_\infty \mathbf{q}^* \right\|_\infty \\
&\leq \frac{\gamma}{1-\gamma} \left\| \mathbf{\Pi}_n \mathbf{q}^* - \boldsymbol{\mathcal{M}}_\infty \mathbf{q}^* \right\|_\infty \qquad \text{by Hölder's inequality.}
\end{aligned}
\tag{45}
$$

Along similar lines with the proof of lemma 4, we obtain:

$$\begin{aligned}
\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty &\leq \frac{\gamma}{1-\gamma}\|\boldsymbol{\Pi}_n\mathbf{q}^* - \boldsymbol{\mathcal{M}}_\infty\mathbf{q}^*\|_\infty \\
&= \frac{\gamma}{1-\gamma}\max_{s\in\mathcal{S}}\left(\boldsymbol{\mathcal{M}}_\infty\mathbf{q}^*(s) - \boldsymbol{\Pi}_n\mathbf{q}^*(s)\right) \\
&\leq \frac{\gamma}{1-\gamma}\left[\max_{s\in\mathcal{S}}\Big(\boldsymbol{\mathcal{M}}_\infty\mathbf{q}_n^A(s) - \boldsymbol{\Pi}_n\mathbf{q}^*(s)\Big) + \delta_n^2\right] && \text{by lemma 4} \\
&\leq \frac{\gamma}{1-\gamma}\max_{s\in\mathcal{S}}\left(\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right)(s) - \boldsymbol{\Pi}_n\mathbf{q}^*(s)\right) \\
&\quad + \frac{\gamma}{1-\gamma}\Big(\delta_n^2 + \frac{c_{\max}}{n}\Big) && \text{by (37)} \\
&\leq \frac{\gamma}{1-\gamma}\left\|\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right) - \boldsymbol{\Pi}_n\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right)\right\|_\infty \\
&\quad + \frac{\gamma}{1-\gamma}\Big(2\delta_n^2 + \frac{2c_{\max}}{n}\Big)
\end{aligned}$$

$$\begin{aligned}
\|\mathbf{q}^{\boldsymbol{\pi}_n} - \mathbf{q}^*\|_\infty &\leq \frac{\gamma}{1-\gamma}\left\|\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right) - \boldsymbol{\Pi}_n\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right)\right\|_\infty \\
&\quad + \frac{\gamma}{1-\gamma}\Big(2\delta_n^2 + \frac{2c_{\max}}{n}\Big) \\
&= \frac{\gamma}{1-\gamma}\left\|\boldsymbol{\mathcal{M}}_\infty\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right) - \boldsymbol{\mathcal{M}}_{\eta n}\left(\mathbf{q}_n^A + \frac{\mathbf{c}_n}{n}\right)\right\|_\infty \\
&\quad + \frac{\gamma}{1-\gamma}\Big(2\delta_n^2 + \frac{2c_{\max}}{n}\Big) \\
&\leq \frac{\gamma}{1-\gamma}\left[\frac{\log(|\mathcal{A}|)}{n\eta} + 2\delta_n^2 + \frac{2c_{\max}}{n}\right] && \text{by lemma 1} \\
&= \frac{\gamma}{1-\gamma}\left[\frac{(1-\gamma)^2\log(|\mathcal{A}|)/\eta + 4L}{n(1-\gamma)^2} + 2\delta_n^2\right] && \text{by (37)} \\
&= \frac{\gamma}{1-\gamma}\left[4\frac{(1-\gamma)^2\log(|\mathcal{A}|)/\eta + 4L}{n(1-\gamma)^4} + 4\gamma^{n-1}\frac{L}{1-\gamma} + 2\sum_{k=1}^{n}\gamma^{n-k}\bar{\varepsilon}_k\right]
\end{aligned}$$

This completes the proof of theorem 3.

References

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmount, Massachusetts, third edition.