Supplementary Material for Non-Stationary Approximate Modified Policy Iteration

A. Proof of Theorem 3

For clarity, we here provide a detailed and complete proof. Throughout this proof we will write P_k (resp. P_*) for the transition kernel P_{π_k} (resp. P_{π_*}) induced by the stationary policy π_k (resp. π_*). We will write T_k (resp. T_*) for the associated Bellman operator. Similarly, we will write $P_{k,\ell}$ for the transition kernel associated with the non-stationary policy $\pi_{k,\ell}$ and $T_{k,\ell}$ for its associated Bellman operator.

For $k \geq 0$ we define the following quantities:

- $b_k = T_{k+1}v_k - T_{k+1,\ell}T_{k+1}v_k$. This quantity which we will call the *residual* may be viewed as a non-stationary analogue of the Bellman residual $v_k - T_{k+1}v_k$.
- $s_k = v_k - v_{\pi_{k,\ell}} - \epsilon_k$. We will call it *shift*, as it measures the shift between the value $v_{\pi_{k,\ell}}$ and the estimate v_k before incurring the error.
- $d_k = v_* - v_k + \epsilon_k$. This quantity, called *distance* thereafter, provides the distance between the k^{th} value function (before the error is added) and the optimal value function.
- $l_k = v_* - v_{\pi_{k,\ell}}$. This is the *loss* of the policy $v_{\pi_{k,\ell}}$. The loss is always non-negative since no policy can have a value greater than or equal to v_*.

The proof is outlined as follows. We first provide a bound on b_k which will be used to express both the bounds on s_k and d_k. Then, observing that $l_k = s_k + d_k$ will allow to express the bound of $\|l_k\|_\infty$ stated by Theorem 3. Our arguments extend those made by Scherrer et al. (2012) in the specific case $\ell = 1$.

We will repeatedly use the fact that since policy π_{k+1} is greedy with respect to v_k, we have

$$\forall \pi', \;\; T_{k+1}v_k \geq T_{\pi'}v_k. \tag{5}$$

For a non-stationary policy $\pi_{k,\ell}$, the induced ℓ-step transition kernel is

$$P_{k,\ell} = P_k P_{k-1} \cdots P_{k-\ell+1}.$$

As a consequence, for any function $f : \mathcal{S} \to \mathbb{R}$, the operator $T_{k,\ell}$ may be expressed as:

$$T_{k,\ell}f = r_k + \gamma P_{k,1}r_{k-1} + \gamma^2 P_{k,2}r_{k-2} + \cdots + \gamma^{\ell-1}P_{k,\ell-1}r_{k-\ell+1} + \gamma^\ell P_{k,\ell}f$$

then, for any function $g : \mathcal{S} \to \mathbb{R}$, we have

$$T_{k,\ell}f - T_{k,\ell}g = \gamma^\ell P_{k,\ell}(f - g) \tag{6}$$

and

$$T_{k,\ell}(f + g) = T_{k,\ell}f + \gamma^\ell P_{k,\ell}(g). \tag{7}$$

The following notation will be useful.

Definition 1 (Scherrer et al. (2012))**.** *For a positive integer n, we define $\mathbb{P}_n$ as the set of discounted transition kernels that are defined as follows:*

1. *for any set of n policies $\{\pi_1, \ldots, \pi_n\}$, $(\gamma P_{\pi_1})(\gamma P_{\pi_2}) \cdots (\gamma P_{\pi_n}) \in \mathbb{P}_n$,*
2. *for any $\alpha \in (0, 1)$ and $P_1, P_2 \in \mathbb{P}_n$, $\alpha P_1 + (1 - \alpha)P_2 \in \mathbb{P}_n$*

With some abuse of notation, we write Γ^n for denoting any element of $\mathbb{P}_n$.

Example 1 (Γ^n notation)**.** *If we write a transition kernel P as $P = \alpha_1\Gamma^i + \alpha_2\Gamma^j\Gamma^k = \alpha_1\Gamma^i + \alpha_2\Gamma^{j+k}$, it should be read as: "There exists $P_1 \in \mathbb{P}_i$, $P_2 \in \mathbb{P}_j$, $P_3 \in \mathbb{P}_k$ and $P_4 \in \mathbb{P}_{j+k}$ such that $P = \alpha_1 P_1 + \alpha_2 P_2 P_3 = \alpha_1 P_1 + \alpha_2 P_4$.".*

We first provide three lemmas bounding the residual, the shift and the distance, respectively.

Lemma 2 (residual bound)**.** *The residual b_k satisfies the following bound:*

$$b_k \leq \sum_{i=1}^{k} \Gamma^{(\ell m+1)(k-i)} x_i + \Gamma^{(\ell m+1)k} b_0$$

where

$$x_k = (I - \Gamma^\ell)\Gamma\epsilon_k.$$

Proof. We have:

$$\begin{aligned}
b_k &= T_{k+1}v_k - T_{k+1,\ell}T_{k+1}v_k \\
&\leq T_{k+1}v_k - T_{k+1,\ell}T_{k-\ell+1}v_k && \{T_{k+1}v_k \geq T_{k-\ell+1}v_k \text{ (5)}\} \\
&= T_{k+1}v_k - T_{k+1}T_{k,\ell}v_k \\
&= \gamma P_{k+1}\left(v_k - T_{k,\ell}v_k\right) \\
&= \gamma P_{k+1}\left((T_{k,\ell})^m T_k v_{k-1} + \epsilon_k - T_{k,\ell}\left((T_{k,\ell})^m T_k v_{k-1} + \epsilon_k\right)\right) \\
&= \gamma P_{k+1}\left((T_{k,\ell})^m T_k v_{k-1} - (T_{k,\ell})^{m+1} T_k v_{k-1} + (I - \gamma^\ell P_{k,\ell})\epsilon_k\right) && \{(7)\} \\
&= \gamma P_{k+1}\left((\gamma^\ell P_{k,\ell})^m \left(T_k v_{k-1} - T_{k,\ell}T_k v_{k-1}\right) + (I - \gamma^\ell P_{k,\ell})\epsilon_k\right) && \{(6)\} \\
&= \gamma P_{k+1}\left((\gamma^\ell P_{k,\ell})^m b_{k-1} + (I - \gamma^\ell P_{k,\ell})\epsilon_k\right).
\end{aligned}$$

Which can be written as

$$b_k \leq \Gamma(\Gamma^{\ell m} b_{k-1} + (I - \Gamma^\ell)\epsilon_k) = \Gamma^{\ell m+1} b_{k-1} + x_k.$$

Then, by induction:

$$b_k \leq \sum_{i=0}^{k-1} \Gamma^{(\ell m+1)i} x_{k-i} + \Gamma^{(\ell m+1)k} b_0 = \sum_{i=1}^{k} \Gamma^{(\ell m+1)(k-i)} x_i + \Gamma^{(\ell m+1)k} b_0.$$

□

Lemma 3 (distance bound)**.** *The distance d_k satisfies the following bound:*

$$d_k \leq \sum_{i=1}^{k} \sum_{j=0}^{mi-1} \Gamma^{\ell j+i-1} x_{k-i} + \sum_{i=1}^{k} \Gamma^{i-1} y_{k-i} + z_k,$$

where

$$y_k = -\Gamma\epsilon_k$$

and

$$z_k = \sum_{i=0}^{mk-1} \Gamma^{k-1+\ell i} b_0 + \Gamma^k d_0.$$

Proof. First expand d_k:

$$\begin{aligned}
d_k &= v_* - v_k + \epsilon_k \\
&= v_* - (T_{k,\ell})^m T_k v_{k-1} \\
&= v_* - T_k v_{k-1} + T_k v_{k-1} - T_{k,\ell} T_k v_{k-1} + T_{k,\ell} T_k v_{k-1} - (T_{k,\ell})^2 T_k v_{k-1} \\
&\quad + (T_{k,\ell})^2 T_k v_{k-1} - \cdots - (T_{k,\ell})^{m-1} T_k v_{k-1} + (T_{k,\ell})^{m-1} T_k v_{k-1} - (T_{k,\ell})^m T_k v_{k-1} \\
&= v_* - T_k v_{k-1} + \sum_{i=0}^{m-1} (T_{k,\ell})^i T_k v_{k-1} - (T_{k,\ell})^{i+1} T_k v_{k-1} \\
&= T_* v_* - T_k v_{k-1} + \sum_{i=0}^{m-1} (\gamma^\ell P_{k,\ell})^i \left(T_k v_{k-1} - T_{k,\ell} T_k v_{k-1} \right) && \{(6)\} \\
&\leq T_* v_* - T_* v_{k-1} + \sum_{i=0}^{m-1} (\gamma^\ell P_{k,\ell})^i b_{k-1} && \{T_k v_{k-1} \geq T_* v_{k-1} \text{ (5)}\} \\
&= \gamma P_* (v_* - v_{k-1}) + \sum_{i=0}^{m-1} (\gamma^\ell P_{k,\ell})^i b_{k-1} && \{(6)\} \\
&= \gamma P_* d_{k-1} - \gamma P_* \epsilon_{k-1} + \sum_{i=0}^{m-1} (\gamma^\ell P_{k,\ell})^i b_{k-1} && \{d_k = v_* - v_k + \epsilon_k\} \\
&= \Gamma d_{k-1} + y_{k-1} + \sum_{i=0}^{m-1} \Gamma^{\ell i} b_{k-1}.
\end{aligned}$$

Then, by induction

$$d_k \leq \sum_{j=0}^{k-1} \Gamma^{k-1-j} \left(y_j + \sum_{p=0}^{m-1} \Gamma^{\ell p} b_j \right) + \Gamma^k d_0.$$

Using the bound on b_k from Lemma 2 we get:

$$\begin{aligned}
d_k &\leq \sum_{j=0}^{k-1} \Gamma^{k-1-j} \left(y_j + \sum_{p=0}^{m-1} \Gamma^{\ell p} \left(\sum_{i=1}^{j} \Gamma^{(\ell m+1)(j-i)} x_i + \Gamma^{(\ell m+1)j} b_0 \right) \right) + \Gamma^k d_0 \\
&= \sum_{j=0}^{k-1} \sum_{p=0}^{m-1} \sum_{i=1}^{j} \Gamma^{k-1-j+\ell p+(\ell m+1)(j-i)} x_i + \sum_{j=0}^{k-1} \sum_{p=0}^{m-1} \Gamma^{k-1-j+\ell p+(\ell m+1)j} b_0 + \Gamma^k d_0 + \sum_{i=1}^{k} \Gamma^{i-1} y_{k-i}.
\end{aligned}$$

First we have:

$$\begin{aligned}
\sum_{j=0}^{k-1} \sum_{p=0}^{m-1} \sum_{i=1}^{j} \Gamma^{k-1-j+\ell p+(\ell m+1)(j-i)} x_i &= \sum_{i=1}^{k-1} \sum_{j=i}^{k-1} \sum_{p=0}^{m-1} \Gamma^{k-1+\ell(p+mj)-i(\ell m+1)} x_i \\
&= \sum_{i=1}^{k-1} \sum_{j=0}^{m(k-i)-1} \Gamma^{k-1+\ell(j+mi)-i(\ell m+1)} x_i \\
&= \sum_{i=1}^{k-1} \sum_{j=0}^{m(k-i)-1} \Gamma^{\ell j+k-i-1} x_i \\
&= \sum_{i=1}^{k-1} \sum_{j=0}^{mi-1} \Gamma^{\ell j+i-1} x_{k-i}.
\end{aligned}$$

Second we have:

$$\sum_{j=0}^{k-1}\sum_{p=0}^{m-1}\Gamma^{k-1-j+\ell p+(\ell m+1)j}b_0=\sum_{j=0}^{k-1}\sum_{p=0}^{m-1}\Gamma^{k-1+\ell(p+mj)}b_0=\sum_{i=0}^{mk-1}\Gamma^{k-1+\ell i}b_0=z_k-\Gamma^k d_0.$$

Hence

$$d_k\le\sum_{i=1}^{k}\sum_{j=0}^{mi-1}\Gamma^{\ell j+i-1}x_{k-i}+\sum_{i=1}^{k}\Gamma^{i-1}y_{k-i}+z_k.$$

□

Lemma 4 (shift bound)**.** *The shift s_k is bounded by:*

$$s_k\le\sum_{i=1}^{k-1}\sum_{j=mi}^{\infty}\Gamma^{\ell j+i-1}x_{k-i}+w_k,$$

where

$$w_k=\sum_{j=mk}^{\infty}\Gamma^{\ell j+k-1}b_0.$$

Proof. Expanding s_k we obtain:

$$\begin{aligned}
s_k&=v_k-v_{\pi_{k,\ell}}-\epsilon_k\\
&=(T_{k,\ell})^mT_kv_{k-1}-v_{\pi_{k,\ell}}\\
&=(T_{k,\ell})^mT_kv_{k-1}-(T_{k,\ell})^\infty T_{k,\ell}T_kv_{k-1} && \{\forall f:\ v_{\pi_{k,\ell}}=(T_{k,\ell})^\infty f\}\\
&=(\gamma^\ell P_{k,\ell})^m\sum_{j=0}^{\infty}(\gamma^\ell P_{k,\ell})^j\left(T_kv_{k-1}-T_{k,\ell}T_kv_{k-1}\right)\\
&=\Gamma^{\ell m}\sum_{j=0}^{\infty}\Gamma^{\ell j}b_{k-1}\\
&=\sum_{j=0}^{\infty}\Gamma^{\ell m+\ell j}b_{k-1}.
\end{aligned}$$

Plugging the bound on b_k of Lemma 2 we get:

$$\begin{aligned}
s_k&\le\sum_{j=0}^{\infty}\Gamma^{\ell m+\ell j}\left(\sum_{i=1}^{k-1}\Gamma^{(\ell m+1)(k-1-i)}x_i+\Gamma^{(\ell m+1)(k-1)}b_0\right)\\
&=\sum_{j=0}^{\infty}\sum_{i=1}^{k-1}\Gamma^{\ell m+\ell j+(\ell m+1)(k-1-i)}x_i+\sum_{j=0}^{\infty}\Gamma^{\ell m+\ell j+(\ell m+1)(k-1)}b_0\\
&=\sum_{j=0}^{\infty}\sum_{i=1}^{k-1}\Gamma^{\ell(j+mi)+i-1}x_{k-i}+\sum_{j=0}^{\infty}\Gamma^{\ell(j+mk)+k-1}b_0\\
&=\sum_{i=1}^{k-1}\sum_{j=mi}^{\infty}\Gamma^{\ell j+i-1}x_{k-i}+\sum_{j=mk}^{\infty}\Gamma^{\ell j+k-1}b_0\\
&=\sum_{i=1}^{k-1}\sum_{j=mi}^{\infty}\Gamma^{\ell j+i-1}x_{k-i}+w_k.
\end{aligned}$$

□

Lemma 5 (loss bound)**.** *The loss l_k is bounded by:*

$$l_k \leq \sum_{i=1}^{k-1} \Gamma^i \left(\sum_{j=0}^{\infty} \Gamma^{\ell j}(I - \Gamma^{\ell}) - I \right) \epsilon_{k-i} + \eta_k,$$

where

$$\eta_k = z_k + w_k = \sum_{i=0}^{mk-1} \Gamma^{k-1+\ell i} b_0 + \Gamma^k d_0 + \sum_{j=mk}^{\infty} \Gamma^{\ell j + k-1} b_0 = \sum_{i=0}^{\infty} \Gamma^{\ell i + k - 1} b_0 + \Gamma^k d_0.$$

Proof. Using Lemmas 3 and 4, we have:

$$\begin{aligned} l_k &= s_k + d_k \\ &\leq \sum_{i=1}^{k-1} \sum_{j=mi}^{\infty} \Gamma^{\ell j + i - 1} x_{k-i} + \sum_{i=1}^{k-1} \sum_{j=0}^{mi-1} \Gamma^{\ell j + i - 1} x_{k-i} + \sum_{i=1}^{k} \Gamma^{i-1} y_{k-i} + z_k + w_k \\ &= \sum_{i=1}^{k-1} \sum_{j=0}^{\infty} \Gamma^{\ell j + i - 1} x_{k-i} + \sum_{i=1}^{k} \Gamma^{i-1} y_{k-i} + \eta_k. \end{aligned}$$

Plugging back the values of x_k and y_k and using the fact that $\epsilon_0 = 0$ we obtain:

$$\begin{aligned} l_k &\leq \sum_{i=1}^{k-1} \sum_{j=0}^{\infty} \Gamma^{\ell j + i - 1} (I - \Gamma^{\ell}) \Gamma \epsilon_{k-i} + \sum_{i=1}^{k-1} \Gamma^{i-1} (-\Gamma) \epsilon_{k-i} - \Gamma^k \epsilon_0 + \eta_k \\ &= \sum_{i=1}^{k-1} \left(\sum_{j=0}^{\infty} \Gamma^{\ell j + i} (I - \Gamma^{\ell}) \epsilon_{k-i} - \Gamma^i \epsilon_{k-i} \right) + \eta_k \\ &= \sum_{i=1}^{k-1} \Gamma^i \left(\sum_{j=0}^{\infty} \Gamma^{\ell j} (I - \Gamma^{\ell}) - I \right) \epsilon_{k-i} + \eta_k. \end{aligned}$$

□

We now provide a bound of η_k in terms of d_0:

Lemma 6.

$$\eta_k \leq \Gamma^k \left(\sum_{i=0}^{\infty} \Gamma^i (\Gamma - I) + I \right) d_0.$$

Proof. First recall that

$$\eta_k = \sum_{i=0}^{\infty} \Gamma^{\ell i + k - 1} b_0 + \Gamma^k d_0.$$

In order to bound η_k in terms of d_0 only, we express b_0 in terms of d_0:

$$\begin{aligned}
b_0 &= T_1 v_0 - (T_1)^\ell T_1 v_0 \\
&= T_1 v_0 - (T_1)^2 v_0 + (T_1)^2 v_0 - \cdots - (T_1)^\ell v_0 + (T_1)^\ell v_0 - (T_1)^{\ell+1} v_0 \\
&= \sum_{i=1}^{\ell} (\gamma P_1)^i (v_0 - T_1 v_0) \\
&= \sum_{i=1}^{\ell} (\gamma P_1)^i (v_0 - v_* + T_* v_* - T_* v_0 + T_* v_0 - T_1 v_0) \\
&\leq \sum_{i=1}^{\ell} (\gamma P_1)^i (v_0 - v_* + T_* v_* - T_* v_0) && \{T_1 v_0 \geq T_* v_0 \ (5)\} \\
&= \sum_{i=1}^{\ell} (\gamma P_1)^i (\gamma P_* - I) d_0.
\end{aligned}$$

Consequently, we have:

$$\begin{aligned}
\eta_k &\leq \sum_{i=0}^{\infty} \Gamma^{\ell i + k - 1} \sum_{j=1}^{\ell} (\gamma P_1)^j (\gamma P_* - I) d_0 + \Gamma^k d_0 \\
&= \sum_{i=0}^{\infty} \Gamma^{\ell i + k} \sum_{j=0}^{\ell-1} (\gamma P_1)^j (\gamma P_* - I) d_0 + \Gamma^k d_0 \\
&= \Gamma^k \left(\sum_{i=0}^{\infty} \Gamma^{\ell i} \sum_{j=0}^{\ell-1} \Gamma^j (\Gamma - I) + I \right) d_0 \\
&= \Gamma^k \left(\sum_{i=0}^{\infty} \Gamma^i (\Gamma - I) + I \right) d_0.
\end{aligned}$$

□

We now conclude the proof of Theorem 3. Taking the absolute value in Lemma 6 we obtain:

$$|\eta_k| \leq \Gamma^k \left(\sum_{i=0}^{\infty} \Gamma^i (\Gamma + I) + I \right) |d_0| = 2 \sum_{i=k}^{\infty} \Gamma^i |d_0|$$

Since l_k is non-negative, from Lemma 5 we have:

$$|l_k| \leq \sum_{i=1}^{k-1} \Gamma^i \left(\sum_{j=0}^{\infty} \Gamma^{\ell j} (I + \Gamma^\ell) + I \right) |\epsilon_{k-i}| + |\eta_k| = 2 \sum_{i=1}^{k-1} \Gamma^i \sum_{j=0}^{\infty} \Gamma^{\ell j} |\epsilon_{k-i}| + 2 \sum_{i=k}^{\infty} \Gamma^i |d_0|. \tag{8}$$

Since $\|v\|_\infty = \max |v|$, $d_0 = v_* - v_0$ and $l_k = v_* - v_{\pi_{k,\ell}}$, we can take the maximum in (8) and conclude that:

$$\left\| v_* - v_{\pi_{k,\ell}} \right\|_\infty \leq \frac{2(\gamma - \gamma^k)}{(1-\gamma)(1-\gamma^\ell)} 2\epsilon + \frac{\gamma^k}{1-\gamma} \|v_* - v_0\|_\infty.$$

B. Proof of Theorem 4

We shall prove the following result.

Lemma 7. *Consider NS-AMPI with parameters $m \geq 0$ and $\ell \geq 1$ applied on the problem of Figure 1, starting from $v_0 = 0$ and all initial policies $\pi_0, \pi_{-1}, \ldots, \pi_{-\ell+2}$ equal to π_*. Assume that at each iteration k, the following error terms are applied, for some $\epsilon \geq 0$:*

$$\forall i, \;\; \epsilon_k(i) = \begin{cases} -\epsilon & \text{if } i = k \\ \epsilon & \text{if } i = k+\ell \\ 0 & \text{otherwise} \end{cases}.$$

Then NS-AMPI can[8] generate a sequence of value-policy pairs that is described below.

For all iterations $k \geq 1$, the policy π_k takes the optimal action in all states but k, that is

$$\forall i \geq 2, \;\; \pi_k(i) = \begin{cases} \rightarrow & \text{if } i = k \\ \leftarrow & \text{otherwise} \end{cases} \tag{9}$$

For all iterations $k \geq 1$, the value function v_k satisfies the following equations:

- *For all $i < k$:*
$$v_k(i) = -\gamma^{(k-1)(\ell m+1)}\epsilon \tag{10.a}$$
- *For all i such that $k \leq i \leq k + ((k-1)m+1)\ell$:*
  - *For $i = k + (qm+p+1)\ell$ with $q \geq 0$ and $0 \leq p < m$ (i.e. $i = k + n\ell$, $n \geq 1$):*
$$v_k(i) = \gamma^{q(\ell m+1)}\left(\frac{\gamma^{\ell(p+1)} - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k-q} + \mathbb{1}_{[p=0]}\epsilon + \sum_{j=1}^{k-q-1}\gamma^{j(\ell m+1)}\left(\frac{\gamma^\ell - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k-q-j} + \epsilon\right)\right) \tag{10.b}$$
  - *For $i = k$:*
$$v_k(k) = v_k(k+\ell) + r_k - 2\epsilon \tag{10.c}$$
  - *For $i = k + q\ell + p$ with $0 \leq q \leq (k-1)m - 1$ and $1 \leq p < \ell$:*
$$v_k(i) = -\gamma^{(k-1)(\ell m+1)}\epsilon \tag{10.d}$$
  - *Otherwise, i.e. when $i = k + (k-1)m\ell + p$ with $1 \leq p < \ell$:*
$$v_k(i) = 0 \tag{10.e}$$
- *For all $i > k + ((k-1)m+1)\ell$*
$$v_k(i) = 0 \tag{10.f}$$

The relative complexity of the different expressions of v_k in Lemma 7 is due to the presence of nested periodic patterns in the shape of the value function along the state space and the horizon. Figures 4 and 5 give the shape of the value function for different values of ℓ and m, exhibiting the periodic patterns. The proof of Lemma 7 is done by recurrence on k.

B.1. Base case $k = 1$

Since $v_0 = 0$, π_1 is the optimal policy that takes $\leftarrow$ in all states as desired. Hence, $(T_{1,\ell})^m T_1 v_0 = 0$ in all states. Accounting for the errors ϵ_1 we have $v_1 = (T_{1,\ell})^m T_1 v_0 + \epsilon_1 = \epsilon_1$. As can be seen on Figures 4 and 5, when $k = 1$ we only need to consider equations (10.b), (10.c), (10.e) and (10.f) since the others apply to an empty set of states.

First, we have
$$v_1(1+\ell) = \epsilon_1(1+\ell) = \epsilon$$

[8] We write here "can" since at each iteration, several policies will be greedy with respect to the current value.

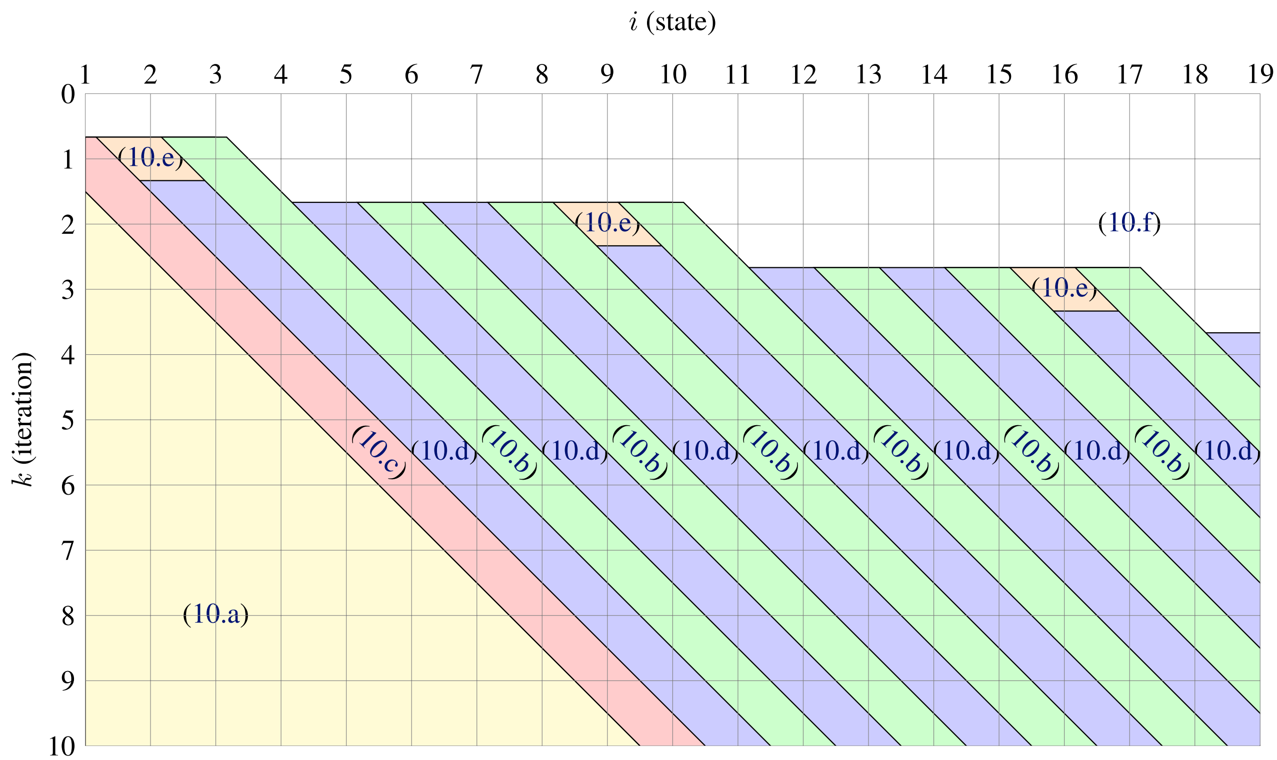

Figure 4. Shape of the value function with $\ell = 2$ and $m = 3$.

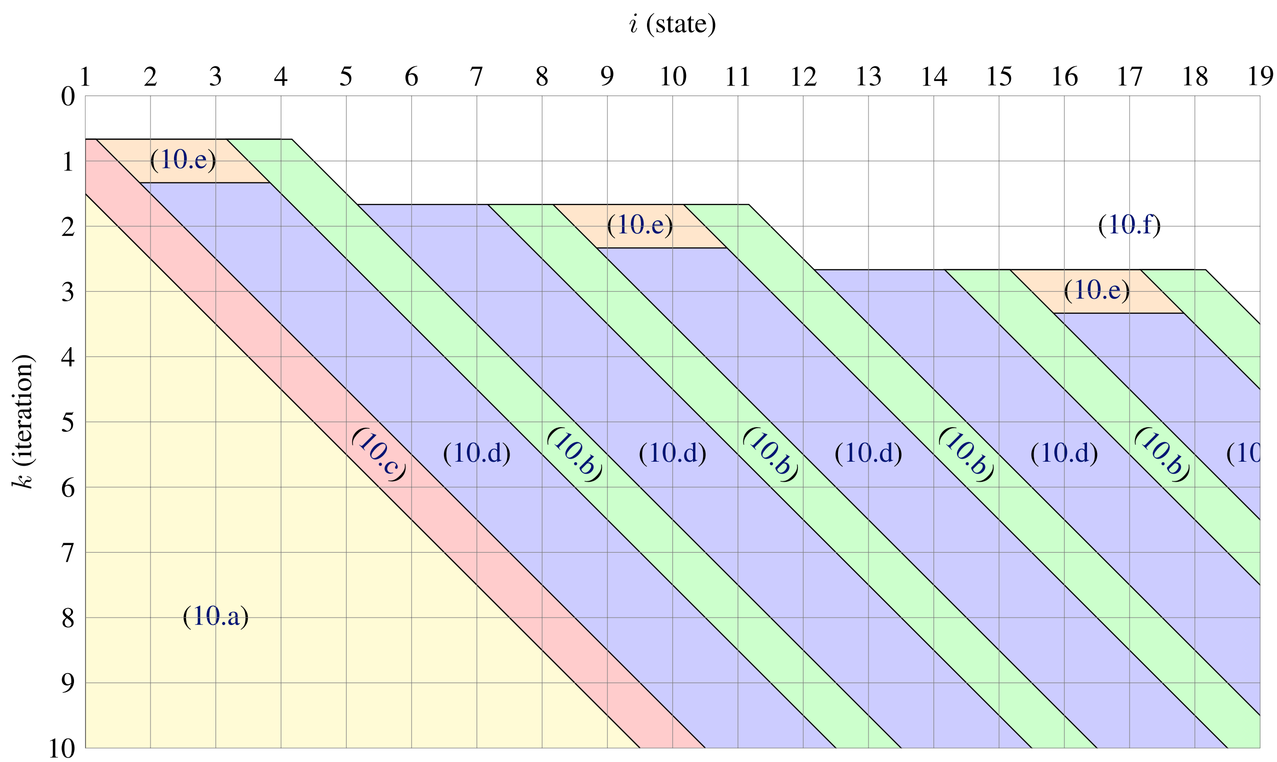

Figure 5. Shape of the value function with $\ell = 3$ and $m = 2$.

which is (10.b) when $q = (k-1) = 0$ and $p = 0$.

Second, we have

$$v_1(1) = \epsilon_1(1) = -\epsilon = \epsilon + 0 - 2\epsilon = v_1(1+\ell) + r_1 - 2\epsilon$$

which corresponds to (10.c).

Third, for $1 \leq p < \ell$ we have

$$v_1(1+p) = \epsilon_1(1+p) = 0$$

corresponding to (10.e).

Finally, for all the remaining states $i > 1 + \ell$, we have

$$v_1(i) = \epsilon_1(i) = 0$$

corresponding to (10.f).

The base case is now proved.

B.2. Induction Step

We assume that Lemma 7 holds for some *fixed* $k \geq 1$, we now show that it also holds for $k+1$.

B.2.1. The policy π_{k+1}

We begin by showing that the policy π_{k+1} is greedy with respect to v_k. Since there is no choice in state 1 is $\rightarrow$, we turn our attention to the other states. There are many cases to consider, each one of them corresponding to one or more states. These cases, labelled from A through F, are summarized as follows, depending on the state i:

(A) $1 < i < k+1$

(B) $i = k+1$

(C) $i = k+1+q\ell+p$ with $1 \leq p < \ell$ and $0 \leq q \leq (k-1)m$

(D) $i = k+1+(qm+p+1)\ell$ with $0 \leq p < m$ and $0 \leq q < k-1$

(E) $i = k+1+((k-1)m+1)\ell$

(F) $i > k+1+((k-1)m+1)\ell$

Figure 6 depicts how those cases cover the whole state space.

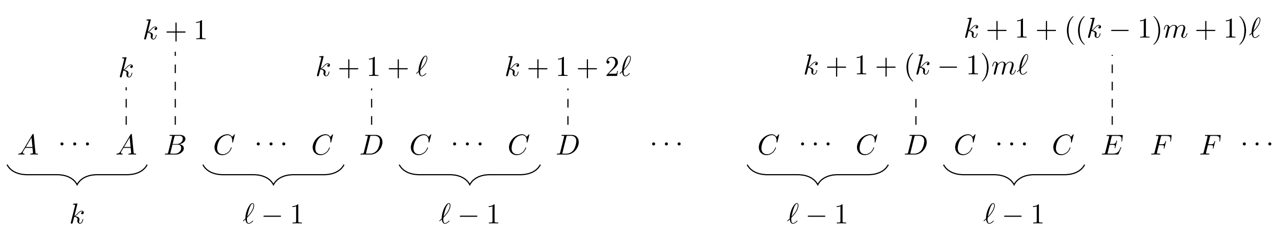

Figure 6. Policy cases, each state is represented by a letter corresponding to a case of the policy π_{k+1}. Starting from 1, state number increase from left to right.

For all states $i > 1$ in each of the above cases, we consider the *action-value functions* $q^{\rightarrow}_{k+1}(i)$ (resp. $q^{\leftarrow}_{k+1}(i)$) of action $\rightarrow$ (resp. $\leftarrow$) defined as:

$$q^{\rightarrow}_{k+1}(i) = r_i + \gamma v_k(i-1) \qquad \text{and} \qquad q^{\leftarrow}_{k+1}(i) = \gamma v_k(i+\ell-1).$$

In case $i = k+1$ (B) we will show that $q^{\rightarrow}_{k+1}(i) = q^{\leftarrow}_{k+1}(i)$ meaning that a policy π_{k+1} greedy for v_k may be either $\pi_{k+1}(k+1) = \rightarrow$ or $\pi_{k+1}(k+1) = \leftarrow$. In all other cases we show that $q^{\rightarrow}_{k+1}(i) < q^{\leftarrow}_{k+1}(i)$ which implies that for those $i \neq k+1$, $\pi_{k+1}(i) = \leftarrow$, as required by Lemma 7.

A: In states $1 < i < k+1$ We have $q^{\rightarrow}_{k+1}(i) = r_i + \gamma v_k(i+\ell-1)$ and $q^{\leftarrow}_{k+1}(i) = \gamma v_k(i-1)$, depending on the value of $i+\ell-1$, which is reached by taking the $\rightarrow$ action, we need to consider two cases:

- Case 1: $i+\ell-1 \neq k$. In this case $v_k(i+\ell-1)$ is described by either (10.a) or (10.d) when $i+\ell-1$ is less than, or greater than k, respectively. In either case we have $v_k(i+\ell-1) = -\gamma^{(k-1)(\ell m+1)}\epsilon = v_k(i-1)$ and hence:

$$q^{\rightarrow}_{k+1}(i) = r_i + \gamma v_k(i+\ell-1) = r_i + \gamma v_k(i-1) < \gamma v_k(i-1) = q^{\leftarrow}_{k+1}(i)$$

  which gives $\pi_{k+1}(i) = \leftarrow$ as desired.

- Case 2: $i+\ell-1 = k$.

$$\begin{aligned}
q^{\rightarrow}_{k+1}(i) &= r_i + \gamma v_k(k) = r_i + \gamma\left(v_k(k+\ell) + r_k - 2\epsilon\right) && \{(10.c)\}\\
&= \gamma\left(\sum_{j=0}^{k-1}\gamma^{j(\ell m+1)}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}}r_{k-j}+\epsilon\right) + r_k - 2\epsilon\right) && \{(10.b)\}\\
&\leq \gamma\left(\sum_{j=0}^{k-1}\gamma^{j(\ell m+1)}\epsilon + r_k - 2\epsilon\right) && \{r_{k-j} \leq 0\}\\
&= \gamma\left(\sum_{j=1}^{k-1}\left(\gamma^{j(\ell m+1)}\epsilon - 2\gamma^j\epsilon\right) - \epsilon\right) && \{r_k = -2\sum_{j=1}^{k-1}\gamma^j\epsilon\}\\
&< -\gamma\epsilon && \{\gamma^{j(\ell m+1)}\epsilon - 2\gamma^j\epsilon < 0\}\\
&< \gamma v_k(i-1) && \{v_k(i-1) = -\gamma^{(k-1)(\ell m+1)}\epsilon \text{ (10.a)}\}\\
&= q^{\leftarrow}_{k+1}(i)
\end{aligned}$$

  giving $\pi_{k+1}(i) = \leftarrow$ as desired.

B: In state $k+1$ Looking at the action value function $q^{\leftarrow}_{k+1}$ in state $k+1$, we observe that:

$$\begin{aligned}
q^{\leftarrow}_{k+1}(k+1) &= \gamma v_k(k) = \gamma\left(r_k - 2\epsilon + v_k(k+\ell)\right) && \{(10.c)\}\\
&= \gamma r_k - 2\gamma\epsilon + \gamma v_k(k+\ell)\\
&= r_{k+1} + \gamma v_k(k+\ell) && \{r_{i+1} = \gamma r_i - 2\gamma\epsilon\}\\
&= q^{\rightarrow}_{k+1}(k+1)
\end{aligned}$$

This means that the algorithm can take $\pi_{k+1}(k+1) = \rightarrow$ so as to satisfy Lemma 7.

C: In states $i = k+1+q\ell+p$ We restrict ourselves to the cases when $1 \leq p < \ell$ and $0 \leq q \leq (k-1)m$. Three cases for the value of q need to be considered:

- Case 1: $0 \leq q < (k-1)m - 1$. We have:

$$\begin{aligned}
q^{\rightarrow}_{k+1}(i) &= r_i + \gamma v_k(k+(q+1)\ell+p)\\
&= r_i + \gamma v_k(k+q\ell+p) && \{\text{(10.d) independent of } q\}\\
&< \gamma v_k(k+q\ell+p) && \{r_i < 0\}\\
&= q^{\leftarrow}_{k+1}(i).
\end{aligned}$$

- Case 2: $q = (k-1)m - 1$

$$\begin{aligned}
q^{\rightarrow}_{k+1}(i) &= r_i + \gamma v_k(k + (q+1)\ell + p) \\
&= r_i + \gamma 0 && \{(10.e)\} \\
&= -2\epsilon \frac{\gamma - \gamma^{k+1+q\ell+p}}{1-\gamma} \\
&= -2\epsilon \left(\frac{\gamma - \gamma^{k+q\ell+p}}{1-\gamma} + \gamma^{k+q\ell+p} \right) \\
&< -\gamma^{k+q\ell+p}\epsilon \\
&= -\gamma^{k+(k-1)\ell m - \ell + p}\epsilon && \{q = (k-1)m - 1\} \\
&< -\gamma^{k+(k-1)\ell m}\epsilon = -\gamma^{(k-1)(\ell m+1)+1}\epsilon && \{p - \ell < 0\} \\
&= \gamma v_k(k + q\ell + p) && \{(10.d)\} \\
&= q^{\leftarrow}_{k+1}(i).
\end{aligned}$$

- Case 3: $q = (k-1)m$

$$\begin{aligned}
q^{\rightarrow}_{k+1}(i) &= r_i + \gamma v_k(k + ((k-1)m+1)\ell + p) \\
&= r_i + \gamma 0 && \{(10.f)\} \\
&= r_i + \gamma v_k(k + ((k-1)m)\ell + p) && \{(10.e)\} \\
&= r_i + \gamma v_k(i-1) \\
&< q^{\leftarrow}_{k+1}(i). && \{r_i < 0\}
\end{aligned}$$

D: In states $i = k + 1 + (qm + p + 1)\ell$ In these states, we have:

$$q^{\leftarrow}_{k+1}(i) = \gamma v_k(k + (qm + p + 1)\ell)$$

$$\begin{aligned}
q^{\rightarrow}_{k+1}(i) &= r_i + \gamma v_k(k + 1 + (qm + p + 1)\ell + \ell - 1) \\
&= r_i + \gamma v_k(k + (qm + p + 2)\ell). \qquad (11)
\end{aligned}$$

As for the right-hand side of (11) we need to consider two cases:

- Case 1: $p + 1 < m$:

In the following, define

$$x_{k,q} = \sum_{j=1}^{k-q-1} \gamma^{j(\ell m+1)} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-j} + \epsilon \right).$$

Then,

$$\begin{aligned}
&q^{\rightarrow}_{k+1}(i) = r_i + \gamma v_k(k + (qm + (p+1) + 1)\ell) \\
&= r_i + \gamma\gamma^{q(\ell m+1)} \left(\frac{\gamma^{\ell(p+2)} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q} + \sum_{j=1}^{k-q-1} \gamma^{j(\ell m+1)} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-j} + \epsilon \right) \right) && \{(10.b)\} \\
&= r_i + \gamma^{q(\ell m+1)+1} \left(\left(\frac{\gamma^{\ell(p+1)} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} - \gamma^{\ell(p+1)} \right) r_{k-q} + x_{k,q} \right) \\
&= r_i - \gamma^{(qm+p+1)\ell+q+1} r_{k-q} + \gamma^{q(\ell m+1)+1} \left(\frac{\gamma^{\ell(p+1)} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q} + x_{k,q} \right) \\
&= r_i - \gamma^{i-k+q} r_{k-q} + \gamma v_k(k + (qm + p + 1)\ell) - \mathbb{1}_{[p=0]}\gamma^{q(\ell m+1)+1}\epsilon && \{(10.b)\} \\
&\leq r_i - \gamma^{i-k+q} r_{k-q} + \gamma v_k(k + (qm + p + 1)\ell) \\
&= r_i - \gamma^{i-k+q} r_{k-q} + q^{\leftarrow}_{k+1}(i). \qquad (12)
\end{aligned}$$

Now, observe that

$$\begin{aligned}
\gamma^{i-k+q} r_{k-q} &= -2\gamma^{i-k+q}\frac{\gamma-\gamma^{k-q}}{1-\gamma}\epsilon \\
&= -2\frac{\gamma^{i-k+q+1}-\gamma^{i}}{1-\gamma}\epsilon \\
&= -2\frac{\gamma-\gamma+\gamma^{i-k+q+1}-\gamma^{i}}{1-\gamma}\epsilon \\
&= -2\frac{\gamma-\gamma^{i}}{1-\gamma}\epsilon - 2\frac{-\gamma+\gamma^{i-k+q+1}}{1-\gamma}\epsilon \\
&= r_i - r_{i-k+q+1}.
\end{aligned}$$

Plugging this back into (12), we get:

$$\begin{aligned}
q_{k+1}^{\rightarrow}(i) &\leq r_i - r_i + r_{i-k+q+1} + q_{k+1}^{\leftarrow}(i) \\
&< q_{k+1}^{\leftarrow}(i). && \{r_{i-k+q+1} < 0\}
\end{aligned}$$

- Case 2: $p+1=m$:

Using the fact that $p+1=m$ implies $\frac{\gamma^{\ell(p+1)}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} = \gamma^{\ell m}$ we have:

$$\begin{aligned}
&q_{k+1}^{\rightarrow}(i) = r_i + \gamma v_k(k+((q+1)m+1)\ell) \\
&= r_i + \gamma\gamma^{(q+1)(\ell m+1)}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-1} + \epsilon + \sum_{j=1}^{k-q-2}\gamma^{j(\ell m+1)}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-j-1} + \epsilon\right)\right) && \{(10.b)\} \\
&= r_i + \gamma\gamma^{(q+1)(\ell m+1)}\left(\sum_{j=0}^{k-q-2}\gamma^{j(\ell m+1)}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-j-1} + \epsilon\right)\right) \\
&= r_i + \gamma\gamma^{q(\ell m+1)}\left(\sum_{j=1}^{k-q-1}\gamma^{j(\ell m+1)}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-j} + \epsilon\right)\right) \\
&= r_i + \gamma\gamma^{q(\ell m+1)}\left(\left(\frac{\gamma^{\ell(p+1)}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} - \gamma^{\ell m}\right) r_{k-q} + \sum_{j=1}^{k-q-1}\gamma^{j(\ell m+1)}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-q-j} + \epsilon\right)\right) \\
&= r_i - \gamma^{q(\ell m+1)+1}\gamma^{\ell m} r_{k-q} + \gamma\left(v_k(k+(qm+p+1)\ell) - \mathbb{1}_{[p=0]}\gamma^{q(\ell m+1)}\epsilon\right) && \{(10.b)\} \\
&\leq r_i - \gamma^{i-k+q} r_{k-q} + \gamma v_k(k+(qm+p+1)\ell) \\
&< q_{k+1}^{\leftarrow}(i),
\end{aligned}$$

where we concluded by observing that this is the same result as (12).

E: In state $i = k+((k-1)m+1)\ell+1$

$$\begin{aligned}
q_{k+1}^{\leftarrow}(i) &= \gamma v_k(i-1) = \gamma v_k(k+((k-1)m+1)\ell) \\
&= \gamma^{(k-1)(\ell m+1)+1}\left(\frac{\gamma^{\ell}-\gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_1 + \epsilon\right) && \{(10.b) \text{ with } q=k-1 \text{ and } p=0\} \\
&= \gamma^{(k-1)(\ell m+1)+1}\epsilon && \{r_1 = 0\} \\
&> r_i && \{r_i < 0\} \\
&= r_i + \gamma v_k(i+\ell-1) && \{v_k(i+\ell+1) = 0 \ \ (10.f)\} \\
&= q_{k+1}^{\rightarrow}(i).
\end{aligned}$$

F: In states $i > k + ((k-1)m+1)\ell + 1$ Following (10.f) we have $v_k(i-1) = v_k(i+\ell-1) = 0$ and hence

$$q^{\leftarrow}_{k+1}(i) = 0 > r_i = q^{\rightarrow}_{k+1}(i).$$

B.2.2. THE VALUE FUNCTION v_{k+1}

In the following we will show that the value function v_{k+1} satisfies Lemma 7. To that end we consider the value of $((T_{k+1,\ell})^m T_{k+1} v_k)(s_0)$ by analysing the trajectories obtained by first following m times $\pi_{k,\ell}$ then π_{k+1} from various starting states s_0.

Given a starting state s_0 and a non stationary policy $\pi_{k+1,\ell}$, we will represent the trajectories as a sequence of triples $(s_i, a_i, r(s_i,a_i))_{i=0,\dots,\ell m}$ arranged in a "trajectory matrix" of ℓ columns and m rows. Each column corresponds to one of the policies $\pi_{k+1}, \pi_k, \dots, \pi_{k+2-\ell}$. In a column labeled by policy π_j the entries are of the form $(s_i, \pi_j(s_i), r(s_i, \pi_j(s_i))$; this layout makes clear which stationary policy is used to select the action in any particular step in the trajectory. Indeed, in column π_j, we have $(s_i, \rightarrow, r_j)$ if and only if $s_i = j$, otherwise each entry is of the form $(s_i, \leftarrow, 0)$. Such a matrix accounts for the first m applications of the operator $T_{k+1,\ell}$. One addional row of only one triple $(s_i, \pi_{k+1}(s_i), r_{\pi_{k+1}}(s_i))$ represents the final application of T_{k+1}. After this triple comes the end state of the trajectory $s_{\ell m+1}$.

$$m = 4 \text{ times} \left\{ \overbrace{\begin{array}{ccc} \pi_4 & \pi_3 & \pi_2 \\ (10,\leftarrow,0) & (9,\leftarrow,0) & (8,\leftarrow,0) \\ (7,\leftarrow,0) & (6,\leftarrow,0) & (5,\leftarrow,0) \\ (4,\rightarrow,r_4) & (6,\leftarrow,0) & (5,\leftarrow,0) \\ (4,\rightarrow,r_4) & (6,\leftarrow,0) & (5,\leftarrow,0) \\ (4,\rightarrow,r_4) & \boxed{6} & \end{array}}^{\ell = 3 \text{ steps}} \right.$$

Figure 7. The trajectory matrix of policy $\pi_{4,\ell}$ starting from state 10 with $m = 4$ and $\ell = 3$.

Example 2. *Figure 7 depicts the trajectory matrix of policy $\pi_{4,\ell} = \pi_4\pi_3\pi_2$ with $m = 4$ and $\ell = 3$. The trajectory starts from state $s_0 = 10$ and ends in state $s_{\ell m+1} = 6$. The $\leftarrow$ action is always taken with reward 0 except when in state 4 under the policy π_4. From this matrix we can deduce that, for any value function v:*

$$\begin{aligned} ((T_{4,\ell})^m T_4 v)(10) &= \gamma^6 r_4 + \gamma^9 r_4 + \gamma^{12} r_4 + \gamma^{13} v(6) \\ &= \gamma^{2\ell} r_4 + \gamma^{3\ell} r_4 + \gamma^{4\ell} r_4 + \gamma^{4\ell+1} v(6) \\ &= \frac{\gamma^{2\ell} - \gamma^{(m+1)\ell}}{1-\gamma^\ell} r_4 + \gamma^{\ell m+1} v(6). \end{aligned}$$

With this in hand, we are going to prove each case of Lemma 7 for v_{k+1}.

In states $i < k+1$ Following m times $\pi_{k+1,\ell}$ and then π_{k+1} starting from these states consists in taking the $\leftarrow$ action $\ell m + 1$ times to eventually finish either in state 1 if $i \le \ell m + 2$ with value

$$v_{k+1}(i) = \gamma^{\ell m+1} v_k(1) + \epsilon_{k+1}(i) = -\gamma^{\ell m+1}\gamma^{(k-1)(\ell m+1)}\epsilon = -\gamma^{k(\ell m+1)}\epsilon$$

or otherwise in state $i - \ell m - 1 < k$ with value

$$v_{k+1}(i) = \gamma^{\ell m+1} v_k(i - \ell m - 1) + \epsilon_{k+1}(i) = -\gamma^{\ell m+1}\gamma^{(k-1)(\ell m+1)}\epsilon = -\gamma^{k(\ell m+1)}\epsilon$$

This matches Equation (10.a) in both cases.

In states $i = k + 1 + (qm + p + 1)\ell$ Consider the states $i = k + 1 + (qm + p + 1)\ell$ with $q \ge 0$ and $0 \le p < m$. Following m times $\pi_{k+1,\ell}$ and then π_{k+1} starting from state i gives the following trajectories:

- when $q = 0$, (*i.e.* $i = k + 1 + (p + 1)\ell$):

$$
\begin{array}{rcccc}
 & \multicolumn{4}{c}{\overbrace{\hphantom{(k+1+(p+1)\ell,\leftarrow,0)\quad(k+(p+1)\ell,\leftarrow,0)\quad\dots\quad(k+(p-1)\ell+2,\leftarrow,0)}}^{\ell \text{ steps}}} \\
 & \pi_{k+1} & \pi_k & \dots & \pi_{k-\ell+2} \\
\multirow{4}{*}{$p+1 \text{ times}\left\{\vphantom{\begin{array}{c}a\\a\\a\\a\end{array}}\right.$} & (k+1+(p+1)\ell, \leftarrow, 0) & (k+(p+1)\ell, \leftarrow, 0) & \dots & (k+p\ell+2, \leftarrow, 0) \\
 & (k+1+p\ell, \leftarrow, 0) & (k+p\ell, \leftarrow, 0) & \dots & (k+(p-1)\ell+2, \leftarrow, 0) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & (k+1+\ell, \leftarrow, 0) & (k+\ell, \leftarrow, 0) & \dots & (k+2, \leftarrow, 0) \\
\multirow{3}{*}{$m-p-1 \text{ times}\left\{\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right.$} & (k+1, \rightarrow, r_{k+1}) & (k+\ell, \leftarrow, 0) & \dots & (k+2, \leftarrow, 0) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & (k+1, \rightarrow, r_{k+1}) & (k+\ell, \leftarrow, 0) & \dots & (k+2, \leftarrow, 0) \\
 & (k+1, \rightarrow, r_{k+1}) & \boxed{k+\ell} & &
\end{array}
$$

Using (10.b) with $q = p = 0$ as our induction hypothesis, this gives

$$
\begin{aligned}
((T_{k+1,\ell})^m T_{k+1} v_k)(i) &= \sum_{j=p+1}^{m} \gamma^{\ell j} r_{k+1} + \gamma^{\ell m+1} v_k(k+\ell) \\
&= \sum_{j=p+1}^{m} \gamma^{\ell j} r_{k+1} + \gamma^{\ell m+1} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_k + \epsilon + \sum_{j=1}^{k-1} \gamma^{j(\ell m+1)} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-j} + \epsilon \right) \right) \\
&= \frac{\gamma^{\ell(p+1)} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k+1} + \sum_{j=1}^{k} \gamma^{j(\ell m+1)} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-j} + \epsilon \right)
\end{aligned}
$$

Accounting for the error term and the fact that $i = k + 1 + \ell \iff p = q = 0$, we get

$$
\begin{aligned}
v_{k+1}(i) &= ((T_{k+1,\ell})^m T_{k+1} v_k)(i) + \mathbb{1}_{[i=k+1+\ell]} \epsilon \\
&= \frac{\gamma^{\ell(p+1)} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k+1} + \mathbb{1}_{[p=0]} \epsilon + \sum_{j=1}^{k} \gamma^{j(\ell m+1)} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k-j} + \epsilon \right)
\end{aligned}
$$

which is (10.b) for $k + 1$ and $q = 0$ as desired.

- when $1 \leq q \leq k$:

In this case we have $i - (\ell m + 1) \geq k + 1$, meaning that $k + 1$, the first state where the $\rightarrow$ action would be available is unreachable (in the sense that the tractory could end in $k + 1$, but no action will be taken there). Consequently the $\leftarrow$ action is taken $\ell m + 1$ times and the system ends in state $i - \ell m - 1 = k + ((q - 1)m + p + 1)\ell$. Therefore, using (10.b) as induction hypothesis and the fact that $i \notin \{k + 1, k + \ell + 1\} \implies \epsilon_{k+1}(i) = 0$, we have:

$$
\begin{aligned}
v_{k+1}(i) &= \gamma^{\ell m+1} v_k(k + ((q-1)m + p + 1)\ell) + \epsilon_{k+1}(i) \\
&= \gamma^{q(\ell m+1)} \left(\frac{\gamma^{\ell(p+1)} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k+1-q} + \mathbb{1}_{[p=0]} \epsilon + \sum_{i=1}^{k-q} \gamma^{i(\ell m+1)} \left(\frac{\gamma^{\ell} - \gamma^{\ell(m+1)}}{1-\gamma^{\ell}} r_{k+1-q-k} + \epsilon \right) \right),
\end{aligned}
$$

which statisfies (10.b) for $k + 1$.

In state $k+1$ Following m times $\pi_{k+1,\ell}$ and then π_{k+1} starting from $k+1$ gives the following trajectory:

$$
\begin{array}{rcccc}
 & \multicolumn{4}{c}{\overbrace{\hphantom{(k+1,\rightarrow,r_{k+1})\ (k+\ell,\leftarrow,0)\ \ldots\ (k+2,\leftarrow,0)}}^{\ell \text{ steps}}} \\
 & \pi_{k+1} & \pi_k & \ldots & \pi_{k-\ell+2} \\
\multirow{3}{*}{m times $\left\{\vphantom{\begin{array}{c}a\\ \vdots \\ a\end{array}}\right.$} & (k+1,\rightarrow,r_{k+1}) & (k+\ell,\leftarrow,0) & \ldots & (k+2,\leftarrow,0) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & (k+1,\rightarrow,r_{k+1}) & (k+\ell,\leftarrow,0) & \ldots & (k+2,\leftarrow,0) \\
 & (k+1,\rightarrow,r_{k+1}) & \boxed{k+\ell} & &
\end{array}
$$

As a consequence, with (10.c) as induction hypothesis we have:

$$
\begin{aligned}
&\left((T_{k+1,\ell})^m T_{k+1} v_k\right)(k+1) = \frac{1-\gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k+1} + \gamma^{\ell m+1} v_k(k+\ell) \\
&= r_{k+1} + \frac{\gamma^\ell - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k+1} + \gamma^{\ell m+1}\left(\frac{\gamma^\ell - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_k + \epsilon + \sum_{j=1}^{k-1} \gamma^{j(\ell m+1)}\left(\frac{\gamma^\ell - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k-j} + \epsilon\right)\right) \\
&= r_{k+1} + \frac{\gamma^\ell - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k+1} + \sum_{j=1}^{k} \gamma^{j(\ell m+1)}\left(\frac{\gamma^\ell - \gamma^{\ell(m+1)}}{1-\gamma^\ell} r_{k-j+1} + \epsilon\right) \\
&= r_{k+1} + v_{k+1}(k+\ell+1) - \epsilon
\end{aligned}
$$

Hence,

$$
\begin{aligned}
v_{k+1}(k+1) &= \left((T_{k+1,\ell})^m T_{k+1} v_k\right)(k+1) + \epsilon_{k+1}(k+1) \\
&= v_{k+1}(k+\ell+1) + r_{k+1} - 2\epsilon,
\end{aligned}
$$

which matches (10.c).

In states $i = k+1+q\ell+p$ For states $i = k+1+q\ell+p$ with $0 \leq q \leq km-1$ and $1 \leq p < \ell$, the policy $\pi_{k+1,\ell}$ always takes the $\leftarrow$ action with either one of the following trajectories

- when $q \geq m$:

$$
\begin{array}{rcccc}
 & \multicolumn{4}{c}{\overbrace{\hphantom{(k+1+(q-m+1)\ell+p,\leftarrow,0)\ (k+q\ell+p,\leftarrow,0)\ \ldots\ (k+(q-m)\ell+p+2,\leftarrow,0)}}^{\ell \text{ steps}}} \\
 & \pi_{k+1} & \pi_k & \ldots & \pi_{k-\ell+2} \\
\multirow{3}{*}{m times $\left\{\vphantom{\begin{array}{c}a\\ \vdots \\ a\end{array}}\right.$} & (k+1+q\ell+p,\leftarrow,0) & (k+q\ell+p,\leftarrow,0) & \ldots & (k+(q-1)\ell+p+2,\leftarrow,0) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & (k+1+(q-m+1)\ell+p,\leftarrow,0) & (k+q\ell+p,\leftarrow,0) & \ldots & (k+(q-m)\ell+p+2,\leftarrow,0) \\
 & (k+1+(q-m)\ell+p,\leftarrow,0) & \boxed{k+(q-m)\ell+p} & &
\end{array}
$$

As a consequence, with (10.d) as induction hypothesis we have:

$$
v_{k+1}(i) = \left((T_{k+1,\ell})^m T_{k+1} v_k\right)(i) = \gamma^{\ell m+1} v_k(k+(q-m)\ell+p) = -\gamma^{\ell m+1}\gamma^{(k-1)(\ell m+1)}\epsilon = -\gamma^{k(\ell m+1)}\epsilon
$$

which satisfies (10.d) in this case.

- when $q < m$:

Assuming that negative states correspond to state 1, where the action is irrelevant, we have the following trajectory:

$$
\begin{array}{r|ccc}
 & \multicolumn{3}{c}{\overbrace{\hspace{8cm}}^{\ell \text{ steps}}} \\
 & \pi_{k+1} & \cdots & \pi_{k-\ell+2} \\
 & (k+1+q\ell+p, \leftarrow, 0) & \ldots & (k+(q-1)\ell+p+2, \leftarrow, 0) \\
q \text{ times} & \vdots & \vdots & \vdots \\
 & (k+1+\ell+p, \leftarrow, 0) & \ldots & (k+p+2, \leftarrow, 0) \\
 & (k+1+p, \leftarrow, 0) & \ldots & (k-\ell+p+2, \leftarrow, 0) \\
 & (k+1-\ell+p, \leftarrow, 0) & \ldots & (k-2\ell+p+2, \leftarrow, 0) \\
m-q \text{ times} & \vdots & \vdots & \vdots \\
 & (k+1-(m-q-1)\ell+p, \leftarrow, 0) & \ldots & (k-(m-q)\ell+p+2, \leftarrow, 0) \\
 & (k+1-(m-q)\ell+p, \leftarrow, 0) & & \boxed{k+(q-m)\ell+p}
\end{array}
$$

In the above trajectory, one can see that only the $\leftarrow$ action is taken (ignoring state 1). Indeed, since we follow the policies $\pi_{k+1}\pi_k, \ldots, \pi_{k-\ell+2}$ the $\rightarrow$ action may only be taken in states $k+1, k, \ldots, k-\ell+2$. When state $k+1$ is reached, the selected action is $\pi_{k-p+1}(k+1)$ which is $\leftarrow$ since $p \geq 1$. The same reasonning applies in the next states $k, \ldots, k-\ell+1$, where $p \geq 1$ prevents to use a policy that would select the $\rightarrow$ action in those states.

Since $p - \ell < 0$ the trajectory always terminates in a state $j < k$ with value $v_k(j) = -\gamma^{(k-1)(\ell m-1)}\epsilon$ as for the $q \geq m$ case, which allows to conclude that (10.d) also holds in this case.

In states $i = k+1+km\ell+p$ Observe that following m times $\pi_{k+1,\ell}$ and then π_{k+1} once amounts to always take $\leftarrow$ actions. Thus, one eventually finishes in state $k+(k-1)m\ell+p \geq k+1$, which, since $\epsilon_k(i) = 0$, gives

$$
v_{k+1}(i) = \left((T_{k+1,\ell})^m T_{k+1} v_k\right)(i) = \gamma^{\ell m+1} v_k(k+(k-1)m\ell+p) = -\gamma^{\ell m+1} 0 = 0,
$$

satisfiying (10.e).

In states $i > k+1+(km+1)\ell$ In these states, the action $\leftarrow$ is taken $\ell m + 1$ times ending up in state $j > k+((k-1)m+1)\ell$, with value $v_k(j) = 0$, from which $v_{k+1}(i) = 0$ follows as required by (10.f).