Improved and Generalized Upper Bounds on the Complexity of Policy Iteration

Bruno Scherrer

INRIA Lorraine, LORIA, MAIA Team

PGMO-COPI'14
Markov Decision Process

(Puterman, 1994; Bertsekas & Tsitsiklis, 1996)

Markov Decision Process (MDP):

- $\mathcal{X} = \{1, 2, \ldots, n\}$ is the finite state space,
- $\mathcal{A} = \{1, 2, \ldots, m\}$ is the finite action space,
- $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the reward function, 
  \hspace{1cm} (r_t = r(x_t, a_t))
- $p : \mathcal{X} \times \mathcal{A} \to \Delta\mathcal{X}$ is the transition function. 
  \hspace{1cm} (x_{t+1} \sim p(\cdot | x_t, a_t))$

Goal: Find a stationary policy $\pi : \mathcal{X} \to \mathcal{A}$ that maximizes the value $v_\pi(x)$ for all $x$:

$$v_\pi(x) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid x_0 = x, \forall t, a_t = \pi(x_t) \right]. \quad (\gamma \in (0, 1))$$
Markov Decision Process (MDP):

- \( \mathcal{X} = \{1, 2, \ldots, n\} \) is the finite state space,
- \( \mathcal{A} = \{1, 2, \ldots, m\} \) is the finite action space,
- \( r : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R} \) is the reward function,
- \( p : \mathcal{X} \times \mathcal{A} \rightarrow \Delta \mathcal{X} \) is the transition function.

\( r_t = r(x_t, a_t) \)
\( x_{t+1} \sim p(\cdot | x_t, a_t) \)

**Goal:** Find a stationary policy \( \pi : \mathcal{X} \rightarrow \mathcal{A} \) that maximizes the value \( v_\pi(x) \) for all \( x \):

\[
v_\pi(x) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \middle| x_0 = x, \ \forall t, \ a_t = \pi(x_t) \right]. \quad (\gamma \in (0, 1))
\]
Bellman Equations and Operators

- For any policy $\pi$, the value $v_\pi$ satisfies:

$$\forall x, \ v_\pi(x) = r(x, \pi(x)) + \gamma \sum_{y \in X} p(y|x, \pi(x)) v_\pi(y) \iff v_\pi = T_\pi v_\pi.$$  

- The optimal value $v_*$ satisfies:

$$\forall x, \ v_*(x) = \max_{a \in A} \left( r(x, a) + \gamma \sum_{y \in X} p(y|x, a) v_*(y) \right) \iff v_* = T v_*.$$  

- $T_\pi : \mathbb{R}^X \to \mathbb{R}^X$ and $T : \mathbb{R}^X \to \mathbb{R}^X$ are $\gamma$-contraction mappings w.r.t. the max norm $\|v\|_\infty = \max_s |v(s)|$.

- $\pi$ is a greedy policy w.r.t. $v$, written $\pi = \mathcal{G} v$, iff

$$\forall x, \ \pi(x) \in \arg \max_{a \in A} \left( r(x, a) + \gamma \sum_{y \in X} p(y|x, a) v(y) \right) \iff T_\pi v = T v.$$
Policy Iteration Scheme

Policy Iteration

\[ \pi_{k+1} \leftarrow \text{switch}(\pi_k, Y_k) \] for some set \( Y_k \) such that \( \emptyset \subsetneq Y_k \subseteq S_{\pi_k} \).

\[
\begin{align*}
    a_\pi &= T v_\pi - v_\pi \geq 0 & \text{advantage} \\
    S_\pi &= \{ i, \ a_\pi(i) > 0 \} & \text{switchable states} \\
    \forall i, \ \text{switch}(\pi, Y)(i) &= \begin{cases} 
        G(v_\pi)(i) & \text{if } i \in Y \\
        \pi(i) & \text{if } i \notin Y.
    \end{cases} & \text{improved policy}
\end{align*}
\]

Lemma (Policy Improvement (Puterman, 1994))

Let \( \pi \) be some non-optimal policy. If \( \pi' = \text{switch}(\pi, Y) \) for some non-empty subset \( Y \) of \( S_\pi \), then \( v_{\pi'} \geq v_\pi \) and there exists at least one state \( i \) such that \( v_{\pi'}(i) > v_\pi(i) \).

Howard’s PI: \( Y_k = S_{\pi_k} \), Simplex-PI: \( Y_k = s_k \) s.t. \( a_{\pi_k}(s_k) = \|a_{\pi_k}\|_\infty \).
Policy Iteration Scheme

Policy Iteration

$$\pi_{k+1} \leftarrow \text{switch}(\pi_k, Y_k) \text{ for some set } Y_k \text{ such that } \emptyset \subsetneq Y_k \subseteq S_{\pi_k}.$$ 

\[
a_\pi = T v_\pi - v_\pi \geq 0 \quad \text{advantage}
\]

\[
S_\pi = \{i, a_\pi(i) > 0\} \quad \text{switchable states}
\]

$$\forall i, \text{ switch}(\pi, Y)(i) = \begin{cases} G(v_\pi)(i) & \text{if } i \in Y \\ \pi(i) & \text{if } i \notin Y. \end{cases} \quad \text{improved policy}$$

Lemma (Policy Improvement (Puterman, 1994))

Let $$\pi$$ be some non-optimal policy. If $$\pi' = \text{switch}(\pi, Y)$$ for some non-empty subset $$Y$$ of $$S_\pi$$, then $$v_{\pi'} \geq v_\pi$$ and there exists at least one state $$i$$ such that $$v_{\pi'}(i) > v_\pi(i)$$.

Howard’s PI: $$Y_k = S_{\pi_k}$$, Simplex-PI: $$Y_k = s_k \text{ s.t. } a_{\pi_k}(s_k) = \|a_{\pi_k}\|_\infty.$$
## Policy Iteration Scheme

### Policy Iteration

\[ \pi_{k+1} \leftarrow \text{switch}(\pi_k, Y_k) \text{ for some set } Y_k \text{ such that } \emptyset \subsetneq Y_k \subseteq S_{\pi_k}. \]

- \[ a_\pi = TV_\pi - v_\pi \geq 0 \] advantage
- \[ S_\pi = \{i, \; a_\pi(i) > 0\} \] switchable states

\[ \forall i, \; \text{switch}(\pi, Y)(i) = \begin{cases} G(v_\pi)(i) & \text{if } i \in Y \\ \pi(i) & \text{if } i \not\in Y. \end{cases} \] improved policy

### Lemma (Policy Improvement (Putereman, 1994))

Let \( \pi \) be some non-optimal policy. If \( \pi' = \text{switch}(\pi, Y) \) for some non-empty subset \( Y \) of \( S_\pi \), then \( v_{\pi'} \geq v_\pi \) and there exists at least one state \( i \) such that \( v_{\pi'}(i) > v_\pi(i) \).

**Howard’s PI:** \( Y_k = S_{\pi_k}, \) **Simplex-PI:** \( Y_k = s_k \) s.t. \[ a_{\pi_k}(s_k) = \|a_{\pi_k}\|_\infty. \]
Complexity of Policy Iteration

How many iterations to converge to an \((\epsilon-)\)optimal policy?

- \(m^n\)
- \(\left\lceil \log \frac{V_{\text{max}}}{\epsilon} \right\rceil \) for Howard’s PI, \(\left\lceil n \log \frac{n V_{\text{max}}}{\epsilon} \right\rceil \) for Simplex-PI
- \(O\left(\frac{2^n}{n}\right)\) for Howard’s PI \((m = 2)\) (Mansour & Singh, 1999)
- \(O\left(\frac{n^2 m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI and Simplex-PI (Ye, 2011)
- \(O\left(\frac{nm}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI (Hansen et al., 2013)
- \(O\left(n^5 m^2 \log^2 n\right)\) for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

\(n = \)number of states, \(m = \)number of actions, \(V_{\text{max}} = \max_{x,a} |r(x,a)| \)

\(\epsilon\)-optimality: \(\|v_* - v_{\pi_k}\|_\infty \leq \epsilon\)
Complexity of Policy Iteration

How many iterations to converge to an (\(\epsilon\)-)optimal policy?

- \(m^n\)
- \(\left\lceil \log \frac{V_{\text{max}}}{\epsilon} \frac{1}{1-\gamma} \right\rceil\) for Howard’s PI,
- \(\left\lceil n \log \frac{nV_{\text{max}}}{\epsilon} \frac{1}{1-\gamma} \right\rceil\) for Simplex-PI
- \(O\left(\frac{2^n}{n}\right)\) for Howard’s PI (\(m = 2\)) (Mansour & Singh, 1999)
- \(O\left(\frac{n^2 m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI and Simplex-PI (Ye, 2011)
- \(O\left(\frac{nm}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI (Hansen et al., 2013)
- \(O\left(n^5 m^2 \log^2 n\right)\) for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

\(n\) = number of states, \(m\) = number of actions, \(V_{\text{max}} = \frac{\max_{x,a} |r(x,a)|}{1-\gamma}\)

\(\epsilon\)-optimality: \(\|v_* - v_{\pi_k}\|_{\infty} \leq \epsilon\)
Complexity of Policy Iteration

How many iterations to converge to an \((\epsilon\text{-})\)optimal policy?

- \(m^n\)
- \(\left\lceil \log \frac{V_{\max}}{\epsilon} \frac{1}{1-\gamma} \right\rceil\) for Howard’s PI,
- \(\left\lceil \frac{n \log n V_{\max}}{\epsilon} \frac{1}{1-\gamma} \right\rceil\) for Simplex-PI
- \(O\left(\frac{2^n}{n}\right)\) for Howard’s PI \((m = 2)\) (Mansour & Singh, 1999)
- \(O\left(\frac{n^2 m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI and Simplex-PI (Ye, 2011)
- \(O\left(\frac{nm}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI (Hansen et al., 2013)
- \(O\left(n^5 m^2 \log^2 n\right)\) for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

\(n = \text{number of states}, \ m = \text{number of actions}, \ V_{\max} = \max_{x,a} |r(x,a)| \frac{1}{1-\gamma}\)

\(\epsilon\text{-}\text{optimality}: \|\nu_* - \nu_{\pi_k}\|_{\infty} \leq \epsilon\)
Complexity of Policy Iteration

How many iterations to converge to an \((\epsilon\text{-})\)optimal policy?

- \(m^n\)
- \(\left\lceil \log \frac{V_{\text{max}}}{\epsilon} \right\rceil \) for Howard’s PI, \(\left\lceil n \log \frac{nV_{\text{max}}}{\epsilon} \right\rceil \) for Simplex-PI

- \(O\left(\frac{2^n}{n}\right)\) for Howard’s PI \((m = 2)\) (Mansour & Singh, 1999)
- \(O\left(\frac{n^2 m \log \left(\frac{n}{1-\gamma}\right)}{1-\gamma}\right)\) for Howard’s PI and Simplex-PI (Ye, 2011)
- \(O\left(\frac{nm \log \left(\frac{n}{1-\gamma}\right)}{1-\gamma}\right)\) for Howard’s PI (Hansen et al., 2013)
- \(O\left(n^5 m^2 \log^2 n\right)\) for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

\(n\) = number of states, \(m\) = number of actions, \(V_{\text{max}} = \max_{x,a} |r(x,a)|\) 
\(\epsilon\text{-}\)optimality: \(\|v_\pi - v_\star\|_\infty \leq \epsilon\)
Complexity of Policy Iteration

How many iterations to converge to an \((\epsilon-)\)optimal policy?

- \(m^n\)
- \(\left\lceil \frac{\log \frac{V_{\text{max}}}{\epsilon}}{1-\gamma} \right\rceil\) for Howard’s PI, \(\left\lceil \frac{n \log \frac{nV_{\text{max}}}{\epsilon}}{1-\gamma} \right\rceil\) for Simplex-PI
- \(O\left(\frac{2^n}{n}\right)\) for Howard’s PI \((m = 2)\) (Mansour & Singh, 1999)
- \(O\left(\frac{n^2m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI and Simplex-PI (Ye, 2011)
- \(O\left(\frac{nm}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI (Hansen et al., 2013)
- \(O\left(n^5m^2 \log^2 n\right)\) for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

\(n = \) number of states, \(m = \) number of actions, \(V_{\text{max}} = \max_{x,a} |r(x,a)| \frac{1}{1-\gamma}\)
\(\epsilon\)-optimality: \(\|v_* - v_{\pi_k}\|_\infty \leq \epsilon\)
Complexity of Policy Iteration

How many iterations to converge to an ($\epsilon$-)optimal policy?

- $m^n$
- $\left\lceil \log \frac{V_{\max}}{\epsilon} \right\rceil$ for Howard's PI, $\left\lceil n \log \frac{nV_{\max}}{\epsilon} \right\rceil$ for Simplex-PI
- $O\left( \frac{2^n}{n} \right)$ for Howard's PI ($m = 2$) (Mansour & Singh, 1999)
- $O\left( \frac{n^2m}{1-\gamma} \log \left( \frac{n}{1-\gamma} \right) \right)$ for Howard's PI and Simplex-PI (Ye, 2011)
- $O\left( \frac{nm}{1-\gamma} \log \left( \frac{n}{1-\gamma} \right) \right)$ for Howard's PI (Hansen et al., 2013)
- $O\left( n^5m^2 \log^2 n \right)$ for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

$n =$ number of states, $m =$ number of actions, $V_{\max} = \frac{\max_{x,a} |r(x,a)|}{1-\gamma}$

$\epsilon$-optimality: $\|v_* - v_{\pi_k}\|_\infty \leq \epsilon$
Complexity of Policy Iteration

How many iterations to converge to an \((\epsilon-)\)optimal policy?

- \(m^n\)
- \(\left\lceil \log \frac{V_{\text{max}}}{\epsilon} \frac{1}{1-\gamma} \right\rceil\) for Howard’s PI, \(\left\lceil n \log \frac{n V_{\text{max}}}{\epsilon} \frac{1}{1-\gamma} \right\rceil\) for Simplex-PI
- \(O\left(\frac{2^n}{n}\right)\) for Howard’s PI \((m = 2)\) (Mansour & Singh, 1999)
- \(O\left(\frac{n^2 m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI and Simplex-PI (Ye, 2011)
- \(O\left(\frac{n m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)\) for Howard’s PI (Hansen et al., 2013)
- \(O\left(n^5 m^2 \log^2 n\right)\) for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

\(n\) = number of states, \(m\) = number of actions, \(V_{\text{max}} = \max_{x,a} |r(x,a)|\frac{1}{1-\gamma}\)

\(\epsilon\)-optimality: \(\|v_* - v_{\pi_k}\|_{\infty} \leq \epsilon\)
Complexity of Policy Iteration

How many iterations to converge to an ($\epsilon$-)optimal policy?

- $m^n$
- $\left\lceil \log \frac{V_{\text{max}}}{\epsilon} \right\rceil$ for Howard’s PI, $\left\lceil n \log \frac{nV_{\text{max}}}{\epsilon} \right\rceil$ for Simplex-PI
- $O\left(\frac{2^n}{n}\right)$ for Howard’s PI ($m = 2$) (Mansour & Singh, 1999)
- $O\left(\frac{n^2m}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)$ for Howard’s PI and Simplex-PI (Ye, 2011)
- $O\left(\frac{nm}{1-\gamma} \log \left(\frac{n}{1-\gamma}\right)\right)$ for Howard’s PI (Hansen et al., 2013)
- $O\left(n^5m^2 \log^2 n\right)$ for Simplex-PI on deterministic MDPs (Post & Ye, 2012)

$n =$number of states, $m =$number of actions, $V_{\text{max}} = \max_{x,a} |r(x,a)|$

$\epsilon$-optimality: $\|v_* - v_{\pi_k}\|_{\infty} \leq \epsilon$
Proofs for $\epsilon$-optimality Bounds

Based on contraction properties.

- **Howard’s PI**: $\|v_* - v_{\pi_{k+1}}\|_\infty \leq \gamma \|v_* - v_{\pi_k}\|_\infty$

  By induction: $\|v_* - v_{\pi_k}\|_\infty \leq \gamma^k \|v_* - v_{\pi_0}\|_\infty \leq \gamma^k V_{\text{max}}$.

  $\|v_* - v_{\pi_k}\|_\infty < \epsilon \iff \gamma^k V_{\text{max}} < \epsilon \iff k \geq \frac{\log V_{\text{max}}}{1-\gamma} > \frac{\log V_{\text{max}}}{1-\gamma}$.

- **Simplex-PI**: $1^T(v_* - v_{\pi_{k+1}}) \leq (1 - \frac{1-\gamma}{n}) 1^T(v_* - v_{\pi_{k+1}})$

  $\|v_* - v_{\pi_k}\|_\infty \leq 1^T(v_* - v_{\pi_k}) \leq \left(1 - \frac{1-\gamma}{n}\right)^k 1^T(v_* - v_{\pi_0}) \leq \left(1 - \frac{1-\gamma}{n}\right)^k n V_{\text{max}}$.

  $\|v_* - v_{\pi_k}\|_\infty < \epsilon \iff \left(1 - \frac{1-\gamma}{n}\right)^k n V_{\text{max}} < \epsilon \iff k \geq \frac{n \log n V_{\text{max}}}{1-\gamma} > \frac{n \log n V_{\text{max}}}{1-\gamma}$.
Proofs for $\epsilon$-optimality Bounds

Based on contraction properties.

- **Howard’s PI:** $\|v^* - v_{\pi_{k+1}}\|_\infty \leq \gamma \|v^* - v_{\pi_k}\|_\infty$

  By induction: $\|v^* - v_{\pi_k}\|_\infty \leq \gamma^k \|v^* - v_{\pi_0}\|_\infty \leq \gamma^k V_{\text{max}}$

  $\|v^* - v_{\pi_k}\|_\infty < \epsilon \iff \gamma^k V_{\text{max}} < \epsilon \iff k \geq \frac{\log \frac{V_{\text{max}}}{\epsilon}}{1 - \gamma} > \frac{\log \frac{V_{\text{max}}}{\epsilon}}{\log \frac{1}{\gamma}}$

- **Simplex-PI:** $\mathbf{1}^T (v^* - v_{\pi_{k+1}}) \leq \left(1 - \frac{1-\gamma}{n}\right) \mathbf{1}^T (v^* - v_{\pi_{k+1}})$

  $\|v^* - v_{\pi_k}\|_\infty \leq \mathbf{1}^T (v^* - v_{\pi_k}) \leq \left(1 - \frac{1-\gamma}{n}\right)^k \mathbf{1}^T (v^* - v_{\pi_0}) \leq \left(1 - \frac{1-\gamma}{n}\right)^k n V_{\text{max}}$

  $\|v^* - v_{\pi_k}\|_\infty < \epsilon \iff \left(1 - \frac{1-\gamma}{n}\right)^k n V_{\text{max}} < \epsilon \iff k \geq \frac{n \log \frac{n V_{\text{max}}}{\epsilon}}{1 - \gamma} > \frac{n \log \frac{n V_{\text{max}}}{\epsilon}}{\log \frac{1}{1 - \frac{1-\gamma}{n}}}$
Proofs for $\epsilon$-optimality Bounds

Based on contraction properties.

- **Howard's PI:** \[ \| v_* - v_{\pi_{k+1}} \|_\infty \leq \gamma \| v_* - v_{\pi_k} \|_\infty \]

  By induction: \[ \| v_* - v_{\pi_k} \|_\infty \leq \gamma^k \| v_* - v_{\pi_0} \|_\infty \leq \gamma^k V_{\text{max}}. \]

  \[ \| v_* - v_{\pi_k} \|_\infty < \epsilon \iff \gamma^k V_{\text{max}} < \epsilon \iff k \geq \frac{\log \frac{V_{\text{max}}}{\epsilon}}{1-\gamma} > \frac{\log \frac{V_{\text{max}}}{\epsilon}}{\log \frac{1}{\gamma}}. \]

- **Simplex-PI:** \[ 1^T(v_* - v_{\pi_{k+1}}) \leq \left( 1 - \frac{1-\gamma}{n} \right) 1^T(v_* - v_{\pi_{k+1}}) \]

  \[ \| v_* - v_{\pi_k} \|_\infty \leq 1^T(v_* - v_{\pi_k}) \leq \left( 1 - \frac{1-\gamma}{n} \right)^k 1^T(v_* - v_{\pi_0}) \leq \left( 1 - \frac{1-\gamma}{n} \right)^k n V_{\text{max}}. \]

  \[ \| v_* - v_{\pi_k} \|_\infty < \epsilon \iff \left( 1 - \frac{1-\gamma}{n} \right)^k n V_{\text{max}} < \epsilon \iff k \geq \frac{n \log \frac{nV_{\text{max}}}{\epsilon}}{1-\gamma} > \frac{n \log \frac{nV_{\text{max}}}{\epsilon}}{\log \frac{1}{1-\gamma}}. \]
Improved Bounds

**Theorem**

*Howard’s PI* terminates after at most \( O\left(\frac{nm}{1-\gamma} \log \left(\frac{1}{1-\gamma}\right)\right) \) iterations.

**Theorem**

*Simplex-PI* terminates after at most \( O\left(\frac{n^2m}{1-\gamma} \log \left(\frac{1}{1-\gamma}\right)\right) \) iterations.

- Better by a factor \( O(\log n) \) than the previously known bounds for *Howard’s PI* (Hansen et al., 2013) and *Simplex-PI* (Ye, 2011).

- **Tightness:** \( n, m \approx \frac{1}{1-\gamma} \) (Fearnley, 2010; Hollanders et al., 2012; Melekokopoglou & Condon, 1994).

- Howard’s PI vs Simplex-PI: similar overall worst-case complexity
Improved Bounds

Theorem

*Howard’s PI* terminates after at most \( O \left( \frac{nm}{1-\gamma} \log \left( \frac{1}{1-\gamma} \right) \right) \) iterations.

Theorem

*Simplex-PI* terminates after at most \( O \left( \frac{n^2m}{1-\gamma} \log \left( \frac{1}{1-\gamma} \right) \right) \) iterations.

- Better by a factor \( O(\log n) \) than the previously known bounds for Howard’s PI (Hansen et al., 2013) and Simplex-PI (Ye, 2011).
- Tightness: \( n, m \, \frac{1}{1-\gamma} \) (Fearnley, 2010; Hollanders et al., 2012; Melekopoglou & Condon, 1994).
- Howard’s PI vs Simplex-PI: similar overall worst-case complexity.
**Improved Bounds**

**Theorem**

*Howard’s PI* terminates after at most \( O \left( \frac{nm}{1-\gamma} \log \left( \frac{1}{1-\gamma} \right) \right) \) iterations.

**Theorem**

*Simplex-PI* terminates after at most \( O \left( \frac{n^2 m}{1-\gamma} \log \left( \frac{1}{1-\gamma} \right) \right) \) iterations.

- Better by a factor \( O(\log n) \) than the previously known bounds for *Howard’s PI* (Hansen et al., 2013) and *Simplex-PI* (Ye, 2011).

- **Tightness:** \( n, m \approx \frac{1}{1-\gamma} \) (Fearnley, 2010; Hollanders et al., 2012; Melekopoglou & Condon, 1994).

- *Howard’s PI vs Simplex-PI:* similar overall worst-case complexity
Improved Bounds

Theorem

*Howard’s PI* terminates after at most \( O \left( \frac{nm}{1-\gamma} \log \left( \frac{1}{1-\gamma} \right) \right) \) iterations.

Theorem

*Simplex-PI* terminates after at most \( O \left( \frac{n^2m}{1-\gamma} \log \left( \frac{1}{1-\gamma} \right) \right) \) iterations.

- Better by a factor \( O(\log n) \) than the previously known bounds for Howard’s PI (Hansen *et al.*, 2013) and Simplex-PI (Ye, 2011).
- **Tightness**: \( n, m? \frac{1}{1-\gamma}? \) (Fearnley, 2010; Hollanders *et al.*, 2012; Melekopoglou & Condon, 1994).
- **Howard’s PI vs Simplex-PI**: similar overall worst-case complexity
Proof for Howard’s PI

**Lemma**

For all pairs of policies \( \pi \) and \( \pi' \),

\[
    v_{\pi'} - v_{\pi} = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_{\pi} - v_{\pi}).
\]

For some state \( s_0 \), (the “worst” state of \( \pi_0 \))

\[
    v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_\infty
    \leq \|v_* - v_{\pi_k}\|_\infty \quad \{\text{Lemma}\}
    \leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_\infty \quad \{\gamma\text{-contraction}\}
    = \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_\infty
    \leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_\infty \quad \{\text{Lemma}\}
    = \frac{\gamma^k}{1 - \gamma} (v_*(s_0) - T_{\pi_0} v_*(s_0)).
\]

Elimination of a non-optimal action:

For all “sufficiently big” \( k \), \( \pi_k(s_0) \) must differ from \( \pi_0(s_0) \).

“sufficiently big”: \( \frac{\gamma^k}{1 - \gamma} < 1 \iff k \geq \left\lceil \log \frac{1}{1 - \gamma} \right\rceil > \left[ \log \frac{1}{1 - \gamma} \right] > \frac{1}{\log \frac{1}{\gamma}} \).

There are at most \( n(m - 1) \) non-optimal actions to eliminate.
Proof for Howard’s PI

Lemma

For all pairs of policies \( \pi \) and \( \pi' \),
\[
v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_\pi - v_\pi).
\]

For some state \( s_0 \), (the “worst” state of \( \pi_0 \))
\[
v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_\infty \\
\leq \|v_* - v_{\pi_k}\|_\infty \quad \{\text{Lemma}\}
\leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_\infty \quad \{\gamma\text{-contraction}\}
= \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_\infty \quad \{\text{Lemma}\}
\leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_\infty. \quad \{\|I - \gamma P_{\pi_0}\|_\infty = \frac{1}{1 - \gamma}\}
= \frac{\gamma^k}{1 - \gamma} (v_*(s_0) - T_{\pi_0} v_*(s_0)).
\]

Elimination of a non-optimal action:
For all “sufficiently big” \( k \), \( \pi_k(s_0) \) must differ from \( \pi_0(s_0) \).

“sufficiently big”:
\[
\frac{\gamma^k}{1 - \gamma} < 1 \iff k \geq \left\lceil \frac{\log \frac{1}{1 - \gamma}}{1 - \gamma} \right\rceil > \left\lceil \frac{\log \frac{1}{1 - \gamma}}{\log \frac{1}{\gamma}} \right\rceil.
\]

There are at most \( n(m - 1) \) non-optimal actions to eliminate.
Proof for Howard’s PI

Lemma

For all pairs of policies $\pi$ and $\pi'$, $v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1} (T_{\pi'} v_\pi - v_\pi)$.

For some state $s_0$, (the “worst” state of $\pi_0$)

$$v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_\infty \leq \|v_* - v_{\pi_k}\|_\infty \leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_\infty$$

$$= \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_\infty$$

$$\leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_\infty.$$

Elimination of a non-optimal action:

For all “sufficiently big” $k$, $\pi_k(s_0)$ must differ from $\pi_0(s_0)$.

“sufficiently big”: $\frac{\gamma^k}{1 - \gamma} < 1 \iff k \geq \left[ \log \frac{1}{1 - \gamma} \right] > \left[ \log \frac{\frac{1}{1 - \gamma}}{\log \frac{1}{\gamma}} \right]$.

There are at most $n(m - 1)$ non-optimal actions to eliminate.
Proof for Howard's PI

Lemma

For all pairs of policies $\pi$ and $\pi'$, $v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_{\pi} - v_\pi)$.

For some state $s_0$, (the “worst” state of $\pi_0$)

$$v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_\infty$$

$$\leq \|v_* - v_{\pi_k}\|_\infty$$

$$\leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_\infty$$

$$= \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_\infty$$

$$\leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_\infty.$$  \{Lemma\}

$$= \frac{\gamma^k}{1 - \gamma} (v_*(s_0) - T_{\pi_0} v_*(s_0)).$$

Elimination of a non-optimal action:

For all “sufficiently big” $k$, $\pi_k(s_0)$ must differ from $\pi_0(s_0)$.

“sufficiently big”:

$$\frac{\gamma^k}{1 - \gamma} < 1 \iff k \geq \left\lceil \log \frac{1}{1 - \gamma} \right\rceil > \left\lceil \log \frac{1}{1 - \gamma} \right\rceil.$$

There are at most $n(m - 1)$ non-optimal actions to eliminate.
Proof for Howard’s PI

Lemma

For all pairs of policies $\pi$ and $\pi'$, \( v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_\pi - v_\pi) \).

For some state $s_0$, (the “worst” state of $\pi_0$)

\[
v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_{\infty} \leq \|v_* - v_{\pi_k}\|_{\infty} \leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_{\infty} = \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_{\infty} \leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_{\infty}.
\]

\[
= \frac{\gamma^k}{1 - \gamma} (v_*(s_0) - T_{\pi_0} v_*(s_0)).
\]

Elimination of a non-optimal action:

For all “sufficiently big” $k$, $\pi_k(s_0)$ must differ from $\pi_0(s_0)$.

“sufficiently big”: \( \frac{\gamma^k}{1 - \gamma} < 1 \iff k \geq \left\lceil \frac{\log \frac{1}{1 - \gamma}}{\log \frac{1}{1 - \gamma}} \right\rceil > \left\lceil \frac{\log \frac{1}{1 - \gamma}}{\log \frac{1}{\gamma}} \right\rceil \).

There are at most $n(m - 1)$ non-optimal actions to eliminate.
Proof for Howard’s PI

Lemma

For all pairs of policies $\pi$ and $\pi'$, $v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_\pi - v_\pi)$.

For some state $s_0$, (the “worst” state of $\pi_0$)

$v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_\infty$

$\leq \|v_* - v_{\pi_k}\|_\infty \quad \{\text{Lemma}\}$

$\leq \gamma^k \|v_{\pi_*} - v_{\pi_0}\|_\infty \quad \{\gamma$-contraction\}$

$= \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_* - T_{\pi_0} v_*)\|_\infty \quad \{\text{Lemma}\}$

$\leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_\infty. \quad \{\| (I - \gamma P_{\pi_0})^{-1}\|_\infty = \frac{1}{1 - \gamma}\}$

$= \frac{\gamma^k}{1 - \gamma} (v_*(s_0) - T_{\pi_0} v_*(s_0))$.

Elimination of a non-optimal action:

For all “sufficiently big” $k$, $\pi_k(s_0)$ must differ from $\pi_0(s_0)$.

“sufficiently big”: $\frac{\gamma^k}{1 - \gamma} < 1 \Leftrightarrow k \geq \left\lceil \log \frac{1}{1 - \gamma} \right\rceil > \left\lceil \log \frac{1}{\log \frac{1}{\gamma}} \right\rceil$.

There are at most $n(m - 1)$ non-optimal actions to eliminate.
Proof for Howard’s PI

Lemma

For all pairs of policies \( \pi \) and \( \pi' \),
\[
v_{\pi'} - v_\pi = (I - \gamma P_{\pi'})^{-1}(T_{\pi'} v_{\pi} - v_\pi).
\]

For some state \( s_0 \), (the “worst” state of \( \pi_0 \))
\[
v_*(s_0) - T_{\pi_k} v_*(s_0) \leq \|v_* - T_{\pi_k} v_*\|_\infty
\]
\[
\leq \|v_* - v_{\pi_k}\|_\infty \quad \{\text{Lemma}\}
\]
\[
\leq \gamma^k \|v_{\pi_k} - v_{\pi_0}\|_\infty \quad \{\gamma\text{-contraction}\}
\]
\[
= \gamma^k \|(I - \gamma P_{\pi_0})^{-1}(v_\pi - T_{\pi_0} v_\pi)\|_\infty \quad \{\text{Lemma}\}
\]
\[
\leq \frac{\gamma^k}{1 - \gamma} \|v_* - T_{\pi_0} v_*\|_\infty. \quad \{\|(I - \gamma P_{\pi_0})^{-1}\|_\infty = \frac{1}{1 - \gamma}\}
\]
\[
= \frac{\gamma^k}{1 - \gamma} (v_*(s_0) - T_{\pi_0} v_*(s_0)).
\]

Elimination of a non-optimal action:
For all “sufficiently big” \( k \), \( \pi_k(s_0) \) must differ from \( \pi_0(s_0) \).

“sufficiently big”: \( \frac{\gamma^k}{1 - \gamma} < 1 \Leftrightarrow k \geq \left\lceil \frac{\log \frac{1}{1 - \gamma}}{\log \frac{1}{\gamma}} \right\rceil > \left\lceil \frac{\log \frac{1}{1 - \gamma}}{\log \frac{1}{\gamma}} \right\rceil \).

There are at most \( n(m - 1) \) non-optimal actions to eliminate.
Structural assumption

\[ x_\pi = 1^T (I - \gamma P_\pi)^{-1} = 1^T \sum_{t=0}^{\infty} (\gamma P_\pi)^t \]

**Assumption**

Let \( \tau_t \geq 1 \) and \( \tau_r \geq 1 \) be the smallest constants such that for all policies \( \pi \) and all states \( i \),

\[
(1 \leq ) x_\pi(i) \leq \tau_t \quad \text{if } i \text{ is transient for } \pi, \text{ and} \quad (1)
\]

\[
\frac{n}{(1 - \gamma)\tau_r} \leq x_\pi(i) \left( \leq \frac{n}{1 - \gamma} \right) \quad \text{if } i \text{ is recurrent for } \pi. \quad (2)
\]

When \( \gamma \) tends to 1:

- \( \tau_t \geq \) expected time to leave the transient states
- \( \frac{1}{\tau_r} \leq \) asymptotic frequency in a recurrent states

Example: in a deterministic MDP, we have \( \tau_t \leq n \) and \( \tau_r \leq n \).
Structural assumption

\[ x_\pi = 1^T (I - \gamma P_\pi)^{-1} = 1^T \sum_{t=0}^{\infty} (\gamma P_\pi)^i \]

### Assumption

Let \( \tau_t \geq 1 \) and \( \tau_r \geq 1 \) be the smallest constants such that for all policies \( \pi \) and all states \( i \),

\[
(1 \leq ) x_\pi (i) \leq \tau_t \quad \text{if } i \text{ is transient for } \pi, \text{ and} \quad (1)
\]

\[
\frac{n}{(1 - \gamma)\tau_r} \leq x_\pi (i) \left( \leq \frac{n}{1 - \gamma} \right) \quad \text{if } i \text{ is recurrent for } \pi. \quad (2)
\]

When \( \gamma \) tends to 1:

- \( \tau_t \geq \) expected time to leave the transient states
- \( \frac{1}{\tau_r} \leq \) asymptotic frequency in a recurrent states

Example: in a deterministic MDP, we have \( \tau_t \leq n \) and \( \tau_r \leq n \).
Structural assumption

\[ x_\pi = 1^T (I - \gamma P_\pi)^{-1} = 1^T \sum_{t=0}^{\infty} (\gamma P_\pi)^i \]

**Assumption**

Let \( \tau_t \geq 1 \) and \( \tau_r \geq 1 \) be the smallest constants such that for all policies \( \pi \) and all states \( i \),

\[
(1 \leq ) x_\pi (i) \leq \tau_t \quad \text{if } i \text{ is transient for } \pi, \text{ and} \tag{1}
\]

\[
\frac{n}{(1 - \gamma)\tau_r} \leq x_\pi (i) \left( \leq \frac{n}{1 - \gamma} \right) \quad \text{if } i \text{ is recurrent for } \pi. \tag{2}
\]

When \( \gamma \) tends to 1:

- \( \tau_t \geq \) expected time to leave the transient states
- \( \frac{1}{\tau_r} \leq \) asymptotic frequency in a recurrent states

Example: in a deterministic MDP, we have \( \tau_t \leq n \) and \( \tau_r \leq n \).
Generalized Bound for Simplex-PI

**Theorem**

*Simplex-PI* terminates after at most $O\left(n^3 m^2 t r \log^2(n t r)\right)$ iterations.

- Generalizes the result of (Post & Ye, 2012) on deterministic MDPs

**Lemma** (Recurrent classes are created often)

After at most $O\left(n^2 m t \log(n t)\right)$ iterations either *Simplex-PI* finishes or a new recurrent class appears.

**Lemma** (New recurrent class $\Rightarrow$ significant improvement)

When *Simplex-PI* moves from $\pi$ to $\pi'$ where $\pi'$ involves a new recurrent class, we have

$$\mathbb{1}^T (v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{\tau r}\right) \mathbb{1}^T (v_{\pi_*} - v_{\pi}).$$

- Second Lemma problematic for an adaptation to Howard’s PI.
Generalized Bound for Simplex-PI

**Theorem**

*Simplex-PI terminates after at most* \( O \left( n^3 m^2 \tau_t \tau_r \log^2 (n \tau_t \tau_r) \right) \) iterations.

- Generalizes the result of (Post & Ye, 2012) on deterministic MDPs

**Lemma (Recurrent classes are created often)**

After at most \( O \left( n^2 m \tau_t \log(n \tau_t) \right) \) iterations either Simplex-PI finishes or a new recurrent class appears.

**Lemma (New recurrent class \( \Rightarrow \) significant improvement)**

When Simplex-PI moves from \( \pi \) to \( \pi' \) where \( \pi' \) involves a new recurrent class, we have

\[
1^T (v_{\pi_*} - v_{\pi'}) \leq \left( 1 - \frac{1}{\tau_r} \right) 1^T (v_{\pi_*} - v_{\pi}).
\]

- Second Lemma problematic for an adaptation to Howard’s PI.
Generalized Bound for Simplex-PI

**Theorem**

*Simplex-PI* terminates after at most $O\left(n^3 m^2 T_t^2 T_r \log^2(n T_t T_r)\right)$ iterations.

- Generalizes the result of (Post & Ye, 2012) on deterministic MDPs

**Lemma (Recurrent classes are created often)**

After at most $O\left(n^2 m T_t \log(n T_t)\right)$ iterations either *Simplex-PI* finishes or a new recurrent class appears.

**Lemma (New recurrent class ⇒ significant improvement)**

When *Simplex-PI* moves from $\pi$ to $\pi'$ where $\pi'$ involves a new recurrent class, we have

$$1^T (v_{\pi^*} - v_{\pi'}) \leq \left(1 - \frac{1}{T_r}\right) 1^T (v_{\pi^*} - v_{\pi}).$$

- Second Lemma problematic for an adaptation to Howard’s PI.
Generalized Bound for Simplex-PI

**Theorem**

*Simplex-PI* terminates after at most $O \left(n^3 m^2 \tau_t \tau_r \log^2(n \tau_t \tau_r)\right)$ iterations.

- Generalizes the result of (Post & Ye, 2012) on deterministic MDPs

**Lemma (Recurrent classes are created often)**

After at most $O \left(n^2 m \tau_t \log(n \tau_t)\right)$ iterations either *Simplex-PI* finishes or a new recurrent class appears.

**Lemma (New recurrent class ⇒ significant improvement)**

When *Simplex-PI* moves from $\pi$ to $\pi'$ where $\pi'$ involves a new recurrent class, we have

$$1^T(v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{\tau_r}\right) 1^T(v_{\pi_*} - v_{\pi}).$$

- Second Lemma problematic for an adaptation to Howard’s PI.
A New Bound

Assumption

The state space $X$ can be partitioned in two sets $\mathcal{T}$ and $\mathcal{R}$ such that for all policies $\pi$, the states of $\mathcal{T}$ are transient and those of $\mathcal{R}$ are recurrent.

Theorem

Under the above assumption, Howard’s PI terminates after at most $O(n m (\tau_t \log(n \tau_t) + \tau_r \log(n \tau_r)))$ iterations and Simplex-PI terminates after at most $O(n^2 m (\tau_t \log(n \tau_t) + \tau_r \log(n \tau_r)))$ iterations.

- Restrictive assumption
- Proof: 1) convergence on $\mathcal{R}$, 2) convergence on $\mathcal{T}$. 
### Assumption

The state space $X$ can be partitioned in two sets $\mathcal{T}$ and $\mathcal{R}$ such that for all policies $\pi$, the states of $\mathcal{T}$ are transient and those of $\mathcal{R}$ are recurrent.

### Theorem

Under the above assumption, Howard’s PI terminates after at most $O(nm(\tau_t \log(n\tau_t) + \tau_r \log(n\tau_r)))$ iterations and Simplex-PI terminates after at most $O(n^2 m(\tau_t \log(n\tau_t) + \tau_r \log(n\tau_r)))$ iterations.

- **Restrictive** assumption
- Proof: 1) convergence on $\mathcal{R}$, 2) convergence on $\mathcal{T}$. 
Conclusion

Contributions:

• \(O(\log n)\) improvement for Howard’s PI and Simplex-PI
• Generalization of the result of (Post & Ye, 2012) for Simplex-PI to stochastic MDPs
• (A new bound for Howard’s PI and Simplex-PI)

Future Work:

• Tightness of the bounds
• What is the complexity of Howard’s PI? (Schmitz, 1985)
  Open problem, even in the deterministic case!
  Best known lower bound: \(O\left(n^2\right)\) (Hansen & Zwick, 2010)
  Best known upper bound: \(O\left(\frac{2^n}{n}\right)\) (m=2) (Mansour & Singh, 1999)
Conclusion

Contributions:

• $O(\log n)$ improvement for Howard’s PI and Simplex-PI
• Generalization of the result of (Post & Ye, 2012) for Simplex-PI to stochastic MDPs
• (A new bound for Howard’s PI and Simplex-PI)

Future Work:

• Tightness of the bounds
• What is the complexity of Howard’s PI? (Schmitz, 1985)
  Open problem, even in the deterministic case!
  Best known lower bound: $O(n^2)$ (Hansen & Zwick, 2010)
  Best known upper bound: $O\left(\frac{2^n}{n}\right)$ (m=2) (Mansour & Singh, 1999)
References I

Neurodynamic Programming.
Athena Scientific.

Fearnley, J. 2010.
Exponential lower bounds for policy iteration.
ICALP’10.
Berlin, Heidelberg: Springer-Verlag.

Lower bounds for howard’s algorithm for finding minimum mean-cost cycles.
Pages 415–426 of: Isaac (1).

Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor.
J. acm, 60(1), 1:1–1:16.

The complexity of policy iteration is exponential for discounted markov decision processes.
In: 51st ieee conference on decision and control (cdc’12).

On the complexity of policy iteration.
Pages 401–408 of: Uai.
References II

On the complexity of the policy improvement algorithm for markov decision processes.  
*Informs journal on computing*, 6(2), 188–192.

Post, I., & Ye, Y. 2012.  
The simplex method is strongly polynomial for deterministic markov decision processes.  

*Markov Decision Processes*.  
Wiley, New York.

How good is howard’s policy improvement algorithm?  
*Zeitschrift für operations research*, 29(7), 315–316.

Ye, Y. 2011.  
The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate.  