Algorithms for MDP Planning

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay
shivaram@cse.iitb.ac.in

August 2019

Overview

1. Value Iteration
2. Linear Programming
3. Policy Iteration
    - Policy Improvement Theorem

Overview

1. Value Iteration
2. Linear Programming
3. Policy Iteration
    - Policy Improvement Theorem

Value Iteration

$V_0 \leftarrow$ Arbitrary, element-wise bounded, n-length vector. $t \leftarrow 0$.
Repeat:
    For $s \in S$**:**
        $V_{t+1}(s) \leftarrow \max_{a \in A} \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \gamma V_t(s') \right)$.
        $t \leftarrow t + 1$.
Until $V_t \approx V_{t-1}$ (up to machine precision).

Value Iteration

$V_0 \leftarrow$ Arbitrary, element-wise bounded, n-length vector. $t \leftarrow 0$.
Repeat:
  For $s \in S$**:**
    $V_{t+1}(s) \leftarrow \max_{a \in A} \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \gamma V_t(s')\right)$.
    $t \leftarrow t + 1$.
Until $V_t \approx V_{t-1}$ (up to machine precision).

Convergence to $V^\star$ guaranteed using a max-norm contraction argument.

Overview

1. Value Iteration
2. Linear Programming
3. Policy Iteration
    - Policy Improvement Theorem

Linear Programming

$$\text{Minimise} \quad \sum_{s \in S} V(s)$$

$$\text{subject to} \quad V(s) \geq \sum_{s' \in S} T(s, a, s')\left(R(s, a, s') + \gamma V(s')\right), \forall s \in S, \forall a \in A.$$

Linear Programming

Minimise $\sum_{s \in S} V(s)$

subject to $V(s) \geq \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \gamma V(s') \right), \forall s \in S, \forall a \in A.$

Let $|S| = n$ and $|A| = k$.

Linear Programming

Minimise $\sum_{s \in S} V(s)$

subject to $V(s) \geq \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \gamma V(s')\right), \forall s \in S, \forall a \in A.$

Let $|S| = n$ and $|A| = k$.

n variables, nk constraints.

Linear Programming

Minimise $\sum_{s \in S} V(s)$

subject to $V(s) \geq \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \gamma V(s')\right), \forall s \in S, \forall a \in A.$

Let $|S| = n$ and $|A| = k$.

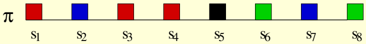
n variables, nk constraints.

Can also be posed as *dual* with nk variables and n constraints.

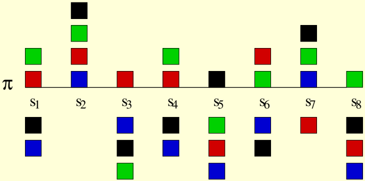
Overview

1. Value Iteration
2. Linear Programming
3. Policy Iteration
    - Policy Improvement Theorem

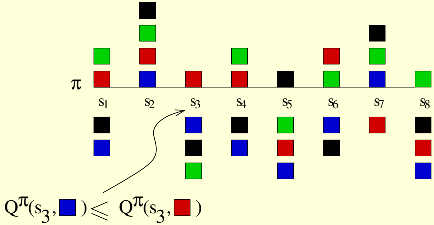
Policy Improvement



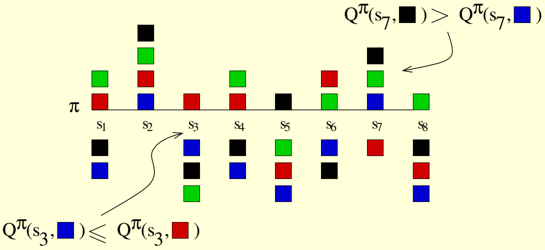
Policy Improvement



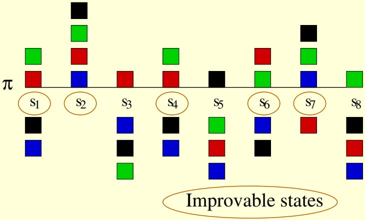
Policy Improvement



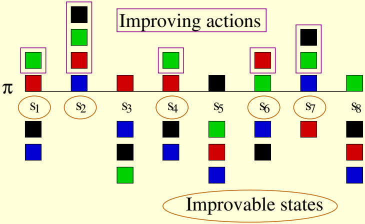
Policy Improvement



Policy Improvement



Policy Improvement



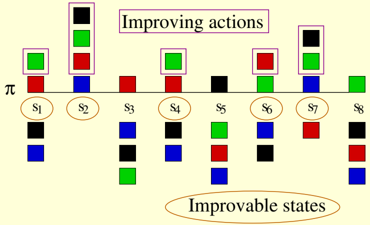
Policy Improvement

Given π,

Pick one or more improvable states, and in them,

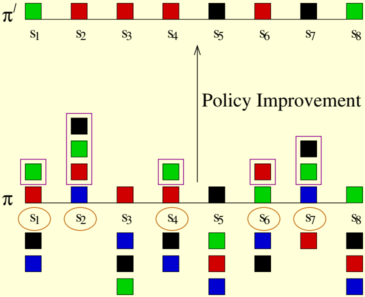
Switch to an arbitrary improving action.

Let the resulting policy be π'.



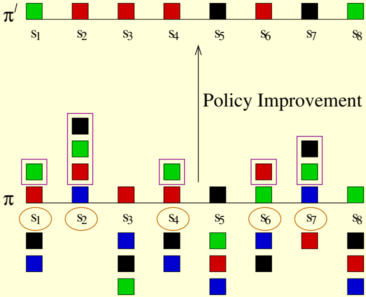
Policy Improvement

Given π,
  Pick one or more improvable states, and in them,
  Switch to an arbitrary improving action.
Let the resulting policy be π'.



Policy Improvement

Given π,
Pick one or more improvable states, and in them,
Switch to an arbitrary improving action.
Let the resulting policy be π'.

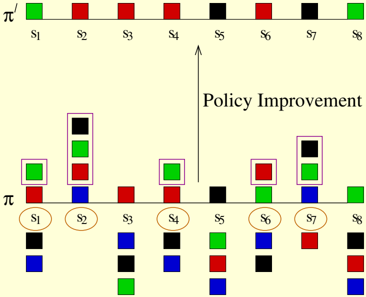


Policy Improvement Theorem:
(1) If π has no improvable states, then it is optimal, else
(2) if π' is obtained as above, then

$$\forall s \in S : V^{\pi'}(s) \geq V^{\pi}(s) \text{ and } \exists s \in S : V^{\pi'}(s) > V^{\pi}(s).$$

Policy Improvement

Given π,
  Pick one or more improvable states, and in them,
  Switch to an arbitrary improving action.
Let the resulting policy be π'.



Policy Improvement Theorem:
(1) If π has no improvable states, then it is optimal, else
(2) if π' is obtained as above, then
$$\forall s \in S : V^{\pi'}(s) \geq V^{\pi}(s) \text{ and } \exists s \in S : V^{\pi'}(s) > V^{\pi}(s).$$

Definitions and Basic Facts

- For $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, we define $X \succeq Y$ if $\forall s \in S : X(s) \geq Y(s)$, and we define $X \succ Y$ if $X \succeq Y$ and $\exists s \in S : X(s) > Y(s)$.

Definitions and Basic Facts

- For $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, we define $X \succeq Y$ if $\forall s \in S : X(s) \geq Y(s)$, and we define $X \succ Y$ if $X \succeq Y$ and $\exists s \in S : X(s) > Y(s)$.

  For policies $\pi_1, \pi_2 \in \Pi$, we define $\pi_1 \succeq \pi_2$ if $V^{\pi_1} \succeq V^{\pi_2}$, and we define $\pi_1 \succ \pi_2$ if $V^{\pi_1} \succ V^{\pi_2}$.

Definitions and Basic Facts

- For $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, we define $X \succeq Y$ if $\forall s \in S : X(s) \geq Y(s)$, and we define $X \succ Y$ if $X \succeq Y$ and $\exists s \in S : X(s) > Y(s)$.

  For policies $\pi_1, \pi_2 \in \Pi$, we define $\pi_1 \succeq \pi_2$ if $V^{\pi_1} \succeq V^{\pi_2}$, and we define $\pi_1 \succ \pi_2$ if $V^{\pi_1} \succ V^{\pi_2}$.

- **Bellman Operator.** For $\pi \in \Pi$, we define $B^{\pi} : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ as follows: for $X : S \to \mathbb{R}$ and $\forall s \in S$,

$$(B^{\pi}(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s') \right).$$

Definitions and Basic Facts

- For $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, we define $X \succeq Y$ if $\forall s \in S : X(s) \geq Y(s)$, and we define $X \succ Y$ if $X \succeq Y$ and $\exists s \in S : X(s) > Y(s)$.

  For policies $\pi_1, \pi_2 \in \Pi$, we define $\pi_1 \succeq \pi_2$ if $V^{\pi_1} \succeq V^{\pi_2}$, and we define $\pi_1 \succ \pi_2$ if $V^{\pi_1} \succ V^{\pi_2}$.

- **Bellman Operator.** For $\pi \in \Pi$, we define $B^\pi : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ as follows: for $X : S \to \mathbb{R}$ and $\forall s \in S$,

$$(B^\pi(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s') \right).$$

- **Fact 1**. For $\pi \in \Pi$, $X : S \to \mathbb{R}$, and $Y : S \to \mathbb{R}$:

$$\text{if } X \succeq Y, \text{ then } B^\pi(X) \succeq B^\pi(Y).$$

Definitions and Basic Facts

- For $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, we define $X \succeq Y$ if $\forall s \in S : X(s) \geq Y(s)$, and we define $X \succ Y$ if $X \succeq Y$ and $\exists s \in S : X(s) > Y(s)$.

  For policies $\pi_1, \pi_2 \in \Pi$, we define $\pi_1 \succeq \pi_2$ if $V^{\pi_1} \succeq V^{\pi_2}$, and we define $\pi_1 \succ \pi_2$ if $V^{\pi_1} \succ V^{\pi_2}$.

- **Bellman Operator.** For $\pi \in \Pi$, we define $B^\pi : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ as follows: for $X : S \to \mathbb{R}$ and $\forall s \in S$,

$$(B^\pi(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s') \right).$$

- **Fact 1**. For $\pi \in \Pi$, $X : S \to \mathbb{R}$, and $Y : S \to \mathbb{R}$:

$$\text{if } X \succeq Y, \text{ then } B^\pi(X) \succeq B^\pi(Y).$$

- **Fact 2**. For $\pi \in \Pi$ and $X : S \to \mathbb{R}$:

$$\lim_{l \to \infty} (B^\pi)^l(X) = V^\pi. \text{ (from Banach's FP Theorem)}$$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$

$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi})$$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.$$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.$$

π has improvable states and policy improvement yields π'

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty}(B^{\pi'})^l(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.$$

π has improvable states and policy improvement yields π'

$$\implies B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty}(B^{\pi'})^l(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.$$

π has improvable states and policy improvement yields π'

$$\implies B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

$$\implies (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$
$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$
$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty}(B^{\pi'})^l(V^{\pi})$$
$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.$$

π has improvable states and policy improvement yields π'

$$\implies B^{\pi'}(V^{\pi}) \succ V^{\pi}$$
$$\implies (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}$$
$$\implies \lim_{l \to \infty}(B^{\pi'})^l(V^{\pi}) \succeq \cdots \succeq (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

Proof of Policy Improvement Theorem

Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$.

π has no improvable states

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi})$$

$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.$$

π has improvable states and policy improvement yields π'

$$\implies B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

$$\implies (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

$$\implies \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \succeq \cdots \succeq (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

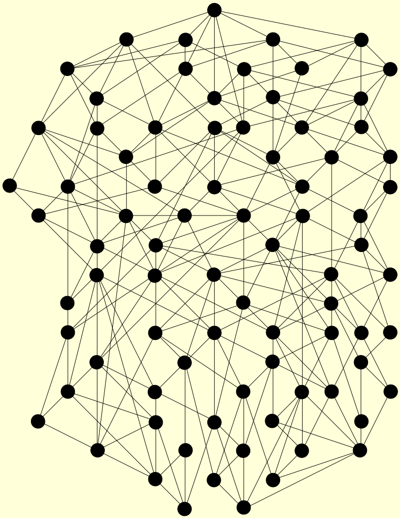
$$\implies V^{\pi'} \succ V^{\pi}.$$

Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi \leftarrow$ PolicyImprovement(π).

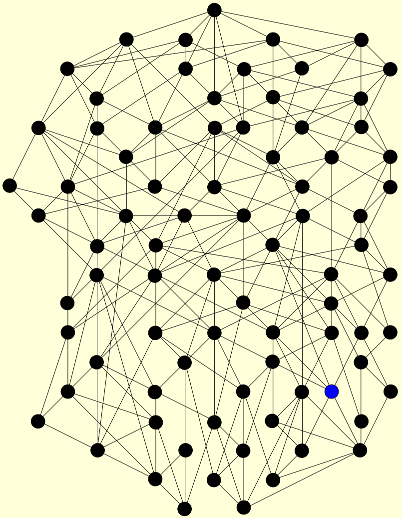
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
$\quad\quad \pi \leftarrow$ PolicyImprovement(π).



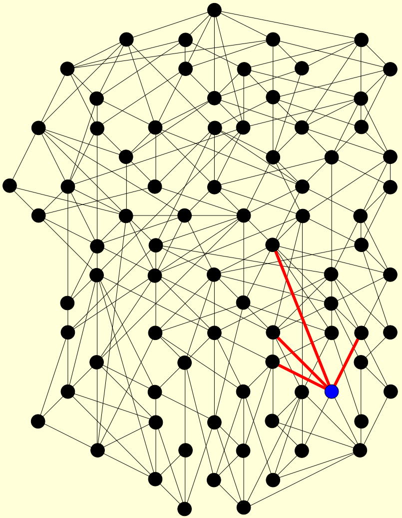
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi \leftarrow$ PolicyImprovement(π).



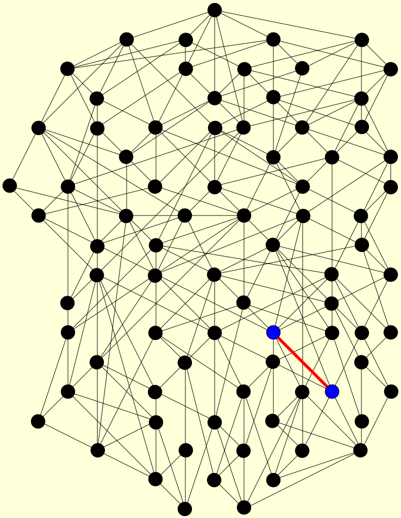
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi \leftarrow$ PolicyImprovement(π).



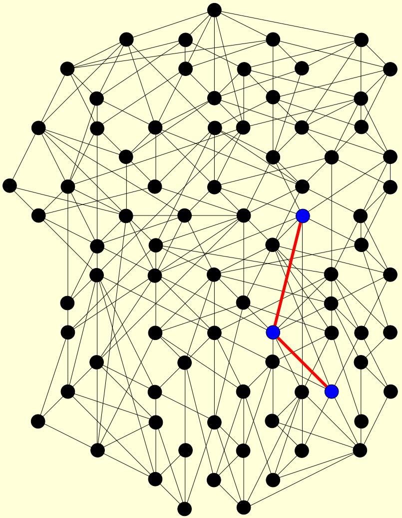
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi \leftarrow$ PolicyImprovement(π).



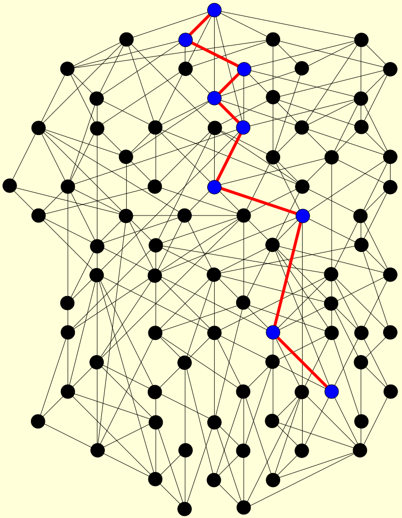
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
$\quad\pi \leftarrow$ PolicyImprovement(π).



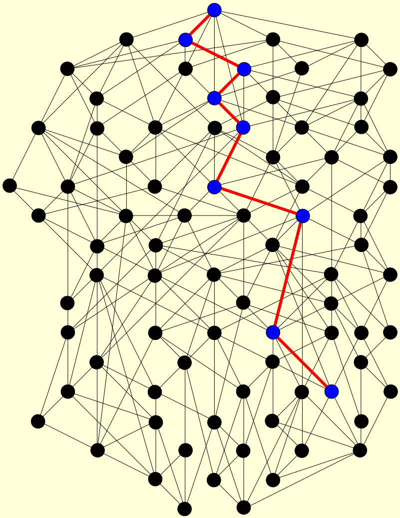
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi \leftarrow$ PolicyImprovement(π).



Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi \leftarrow$ PolicyImprovement(π).



Number of iterations depends on switching strategy. Current bounds quite loose.