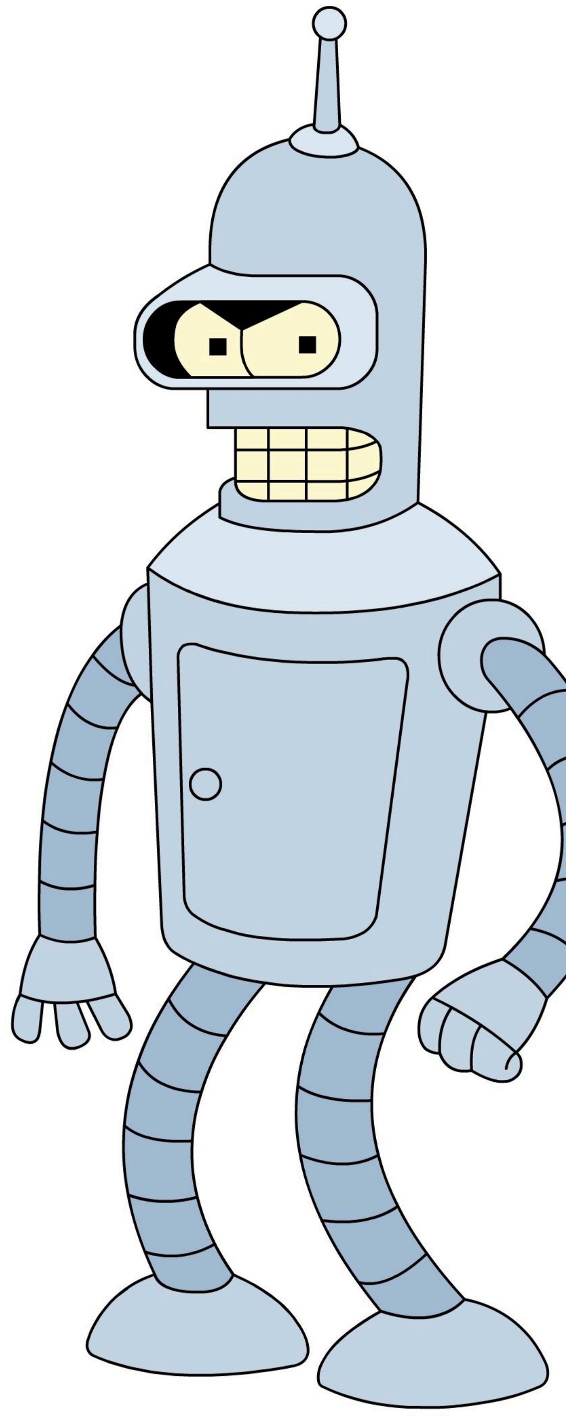
Reinforcement Learning

Section 2: Dynamic Programming Analysis



Sergei Laktionov
slaktionov@hse.ru
LinkedIn

Bellman Operators

Bellman expectation operator for $V(s)$:

$$[\mathcal{T}^{\pi}V](s) = \mathbb{E}_{r,s'|s,a\sim\pi(.|s)}[r + \gamma V(s')]$$

Bellman expectation operator for $Q(s, a)$:

$$[\mathcal{T}^{\pi}Q](s, a) = \mathbb{E}_{r,s'|s,a}\left[r + \gamma \mathbb{E}_{a'\sim\pi(.|s)}[Q(s', a')]\right]$$

Bellman optimality operator for $V(s)$:

$$[\mathcal{T}V](s) = \max_{a} \mathbb{E}_{r,s'|s,a}[r + \gamma V(s')]$$

Bellman optimality operator for $Q(s, a)$:

$$[\mathcal{T}Q](s, a) = \mathbb{E}_{r,s'|s,a}\left[r + \gamma \max_{a'} Q(s', a')\right]$$

Dynamic Programming Algorithms

Assume that $\mathcal{S}, \mathcal{A}$ are finite.

Policy Iteration

1. Initialise π, V; $V(s) = 0$ if s is terminal
2. Policy Evaluation:
   Apply $\mathcal{T}^\pi V$ as an update rule for each s until convergence or solve a system of linear equations for V^π.
3. Policy Improvement:
   calculate $Q(s, a) = \mathbb{E}_{r,s'|s,a}[r + \gamma V(s')]$
   update π greedily w.r.t. $Q(s, a)$
4. Repeat 2-3 until policy stabilisation

Value Iteration

1. Initialise V; $V(s) = 0$ if s is terminal
2. Apply $\mathcal{T} V$ for each s until convergence
3. Calculate $Q(s, a) = \mathbb{E}_{r,s'|s,a}[r + \gamma V(s')]$
   Assign π to the greedy policy w.r.t. $Q(s, a)$

Value Iteration Convergence

Contraction: $||\mathcal{T}V - \mathcal{T}U||_\infty \leq \gamma ||V - U||_\infty$

$|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$

Proof: $|[\mathcal{T}V](s) - [\mathcal{T}U](s)| = |\max_a \mathbb{E}_{r,s'|s,a}\left[r + \gamma V(s')\right] - \max_a \mathbb{E}_{r,s'|s,a}\left[r + \gamma U(s')\right]| \leq$

$\leq \gamma \max_a |\mathbb{E}_{s'|s,a}\left[V(s') - U(s')\right]| \leq \gamma \max_{s'} |V(s') - U(s')| \leq \gamma ||V - U||_\infty$

Thus, $||\mathcal{T}V - \mathcal{T}U||_\infty \leq \gamma ||V - U||_\infty$

Value Iteration Convergence

Contraction: $||\mathscr{T}V - \mathscr{T}U||_\infty \leq \gamma ||V - U||_\infty$

$|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$

Proof: $|[\mathscr{T}V](s) - [\mathscr{T}U](s)| = |\max_a \mathbb{E}_{r,s'|s,a}\left[r + \gamma V(s')\right] - \max_a \mathbb{E}_{r,s'|s,a}\left[r + \gamma U(s')\right]| \leq$

$\leq \gamma \max_a |\mathbb{E}_{s'|s,a}\left[V(s') - U(s')\right]| \leq \gamma \max_{s'} |V(s') - U(s')| \leq \gamma ||V - U||_\infty$

Thus, $||\mathscr{T}V - \mathscr{T}U||_\infty \leq \gamma ||V - U||_\infty$

Convergence:

By Banach fixed point theorem, $V_k = \mathscr{T}V_{k-1} \rightarrow V^*$ s.t. $V^* = \mathscr{T}V^*$ and V^* is unique.

Policy Iteration Convergence

Monotonicity: For all V, U if $V(s) \leq U(s)\ \forall s \in \mathcal{S}$ then $(\mathcal{T}^\pi V)(s) \leq (\mathcal{T}^\pi U)(s)\ \forall s \in \mathcal{S}$

Proof: $(\mathcal{T}^\pi V)(s) = \mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma V(s')\right] \leq \mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma U(s')\right] = (\mathcal{T}^\pi U)(s)$

Policy Iteration Convergence

Contraction: $||\mathcal{T}^\pi V - \mathcal{T}^\pi U||_\infty \leq \gamma ||V - U||_\infty$

Proof: $|[\mathcal{T}^\pi V](s) - [\mathcal{T}^\pi U](s)| = |\mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma V(s')\right] - \mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma U(s')\right]| =$

$= |\mathbb{E}_{r,s'|s,a=\pi(s)}\left[\gamma V(s') - \gamma U(s')\right]| \leq \gamma \max_{s'} |V(s') - U(s')| = \gamma ||V - U||_\infty$

Thus, $||\mathcal{T}^\pi V - \mathcal{T}^\pi U||_\infty \leq \gamma ||V - U||_\infty$

Policy Iteration Convergence

Contraction: $||\mathcal{T}^{\pi}V - \mathcal{T}^{\pi}U||_{\infty} \leq \gamma ||V - U||_{\infty}$

Proof: $|[\mathcal{T}^{\pi}V](s) - [\mathcal{T}^{\pi}U](s)| = |\mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma V(s')\right] - \mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma U(s')\right]| =$

$= |\mathbb{E}_{r,s'|s,a=\pi(s)}\left[\gamma V(s') - \gamma U(s')\right]| \leq \gamma \max_{s'} |V(s') - U(s')| = \gamma ||V - U||_{\infty}$

Thus, $||\mathcal{T}^{\pi}V - \mathcal{T}^{\pi}U||_{\infty} \leq \gamma ||V - U||_{\infty}$

Policy evaluation convergence:

By Banach fixed point theorem, $V_k = \mathcal{T}^{\pi}V_{k-1} \rightarrow V^{\pi}$ s.t. $V^{\pi} = \mathcal{T}^{\pi}V^{\pi}$ and V^{π} is unique.

Policy Iteration Convergence

Policy Improvement Step: $\pi_{k+1}(s) = argmax_a Q^{\pi_k}(s, a)$

Policy Improvement Theorem:

Let π, π' be two policies s.t. $Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s)\ \forall s \in \mathcal{S}$. Then $V^{\pi'} \geq V^{\pi}$.

Proof:

Policy Iteration Convergence

Policy Improvement Step: $\pi_{k+1}(s) = argmax_a Q^{\pi_k}(s, a)$

Policy Improvement Theorem:

Let π, π' be two policies s.t. $Q^\pi(s, \pi'(s)) \geq V^\pi(s)\ \forall s \in \mathcal{S}$. Then $V^{\pi'} \geq V^\pi$.

Proof: $V^\pi(s) \leq Q^\pi(s, \pi'(s)) = \mathbb{E}_{r,s'|s,\pi'(s)}[r + \gamma V^\pi(s')] \leq \mathbb{E}_{r,s'|s,\pi'(s)}[r + \gamma Q^\pi(s', \pi'(s'))] =$

$= \mathbb{E}_{r,s',r',s''|s,\pi'(s)}[r + \gamma r' + \gamma^2 V^\pi(s'')] \leq \ldots \leq V^{\pi'}(s).$

Policy Iteration Convergence

Policy Improvement Step: $\pi_{k+1}(s) = argmax_a Q^{\pi_k}(s, a)$

Policy Improvement Theorem:

Let π, π' be two policies s.t. $Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s)\ \forall s \in \mathcal{S}$. Then $V^{\pi'} \geq V^{\pi}$.

Proof: $V^{\pi}(s) \leq Q^{\pi}(s, \pi'(s)) = \mathbb{E}_{r,s'|s,\pi'(s)}[r + \gamma V^{\pi}(s')] \leq \mathbb{E}_{r,s'|s,\pi'(s)}[r + \gamma Q^{\pi}(s', \pi'(s'))] =$

$= \mathbb{E}_{r,s',r',s''|s,\pi'(s)}[r + \gamma r' + \gamma^2 V^{\pi}(s'')] \leq \ldots \leq V^{\pi'}(s)$.

Then we get the sequence $\pi_0 \leq \pi_1 \leq \ldots \leq \pi_k \leq \ldots$. Since finite MDP has finite number of deterministic policies the process stabilises i.e. $\exists k : \pi_k = \pi_{k+1}$.

So $\pi_k(s) = argmax_a Q^{\pi_k}(s, a) \rightarrow V^{\pi_k}(s) = \max_a \mathbb{E}_{r,s'|s,a}[r + \gamma V^{\pi_k}(s')] \rightarrow \pi_k = \pi^*$,

as V^{π_k} satisfies Bellman Optimality Equation.