A Convergent Form of Approximate Policy Iteration

Monica Patel (260728093)

March 17, 2017

1 Synopsis of Paper

- The paper presents convergence proof of 'Approximate' policy iteration method.
- It also provides intuitive understanding about why there is lack of convergence in such methods and explains conditions in which the convergence and achieved and that why those conditions are necessary.

2 Problem with Approximation of State-Value function

- Studies in many papers show that RL algorithms such as Q learning, Sarsa, and approximate policy iteration can diverge or cycle without convergence when combined with generalizing function approximators.
- One reason for this can be that, even if one assumes that the agent's environment is Markovian, the problem is Non-Markovian from agent's point of view.
- Another reason is, discontinuous agent's behavior and use of approximator with it. Meaning, if any kind of state-action value approximator is used, the learned values depend upon frequency of vising that state-action pair. And if agent's behavior is discontinuous, slight change in value estimates may result in radical change in the behavior.
- This can be avoided by making agent's policy continuous function of its value, which will ensure that small changes in action-value results in small change in behavior.
- But research have shown that continuity of the agent's behavior leads to fixed point, only to which algorithm can converge.

3 Intuitive Idea used in Paper

- The similar idea, that agent's behavior should not change drastically with value estimate is used in the paper.
- Paper studies the form of approximate policy iteration, in which at each iteration:
  - Sarsa updating is used to learn weights of linear approximation.
  - Policy Improvement operator is used to determine new policy based on learned action-values
- Paper shows that, if the Policy Improvement operator is epsilon-soft and Lipschitz continuous in action value (everywhere differentiable and bound by a constant) and that constant is not too large, then generated policies are guaranteed to converge.

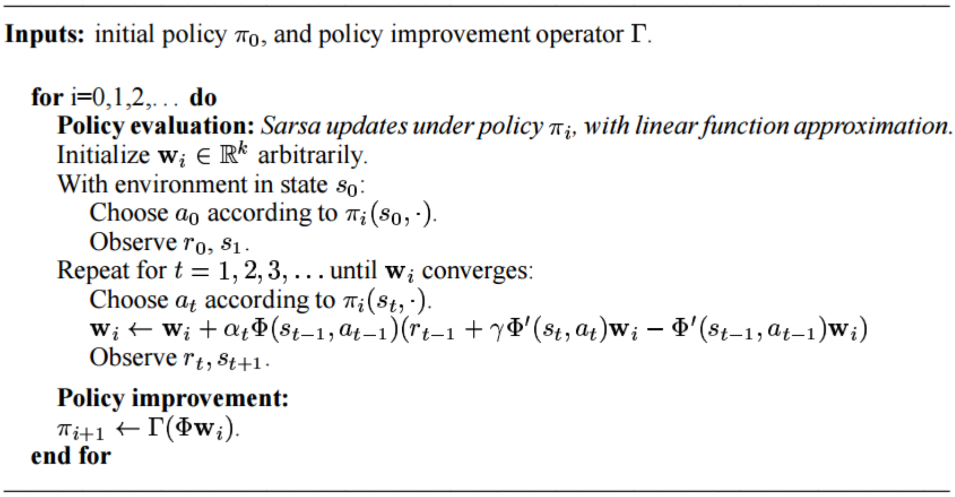
4 Markov Decision Process –> Notation

- Infinite horizon discounted Markov decision process.
- Finite state-set S cardinality m, finite action set A cardinality n.
- Immediate reward: if in state s and choses action a –> r_s^a and process transitions to next state with probability –> $p^a_{s,s'}$
- r is length mn vector of expected immediate reward following each state-action pair.
- probability that agent chooses a when in state s –> $\pi(s, a)$. Policy π is called ϵ-soft if π(s,a) $>= \epsilon$. Policy π can be viewed as element of R^{mn}
- The approximate policy iteration algorithm learns linear approximation to the action value function. The state action pair is represented by length ***k*** feature vector. weight vector is ***w***, therefore, $\hat{Q}(s, a) = \phi'(s, a)w$

5 Assumptions made for the proof:

1. Under any policy π, the Markov decision process behaves as the irreducible, aperiodic Markov chain over the state set S, meaning for a policy π for markov chain it is possible to get from any state to any state and does not return to particular state after every particular amount of steps.
2. Columns of ϕ are linearly independent.

6 The version of Approximate Policy iteration studied in Paper



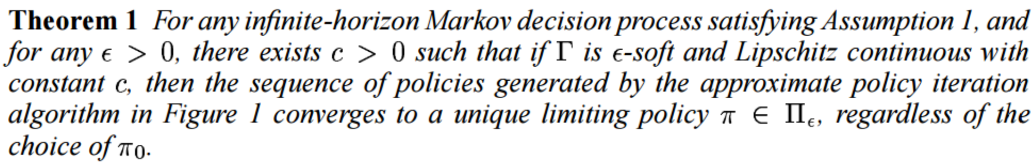
```
Inputs: initial policy $\pi_0$, and policy improvement operator $\Gamma$.

for i=0,1,2,... do
    Policy evaluation: Sarsa updates under policy $\pi_i$, with linear function approximation.
    Initialize $\mathbf{w}_i \in \mathbb{R}^k$ arbitrarily.
    With environment in state $s_0$:
        Choose $a_0$ according to $\pi_i(s_0, \cdot)$.
        Observe $r_0$, $s_1$.
    Repeat for $t = 1, 2, 3, \ldots$ until $\mathbf{w}_i$ converges:
        Choose $a_t$ according to $\pi_i(s_t, \cdot)$.
        $\mathbf{w}_i \leftarrow \mathbf{w}_i + \alpha_t \Phi(s_{t-1}, a_{t-1})(r_{t-1} + \gamma \Phi'(s_t, a_t)\mathbf{w}_i - \Phi'(s_{t-1}, a_{t-1})\mathbf{w}_i)$
        Observe $r_t$, $s_{t+1}$.
    Policy improvement:
    $\pi_{i+1} \leftarrow \Gamma(\Phi \mathbf{w}_i)$.
end for
```

7 Previous work and work done in paper

1. Previous Work: Bertsekas and Tsitsiklis
    - In approximate policy iteration algorithm, the policy evaluation step is not exact. $V^{\pi i}$ is approximated by weighted linear combination of state feature. Weights are determined by Monte-carlo of TD(λ) learning rules. And policy improvement step is same as normal policy iteration.
2. Work in Paper:

- Sarsa update rule is used to learn the weights of the linear value function.
- Action-value function is used instead of state value.
- Generic policy improvement operator is assumed which maps $Q \in R^{mn}$ to stochastic policy.

8 Theorem in Paper and its Intuitive Notion:

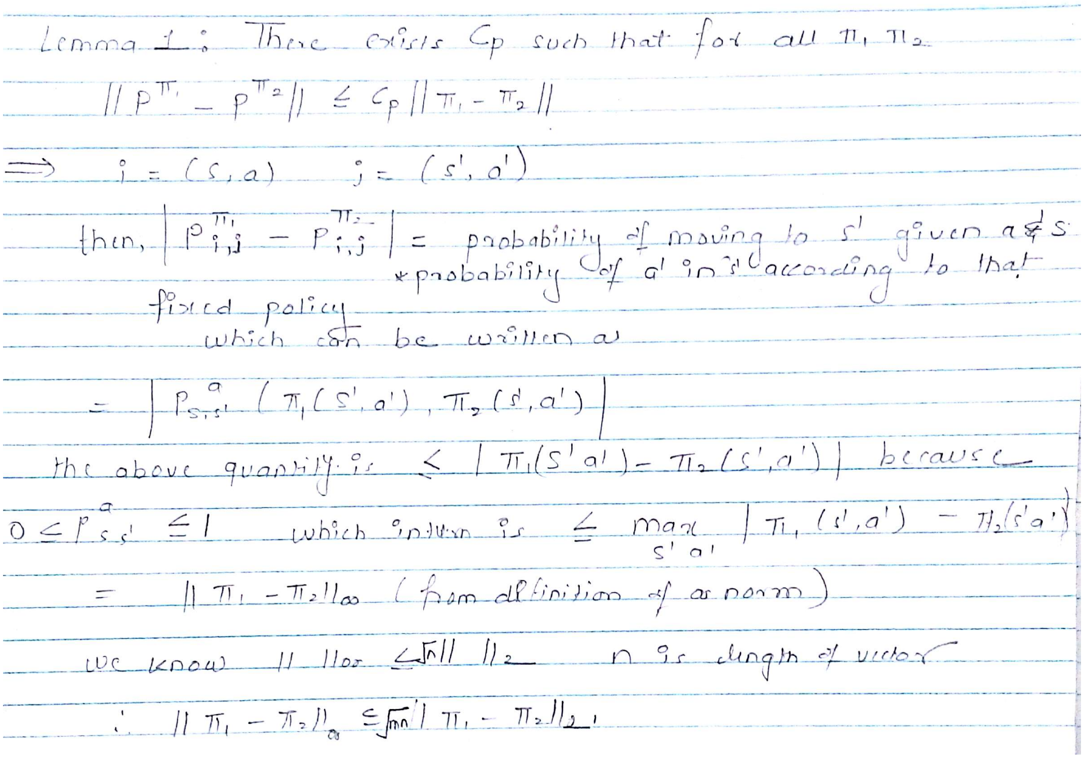


Theorem 1 *For any infinite-horizon Markov decision process satisfying Assumption 1, and for any $\epsilon > 0$, there exists $c > 0$ such that if Γ is ϵ-soft and Lipschitz continuous with constant c, then the sequence of policies generated by the approximate policy iteration algorithm in Figure 1 converges to a unique limiting policy $\pi \in \Pi_\epsilon$, regardless of the choice of π_0.*

Meaning–> if the behavior of the agent does not change greatly with change in action value estimate, the we can get guaranteed.

- The theorem states simple condition under which a form of model-free reinforcement learning control based on approximating value functions converges for a general class of problems.
- It offers no guarantee on the quality of the policy to which the algorithm converges. Meaning the final policy may not be the optimal policy for the problem but will achieve goal.
- It does not find the particular constant c for which convergence is achieved but only says that convergence exists for some c.

9 Lemmas and proofs



Lemma 1: There exists C_p such that for all π_1, π_2

$$\| P^{\pi_1} - P^{\pi_2} \| \leq C_p \| \pi_1 - \pi_2 \|$$

$\Longrightarrow$ $i = (s, a)$ $j = (s', a')$

then, $\left| P^{\pi_1}_{i,j} - P^{\pi_2}_{i,j} \right|$ = probability of moving to s' given a & s * probability of a' in s' according to that fixed policy

which can be written as

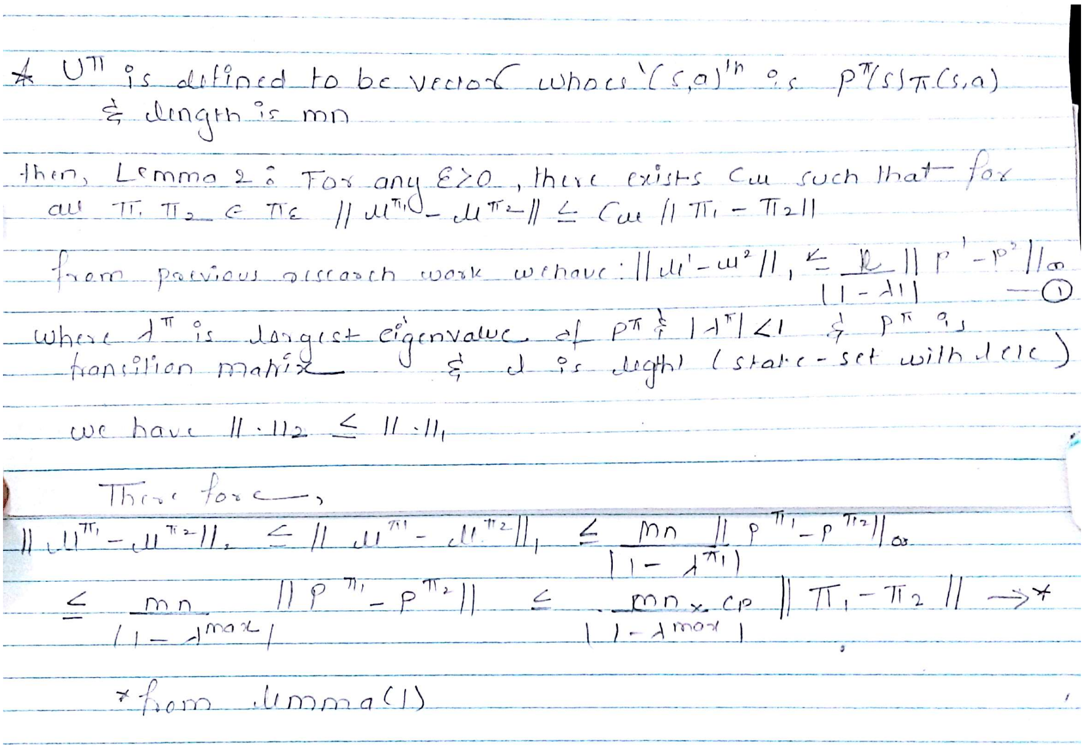
$$= \left| P^{a}_{s,s'} \left(\pi_1(s', a'), \pi_2(s', a') \right) \right|$$

the above quantity is $\leq | \pi_1(s' a') - \pi_2(s', a') |$ because $0 \leq P^{a}_{s s'} \leq 1$ which inturn is $\leq \max_{s' a'} | \pi_1(s', a') - \pi_2(s' a') |$

$= \| \pi_1 - \pi_2 \|_\infty$ (from definition of ∞ norm)

we know $\| \, \|_\infty \leq \sqrt{n} \| \, \|_2$ n is length of vector

$\therefore \| \pi_1 - \pi_2 \|_\infty \leq \sqrt{mn} \| \pi_1 - \pi_2 \|_2$,



* U^{π} is defined to be vector whose $(s,a)^{th}$ is $p^{\pi}(s)\pi(s,a)$ & length is mn

then, Lemma 2 : For any $\varepsilon > 0$, there exists C_{u} such that for all $\pi_1\ \pi_2 \in \pi_{\varepsilon}$ $\| \mu^{\pi_1} - \mu^{\pi_2} \| \leq C_{u} \| \pi_1 - \pi_2 \|$

from previous research work we have : $\| \mu^{1} - \mu^{2} \|_{1} \leq \dfrac{k}{|1 - \lambda|} \| p^{1} - p^{2} \|_{\infty}$ — ①

where λ^{π} is largest eigenvalue of p^{π} & $|\lambda^{\pi}| < 1$ & p^{π} is transition matrix & d is length (state-set with d(c))

we have $\| \cdot \|_{2} \leq \| \cdot \|_{1}$

There fore,

$$\| \mu^{\pi_1} - \mu^{\pi_2} \|_{2} \leq \| \mu^{\pi_1} - \mu^{\pi_2} \|_{1} \leq \frac{mn}{|1 - \lambda^{\pi_1}|} \| p^{\pi_1} - p^{\pi_2} \|_{\infty}$$

$$\leq \frac{mn}{|1 - \lambda^{max}|} \| p^{\pi_1} - p^{\pi_2} \| \leq \frac{mn \times C_{p}}{|1 - \lambda^{max}|} \| \pi_1 - \pi_2 \| \rightarrow *$$

* from lemma(1)

Finding weights: —

we define D^{π} as matrix whose diagonal is μ^{π}

from previous research we know

$$\underbrace{\phi' D^{\pi} (I - \gamma P^{\pi}) \phi w}_{\substack{\Downarrow \\ A^{\pi}}} = \underbrace{\phi' D^{\pi} r}_{\substack{\Downarrow \\ b^{\pi}}} \quad — ①$$

$A^{\pi} \Rightarrow$ invertible therefore $w^{\pi} = A^{\pi^{-1}} b^{\pi}$ from ①

Lemma 3 : There exist C_b & C_A such that for π^1 π^2

$\| b^{\pi_1} - b^{\pi_2} \| \le C_b \| \pi_1 - \pi_2 \|$ and $\| A^{\pi_1} - A^{\pi_2} \| \le C_A \| \pi_1 - \pi_2 \|$

we have $\| x y \| \le \| x \| \| y \|$

Just by substituting value for I^{st} we can see

$\| \phi' (D^{\pi_1} - D^{\pi_2}) r \| \le \| \phi' \| \, \| D^{\pi_1} - D^{\pi_2} \| \, \| r \|$

we have $\| D^{\pi_1} - D^{\pi_2} \| \le \| \mu^{\pi_1} - \mu^{\pi_2} \|$ — ②

$\therefore$ from 2 & lemma 2 :

$\| \phi' (D^{\pi_1} - D^{\pi_2}) r \| \le \| \phi' \| \, C_\mu \, \| \pi_1 - \pi_2 \| \, \| r \|$

for second part $\Rightarrow$ substituting values, expanding & re-arranging

$\| A^{\pi_1} - A^{\pi_2} \| \le \| \phi' \| \, \| D^{\pi_1} - D^{\pi_2} - \gamma D^{\pi_1} (P^{\pi_1} - P^{\pi_2}) - \gamma (D^{\pi_1} - D^{\pi_2}) P^{\pi_2} \| \, \| \phi \|$

using norm & inequality properties

$\le \| \phi' \| (\| D^{\pi_1} - D^{\pi_2} \| + \gamma \| D^{\pi_1} \| \, \| P^{\pi_1} - P^{\pi_2} \| + \gamma \| D^{\pi_1} - D^{\pi_2} \| \, \| P^{\pi_2} \|) \| \phi \|$

$\le ((1 + \gamma) C_\mu + \gamma C_p) \| \phi' \| \, \| \phi \| \, \| \pi_1 - \pi_2 \|$

from lemma 1 & 2 $\| D^{\pi} \| < 1$ & $\| P^{\pi} \| = 1$

Lemma 4 : For any $\varepsilon > 0$ there exists C_ω such that $\| \omega^\pi \| \leq C_\omega$ for all $\pi \in \Pi_\varepsilon$

Lemma 5 : For any $\varepsilon > 0$ there exists $C_g > 0$ such that for all $\pi \in \Pi_\varepsilon$ $g(A^\pi) \geq C_g$

Lemma 6 : For any $\varepsilon > 0$ there exists $C_{\omega 2}$ such that for all $\pi_1, \pi_2 \in \Pi_\varepsilon$ $\| \omega^{\pi_1} - \omega^{\pi_2} \| \leq C_{\omega 2} \| \pi_1 - \pi_2 \|$

we saw $A^{\cancel{\omega}\pi_1} \omega^\pi = b^{\pi_1}$

Thus :

$$A^{\pi_1} \omega^{\pi_1} - A^{\pi_2} \omega^{\pi_2} = b^{\pi_1} - b^{\pi_2}$$

addin ω^{π_2} & substracting ω^{π_2} & rearranging we have

$$A^{\pi_1} (\omega^{\pi_1} - \omega^{\pi_2}) + (A^{\pi_1} - A^{\pi_2}) \omega^{\pi_2} = b^{\pi_1} - b^{\pi_2}$$

rearranging & taking norm both side

$$\| A^{\pi_1} (\omega^{\pi_1} - \omega^{\pi_2}) \| \leq \| b^{\pi_1} - b^{\pi_2} \| + \| A^{\pi_1} - A^{\pi_2} \| \| \omega^{\pi_2} \|$$

substituting using lemma 3 & 5 & 4

$$C_g \| \omega^{\pi_1} - \omega^{\pi_2} \| \leq C_b \| \pi_1 - \pi_2 \| + C_\omega C_A \| \pi_1 - \pi_2 \|$$

$$\Rightarrow \| \omega^{\pi_1} - \omega^{\pi_2} \| = C_g^{-1} (C_b + C_\omega C_A) \| \pi_1 - \pi_2 \|$$

Contraction Argument & proof :-

we have observed :

$$\| \Gamma \hat{Q}^{\pi_1} - \Gamma \hat{Q}^{\pi_2} \| \leq c \| \hat{Q}^{\pi_1} - \hat{Q}^{\pi_2} \|$$

substituting from definition & lemma 6

$$= c \| \phi \, \omega^{\pi_1} - \omega^{\pi_2} \| \leq c \| \phi \| C_{\omega 2} \| \pi_1 - \pi_2 \|$$

if $c < \| \phi \|^{-1} C_{\omega_2}^{-1}$ then for $\beta \in [0, 1)$

we have $\| \Gamma \hat{Q}^{\pi_1} - \Gamma \hat{Q}^{\pi_2} \| \leq \beta \| \pi_1 - \pi_2 \|$

Thus each step of the approximate policy Iteration is Contraction .

$\therefore$ By contraction Mapping theorem

$\pi \rightarrow \Gamma(\hat{Q}^\pi)$

$\therefore$ policies converges to fixed point .

10 Conclusion

- The magnitude of the constant that ensures convergence depends on the model of the environment and on properties of the feature representation. Higher c can be chosen to begin with and if the contraction property fails then c should be reduced.
- It is possible that convergence could be obtained with much higher values of than are suggested by the bound in the proof of Theorem 1.
- There is no theoretical guarantees on the quality of solutions found.