Safe Policy Iteration

Matteo Pirotta, Marcello Restelli
Alessio Pecorino, Daniele Calandriello

Department of Electronic, Information and Bioengineering
Politecnico di Milano, Italy

International Conference on Machine Learning, 2013
Atlanta, Georgia

June 17, 2013
Approximate Policy Iteration (API)

Requirements:
- Transition samples, typically in the form: \( DS = \left\{ s_i, a_i, Q_i^\pi \right\}_{i=1}^{N} \)
- Approximate representation of the policy

Approximate value prediction problem \( \hat{Q} \)

Policy evaluation

Policy improvement

Search in a restricted policy space
Approximate Policy Iteration (API)

Policy evaluation

Approximate value prediction problem $\hat{Q}$

Requirements:

- Transition samples, typically in the form: $DS = \left\{ s_i, a_i, Q_i^\pi \right\}_{i=1}^N$
- Approximate representation of the policy

Question: What happens to the guarantees of PI approaches?
Lost of convergence guarantees

**Policy oscillation:** process is trapped in repeating the same sequence of policies $\pi_i$. The learning process may degrade and diverge. [Bertsekas, 2010]
Question: How can we overcome the policy oscillation problem?
Question: How can we overcome the policy oscillation problem?

1. Bound the policy performance of a generic policy $\pi'$ w.r.t. an other policy $\pi$.
2. Maximize the bound on the policy performance improvement at each iteration.

Related works:

- **DPI** [Lagoudakis and Parr, 2003b, Fern et al., 2006, Lazaric et al., 2010];
- **CPI** [Kakade and Langford, 2002, Kakade, 2003].
Outline

1. Fundamentals
2. Bound Derivation
3. Algorithms
4. Experiments
5. Conclusions
Investigation of API issue

Lemma

For any stationary policies $\pi$ and $\pi'$ of an infinite horizon Markov Decision Process (MDP) $M$ and any starting state distribution $\mu$:

$$
J_{\mu}^{\pi'} - J_{\mu}^{\pi} = d_{\mu}^{\pi'} T_{\pi'} A_{\pi'}^{\pi'}.
$$

The search of optimal policy $\pi^*$ can become a maximization problem if $d_{\mu}^{\pi'}$ is known for all $\pi'$.

Issue: API exploits the distribution of the greedy policy to compute the improvement

- critical because $d_{\mu}^{\pi^+} \neq d_{\mu}^{\pi'}$
Question: is it possible to bound the distance between $d^\pi_{\mu}$ and $d^{\pi'}_{\mu}$ knowing a similarity measure between the policies?
Question: is it possible to bound the distance between $d_{\mu}^\pi$ and $d_{\mu}^{\pi'}$ knowing a similarity measure between the policies?

Answer: We have been able to provide a bound on the $L_1$–norm of the difference of the $\gamma$–discounted future state distributions under initial state distribution $\mu$:

$$\left\| d_{\mu}^{\pi'} - d_{\mu}^\pi \right\|_1 \leq \frac{\gamma}{(1 - \gamma)^2} \left\| \pi' - \pi \right\|_{\infty} .$$

Results: Similar policies have similar future state distributions.
Question: How is it possible to exploit previous relationships?

\[ J_{\mu}^{\pi'} - J_{\mu}^{\pi} = d_{\mu}^{\pi'} \mathbf{A}_{\pi'} \]
Question: How is it possible to exploit previous relationships?

\[
J_{\pi'}^\mu - J_\mu^\pi = d_\mu^{\pi'} A_\pi^{\pi'}
\]

exact improvement

\[
= d_\mu^{\pi'} A_\pi^{\pi'} \pm d_\mu^{\pi b} A_\pi^{\pi'}
\]

\[
= d_\mu^{\pi b} A_\pi^{\pi'} + \left( d_\mu^{\pi'} - d_\mu^{\pi b} \right) A_\pi^{\pi'}
\]

\[\|d_\mu^{\pi'} - d_\mu^{\pi b}\|_2 \]

[Haviv and Heyden, 1984]
Lower Bound on Policy Performance

**Question:** How is it possible to exploit previous relationships?

\[
J_{\pi}^{\prime} - J_{\mu}^{\pi} = d_{\mu}^{\pi} T A_{\pi}^{\pi} \\
= d_{\mu}^{\pi} T A_{\pi}^{\pi} \pm d_{b}^{\pi} T A_{\pi}^{\pi} \\
= d_{\mu}^{\pi} T A_{\pi}^{\pi} + \left( d_{\mu}^{\pi} T - d_{\mu}^{\pi} b T \right) A_{\pi}^{\pi} \\
\geq d_{\mu}^{\pi} b T A_{\pi}^{\pi} - \left\| d_{\mu}^{\pi} - d_{\mu}^{\pi} b \right\|_1 \frac{\Delta A_{\pi}^{\pi'}}{2} \\
\text{[Haviv and Heyden, 1984]}
\]
Lower Bound on Policy Performance

Theorem

For any stationary policies $\pi$ and $\pi'$ and any starting state distribution $\mu$, given any baseline policy $\pi_b$, the difference between the performance of $\pi'$ and the one of $\pi$ can be lower bounded as follows:

$$
\begin{align*}
J^{\pi'}_{\mu} - J^{\pi}_{\mu} &\geq d_{\mu}^{T} A_{\pi}^{\pi'} - \frac{\gamma}{(1 - \gamma)^{2}} \left\| \pi' - \pi_b \right\|_{\infty} \frac{\Delta A_{\pi}^{\pi'}}{2} \\
\end{align*}
$$

where $\Delta A_{\pi}^{\pi'} = \max_{i, j \in \{1, 2, \ldots, |S|\}} |A_{\pi}^{\pi'}(i) - A_{\pi}^{\pi'}(j)|$.

Denotes a trade–off between

- high improvement (positive contribution)
- small policy difference (negative contribution)
Outline

1. Fundamentals
2. Bound Derivation
3. Algorithms
4. Experiments
5. Conclusions
Unique–parameter Safe Policy Iteration (USPI)

**Convex update rule**

\[ \pi' = \alpha \pi + (1 - \alpha) \overline{\pi}, \]

- \( \alpha \) is a trade-off between a full update toward \( \overline{\pi} \) (\( \alpha = 1 \)) and a conservative update in favor of current policy \( \pi \) (\( \alpha = 0 \))
- \( \| \pi' - \pi \|_\infty = \alpha \| \overline{\pi} - \pi \|_\infty \)
- \( \overline{\pi} \) is the greedy policy
- \( \pi_b = \pi \)

**Answer:** exploit the bound to select the \( \alpha \) that guarantees the maximum policy improvement

\[ \alpha = \arg \max_{t \in [0,1]} J^{\pi'}_{\mu}(t) - J^{\pi}_{\mu} \]
\[ J_{\pi'}^\mu - J_{\pi}^\mu \geq \alpha d_{\mu}^{\pi} A_{\pi}^{\pi} - \alpha^2 \frac{\gamma}{(1 - \gamma)^2} \| \pi - \bar{\pi} \|_{\infty} \Delta A_{\pi}^{\pi}. \]
Guaranteed improvement:

\[ J_{\mu}^{\pi'} - J_{\mu}^{\pi} \geq \alpha \mathbf{d}_{\mu}^{\pi} \mathbf{A}^{\pi} - \alpha^2 \frac{\gamma}{(1 - \gamma)^2} \|\pi - \pi\|_{\infty} \frac{\Delta \mathbf{A}_{\pi}^{\pi}}{2}. \]
Multiple–parameter Safe Policy Iteration (MSPI)

- Convex combination update rule with multiple parameters

\[
\pi'(a|s) = \alpha(s) \pi(a|s) + (1 - \alpha(s)) \pi(a|s), \quad \forall s \in S, \forall a \in A
\]

where

\[
\alpha(s) \in [0, 1] \quad \forall s \in S
\]

- Simplified bound

\[
J_{\mu}' - J_{\mu} \geq d_{\mu}^T A_{\pi}' - \frac{\gamma}{(1 - \gamma)^2} \|\pi' - \pi\|^2_{\infty} \frac{\|q^\pi\|_{\infty}}{2}
\]

- No close form solution for the bound, iterative approach with complexity \(O(S \log S)\)
\[
J'_{\mu} - J_{\mu} \geq \mathbb{E}_{s \sim d_{\mu}^\pi(\cdot)} \left[ \alpha(s) A_{\pi}^\pi(s) \right] - \frac{\gamma}{(1 - \gamma)^2} \max_{s \in \mathcal{S}} \left( \alpha(s) \| \pi(\cdot | s) - \pi(\cdot | s) \|_1 \right)^2 \frac{\| q^\pi \|_{\infty}}{2}
\]

- If \( A_{\pi}^\pi(s) \leq 0 \) then \( \alpha(s) = 0 \)
- If \( A_{\pi}^\pi(s) > 0 \), define \( \mathcal{S}_{\pi}^\pi = \{ s \mid A_{\pi}^\pi(s) > 0 \} \). Then

\[
\Lambda = \max_{s \in \mathcal{S}_{\pi}^\pi} \left( \alpha(s) \| \pi(\cdot | s) - \pi(\cdot | s) \|_1 \right)
\]

As a consequence we can impose

\[
\alpha(s) \leq \min \left( 1, \frac{\Lambda}{\| \pi(\cdot | s) - \pi(\cdot | s) \|_1} \right)
\]
\[ J_{\mu'}^\pi - J_{\mu}^\pi \geq \mathbb{E}_{s \sim d_{\mu}^\pi(\cdot)} \left[ \alpha(s)A_{\pi}^\pi(s) \right] - \frac{\gamma}{(1 - \gamma)^2} \max_{s \in S} \left( \alpha(s) \| \pi(\cdot | s) - \pi(\cdot | s) \|_1 \right)^2 \frac{\| q^\pi \|_\infty}{2} \]

- If \( A_{\pi}^\pi(s) \leq 0 \) then \( \alpha(s) = 0 \)
- If \( A_{\pi}^\pi(s) > 0 \), define \( S_{\pi}^\pi = \{ s | A_{\pi}^\pi(s) > 0 \} \). Then

\[ \Lambda = \max_{s \in S_{\pi}^\pi} \left( \alpha(s) \| \pi(\cdot | s) - \pi(\cdot | s) \|_1 \right) \]

As a consequence we can impose

\[ \alpha(s) \leq \min \left( 1, \frac{\Lambda}{\| \pi(\cdot | s) - \pi(\cdot | s) \|_1} \right) \]

**Question:** How is possible to compute \( \Lambda \)?
Multiple–parameter Safe Policy Iteration (MSPI)

\[ B(\Lambda) = \sum_{s \in S_\pi} \min \left( 1, \frac{\Lambda}{\| \pi(\cdot|s) - \pi(\cdot|s) \|_1} \right) d^\pi_\mu(s) A^\pi_\mu(s) - \Lambda^2 \frac{\gamma}{(1 - \gamma)^2} \frac{\| q^\pi \|_\infty}{2} \]

Magnitude of the bound

Magnitude of the gradient

\[ \Lambda_1 \quad \Lambda_2 \quad \Lambda^* \quad \Lambda_3 \quad \Lambda_4 \quad \Lambda_5 \quad \Lambda_6 \]
Multiple–parameter Safe Policy Iteration (MSPI)

\[ B(\Lambda) = \sum_{s \in S_\pi^\pi} \min\left(1, \frac{\Lambda}{\|\pi(\cdot|s) - \pi(\cdot|s)\|_1}\right) d_\mu^\pi(s)A_\pi^\pi(s) - \Lambda^2 \frac{\gamma}{(1 - \gamma)^2} \frac{\|q_\pi\|_\infty}{2} \]
Considerations

Algorithm properties

1. The policy performance is improved at every iteration;
2. Stopping condition can be directly based on the performance of the policy $\pi$ w.r.t. the optimal $\pi^*$;
3. Incorporate exploration into the algorithm;
4. Convergence to the optimal policy.
Approximate Scenario

**Question:** When the model is unavailable?

Approximate computation of the involved quantities, we want to achieve with a certain probability $1 - \delta$ (statistical framework)

- guarantees on the terminal policy
- finite sample bound.

**Result**

With probability $1 - \delta$, SPI terminates with a policy $\pi$ such that

$$\forall \pi^+ \in \Pi, \quad A_{\pi^+} \leq \epsilon.$$ 

where $\Pi$ is a free to be designed policy space.

The condition can be translated into a bound on the policy performance of $\pi$ w.r.t. $\pi^*$:

$$J^{\pi^*} - J^\pi \leq K(\epsilon)$$
Finite sample complexity

\[ DS = \left\{ s_i, a_i, Q_i^\pi(s, a) \right\}_{i=1}^{N(\Pi, \mu, \epsilon, \delta)} \]

Affect the performance bound

Transition number

\[ N_{tot} = N_{iterations} \cdot N_{samples \ per \ iteration} \cdot N_{transitions \ per \ sample} \]

\[ N_{iterations} = O \left( \frac{1}{(1 - \gamma)^2 \epsilon^2} \right) \]

\[ N_{samples \ per \ iteration} = O \left( \frac{|A|^2}{\epsilon^2} \left( \log(2|\Pi|) + \log \frac{1}{(1 - \gamma)^2 \epsilon^2 \delta} \right) \right) \]

\[ N_{transitions \ per \ sample} = \log_{\gamma} \left( \frac{\epsilon}{2} \right) \]
Outline

1. Fundamentals
2. Bound Derivation
3. Algorithms
4. Experiments
5. Conclusions
**Approximate setting – Chain Walk**

**Chain Walk Domain** [Lagoudakis and Parr, 2003a]

Approximate policy evaluation through linear architecture

---

**Graph:**
- **x-axis:** Iterations
- **y-axis:** Performance
- **Legend:**
  - PI
  - CPI
  - MSPI
  - USPI

---

 Pirotta, Restelli, Pecorino, Calandriello – Safe Policy Iteration

June 17, 2013
Approximate setting – BlackJack

BlackJack [Dutech et al., 2005]

Approximate value prediction and limited policy space

---

Performance vs. Iterations

- aMSPI
- aUSPI
- aCPI
- aPI

π₇
πₛ
Conclusions:

- Significant theoretical contributes
- Safe approaches design
- Policy oscillation prevention
- Strong performance guarantees
- Instructional experiments development

Future works:

- Theoretical refinements
- Experiments extension
- Model-free approach development
Thank you for your attention


On the sample complexity of reinforcement learning.

Approximately optimal approximate reinforcement learning.

Least-squares policy iteration.

Reinforcement learning as classification: Leveraging modern classifiers.


Analysis of a classification-based policy iteration algorithm.