**Problem**

- Classical API approaches may generate a policy \( \pi_{n+1} \) that performs worse than the previous policy \( \pi_n \).
- This undesired improvement may lead to the policy oscillation phenomenon that can prevent convergence to the optimal policy and degrade the learning process.
- Our “safe” approach tries to overcome this issue by visiting a sequence of policies with monotonic improving performance. Following this approach, the policy is constrained to improve overtime and, as a consequence, the degradation of the policy performance between consecutive iteration is prevented.

**Contributions**

1. **Theoretical contribution.** We introduce a new, more general lower bound to the policy improvement of an arbitrary policy compared to another policy based on the ability to bound the distance between the future state distributions.
2. **Algorithmic contribution.** We define two approximate policy–iteration algorithms whose policy improvement moves toward the estimated greedy policy by maximizing the policy improvement bounds.
3. **Empirical contribution.** We report results on a simple chain walk and Blackjack domains that confirm the main theoretical findings.

**Theoretical Bounds**

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

\[
J_\mu' - J_\mu = d_\mu^{-1/2} A_{\pi'}
\]

The search of optimal policy \( \pi^* \) can become a maximization problem if \( d_\mu' \) is known for all \( \pi' \). API approaches do not have access to \( d_\mu' \) but exploit the distribution of the greedy policy: \( d_\mu' \neq d_\mu' \).

Given two stationary policies \( \pi \) and \( \pi' \) for an infinite horizon MDP \( M \), the \( L_1 \)-norm of their \( \gamma \)-discounted distribution can be upper bounded as:

\[
\|d_{\pi'} - d_{\pi}\|_1 \leq \frac{\gamma}{(1 - \gamma)^2} \|\pi' - \pi\|_\infty,
\]

i.e., similar policies have similar future state distribution.

Merging previous results is possible to derive a lower bound on policy performance

\[
J_\mu' - J_\mu \geq \min_{\pi, \pi'} \|A_{\pi'}(i) - A_{\pi}(i)\|
\]

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

**Finite Sample Analysis**

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

\[
\forall n \in \mathbb{N}, \quad d_\mu'^{1/2} A_{\pi} \leq \epsilon,
\]

where \( \epsilon \) is a free to be designed policy space.

The number of transitions required to obtain the \( \epsilon \)-accurate estimate is

\[
N_{\text{itr}} = \frac{1}{\epsilon} \left( \log(2\pi) + \log \left( \frac{1}{\epsilon(1 - \gamma)^2} \right) \right)
\]

If the same target policy \( \pi \) is used at each iteration, the approximate version of the algorithms terminate after \( O \left( \frac{1}{\epsilon^2(1 - \gamma)^2} \right) \).

**Results**

The Chain–Walk domain

State: \( S = 1, \ldots, 4 \)

Actions: \( A = \{\text{left}, \text{right}\} \)

Dynamics: for action \( \text{right} \):

\[
x_{t+1} = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{otherwise}
\end{cases}
\]

- Approximate policy evaluation;
- Linear architecture.

**Unique–parameter Safe Policy Improvement (USPI)**

**Multiple–parameter Safe Policy Improvement (MSPI)**

Convex update rule: state–dependent parameters

\[
J_\mu' - J_\mu \geq \alpha A_{\pi'} \cdot \sqrt{\frac{\gamma}{(1 - \gamma)^2} \|\pi' - \pi\|_\infty} \Delta A_{\pi'}^2.
\]

Guaranteed Improvement with \( \alpha = \min \left( 1, \frac{1 - \gamma}{\sqrt{\pi' - \pi}} \right) \).

Convex update rule: state–dependent parameters

\[
J_\mu' - J_\mu \geq E_{\pi'}(s) \left[ A_{\pi'}(s) \right] - \frac{\gamma}{(1 - \gamma)^2} \max_{i \in S} \left( \pi(s) - \pi(s) \right)^2 \|q_i\|_\infty \cdot \Delta A_{\pi'}^2.
\]

No closed–form solution but an iterative solution in order to find the optimal value \( \pi' \) that maximize

\[
B(A) = \sum \min \left( 1, \frac{A}{\pi(s) - \pi(s)} \right) d_{\pi'}(A) \Delta A_{\pi'}^2.
\]