Safe Policy Iteration – Supplementary Material

Matteo Pirotta MATTEO.PIROTTA@POLIMI.IT
Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Alessio Pecorino ALESSIO.PECORINO@MAIL.POLIMI.IT
Daniele Calandriello DANIELE.CALANDRIELLO@MAIL.POLIMI.IT
Dept. Elect., Inf., and Bioeng., Politecnico di Milano, piazza Leonardo da Vinci 32, I-20133, Milan, ITALY

Abstract

This document provides additional material to the main paper. In particular, it provides: 1) the complete set of theorems, lemmas and corollaries with the relative proofs; 2) additional experiment in chain walk and BlackJack domains; 3) a detailed analysis of the performances in terms of computational time.

1. Proofs

In this section, we will prove the lemmas, theorems, and corollaries stated in our paper.

Lemma 3.1 Let π and π' be two stationary policies for an infinite horizon MDP M with state transition matrix $\mathbf{P}$. The L_1–norm of the difference between their γ–discounted future state distributions under starting state distribution μ can be upper bounded as follows:

$$\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi}\right\|_1 \leq \frac{\gamma}{1-\gamma}\left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty.$$

Proof To prove the lemma we rewrite the difference $\mathbf{d}_\mu^{\pi'^{\mathsf{T}}} - \mathbf{d}_\mu^{\pi^{\mathsf{T}}}$ as follows:

$$\begin{aligned}\mathbf{d}_\mu^{\pi'^{\mathsf{T}}} - \mathbf{d}_\mu^{\pi^{\mathsf{T}}} &= \gamma\mathbf{d}_\mu^{\pi'^{\mathsf{T}}}\mathbf{P}^{\pi'} - \gamma\mathbf{d}_\mu^{\pi^{\mathsf{T}}}\mathbf{P}^{\pi} \\ &= \gamma\left(\mathbf{d}_\mu^{\pi'^{\mathsf{T}}} - \mathbf{d}_\mu^{\pi^{\mathsf{T}}}\right)\mathbf{P}^{\pi'} + \gamma\mathbf{d}_\mu^{\pi^{\mathsf{T}}}\left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right) \\ &= \gamma\mathbf{d}_\mu^{\pi^{\mathsf{T}}}\left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right)\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1},\end{aligned}$$

where the last equality follows from the convergence of Neumann series. It is worth to notice that the inverse of matrix $I - \gamma\mathbf{P}^{\pi'}$ exists for any $\gamma < 1$. From this equation, the bound on the L_1–norm follows:

$$\begin{aligned}\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi}\right\|_1 &= \left\|\mathbf{d}_\mu^{\pi'^{\mathsf{T}}} - \mathbf{d}_\mu^{\pi^{\mathsf{T}}}\right\|_\infty \\ &\leq \gamma\left\|\mathbf{d}_\mu^{\pi^{\mathsf{T}}}\right\|_\infty \left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty \\ &= \frac{\gamma}{1-\gamma}\left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty\end{aligned}$$

□

Remark 1 (Undiscounted case). When the MDP is undiscounted ($\gamma = 1$), the γ-discounted future state distributions are replaced by the steady–state distributions d^π and $d^{\pi'}$. The L_1–norm bound between these distributions can be derived in a similar way, but the matrix $\mathbf{I} - \gamma\mathbf{P}^{\pi'}$ is singular when $\gamma = 1$. To overcome this problem, $(\mathbf{I} - \mathbf{P}^{\pi'})^{-1}$ can be replaced by the fundamental matrix of the corresponding Markov chain: $(\mathbf{I} - \mathbf{P}^{\pi'} + \overline{\mathbf{P}^{\pi'}})^{-1}$, where $\overline{\mathbf{P}^{\pi'}} = \mathbf{e}\cdot\mathbf{d}^{\pi'^{\mathsf{T}}}$ is the long-term limiting matrix. An overview of bounds for $\left\|\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}\right\|$ can be found in (**?**).

Corollary 3.2 Let π and π' two stationary policies for an infinite horizon MDP M. The L_1–norm of the difference between their γ–discounted future state distributions under starting state distribution μ can be upper bounded as follows:

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

Proof This Corollary follows from Lemma 3.1, from Neumann series expansion of the inverse and from $\|\mathbf{P}\|_\infty = 1$.

$$\begin{aligned}\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi}\right\|_1 &\leq \frac{\gamma}{1-\gamma}\left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty \\ &\leq \frac{\gamma}{1-\gamma}\left\|\mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi}\right\|_\infty \|\mathbf{P}\|_\infty \sum_{t=0}^{\infty}\gamma^t\left\|\mathbf{P}^{\pi'}\right\|_\infty^t \\ &= \frac{\gamma}{(1-\gamma)^2}\left\|\mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi}\right\|_\infty\end{aligned}$$

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

□

Lemma 3.3 (**?**)
For any stationary policies π and π' and any starting state distribution μ:

$$J_\mu^{\pi'} - J_\mu^{\pi} = \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{A}_\pi^{\pi'}.$$

Proof

$$\begin{aligned} J_\mu^{\pi'} &= \boldsymbol{\mu}^{\mathrm{T}} \mathbf{v}^{\pi'} = \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{r}^{\pi'} \\ &= \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{r}^{\pi'} + \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{v}^{\pi} - \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{v}^{\pi} \\ &= \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{r}^{\pi'} + \left(\boldsymbol{\mu}^{\mathrm{T}} + \gamma \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{P}^{\pi'}\right) \mathbf{v}^{\pi} - \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{v}^{\pi} \\ &= \mathbf{d}_\mu^{\pi'\mathrm{T}} \left(\mathbf{r}^{\pi'} + \gamma \mathbf{P}^{\pi'} \mathbf{v}^{\pi} - \mathbf{v}^{\pi}\right) + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{v}^{\pi} \\ &= \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{A}_\pi^{\pi'} + J_\mu^{\pi} \end{aligned}$$

□

Theorem 3.5 For any stationary policies π and π' and any starting state distribution μ, given any baseline policy π_b, the difference between the performance of π' and the one of π can be lower bounded as follows:

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \mathbf{d}_\mu^{\pi_b\mathrm{T}} \mathbf{A}_\pi^{\pi'} - \frac{\gamma}{(1-\gamma)^2} \left\| \boldsymbol{\Pi}^{\pi'} - \boldsymbol{\Pi}^{\pi_b} \right\|_\infty \frac{\Delta \mathbf{A}_\pi^{\pi'}}{2}.$$

Proof The proof can be easily obtained from Lemma 3.1:

$$\begin{aligned} J_\mu^{\pi'} - J_\mu^{\pi} &= \mathbf{d}_\mu^{\pi'\mathrm{T}} \mathbf{A}_\pi^{\pi'} \\ &= \mathbf{d}_\mu^{\pi_b\mathrm{T}} \mathbf{A}_\pi^{\pi'} + \left(\mathbf{d}_\mu^{\pi'\mathrm{T}} - \mathbf{d}_\mu^{\pi_b\mathrm{T}}\right) \mathbf{A}_\pi^{\pi'} \\ &\geq \mathbf{d}_\mu^{\pi_b\mathrm{T}} \mathbf{A}_\pi^{\pi'} - \left\| \mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi_b} \right\|_1 \frac{\Delta \mathbf{A}_\pi^{\pi'}}{2}, \end{aligned}$$

where the last inequality follows from Lemma 3.4 since $\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi_b}$ is a zero–mean vector. The theorem is proved by replacing $\left\| \mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi_b} \right\|_1$ with the bound in Corollary 3.2. □

Corollary 3.6 For any stationary policies π and π' and any starting state distribution μ, the difference between the performance of π' and the one of π can be lower bounded as follows:

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \mathbf{d}_\mu^{\pi\mathrm{T}} \mathbf{A}_\pi^{\pi'} - \frac{\gamma}{(1-\gamma)^2} \left\| \boldsymbol{\Pi}^{\pi'} - \boldsymbol{\Pi}^{\pi} \right\|_\infty^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2}.$$

Proof The proof comes from a lower bound to the bound in Theorem 1 when $\pi_b = \pi$. Such lower bound involves the upper bound of $\frac{\Delta \mathbf{A}_\pi^{\pi'}}{2}$:

$$\begin{aligned} \frac{\Delta \mathbf{A}_\pi^{\pi'}}{2} &\leq \left\| \mathbf{A}_\pi^{\pi'} \right\|_\infty = \left\| \left(\boldsymbol{\Pi}^{\pi'} - \boldsymbol{\Pi}^{\pi}\right) Q^\pi \right\|_\infty \\ &= \max_s \left((\pi'(\cdot|s) - \pi(\cdot|s)) \cdot Q^\pi(s,\cdot) \right) \\ &\leq \max_s \left(\left\| \pi'(\cdot|s) - \pi(\cdot|s) \right\|_1 \frac{\Delta Q^\pi(s,\cdot)}{2} \right) \\ &\leq \left\| \boldsymbol{\Pi}^{\pi'} - \boldsymbol{\Pi}^{\pi} \right\|_\infty \frac{\|\mathbf{q}^\pi\|_\infty}{2} \end{aligned}$$

□

Corollary 4.1 If $\mathbb{A}_{\pi,\mu}^{\overline{\pi}} \geq 0$, then, using $\alpha^* = \frac{(1-\gamma)^2 \mathbb{A}_{\pi,\mu}^{\overline{\pi}}}{\gamma \|\boldsymbol{\Pi}^{\overline{\pi}} - \boldsymbol{\Pi}^{\pi}\|_\infty \Delta \mathbf{A}_\pi^{\overline{\pi}}}$, we set $\alpha = \min(1, \alpha^*)$, so that when $\alpha^* \leq 1$ we can guarantee the following policy improvement:

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \frac{(1-\gamma)^2 {\mathbb{A}_{\pi,\mu}^{\overline{\pi}}}^2}{2\gamma \left\| \boldsymbol{\Pi}^{\overline{\pi}} - \boldsymbol{\Pi}^{\pi} \right\|_\infty \Delta \mathbf{A}_\pi^{\overline{\pi}}},$$

and when $\alpha^* > 1$, we perform a full update towards the target policy $\overline{\pi}$ with a policy improvement equal to the one specified in Theorem 1.

Proof Setting $\pi_b = \pi$ and $\pi' = \alpha\overline{\pi} + (1-\alpha)\pi$, we can rewrite the bound in Theorem 1 as:

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \alpha \mathbb{A}_{\pi,\mu}^{\overline{\pi}} - \alpha^2 \frac{\gamma}{(1-\gamma)^2} \left\| \boldsymbol{\Pi}^{\overline{\pi}} - \boldsymbol{\Pi}^{\pi} \right\|_\infty \frac{\Delta \mathbf{A}_\pi^{\overline{\pi}}}{2}.$$

α^* is the value of α that maximizes this bound. □

Corollary 4.2 Let $\mathcal{S}_\pi^{\overline{\pi}}$ be the subset of states where the advantage of policy $\overline{\pi}$ over policy π is positive: $\mathcal{S}_\pi^{\overline{\pi}} = \{s \in \mathcal{S} | A_\pi^{\overline{\pi}}(s) > 0\}$.
The bound in Corollary 3.6 is optimized by taking $\alpha(s) = 0$, $\forall s \notin \mathcal{S}_\pi^{\overline{\pi}}$ and $\alpha(s) = \min\left(1, \frac{\overline{\Upsilon}^*}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right)$, $\forall s \in \mathcal{S}_\pi^{\overline{\pi}}$, where $\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 = \sum_{a \in \mathcal{A}} |\overline{\pi}(a|s) - \pi(a|s)|$ and $\overline{\Upsilon}^*$ is the value that maximizes the following function:

$$B(\overline{\Upsilon}) = \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \min\left(1, \frac{\overline{\Upsilon}}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right) d_\mu^\pi \mathbf{A}_\pi^{\overline{\pi}} - \overline{\Upsilon}^2 \frac{\gamma}{(1-\gamma)^2} \frac{\|\mathbf{q}^\pi\|_\infty}{2}$$

Proof Given a state s with negative advantage, the larger is $\alpha(s)$ the lower will be the bound on the policy improvement as stated in Corollary 3.6, so the optimal choice for these states is to set $\alpha(s) = 0$. Given the set of states with positive advantages $\mathcal{S}_\pi^{\overline{\pi}}$, we introduce $\overline{\Upsilon} = \max_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \alpha(s) \|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1$. For all the states in $\mathcal{S}_\pi^{\overline{\pi}}$, the first term of the bound from Corollary 3.6 would be maximized by setting $\alpha(s) = 1$. On the other hand, from the definition of $\overline{\Upsilon}$, we have the following constraint: $\alpha(s) \leq \frac{\overline{\Upsilon}}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}$. So, given

a value of $\overline{\Upsilon}$ the optimal value for the coefficient of any state $s \in \mathcal{S}_\pi^{\overline{\pi}}$ is $\min\left(1, \frac{\overline{\Upsilon}}{\|\overline{\pi}(\cdot|s)-\pi(\cdot|s)\|_1}\right)$. Function $B(\overline{\Upsilon})$ is obtained by using the previous definitions in the bound from Corollary 3.6. As a result, the optimization of the bound over the set of $|\mathcal{S}|$ coefficients $\alpha(s)$ has been translated into the maximization of the univariate function $B(\overline{\Upsilon})$. □

Theorem 5.1 If the same target policy $\overline{\pi}$ is used at each iteration, *aUSPI* and *aMSPI* terminates after $O\left(\frac{1}{(1-\gamma)^2\epsilon}\right)$.

Proof At each iteration of aUSPI $\hat{\mathbb{A}}_{\pi,\mu}^{\overline{\pi}} > \frac{2\epsilon}{3(1-\gamma)}$, that (since $\hat{\mathbb{A}}_{\pi,\mu}^{\overline{\pi}}$ is an $\frac{\epsilon}{3(1-\gamma)}$–accurate estimate of $\mathbb{A}_\pi^{\overline{\pi}}$) implies that $\mathbb{A}_{\pi,\mu}^{\overline{\pi}} > \frac{\epsilon}{3(1-\gamma)}$. From Corollary 4.1, it is ease to derive a lower bound to the policy improvement for the i–th iteration of the aUSPI algorithm:

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \frac{(1-\gamma)\epsilon^2}{18\gamma \left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty^2}.$$

Since policy performances range in the interval $[0, \frac{1}{1-\gamma}]$, an upper bound to the number N of iterations of aUSPI can be computed from the following inequality:

$$\frac{(1-\gamma)\epsilon^2}{18\gamma} \sum_{i=0}^{N} \frac{1}{\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty^2} \leq \frac{1}{1-\gamma} \qquad (1)$$

To solve such inequality for N, we need to analyze the value of $\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty$, that in aUSPI can be rewritten as follows:

$$\begin{aligned}
&\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty \\
&\quad = \left\|\mathbf{\Pi}^{\overline{\pi}} - \left(\alpha_{i-1}\mathbf{\Pi}^{\overline{\pi}} + (1-\alpha_{i-1})\mathbf{\Pi}^{\pi_{i-1}}\right)\right\|_\infty \\
&\quad = (1-\alpha_{i-1})\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_{i-1}}\right\|_\infty,
\end{aligned}$$

where $\alpha_{i-1} = \frac{(1-\gamma)^2\epsilon}{3\gamma\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_{i-1}}\|_\infty^2}$. By replacing the value of α_{i-1} in the previous equation, we get the following recursive equation:

$$\begin{aligned}
\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty &= \left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_{i-1}}\right\|_\infty \\
&\quad - \frac{(1-\gamma)^2\epsilon}{3\gamma\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_{i-1}}\right\|_\infty},
\end{aligned}$$

where we pessimistically assume that $\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_0}\right\|_\infty = 2$. Unfortunately, the above equation does not have closed-form solution. However, we can consider the following upper bound:

$$\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty \leq 2 - \frac{(1-\gamma)^2\epsilon}{6\gamma}(i-1).$$

Such upper bound allows us to lower bound the summation in inequality 1:

$$\begin{aligned}
\sum_{i=0}^{N} \frac{1}{\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi_i}\right\|_\infty^2} &\geq \int_0^N \frac{1}{(2 - \frac{(1-\gamma)^2\epsilon}{6\gamma}(x-1))^2} dx \\
&= \frac{36\gamma^2(1-\gamma)N}{(12\gamma + \epsilon(1-\gamma)^2)(12\gamma + (1-N)\epsilon(1-\gamma)^2)}.
\end{aligned}$$

Replacing the above expression in inequality 1 and solving for N we get:

$$N \leq \frac{(12\gamma + (1-\gamma)^2\epsilon)^2}{(1-\gamma)^2\epsilon(12\gamma + (1-\gamma^2)\epsilon)} = O\left(\frac{1}{(1-\gamma)^2\epsilon}\right).$$

The proof is completed by observing that aMSPI produces improvements that are never worse than aUSPI. □

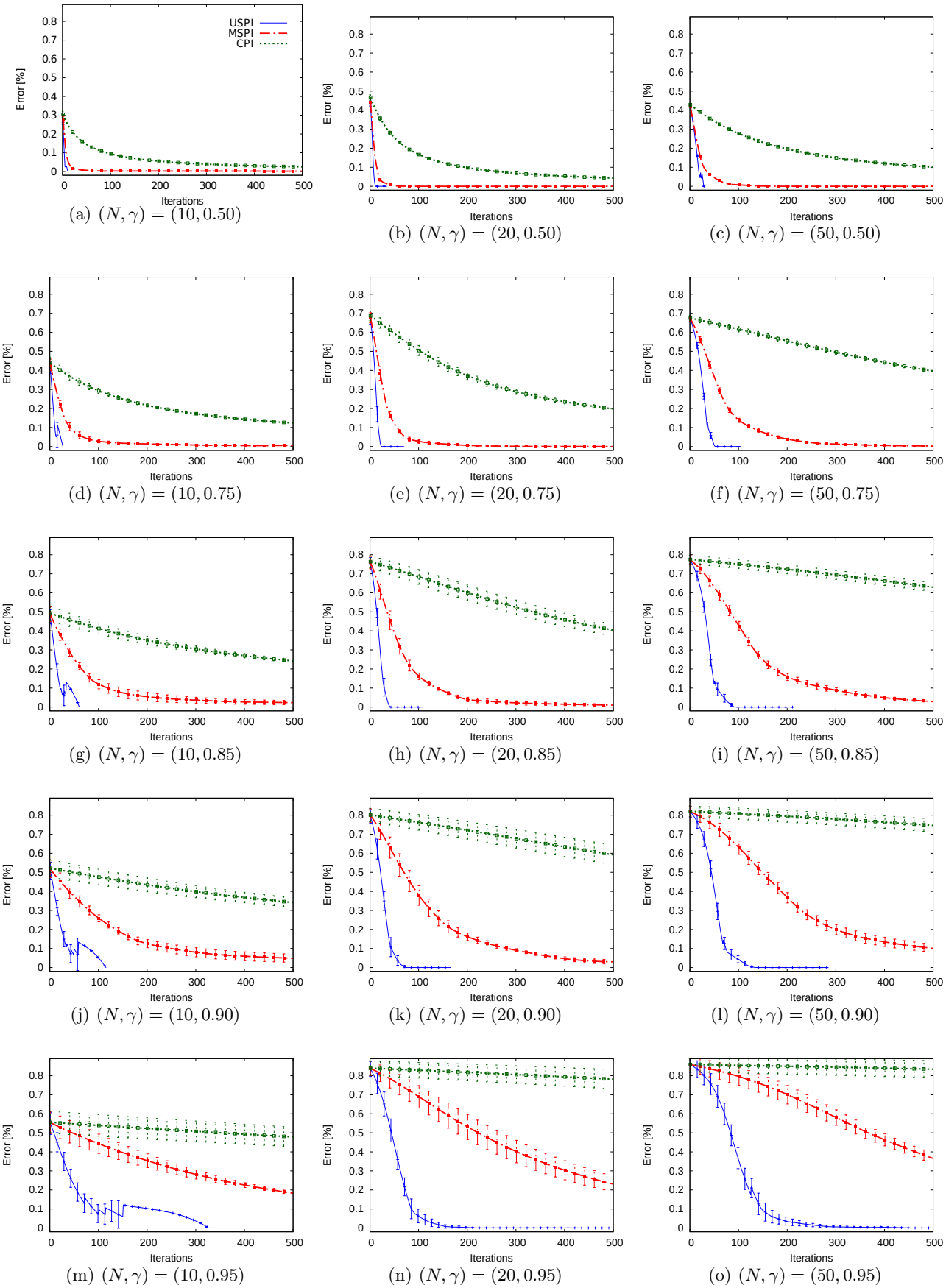
2. Additional Experiments

2.1. Chain Walk Domain

It is interesting to compare the performance of the different algorithms using as benchmark the environment defined by (**?**). Such MDP is defined as chain walk domain which is modeled as a N-state chain (numbered from 1 to N). Chain is traversed performing two actions, "left" (L) and "right" (R). Each action induces a transition into the associated direction and to the opposite one with probability p and $1-p$ (in this experiments p is set to 0.9). Reward +1 is assigned only when the agent enters one of the two states located at a distance of $N/4$ from the boundaries, otherwise the reward is 0. The starting state distribution D is assumed uniform over state space in any configuration.

2.1.1. Exact Settings

In this model-based domain, we have analyzed the performance of the proposed algorithms w.r.t. the CPI approach. The analysis is performed for different state space dimensions and for different values of the discount factor γ. Algorithms are tested over multiple runs, in particular 10 runs are performed starting from random policies. Figure 1 shows the behavior of the algorithms in term of distance between the performance of the policy at iteration i and the optimal performance. It can be seen that the CPI is always outperformed by the MSPI and USPI. At the same time the USPI achieves a significant higher learning behavior than MSPI, that leads to faster convergence to the optimal performance.



(a) $(N, \gamma) = (10, 0.50)$ (b) $(N, \gamma) = (20, 0.50)$ (c) $(N, \gamma) = (50, 0.50)$

(d) $(N, \gamma) = (10, 0.75)$ (e) $(N, \gamma) = (20, 0.75)$ (f) $(N, \gamma) = (50, 0.75)$

(g) $(N, \gamma) = (10, 0.85)$ (h) $(N, \gamma) = (20, 0.85)$ (i) $(N, \gamma) = (50, 0.85)$

(j) $(N, \gamma) = (10, 0.90)$ (k) $(N, \gamma) = (20, 0.90)$ (l) $(N, \gamma) = (50, 0.90)$

(m) $(N, \gamma) = (10, 0.95)$ (n) $(N, \gamma) = (20, 0.95)$ (o) $(N, \gamma) = (50, 0.95)$

Figure 1. Error trend of policy $\mathbf{\Pi}_i^\pi$ w.r.t. the optimal performance $J_\mu^{\pi^*}$ in different N-states chain walk domains. 99% confidence interval bars are shown.

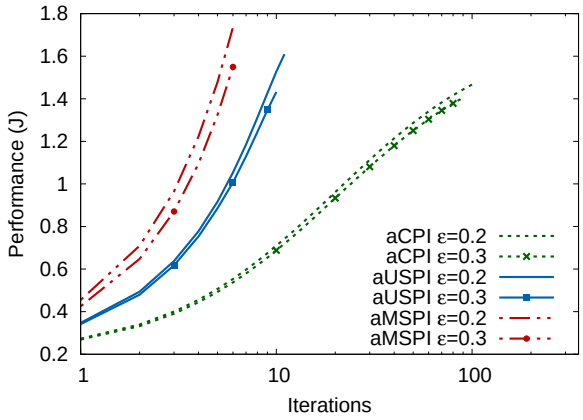


Figure 2. Approximate policy performance J_μ^π of aCPI, aMSPI and aUSPI in a 4-states chain walks with ϵ equals to 0.2 and 0.3.

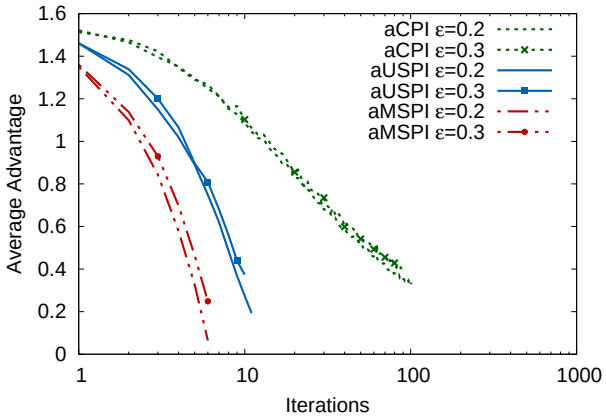


Figure 3. Approximate average advantage $\mathbb{A}_{\pi,\mu}^{\pi'}$ of aCPI, aMSPI and aUSPI in a 4-states chain walks with ϵ equals to 0.2 and 0.3.

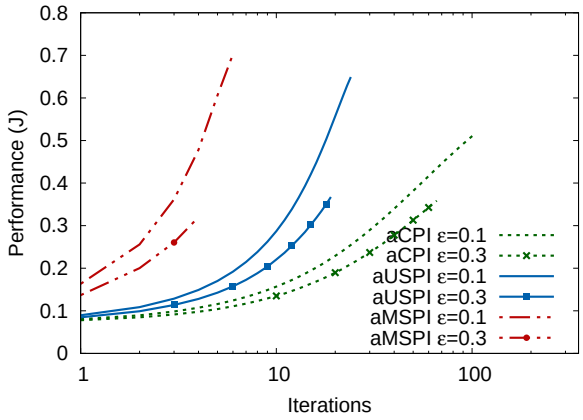


Figure 4. Approximate policy performance J_μ^π of aCPI, aMSPI and aUSPI in a 10-states chain walks with ϵ equals to 0.1 and 0.3.

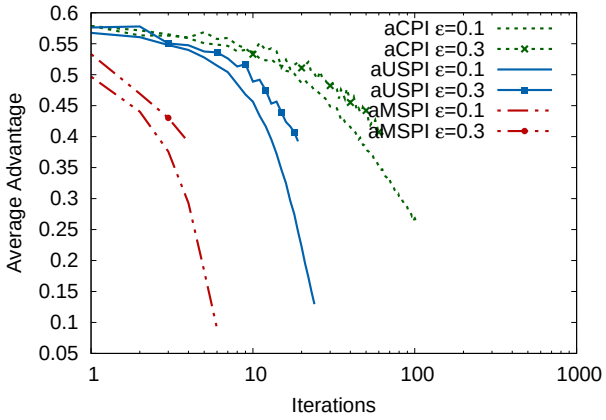


Figure 5. Approximate average advantage $\mathbb{A}_{\pi,\mu}^{\pi'}$ of aCPI, aMSPI and aUSPI in a 10-states chain walks with ϵ equals to 0.1 and 0.3.

2.1.2. Approximate Settings

The analysis in exact environments is not of practical interest. To give a complete overview of the performance of the algorithms, we have moved to an approximate framework. We consider the error induced by the estimation of the value function via a set of samples $\left\{s_i, a_i, Q_i^\pi\right\}_{i=1\ldots N_s}$. The experiments reported in the article are here extended in the scenario of 4-states and 10–states chain walk with approximation error ϵ equal to 0.1, 0.2 and 0.3. Results in the 4-states chain walk are reported in Figure 2 with the corresponding average advantage per iteration (see Figure 3), whereas Figure 5 and 4 show the average advantage $\mathcal{A}_{\pi,\mu}^{\pi'}$ estimated at each iteration and the performance of the new policy π' in a 10–states chain walk, respectively. It is notable that aUSPI and aMSPI perform similarly, whereas aCPI shows a low learning trend.

2.2. BlackJack Domain

The BlackJack is a card game where the player attends to beat the dealer by obtaining a total score greater than the dealer's one without exceeding 21. Each card counts as its numerical value (2 through 10) except for aces and figures. The Jack, Queen and King are worth 10, whereas the ace may value as either 1 or 11. The value of the ace is hand such that it produces the highest value equal to or less than 21. An hand is called *soft* when the ace is counted as 11. The set of cards is composed by 6 decks each one is a standard 52–cards deck.

At the beginning of the game the dealer deals two cards to each player, including himself. One card is faced up and the other is faced down. The player checks his two cards and chooses to receiver a new card (*hit*) or to stop (*stand*). The player may ask for more cards

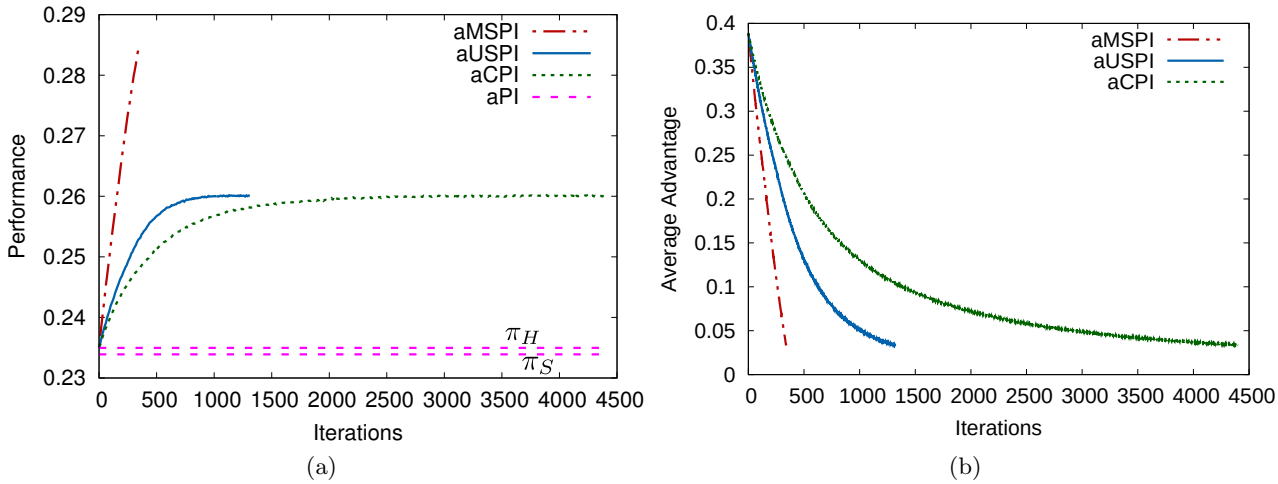


Figure 6. BlackJack performances. The underling domain consists in a BlackJack game with discount factor of 0.8. Figure (a) shows the performance of the algorithms. While aPI oscillates between policy π_H and policy π_S (figure reports only the performance of the two policies for seek of clarity), aMSPI, aUSPI and aCPI converge towards policies that outperform both π_H and π_S. Figure (b) reports the average advantage $\mathbb{A}^{\overline{\pi}}_{\pi,\mu}$ observed by the algorithms.

as long as he does not *bust*, i.e., the sum of the card values does not overcome 21. When all the players go bust or stops, is the turn of the dealer.

In this work, we consider a simplified version of the blackjack game by removing advanced actions as "doubling", "splitting", etc. The game is composed by a player and a dealer. The state of the game is defined by three components: the sum of the cards of the player (2 to 20), the dealer's faced-up card (1 to 10) and the soft hand flag. The player is forced to play "stand" action on blackjack and on 21. Moreover the soft hand flag is irrelevant when player's value is greater than 11. As a consequence, the cardinality of the state space is 260. The rewards assigned to the player are $+1$ for winning ($+1.5$ for blackjack), -1 for loosing and 0 for every hit. Rewards have been scaled to fit the interval $[0, 1]$.

To evaluate the performance of the algorithms we have exploited the simplified BlackJack model with discount factor equal to 0.8 and "stands on soft 17" strategy for the dealer. The evaluation measure is the estimated player edge, i.e., the average reward over multiple runs. We have been able to define a configuration where an approximate policy iteration, using a sample–based policy evaluation step and an exact improvement, oscillates between two non optimal policies. This configuration has been obtained by limit the policy space to two policies: both the policies select the best action (H) when player's value is equal to 20 and opposite actions for the other states (π_S selects S and π_H selects H). States with dealer's values equal to 9 and 10 are treated in an opposite way: policy π_S selects H and policy π_H chooses S. To summarize, the policies are defined according to the following rules:

$$\pi_S = \begin{cases} H, & \text{if } \textit{player's value} \text{ is } 20 \\ S, & \text{if } \textit{player's value} \text{ is less than } 20 \\ H, & \text{if } \textit{dealer's value} \text{ is } 9 \text{ or } 10 \end{cases}$$

$$\pi_H = \begin{cases} H, & \text{if } \textit{player's value} \text{ is } 20 \\ H, & \text{if } \textit{player's value} \text{ is less than } 20 \\ S, & \text{if } \textit{dealer's value} \text{ is } 9 \text{ or } 10 \end{cases}$$

Policy π_H has been chosen as initial policy. Figure 6 reports the performance of the policies obtained by aPI, aCPI, aUSPI and aMSPI algorithms using an approximation error ϵ of 0.01 and a estimation probability δ of 0.1. While aPI oscillates between π_H and π_S, other algorithms do not get stuck and converge towards better policies. aMSPI outperforms both aUSPI and aCPI.

It is worth to underline that, in this highly stochastic domain, the aMSPI is able to exploit the flexibility given by the multiple convex coefficients and to converge faster than aUSPI and aCPI.

3. Time Data

In the paper we have stated a brief comment about the execution times of our algorithms. Here, it can be found more details. Algorithms with single learning parameter (aCPI and aUSPI) share the same per–iteration computational complexity but the learning

Table 1. Per–iteration time complexity (sample mean ± standard deviation of the mean estimation) in approximate settings in 4-states (a) and 10-states chain walk (b). The time required for the improvement step is shown, the rightmost column (t_s) reports the time required for the generation of the samples (more than 99% of the overall time). Results have been averaged over 20 runs for all the algorithms. Initial policies have been chosen at random. Tests have been performed using single threaded algorithms on an Intel®Xeon®Processor E5345@2.33GHz. Table (c) presents the computational time (averaged over more than 300 samples) required by experiments in the BlackJack domain. Tests have been performed using 2 threads on a server architecture composed by 4 Intel®Xeon®Processor E5345@2.33GHz. The evaluation time dominates the improvement time, the time required for the samples generation is more than the 99% of the overall time.

(a)

γ	ϵ	aCPI $[ms]$	aUSPI $[ms]$	aMSPI $[ms]$	t_s $[ms]$
0.5	0.1	3.50 ± 0.01	3.58 ± 0.03	3.57 ± 0.02	562.03 ± 0.59
0.5	0.2	0.92 ± 0.01	0.95 ± 0.01	0.99 ± 0.02	116.04 ± 0.08
0.65	0.1	3.68 ± 0.01	3.70 ± 0.01	3.76 ± 0.03	927.08 ± 0.99
0.65	0.2	1.00 ± 0.01	1.03 ± 0.01	1.05 ± 0.01	180.40 ± 0.30

(b)

γ	ϵ	aCPI $[ms]$	aUSPI $[ms]$	aMSPI $[ms]$	t_s $[ms]$
0.5	0.1	5.35 ± 0.04	5.38 ± 0.04	5.50 ± 0.04	826.74 ± 0.91
0.5	0.2	1.40 ± 0.02	1.42 ± 0.02	1.46 ± 0.03	173.42 ± 0.15
0.65	0.1	5.28 ± 0.37	4.98 ± 0.07	5.93 ± 0.88	1342.85 ± 1.27
0.65	0.2	2.18 ± 0.77	1.43 ± 0.02	1.46 ± 0.02	268.72 ± 0.36

(c)

γ	ϵ	aCPI $[ms]$	aUSPI $[ms]$	aMSPI $[ms]$	t_s $[s]$
0.8	0.01	11.19 ± 0.14	9.88 ± 0.19	17.08 ± 0.15	75.09 ± 0.13

rate of the aCPI is sensibly lower than the aUSPI one. As a consequence, the aCPI requires, in general, a higher number of iterations to converge resulting in a higher overall computational time (refer to Section 6 in the article). The aMSPI algorithm has a higher per–iteration computational complexity w.r.t. aUSPI and aCPI. However, the time needed for the improvement step is dominated by the time required for the sample–based evaluation. As a consequence, the additional computational effort required by aMSPI in not significant when an approximate scenario in faced. This considerations are supported by the experiments (see Table 1).

References

Cho, G.E. and Meyer, C.D. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335(1-3):137–150, 2001.

Kakade, S.M. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of ICML*, pp. 267–274, 2002.

Koller, Daphne and Parr, Ronald. Policy Iteration for Factored MDPs. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 326–334, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-709-9.