# On the Complexity and Convergence of Approximate Policy Iteration Schemes

Arjun Subramonian, Shree Kesava Narayan Prasanna, Nikil Roashan Selvam, Justin Yi
University of California, Los Angeles

## INTRODUCTION

Policy iteration is dynamic programming approach to solving Markov Decision Processes (MDPs), which are fundamentally control problems. The key reason for convergence of Policy Iteration (PI) is the policy improvement property, which guarantees that the new policy at the end of an iteration is at least as good as the policy in the previous iteration, and that the lack of optimality of the previous policy guarantees a strict improvement in the current iteration. While PI is guaranteed to converge, the time complexity of PI leads us to naturally consider approximations in the PI algorithm.

We aim to carry out a survey of various Approximate Policy Iteration schemes and results from literature. First, we start with *On the Convergence of Approximate and Regularized Policy Iteration Schemes* by Smirnova and Dohmatob, which establishes convergence rates for the classic Approximate Modified Policy Iteration (AMPI) and Regularized Modified Policy Iteration (regMPI). The key idea behind the technique of AMPI is to allow approximations in both the Policy Evaluation and Policy Improvement steps of the classic Policy Iteration algorithm. We then consider Scherrer's *Approximate Policy Iteration Schemes: A Comparison*, which introduces us to multiple variations of API including Conservative Policy Iteration(CPI) (Kakade, Langford, 2002), Policy Search by Dynamic Programming algorithm (Bagnell etal., 2003) to the infinite-horizon case (PSDP$\infty$), and Non-Stationary Policy Iteration (NSPI(m))(Scherrer, Lesner, 2012). In our work, we focus on the performance bound with respect to the per-iteration error of CPI, which at each step generates a stochastic mixture of all the policies that are returned by the approximately greedy operator that is used to generate new policies at each step. Scherrer proves that CPI's performance is arbitrarily better than that of API at a cost exponential in $\frac{1}{\epsilon}$ relative to number of iterations. Lastly, we turn our attention to Perkins' and Precup's *A Convergent Form of Approximation Policy Iteration*, which provides a more general treatment of convergence criteria for API. Notably, this paper makes no assumptions on the policy improvement operator $\Gamma$, but guarantees convergence of Approximate Policy Iteration if $\Gamma$ is both $\epsilon$-soft and Lipschitz continuous.

There has been a lot of other interesting and influential work in this area that we are unable to delve into greater detail in this paper due to space limitations. Bertsekas's *Approximate policy iteration: a survey and some new methods* discusses the role of policy oscillation and its effect on performance guarantees. Bertsekas proves that although policy evaluation may result in policy oscillation when using projected equation , there is no such oscillation or chattering when using aggregation methods. *Error Propagation for Approximate Policy and Value Iteration* by Farahmand, Munos, and Szepesvari proves that the performance loss is influenced significantly more by the approximation error in the later iterations of API than by the errors in earlier iterations, and that the effect of the error terms in earlier iterations decays exponentially fast.

In our work, we provide theoretical proof sketches involved in the analysis of the complexity bounds, convergence guarantees, and rates of convergence for the approximate policy iteration algorithms presented in the surveyed papers. By doing so, we will develop a comprehensive overview of the theory behind these algorithms and identify their similarities and differences.

## On the Convergence of Approximate and Regularized Policy Iteration Schemes

This paper by Smirnova and Dohmatob covers numerous MPI schemes, namely Approximate MPI (AMPI) and Regularized MPI (reg-MPI) (which they posit is a subclass of AMPI). The authors provide "sufficient conditions for convergence of a large class of regularized dynamic programming algorithms". The paper provides "explicit convergence rates to the optimality depending on the decrease rate of the regularization parameter".

$G(v) = \{\pi : T_\pi v = Tv\}$. The Modified Policy Iteration algorithm can be defined as follows

$$\begin{cases} \pi_{t+1} \in G(v_t) & \text{Policy Improvement} \\ \mathrm{v}_{t+1} = (T_{\pi_{t+1}})^m v_t & \text{Policy Evaluation} \end{cases}$$

Approximate MPI (AMPI) is a generalization of MPI, allowing for errors in the evaluation and improvement steps. The AMPI algorithm is defined as follows

$$\begin{cases} \pi_{t+1} \in G_{\epsilon'_{t+1}}(v_t) & \text{Policy Improvement} \\ \mathrm{v}_{t+1} = (T_{\pi_{t+1}})^m v_t + \epsilon_{t+1} & \text{Policy Evaluation} \end{cases}$$

$\epsilon_t$ and $\epsilon'_t$ are associated with errors in the policy evaluation and the policy improvement steps respectively.

Smirnova and Dohmatob theorize bounds on the convergence rate of an AMPI scheme to the optimal greedy policy, based on assumptions on error bounds. Suppose the error sequences $(||\epsilon_N||_\infty)_N$ and $(||\epsilon'_N||_\infty)_N$ satisfy $||\epsilon_N||_\infty + ||\epsilon'_N||_\infty \leq Cr_N$ for some constant $C > 0$ and a sequence $r_N \to 0$. Once the errors are so bounded, the convergence of the AMPI scheme, the authors posit, is guaranteed.

This has a profound impact on the viability of control problems. One can tell from bounds on errors whether certain problems are worth pursuing. If errors are bounded as prescribed above, one is guaranteed to obtain the optimal greedy policy. Time and resources can then be allocated to performing the MPI. If errors are not bounded as necessary, one is less certain of convergence, and pursuing such a project may not be worth the limited resources available in practice.

## KEY THEOREMS

**Theorem 1** (AMPI Convergence) *Suppose the error sequences $(||\epsilon_N||_\infty)_N$ and $(||\epsilon'_N||_\infty)_N$ satisfy $||\epsilon_N||_\infty + ||\epsilon'_N||_\infty \leq Cr_N$ for some constant $C > 0$ and a sequence $r_N \to 0$. Then, the AMPI scheme, as defined in the paper, converges to the optimal greedy policy of the exact MPI.*
*Furthermore, the limits $\underline{\rho} := \underline{\lim}\, r_N/r_{N-1}$ and $\overline{\rho} := \overline{\lim}\, r_N/r_{N-1}$. The following bounds hold*
*(A)* ***Slow Convergence.*** *If $\underline{\rho} > \gamma$, then*

$$||v_N - v^*||_\infty = \mathcal{O}(r_N)$$

*(B)* ***(Almost) linear convergence.*** *If $\overline{\rho} \leq \gamma$, then*

$$||v_N - v^*||_\infty = \begin{cases} \mathcal{O}(\gamma^N), & if\, \overline{\rho} < \gamma \\ \mathcal{O}(N\gamma^N), & if\, \overline{\rho} = \gamma \end{cases}$$

**Theorem 2** *At each $k < k^*$ of CPI, the expected loss satisfies:*

$$\mu(v_{\pi_*} - v_{\pi_k}) \leq \frac{C^{(1,0)}}{(1-\gamma)^2} \sum_{i=1}^{k} \alpha_i \epsilon_i + e^{\{-(1-\gamma)\sum_{i=1}^{k} \alpha_i\}} V_{max}$$

*Here, the concentrability constant $C^{(1,0)} = (1-\gamma)\sum_{i=0}^{\infty} \gamma^i c(i)$, where $c(i)$ is the smallest coefficient in $[1,\infty) \cup \{\infty\}$ such that for all $i$ and all sets of deterministic stationary policies $\pi_1, \pi_2, \ldots, \pi_i, \mu P_{\pi_1} P_{\pi_2} \cdots P_{\pi_i} \leq c(i)\nu$.*

**Corollary 1** *The smallest (random) iteration $k^\dagger$ such that $\frac{\log \frac{V_{max}}{\epsilon}}{1-\gamma} \leq \sum_{i=1}^{k^\dagger} \alpha_i \leq \frac{\log \frac{V_{max}}{\epsilon}+1}{1-\gamma}$ is such that $k^\dagger \leq \frac{12\gamma V_{max} \log \frac{V_{max}}{\epsilon}}{\epsilon(1-\gamma)^2}$ and the policy $\pi_{k^\dagger}$ satisfies:*

$$\mu(v_{\pi_*} - v_{\pi_{k^\dagger}}) \leq \left(\frac{C^{(1,0)}(\sum_{i=1}^{k^\dagger} \alpha_i)}{(1-\gamma)^2} + 1\right)\epsilon \leq \left(\frac{C^{(1,0)}(\log \frac{V_{max}}{\epsilon} + 1)}{(1-\gamma)^3} + 1\right)\epsilon$$

**Theorem 3** *For any infinite-horizon Markov decision process behaving as an irreducible, aperiodic Markov chain over the states, and for any $\epsilon > 0$, there exists $c > 0$ such that if $\Gamma$ is $\epsilon$-soft and Lipschitz continuous with constant c, then the sequence of policies generated by the approximate policy iteration algorithm converges to a unique limiting policy $\pi \in \Pi_\epsilon$, regardless of initial policy $\pi_0$.*

## Approximate Policy Iteration Schemes: A Comparison

*Approximate Policy Iteration Schemes: A Comparison* by Scherrer analyzes the performance bounds with respect to the per-iteration error $\epsilon$ of several approximate variations of the Policy Iteration algorithm for the infinite-horizon discounted optimal control problem. These variations include Approximate Policy Iteration (API) (Bertsekas, Tsitsiklis, 1996) and Conservative Policy Iteration (CPI) (Kakade, Langford, 2002).

Each of these approximate policy iteration variations implements an approximate greedy operator that returns a policy $\pi$ that is $(\epsilon, \nu)$-approximately greedy with respect to $v$ in the sense that $\nu(\max_{\pi'} T_{\pi'} v - T_\pi v) \leq \epsilon$. In API, at each iteration $k$, the algorithm switches to the policy that is approximately greedy with respect to the value of the previous policy for some distribution $\nu$. CPI is similar to API except the distribution used in the approximate greedy operator is the discounted cumulative occupancy measure induced by $\pi_k$ when starting from $\nu$. It also employs a variable step size when updating the policy, which allows for a stochastic mixture of all policies returned by successive calls to the approximate greedy operator.

Scherrer proves that CPI's performance bound is arbitrarily better than that of API at a cost exponential in $\frac{1}{\epsilon}$ relative to number of iterations. Our contribution is elaborating on the proof sketch presented by Scherrer, filling in, proving, and providing high-level commentary on the many necessary details and intermediate lemmas and steps omitted by Scherrer.

We first present a detailed proof of **Theorem 2**, which describes the bound on the expected loss $\mathbb{E}_{s\sim\mu}[v_{\pi_*}(s) - v_\pi(s)] = \mu(v_{\pi_*} - v_\pi)$ of using the (possibly stochastic or non-stationary) policy $\pi$ output by CPI instead of the optimal policy $\pi_*$ from some initial distribution $\mu$ of interest as a function of an upper bound $\epsilon$ on all errors $(\epsilon_k)$. At a high level, we achieve this by first bounding the difference between the state value functions corresponding to policies from consecutive iterations of CPI. Then, we derive a recurrence inequality that, for an arbitrary iteration $k$ of CPI, relates $v_{\pi_*} - v_{\pi_{k+1}}$ to $v_{\pi_*} - v_{\pi_k}$. Thereby, we inductively construct an explicit inequality that bounds $v_{\pi_*} - v_{\pi_k}$ from above. Clearly, $e^{\{-(1-\gamma)\sum_{i=1}^{k} \alpha_i\}} V_{max}$ tends to 0 exponentially quickly in the number of iterations.

We can make use of **Theorem 2** to prove **Corollary 1**. Conceptually, **Corollary 1** shows that CPI has a performance bound with the coefficient $C^{(1,0)}$ of API in a number of iterations $O(\frac{\log \frac{1}{\epsilon}}{\epsilon})$. However, API only requires $O(\log \frac{1}{\epsilon})$ iterations to achieve this performance bound, so CPI is still exponentially slower than API.

## A Convergent Form of Approximation Policy Iteration

Perkins' and Precup's *A Convergent Form of Approximation Policy Iteration* explores the conditions in which convergence of approximate policy iteration is guaranteed, namely when a so called policy improvement operator is both $\epsilon$-soft and Lipschitz continuous in action values **Theorem 3**. This method, after learning weights of a linear approximation to the action-state function via Sarsa updating introduces a notion of a policy improvement operator, $\Gamma$, that maps this function to a stochastic policy.

A policy $\pi$ is called *$\epsilon$-soft* if $\pi(s,a) \geq \epsilon$ for all $s$ and $a$. For any $\epsilon > 0$ let $\Pi_\epsilon$ denote the set of $\epsilon$-soft policies. We observe that $\Pi_\epsilon$ can be thought of as a compact subset of $R^{mn}$.

A generic policy improvement operator $\Gamma$, which maps every $Q \in R^{mn}$ to a stochastic policy. $\Gamma$ is *Lipschitz continuous with constant* $c$ if $\forall Q_1, Q_2 \in R^{mn}$, $\Gamma(Q_1) - \Gamma(Q_2) \leq c(Q_1) - (Q_2)$. $\Gamma$ is $\epsilon$-soft if, $\forall Q \in R^{mn}$, $\Gamma(Q)$ is $\epsilon$-soft.

If the behavior of the agent does not change too greatly in response to changes in its actions value estimates, then convergence is guaranteed. While this presents a general statement of condition and consequence for convergence, it merely posits the existence of such convergent behavior, with little to be said for the quality of the convergent policy in question.

## CONCLUSIONS AND FUTURE DIRECTIONS

Approximate policy iteration seems to be extremely promising, especially for the computation advantage that it provides. However, unsurprisingly, we lose the usual converge guarantees that we have, and it is often unclear whether an algorithm converges, and even if it does, whether or not it converges to the optimal policy.

One major contribution by Smirnova and Dohmatob in their *On the Convergence of Approximate and Regularized Policy Iteration Schemes* is the guarantee of convergence of the AMPI algorithm solely based on errors in each step of policy iteration, in both the policy evaluation and policy improvement steps.

Scherrer, in his *Approximate Policy Iteration Schemes: A Comparison*, analyzes CPI, which at each step generates a stochastic mixture of all the policies that are returned by the approximate greedy operator, and provides useful performance bounds and number of iterations (relative to API) for the same. However, Scherrer assumes the existence of a reasonable $\epsilon$-approximate greedy operator. This might not always be the case as there different optimization problems might have different underlying structures which might make it computationally difficult to have an $\epsilon$-approximate greedy operator. One potential future direction of research is further unpacking the $\epsilon$-part of the approximate greedy operator. Furthermore, Scherrer analyzes the expected loss on each iteration of CPI as a weighted $l_1$-norm of the loss $v_{\pi_*} - v_\pi$. He admits that it's possible to consider the weighted $l_p$ norm (for $p \geq 2$) of the loss but does not do so in his paper. Carrying out the analysis presented in Section 3.2 for a weighted $l_p$ norm (for $p \geq 2$) of the loss $v_{\pi_*} - v_\pi$ (and for other approximate policy iteration algorithms) is another promising direction of research.

Perkins' and Precup's *A Convergent Form of Approximation Policy Iteration* paper considers online Sarsa for the policy evaluation step with linear function approximation, but makes no assumptions on the policy improvement operator $\Gamma$. However, it remarkably still guarantees convergence of Approximate Policy Iteration if $\Gamma$ is both $\epsilon$-soft and Lipschitz continuous. The authors hope this type of guarantee can be extended to similar algorithms such as online Sarsa or Sarsa(lambda), and we expect this to be fairly straightforward. The more interesting aspect of this paper is the fact that it serves as a starting point for refuting one major drawback of most reinforcement learning algorithms that approximate value functions: convergence. Future research could delve deeper to establish not just convergence criteria for such algorithms, but also potentially optimality guarantees by possibly considering stronger assumptions on the operators being used.

A natural future direction for us is to further explore Bertsekas's *Approximate policy iteration: a survey and some new methods* and *Error Propagation for Approximate Policy and Value Iteration* by Farahmand, Munos, and Szepesvari in greater detail and present in-depth proofs, as we could not do so here due to space limitations. We hope that this survey would provide a comprehensive overview of current approximate policy iteration schemes, and serve as a valuable resource for future research in this area.

## ACKNOWLEDGEMENTS

We would like to thank Prof. Lin Yang and TAs Joris Kenanian, Qiujing Lu, and Chinmay Talegaonkar for teaching us concepts that were essential in understanding and summarizing the papers we chose to cover. We thank them for their support and for being open to our questions and concerns.

We would like to thank the Henry Samueli School of Engineering and Applied Science for scheduling ECE239AS: Introduction to Reinforcement Learning, which made this endeavor possible.