On the Complexity and Convergence of Approximate Policy Iteration Schemes

Arjun Subramonian
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 95014
arjunsub@ucla.edu

Shree Kesava Narayan Prasanna
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 95014
kesavapr@ucla.edu

Nikil Roashan Selvam
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 95014
nikilrselvam@ucla.edu

Justin Yi
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 95014
joostinyi00@g.ucla.edu

Abstract

Policy iteration (PI) is a popular method of converging upon an optimal policy in a Markov Decision Process. More specifically, we consider performing policy iteration with some bounded error at each step. This leads to the class of Approximate Policy Iteration Algorithms.
In this project, we survey relevant literature in approximate policy iteration, and provide theoretical proof sketches involved in the analysis of the complexity bounds, convergence guarantees, and rates of convergence for various approximate policy iteration algorithms. We conclude with promising future directions of research in analyzing the performance of approximate policy iteration schemes.

1 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. The following sections of the paper summarize the literature we are surveying and detail their role in our project. 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∞), and Non-Stationary Policy Iteration (NSPI(m))(Scherrer, Lesner, 2012). In this paper, 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 approximate greedy operator

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

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; we elaborate 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. 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 Γ, but guarantees convergence of Approximate Policy Iteration if Γ is both ϵ-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* provides a broad summary of existing policy evaluation methods including matrix inversion methods, such as least-squares temporal difference (LSTD), and iterative methods, such as least-squares policy evaluation (LSPE). The paper also 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 quantifies the change in quality of the final policy induced by the approximation/Bellman error in each step of API. The authors prove 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 this paper, we will 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.

2 Notation and Terminology

A Markov Decision Process (MDP) is defined as a set (S, A, P, r, γ), where S is a possibly infinite state space, A is a finite action space, $P(s'|s, a)$, for all (s, a) is a known state transition probability matrix on S, $r : S \to [-R_{max}, R_{max}]$ is a reward function bounded by R_{max}, and $\gamma \in (0, 1)$ is a discount factor.

A policy π is a map $\pi : S \to A$ that maps states to actions. In this paper, we usually consider a stationary deterministic policy (i.e. a non-stochastic policy that doesn't change over the course of an episode). We then denote $P_\pi(s'|s) = P(s'|s, \pi(s))$. In some cases, we consider stochastic policies, which map states to a distribution over the action space.

The return from time step t, denoted G_t, is the total discounted sum of rewards starting from time step t. It is given by $g_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$. The state value function $v_\pi(s)$ is the expected discounted return if the agent follows policy π from state s,and is given by $v_\pi(s) = \mathbb{E}[g_t|s_t = s]$. The state-action value function $q^\pi(s, a)$ is the expected discounted return if the agent takes action a from state s and follows policy π thereafter, and is given by $q_\pi(s, a) = \mathbb{E}[g_t|s_t = s, a_t = a]$

Additionally, v_π, and is the fixed point of the linear Bellman operator associated with π: $T_\pi : v \to r + \gamma P_\pi v$. The Bellman optimality operator is defined as $T : v \to \max_\pi T_\pi v$, with unique fixed point $v_* = \max_\pi T_\pi v$, which is the optimal value.

A policy π is greedy with respect to v if $T_\pi v = Tv$. We denote the set of all such greedy policies Gv. A policy π_* is optimal with $v_{\pi_*} = v_*$ if and only if $\pi_* \in Gv_* \iff T_{\pi_*} v_* = v_*$.

The approximate greedy operator G_ϵ, which is employed by numerous approximate policy iteration schemes is defined as $G_\epsilon : (\nu, v : S \to \mathbb{R}) \to \pi$ such that $\nu(Tv - T_piv) = \nu(\max_{\pi'} T_{\pi'} - T_\pi v) \leq \epsilon$. ν is a probability distribution over S and $\forall x, \nu x = \mathbb{E}_{s\sim\nu}[x(s)]$. In many cases, π is obtained by l_p regression of the state-value function.

We also consider a generic policy improvement operator, Γ, which maps every state- action value function q to a stochastic policy. Stochastic policies include greedy policies, ϵ-greedy policies, or policies with action selection probabilities based on the softmax function.

A policy improvement operator Γ is Lipschitz continuous with constant c if, for all state-action value functions q_1 and q_2, $\|\Gamma(q_1) - \Gamma(q_2)\| \leq c\,\|(q_1) - (q_2)\|$ where $\|\|$denotes the Euclidean norm. The operator Γ is deifned to be ϵ-soft if, for all state-action value functions q, $\Gamma(q)$ is ϵ-soft.

3 Approximate Policy Iteration Schemes

3.1 On the Convergence of Approximate and Regularized Policy Iteration Schemes

We consider an infinite-horizon discounted MDP (S, A, P, r, γ), where S is a possibly infinite state space, A is a finite action space, $P(ds'|s, a)$, for all (s, a) is a known state transition probability matrix on S, $r : S \to [-R_{max}, R_{max}]$ is a reward function bounded by R_{max}, and $\gamma \in (0, 1)$ is a discount factor. In addition, we consider a stationary deterministic policy $\pi : S \to A$ that maps states to actions (i.e. a non-stochastic policy that doesn't change over the course of an episode). Hence, we denote $P_\pi(ds'|s) = P(ds'|s, \pi(s))$. Additionally, $v_\pi = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t)|s_0 = s, s_{t+1} \sim P_\pi(\cdot|s_t)]$, and is the fixed point of the linear Bellman operator associated with π: $T_\pi : v \to r + \gamma P_\pi v$. The Bellman optimality operator is defined as $T : v \to \max_\pi T_\pi v$, with unique fixed point $v_* = \max_\pi T_\pi v_*$, which is the optimal value.

We define the set G as follows

$$G(v) := argmax_\pi T_\pi v$$

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

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

In other words, MPI can be decomposed into two steps: *policy evaluation* and *policy improvement*. In policy evaluation, the value function of the current policy is (partially) calculated. Note that T_{π_t} is applied m times, for $1 \leq m \leq \infty$, using the current policy π_t. In policy improvement, the new greedy policy is extracted from the updated (approximate) value function of the current policy. This process is repeated until convergence.

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

$$\begin{cases} \pi_{t+1} \in G_{\epsilon'_{t+1}}(v_t) & \text{Policy Improvement} \\ \mathrm{v}_{t+1} = (T_{\pi_{t+1}})^m v_t + \epsilon_{t+1} & \text{Policy Evaluation} \end{cases} \tag{2}$$

Here, $\epsilon_t, \epsilon'_t \in \mathbb{R}^{|S|}$. ($|S|$ is loosely interpreted as the number of states used in the approximation scheme.) ϵ_t and ϵ'_t are associated with errors in the policy evaluation and the policy improvement steps respectively. Section 4 of [5] outlines the various possible reasons for error injection, based on the particular AMPI algorithm employed. $G_{\epsilon'}$ is defined as: $\pi \in G_{\epsilon'}(v) \iff \forall \pi' \quad T^{\pi'} v \leq T^\pi v + \epsilon'$. This means that any $\pi \in G_{\epsilon'}(v)$ is *approximately greedy*.

We now dive into an analysis of the convergence of the above AMPI scheme. The proof assumes the following lemma. We omit a proof of the lemma itself for brevity. A detailed proof of the lemma can be found in Theorem 7 of [5].

Lemma 1 (AMPI Error Propagation). *For any initial value function V_0 and $m \geq 1$, consider the AMPI scheme (2). Then*

$$||v_N - v^*||_\infty \leq \frac{2}{1-\gamma}(E_N + \gamma^N ||v_0 - v^*||_\infty), \tag{3}$$

$$\text{where } E_N := \sum_{t=1}^{N-1} \gamma^{N-t}(||\epsilon_t||_\infty + ||\epsilon'_t||_\infty). \tag{4}$$

Note that, in (3), $\gamma^N ||v_0 - v^*||_\infty \to 0$ as $N \to \infty$ (as $\gamma \in [0,1)$). Then, it is sufficient to analyze E_N to gain insight on how v_N approaches v^*.

We first make some observations regarding bounds on sequences of numbers. Let $r_1, r_2, ..., r_t, ...$ be a sequence of positive real numbers. Let $\underline{\rho} :=$ lim inf$\frac{r_N}{r_{N-1}}$. Similarly, $\overline{\rho} :=$ lim sup$\frac{r_N}{r_{N-1}}$. Let the sum $S_N := \sum_{t=0}^{N-1} \theta^{N-t} r_t$, for $N \geq 1$. Here, $\theta \in [0,1)$.

Case 1: $\underline{\rho} \geq \theta$, then, lim inf$\frac{r_N}{r_{N-1}} \geq \theta$. The limit inferior of a sequence is defined as $lim\, inf\, x_n := lim_{n \to \infty}(inf_{m \geq n}\, x_m)$. So, given that the infinimum (greatest lower bound) of the set $\{x_{t \geq n}\}$ is $\underline{\rho}$ (where $x_t = r_t / r_{t-1}$), then for large enough n and $t \leq N$, we have

$$x_N \geq \underline{\rho}, \quad x_{N-1} \geq \underline{\rho}, \quad ... \quad x_t \geq \underline{\rho} \quad \Rightarrow \quad r_N \geq \underline{\rho} r_{N-1} \geq ... \geq \underline{\rho}^{N-t} r_t$$

So, $r_t \leq r_N \underline{\rho}^{-(N-t)}$. So, for large N:

$$\begin{aligned} S_N := \sum_{t=1}^{N-1} \theta^{N-t} r_t \leq \sum_{t=1}^{N-1} \theta^{N-t} r_N \underline{\rho}^{-(N-t)} &= r_N \sum_{t=1}^{N-1} (\theta/\underline{\rho})^{(N-t)} \\ &= r_N (\theta/\underline{\rho}) \frac{1 - (\theta/\underline{\rho})^N}{1 - (\theta/\underline{\rho})} \text{ (sum of a geometric series)} \\ &= r_N \frac{\theta(1 - (\theta/\underline{\rho})^N)}{\underline{\rho} - \theta} \leq r_N \frac{\theta}{\underline{\rho} - \theta}. \quad \text{So, } S_N = \mathcal{O}(r_N) \end{aligned} \tag{5}$$

Call the above result **Lemma 2.1**.

Case 2: $\overline{\rho} < \theta$, then, lim sup$\frac{r_N}{r_{N-1}} < \theta$. The limit superior of a sequence is defined as $lim\, sup\, x_n := lim_{n \to \infty}(sup_{m \geq n}\, x_m)$. So, given that the supremum (lowest upper bound) of the set $\{x_{t \geq n}\}$ is $\overline{\rho}$ (where $x_t = r_t / r_{t-1}$), then for large enough n and $t \leq N$, we have

$$x_t \leq \overline{\rho}, \quad x_{t-1} \leq \overline{\rho}, \quad ... \quad x_1 \leq \overline{\rho} \quad \Rightarrow \quad r_t \leq \overline{\rho} r_{t-1} \leq ... \leq \overline{\rho}^{t-1} r_1 \tag{6}$$

So, $r_t \leq r_1 \overline{\rho}^{(t-1)}$. So, for large N:

$$S_N := \sum_{t=1}^{N-1} \theta^{N-t} r_t \leq \sum_{t=1}^{N-1} \theta^{N-t} r_1 \overline{\rho}^{(t-1)} = r_1 \theta^{N-1} \sum_{t=0}^{N-2} (\overline{\rho}/\theta)^t \tag{7}$$

$$= r_1 \theta^{N-1} \frac{1 - (\overline{\rho}/\theta)^{N-1}}{1 - (\overline{\rho}/\theta)} \text{ (sum of a geometric series)}$$

$$= r_1 \theta^{N-1} \frac{\theta(1 - (\overline{\rho}/\theta)^{N-1})}{\theta - \overline{\rho}} \leq r_1 \theta^{N-1} \frac{\theta}{\theta - \overline{\rho}}. \quad \text{So, } S_N = \mathcal{O}(\theta^N) \tag{8}$$

Call the above result **Lemma 2.2**.

Case 3: $\overline{\rho} = \theta$. Then, the inequality (6) still holds. Further, in Eq (7), $(\overline{\rho}/\theta) = 1$. So,

$$S_N = r_1 \theta^{N-1}(N-1) = \mathcal{O}(N\theta^N) \tag{9}$$

Call the above result **Lemma 2.3**.

Using the above observations (**Lemma 2** in all) and **Lemma 1**, we seek to prove the following theorem.

Theorem 1 (AMPI Convergence) *Suppose the error sequences* $(||\epsilon_N||_\infty)_N$ *and* $(||\epsilon'_N||_\infty)_N$ *satisfy* $||\epsilon_N||_\infty + ||\epsilon'_N||_\infty \leq C r_N$ *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*

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

Proof of convergence (without explicit asymptotic bounds).
As $r_t \to 0$, for some $\delta > 0$, we have, for large enough t, $r_t \leq \delta$ (this comes from the definition of a convergent sequence). Therefore, for sufficiently large N, we have

$$\begin{aligned} E_N &:= \sum_{t=1}^{N-1} \gamma^{N-t}(||\epsilon_t||_\infty + ||\epsilon_t'||_\infty) \leq \sum_{t=1}^{N-1} \gamma^{N-t} C r_t \text{ (from constraints on errors in Theorem 1)} \\ &\leq C \sum_{t=1}^{N-1} \gamma^{N-t}\delta \text{ (for large enough N)} \leq C\delta \sum_{t=1}^{\infty} \gamma^t = C\frac{\gamma}{1-\gamma}\delta \text{ (some positive constant)} \end{aligned}$$

Therefore, as $N \to \infty$, $E_N \leq \delta$ for some $\delta > 0$. So, $E_N \to 0$ as $N \to \infty$. Then, from the bounds (3) in Lemma 1, $||v_N - v^*||_\infty \to 0$ as $N \to \infty$, and so, v_N converges to v^*.

Proof of asymptotic bounds.
(A): $\underline{\rho} :=$ lim inf$\frac{r_N}{r_{N-1}} \geq \gamma$. So, for large N:

$$\begin{aligned} E_N &:= \sum_{t=1}^{N-1} \gamma^{N-t}(||\epsilon_t||_\infty + ||\epsilon_t'||_\infty) \leq C \sum_{t=1}^{N-1} \gamma^{N-t} r_t \text{ (from constraints on errors in Theorem 1)} \\ &= \mathcal{O}(r_N) \text{ (from (5) of } \textbf{Lemma 2.1}\text{, where } \theta = \gamma) \end{aligned}$$

(B): $\overline{\rho} :=$ lim sup$\frac{r_N}{r_{N-1}} < \gamma$. So, for large N:

$$\begin{aligned} E_N &:= \sum_{t=1}^{N-1} \gamma^{N-t}(||\epsilon_t||_\infty + ||\epsilon_t'||_\infty) \leq C \sum_{t=1}^{N-1} \gamma^{N-t} r_t \text{ (from constraints on errors in Theorem 1)} \\ &= \mathcal{O}(\gamma^N) \text{ (from (8) of } \textbf{Lemma 2.2}\text{, where } \theta = \gamma) \end{aligned}$$

(C): $\overline{\rho} :=$ lim sup$\frac{r_N}{r_{N-1}} = \gamma$. So, for large N:

$$\begin{aligned} E_N &:= \sum_{t=1}^{N-1} \gamma^{N-t}(||\epsilon_t||_\infty + ||\epsilon_t'||_\infty) \leq C \sum_{t=1}^{N-1} \gamma^{N-t} r_t \text{ (from constraints on errors in Theorem 1)} \\ &= \mathcal{O}(N\gamma^N) \text{ (from (9) of } \textbf{Lemma 2.3}\text{, where } \theta = \gamma) \end{aligned}$$

□

3.2 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 ϵ 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), Conservative Policy Iteration (CPI) (Kakade, Langford, 2002), Policy Search by Dynamic Programming algorithm (Bagnell et al., 2003) to the infinite-horizon case (PSDP$_\infty$), and Non-Stationary Policy Iteration (NSPI(m)) (Scherrer, Lesner, 2012). Scherrer also introduces a hierarchy of concentrability constants (defined below) useful in comparing the performance bounds of the variations. Each of these approximate policy iteration variations implements an approximate greedy operator that takes as input a distribution ν and a state-value function $v : S \to \mathbb{R}$ and returns a policy π that is (ϵ, ν)-approximately greedy with respect to v in the sense that $\nu(\max_{\pi'} T_{\pi'}v - T_\pi v) \leq \epsilon$, where $T_\pi : v \to r + \gamma P_\pi v$ is the linear Bellman operator associated with π.

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 ν. It is equivalent to exact policy iteration if the maximum approximation error ϵ is 0. CPI is similar to API except the distribution used in the approximate greedy operator is the discounted cumulative occupancy measure induced by π_k when starting from ν. 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.

In Scherrer's comparisons of the performance bounds of these algorithms, he focuses on the concentrability constants involved (defined below) and the number of iterations required. 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. In this section, we delve into this proof, which is included as Section C of the Supplementary Material. 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, and correcting his numerous typos.

First, we define some notation. We consider an infinite-horizon discounted MDP (S, A, P, r, γ), as defined in section 3.1. A policy π is greedy with respect to v if $T_\pi v = Tv$. Denote the set of all such greedy policies Gv. A policy π_* is optimal with $v_{\pi_*} = v_*$ if and only if $\pi_* \in Gv_* \iff T_{\pi_*} v_* = v_*$.

We now define the approximate greedy operator G_ϵ employed by numerous approximate policy iteration schemes. $G_\epsilon : (\nu, v : S \to \mathbb{R}) \to \pi$ such that $\nu(Tv - T_\pi v) = \nu(\max_{\pi'} T_{\pi'} - T_\pi v) \le \epsilon$. ν is a probability distribution over S and $\forall x, \nu x = \mathbb{E}_{s\sim\nu}[x(s)]$. In many cases, π is obtained via l_p value function approximation.

We consider Conservative Policy Iteration (CPI), first proposed by Kakade and Langford in 2002. We define CPI like so:

$$\pi_{k+1} \leftarrow (1 - \alpha_{k+1})\pi_k + \alpha_{k+1} G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})$$

$d_{\pi_k,\nu} = (1-\gamma)\nu(I - \gamma P_{\pi_k})^{-1}$, which is called the discounted cumulative occupancy measure induced by π_k when starting from ν. Furthermore, α_k is a stepsize used to generate a stochastic mixture of policies returned by successive calls to the approximate greedy operator, hence the name "conservative." α_k should be implemented with a line-search mechanism or be fixed to a small value α, such that α_{k+1} leads to an improvement of the expected value of the policy. The second variation is called CPI(α).

We will now present a proof of the performance bound with respect to the per-iteration error ϵ of CPI and the corresponding number of iterations required to achieve this bound. Furthermore, we provide commentary contrasting CPI's performance bound and number of iterations with those of API. We begin with the following theorem, 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 π output by CPI instead of the optimal policy π_* from some initial distribution μ of interest as a function of an upper bound ϵ on all errors (ϵ_k). Notably, the expected loss is a weighted l_1-norm of the loss $v_{\pi_*} - v_\pi$; Scherrer admits that it's possible to consider the weighted l_p norm (for $p \ge 2$) but does not do so in his paper. Furthermore, this theoretical guarantee involves the definition of concentrability constants that relate the distribution μ and the distribution ν used by CPI.

Theorem 1. *At each $k < k^*$ of CPI, the expected loss satisfies:*

$$\mu(v_{\pi_*} - v_{\pi_k}) \le \frac{C^{(1,0)}}{(1-\gamma)^2} \sum_{i=1}^{k} \alpha_i \epsilon_i + e^{\{-(1-\gamma)\sum_{i=1}^{k} \alpha_i\}} V_{max} \tag{10}$$

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 $\pi_1, \pi_2, \ldots, \pi_i, \mu P_{\pi_1} P_{\pi_2} \cdots P_{\pi_i} \le c(i)\nu$. Furthermore, $\mu(v_{\pi_*} - v_\pi) = E_{s\sim\mu}[v_{\pi_*}(s) - v_\pi(s)]$ is the expected loss of using the (possibly stochastic or non-stationary) policy π output by CPI instead of the optimal policy π_* from some initial distribution μ of interest. We hope to express this expected loss as a function of an upper bound ϵ on all errors ϵ_k.

We know that, in CPI, $\pi_{k+1} \leftarrow (1 - \alpha_{k+1})\pi_k + \alpha_{k+1} G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})$. Hence,
$T_{\pi_{k+1}} v_{\pi_k} = r + \gamma P_{\pi_{k+1}} v_{\pi_k}$
$= r + \gamma[(1 - \alpha_{k+1})P_{\pi_k} + \alpha_{k+1} P_{G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})}] v_{\pi_k}$

$= (1-\alpha_{k+1})r + \alpha_{k+1}r + (1-\alpha_{k+1})[\gamma P_{\pi_k} v_{\pi_k}] + \alpha_{k+1}[\gamma P_{G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})} v_{\pi_k}]$
$= (1-\alpha_{k+1})[r + \gamma P_{\pi_k} v_{\pi_k}] + \alpha_{k+1}[r + \gamma P_{G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})} v_{\pi_k}]$
$= (1-\alpha_{k+1}) T_{\pi_k} v_{\pi_k} + \alpha_{k+1} T_{G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})} v_{\pi_k}$.
By the fixed-point property of the linear Bellman operator, $T_{\pi_k} v_{\pi_k} = v_{\pi_k}$. Furthermore, we denote $\pi'_{k+1} = G_{\epsilon_{k+1}}(d_{\pi_k,\nu}, v_{\pi_k})$. So, in summary, $T_{\pi_{k+1}} v_{\pi_k} = (1-\alpha_{k+1}) v_{\pi_k} + \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k}$.

Next, we denote the error on the $(k+1)$-th iteration from applying the approximate greedy operator $e_{k+1} = \max_{\pi'} T_{\pi'} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k}$. Then, we can deduce $v_{\pi_*} - v_{\pi_{k+1}}$
$= v_{\pi_*} - T_{\pi_{k+1}} v_{\pi_{k+1}}$ (by the fixed-point property of the linear Bellman operator)
$= v_{\pi_*} - T_{\pi_{k+1}} v_{\pi_k} + T_{\pi_{k+1}} v_{\pi_k} - T_{\pi_{k+1}} v_{\pi_{k+1}}$
$= v_{\pi_*} - (1-\alpha_{k+1}) v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k} + (r + \gamma P_{\pi_{k+1}} v_{\pi_k}) - (r + \gamma P_{\pi_{k+1}} v_{\pi_{k+1}})$
$= v_{\pi_*} - (1-\alpha_{k+1}) v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k} + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$
$= (1-\alpha_{k+1}) v_{\pi_*} + \alpha_{k+1} T_{\pi_*} v_{\pi_*} - (1-\alpha_{k+1}) v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k} + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$
(by the fixed-point property of the linear Bellman operator)
$= (1-\alpha_{k+1})(v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1}(T_{\pi_*} v_{\pi_*} - T_{\pi_*} v_{\pi_k}) + \alpha_{k+1}(T_{\pi_*} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k}) + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$
$\leq (1-\alpha_{k+1})(v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1}(T_{\pi_*} v_{\pi_*} - T_{\pi_*} v_{\pi_k}) + \alpha_{k+1}(\max_{\pi'} T_{\pi'} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k}) + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$ (because $\max_{\pi'} T_{\pi'} v_{\pi_k} \geq T_{\pi_*} v_{\pi_k}$)
$= (1-\alpha_{k+1})(v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1}(r + \gamma P_{\pi_*} v_{\pi_*} - r - \gamma P_{\pi_*} v_{\pi_k}) + \alpha_{k+1} e_{k+1} + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$
$= (1-\alpha_{k+1}) I (v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1} \gamma P_{\pi_*}(v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1} e_{k+1} + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$
$= [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1} e_{k+1} + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$.

Now, by manipulating the linear Bellman operator, we get that $v_{\pi_{k+1}} = r + \gamma P_{\pi_{k+1}} v_{\pi_{k+1}}$
$\implies v_{\pi_{k+1}} - \gamma P_{\pi_{k+1}} v_{\pi_{k+1}} = r$
$\implies (I - \gamma P_{\pi_{k+1}}) v_{\pi_{k+1}} = r$
$\implies v_{\pi_{k+1}} = (I - \gamma P_{\pi_{k+1}})^{-1} r \geq 0$

Therefore, $v_{\pi_k} - v_{\pi_{k+1}} = (I - \gamma P_{\pi_{k+1}})^{-1}(I - \gamma P_{\pi_{k+1}}) v_{\pi_k} - (I - \gamma P_{\pi_{k+1}})^{-1} r$
$= (I - \gamma P_{\pi_{k+1}})^{-1}(v_{\pi_k} - \gamma P_{\pi_{k+1}} v_{\pi_k} - r)$
$= (I - \gamma P_{\pi_{k+1}})^{-1}(T_{\pi_k} v_{\pi_k} - T_{\pi_{k+1}} v_{\pi_k})$
$= (I - \gamma P_{\pi_{k+1}})^{-1}(T_{\pi_k} v_{\pi_k} - (1-\alpha_{k+1}) v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k})$
$= (I - \gamma P_{\pi_{k+1}})^{-1}(T_{\pi_k} v_{\pi_k} - (1-\alpha_{k+1}) T_{\pi_k} v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k})$
$= (I - \gamma P_{\pi_{k+1}})^{-1}(\alpha_{k+1} T_{\pi_k} v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k})$
$= (I - \gamma P_{\pi_{k+1}})^{-1} \alpha_{k+1}(T_{\pi_k} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k})$
$\leq (I - \gamma P_{\pi_{k+1}})^{-1} \alpha_{k+1}(\max_{\pi'} T_{\pi'} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k})$
$= (I - \gamma P_{\pi_{k+1}})^{-1} \alpha_{k+1} e_{k+1}$.
This result bounds the difference between the state value functions corresponding to policies from consecutive iterations of CPI.

Integrating this result into the previously-obtained result, we show $v_{\pi_*} - v_{\pi_{k+1}}$
$\leq [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1} e_{k+1} + \gamma P_{\pi_{k+1}}(v_{\pi_k} - v_{\pi_{k+1}})$
$\leq [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1} e_{k+1} + \gamma P_{\pi_{k+1}}(I - \gamma P_{\pi_{k+1}})^{-1} \alpha_{k+1} e_{k+1}$
$= [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + [I + \gamma P_{\pi_{k+1}}(I - \gamma P_{\pi_{k+1}})^{-1}] \alpha_{k+1} e_{k+1}$
$= [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + [(I - \gamma P_{\pi_{k+1}})(I - \gamma P_{\pi_{k+1}})^{-1} + \gamma P_{\pi_{k+1}}(I - \gamma P_{\pi_{k+1}})^{-1}] \alpha_{k+1} e_{k+1}$
$= [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + [(I - \gamma P_{\pi_{k+1}} + \gamma P_{\pi_{k+1}})(I - \gamma P_{\pi_{k+1}})^{-1}] \alpha_{k+1} e_{k+1}$
$= [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + [I(I - \gamma P_{\pi_{k+1}})^{-1}] \alpha_{k+1} e_{k+1}$
$= [(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1}(I - \gamma P_{\pi_{k+1}})^{-1} e_{k+1}$

We now have a recurrence inequality that, for an arbitrary iteration k of CPI, relates $v_{\pi_*} - v_{\pi_{k+1}}$ to $v_{\pi_*} - v_{\pi_k}$. Hence, we can inductively construct an explicit inequality that bounds $v_{\pi_*} - v_{\pi_k}$ from above. First, we define the matrix $Q_k = (1-\alpha_k) I + \alpha_k \gamma P_{\pi_*}$, which is clearly non-negative, since the quantities being multiplied and added $(1-\alpha_k), I, \alpha_k, \gamma, P_{\pi_*}$ are all non-negative. (Importantly, Q_k being non-negative is what allows us to preserve the direction of the inequality in the inductive

steps below.)
$v_{\pi_*} - v_{\pi_k} \le [(1-\alpha_k)I + \alpha_k \gamma P_{\pi_*}](v_{\pi_*} - v_{\pi_{k-1}}) + \alpha_k (I - \gamma P_{\pi_k})^{-1} e_k$
$= Q_k(v_{\pi_*} - v_{\pi_{k-1}}) + \alpha_k (I - \gamma P_{\pi_k})^{-1} e_k$
$\le Q_k(Q_{k-1}(v_{\pi_*} - v_{\pi_{k-2}}) + \alpha_{k-1}(I - \gamma P_{\pi_{k-1}})^{-1} e_{k-1}) + \alpha_k (I - \gamma P_{\pi_k})^{-1} e_k$
$\le Q_k(Q_{k-1}(Q_{k-2}(v_{\pi_*} - v_{\pi_{k-3}}) + \alpha_{k-2}(I - \gamma P_{\pi_{k-2}})^{-1} e_{k-2}) + \alpha_{k-1}(I - \gamma P_{\pi_{k-1}})^{-1} e_{k-1}) + \alpha_k (I - \gamma P_{\pi_k})^{-1} e_k$
$= Q_k Q_{k-1} Q_{k-2}(v_{\pi_*} - v_{\pi_{k-3}}) + Q_k Q_{k-1} \alpha_{k-2}(I - \gamma P_{\pi_{k-2}})^{-1} e_{k-2} + Q_k \alpha_{k-1}(I - \gamma P_{\pi_{k-1}})^{-1} e_{k-1} + \alpha_k (I - \gamma P_{\pi_k})^{-1} e_k$
$\le (\prod_{i=0}^{k-1} Q_{k-i})(v_{\pi_*} - v_{\pi_0}) + \sum_{i=0}^{k-1} (\prod_{j=0}^{i-1} Q_{k-j}) \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i}$ (by induction)
Note: We define $\prod_{j=0}^{-1} Q_{k-j} = I$.

Let's now manipulate the $(\prod_{i=0}^{k-1} Q_{k-i})(v_{\pi_*} - v_{\pi_0})$ term. We assume the initialization $v_{\pi_0} = 0$. Then, $Q_1(v_{\pi_*} - v_{\pi_0})$
$= Q_1 v_{\pi_*}$
$= ((1-\alpha_1)I + \alpha_1 \gamma P_{\pi_*}) v_{\pi_*}$
$= (1-\alpha_1) v_{\pi_*} + \alpha_1 \gamma (P_{\pi_*} v_{\pi_*})$
$= (1-\alpha_1) v_{\pi_*} + \alpha_1 \gamma \frac{v_{\pi_*} - r}{\gamma}$ (by the fixed-point property of the linear Bellman operator, $v_{\pi_*} = r + \gamma P_{\pi_*} v_{\pi_*}$)
$\le (1-\alpha_1) v_{\pi_*} + \alpha_1 \gamma v_{\pi_*}$
$= (1 - \alpha_1 + \gamma \alpha_1) v_{\pi_*}$
$= (1 - (1-\gamma)\alpha_1) v_{\pi_*}$
We denote $\beta_i = 1 - (1-\gamma)\alpha_i$. So, $Q_1(v_{\pi_*} - v_{\pi_0}) \le \beta_1 v_{\pi_*}$. By similar reasoning, $Q_2 Q_1 (v_{\pi_*} - v_{\pi_0}) \le Q_2(\beta_1 v_{\pi_*}) = \beta_1(Q_2 v_{\pi_*}) \le \beta_1 \beta_2 v_{\pi_*}$. Hence, by induction, $(\prod_{i=0}^{k-1} Q_{k-i})(v_{\pi_*} - v_{\pi_0}) \le (\prod_{i=1}^{k} \beta_i) v_{\pi_*} \le (\prod_{i=1}^{k} \beta_i) V_{\text{max}}$. Lastly, we denote $\delta_k = \prod_{i=1}^{k} \beta_i$. Therefore, $(\prod_{i=0}^{k-1} Q_{k-i})(v_{\pi_*} - v_{\pi_0}) \le \delta_k V_{\text{max}}$.

Plugging this result back into our previously-obtained inequality, $v_{\pi_*} - v_{\pi_k}$
$\le (\prod_{i=0}^{k-1} Q_{k-i})(v_{\pi_*} - v_{\pi_0}) + \sum_{i=0}^{k-1} (\prod_{j=0}^{i-1} Q_{k-j}) \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i}$
$\le \delta_k V_{\text{max}} + \sum_{i=0}^{k-1} (\prod_{j=0}^{i-1} Q_{k-j}) \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i}$

We now turn our efforts to re-expressing $\prod_{j=0}^{i-1} Q_{k-j}$. Denote the ordered set $\mathcal{N}_{i,k} = \{k > \cdots > k-i+1\}$, which contains exactly i elements. Furthermore, we define the operator $\mathcal{P}_j(\mathcal{N}_{i,k})$ to return the set of ordered subsets of $\mathcal{N}_{i,k}$ of size j such that each element of $\mathcal{P}_j(\mathcal{N}_{i,k})$ preserves the ordering of $\mathcal{N}_{i,k}$.

We first consider the case $i = 0$, in which $\prod_{j=0}^{i-1} Q_{k-j} = \prod_{j=0}^{-1} Q_{k-j} = I$
$= \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j$
$= \sum_{j=0}^{0} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{0,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{0,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j$
$= \sum_{I \in \{\{\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{\} \setminus I} (1-\alpha_n)) I$
$= (\prod_{n \in \{\}} \alpha_n)(\prod_{n \in \{\}} (1-\alpha_n)) I$
$= 1 \cdot 1 \cdot I$
$= I.$
Next, we consider the case $i = 1$, in which $\prod_{j=0}^{i-1} Q_{k-j} = \prod_{j=0}^{0} Q_{k-j} = Q_k$
$= \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j$
$= \sum_{j=0}^{1} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{1,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{1,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j$
$= \sum_{I \in \{\{\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{k\} \setminus I} (1-\alpha_n)) I + \sum_{I \in \{\{k\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{k\} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})$
$= (\prod_{n \in \{\}} \alpha_n)(\prod_{n \in \{k\}} (1-\alpha_n)) I + (\prod_{n \in \{k\}} \alpha_n)(\prod_{n \in \{\}} (1-\alpha_n))(\gamma P_{\pi_*})$
$= 1 \cdot (1-\alpha_k) \cdot I + \alpha_k \cdot 1 \cdot (\gamma P_{\pi_*})$
$= (1-\alpha_k) I + \alpha_k \gamma P_{\pi_*}$
$= Q_k.$
Subsequently, we consider the case $i = 2$, in which $\prod_{j=0}^{i-1} Q_{k-j} = \prod_{j=0}^{1} Q_{k-j} = Q_k Q_{k-1}$
$= \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j$

$$
\begin{aligned}
&= \sum_{j=0}^{2} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{2,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{2,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j \\
&= \sum_{I \in \{\{\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{k,k-1\} \setminus I} (1-\alpha_n)) I \\
&+ \sum_{I \in \{\{k-1\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{k,k-1\} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*}) \\
&+ \sum_{I \in \{\{k\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{k,k-1\} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*}) \\
&+ \sum_{I \in \{\{k,k-1\}\}} (\prod_{n \in I} \alpha_n)(\prod_{n \in \{k,k-1\} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^2 \\
&= (\prod_{n \in \{\}} \alpha_n)(\prod_{n \in \{k,k-1\}} (1-\alpha_n)) I \\
&+ (\prod_{n \in \{k-1\}} \alpha_n)(\prod_{n \in \{k\}} (1-\alpha_n))(\gamma P_{\pi_*}) \\
&+ (\prod_{n \in \{k\}} \alpha_n)(\prod_{n \in \{k-1\}} (1-\alpha_n))(\gamma P_{\pi_*}) \\
&+ (\prod_{n \in \{k,k-1\}} \alpha_n)(\prod_{n \in \{\}} (1-\alpha_n))(\gamma P_{\pi_*})^2 \\
&= 1 \cdot (1-\alpha_k)(1-\alpha_{k-1}) \cdot I \\
&+ \alpha_{k-1} \cdot (1-\alpha_k) \cdot (\gamma P_{\pi_*}) \\
&+ \alpha_k \cdot (1-\alpha_{k-1}) \cdot (\gamma P_{\pi_*}) \\
&+ \alpha_k \alpha_{k-1} \cdot 1 \cdot (\gamma P_{\pi_*})^2 \\
&= [(1-\alpha_k) I + \alpha_k \gamma P_{\pi_*}][(1-\alpha_{k-1}) I + \alpha_{k-1} \gamma P_{\pi_*}] \\
&= Q_k Q_{k-1}.
\end{aligned}
$$

Hence, by induction, it can be shown that $\prod_{j=0}^{i-1} Q_{k-j} = \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j$. In summary, $v_{\pi_*} - v_{\pi_k} \leq \delta_k V_{\max} + \sum_{i=0}^{k-1} (\sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j) \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i}$.

Recall that for an arbitrary policy π, $d_{\pi,\nu} = (1-\gamma)\nu(I - \gamma P_\pi)^{-1}$, which is called the discounted cumulative occupancy measure induced by π when starting from ν. Furthermore, by definition, $c(i)$ is the smallest coefficient in $[1, \infty) \cup \{\infty\}$ such that for all i and all sets of deterministic stationary policies $\pi_1, \pi_2, \ldots, \pi_i, \mu P_{\pi_1} P_{\pi_2} \cdots P_{\pi_i} \leq c(i)\nu$.

$$
\begin{aligned}
&\mu(v_{\pi_*} - v_{\pi_k}) \leq \mu(\delta_k V_{\text{max}} + \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n))(\gamma P_{\pi_*})^j \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i}) \\
&= \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^j \mu P_{\pi_*}^j \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i} + \delta_k V_{\text{max}} \\
&\leq \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^j c(j) \nu (I - \gamma P_{\pi_{k-i}})^{-1} \alpha_{k-i} e_{k-i} + \delta_k V_{\text{max}} \\
&= \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^j c(j)(1-\gamma)\nu(I - \gamma P_{\pi_{k-i}})^{-1} \alpha_{k-i} e_{k-i} + \delta_k V_{\text{max}} \\
&\leq \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^j c(j) d_{\pi_{k-i},\nu} \alpha_{k-i} e_{k-i} + \delta_k V_{\text{max}} \\
&\leq \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^j c(j) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}}
\end{aligned}
$$

(because, as per the definition of the greedy operator, for CPI, π_{k-i} is such that $d_{\pi_{k-i},\nu}(\max_{\pi'} T_{\pi'} - T_{\pi_{k-1}} v) = d_{\pi_{k-i},\nu} e_{k-i} \leq \epsilon_{k-i}$)

$$
\begin{aligned}
&\leq \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{l=j}^{\infty} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^l c(l) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}} \\
&\leq \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{l=0}^{\infty} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \gamma^l c(l) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}} \\
&= \frac{1}{1-\gamma} (\sum_{l=0}^{\infty} \gamma^l c(l)) \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} (\prod_{n \in I} \alpha_n)(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}} \\
&= \frac{1}{1-\gamma} (\sum_{l=0}^{\infty} \gamma^l c(l)) \sum_{i=0}^{k-1} (\prod_{j \in \mathcal{N}_{i,k}} 1 - \alpha_j + \alpha_j) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}} \\
&= \frac{1}{1-\gamma} (\sum_{l=0}^{\infty} \gamma^l c(l)) \sum_{i=0}^{k-1} \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}} \\
&= \frac{1}{1-\gamma} \cdot \frac{C^{(1,0)}}{1-\gamma} \cdot \sum_{i=0}^{k-1} \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}} \\
&= \frac{C^{(1,0)}}{(1-\gamma)^2} \sum_{i=0}^{k-1} \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\text{max}}
\end{aligned}
$$

By the fact that $\forall x \in (0,1), \log(1-x) \leq -x$, we know that $\log \delta_k = \log \prod_{i=1}^{k} \beta_i = \sum_{i=1}^{k} \log \beta_i = \sum_{i=1}^{k} \log(1 - \alpha_i(1-\gamma)) \leq -(1-\gamma) \sum_{i=1}^{k} \alpha_i$. Because log increases monotonically on its domain, this implies that $\delta_k \leq e^{\{-(1-\gamma)\sum_{i=1}^{k} \alpha_i\}}$. Hence, in conclusion, we prove the key statement of **Theorem 1**.

$$\mu(v_{\pi_*} - v_{\pi_k}) \leq \frac{C^{(1,0)}}{(1-\gamma)^2}\sum_{i=1}^{k}\alpha_i\epsilon_i + e^{\{-(1-\gamma)\sum_{i=1}^{k}\alpha_i\}}V_{max} \tag{11}$$

Visibly, we now have a 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 π output by CPI instead of the optimal policy π_* from some initial distribution μ of interest as a function of an upper bound ϵ on all errors (ϵ_k). Furthermore, clearly, $e^{\{-(1-\gamma)\sum_{i=1}^{k}\alpha_i\}}V_{max}$ tends to 0 exponentially quickly in the number of iterations. We can now make use of **Theorem 1** to prove the following corollary. Denote ϵ such that $\forall i, \epsilon_i \leq \epsilon$.

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:*

$$\mu(v_{\pi_*} - v_{\pi_{k^\dagger}}) \leq (\frac{C^{(1,0)}(\sum_{i=1}^{k^\dagger}\alpha_i)}{(1-\gamma)^2} + 1)\epsilon \leq (\frac{C^{(1,0)}(\log\frac{V_{max}}{\epsilon}+1)}{(1-\gamma)^3} + 1)\epsilon \tag{12}$$

In the analysis of CPI, Kakade and Langford (2002) show that the learning steps that ensure a nice performance guarantee for CPI satisfy $\forall i \leq k^\dagger, \alpha_i \geq \frac{(1-\gamma)\epsilon}{12\gamma V_{max}}$. (Unfortunately, we do not have space for this proof in this paper.) Hence, $\sum_{i=1}^{k^\dagger}\alpha_i \geq \sum_{i=1}^{k^\dagger}\frac{(1-\gamma)\epsilon}{12\gamma V_{max}} = k^\dagger\frac{(1-\gamma)\epsilon}{12\gamma V_{max}}$. Clearly, $k^\dagger\frac{(1-\gamma)\epsilon}{12\gamma V_{max}}$ is a lower bound on $\sum_{i=1}^{k^\dagger}\alpha_i$. Therefore, the smallest random iteration $k^\dagger$ such that it is possible $\sum_{i=1}^{k^\dagger}\alpha_i \geq k^\dagger\frac{(1-\gamma)\epsilon}{12\gamma V_{max}} \geq \frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma}$ satisfies $k^\dagger \leq \frac{12\gamma V_{max}}{(1-\gamma)\epsilon}\cdot\frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma} = \frac{12\gamma V_{max}\log\frac{V_{max}}{\epsilon}}{\epsilon(1-\gamma)^2}$. Furthermore, $\forall i, \alpha_i \leq 1 < \frac{1}{1-\gamma}$. The width of the interval $[\frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma}, \frac{\log\frac{V_{max}}{\epsilon}+1}{1-\gamma}]$ is $\frac{\log\frac{V_{max}}{\epsilon}+1}{1-\gamma} - \frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma} = \frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma} + \frac{1}{1-\gamma} - \frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma} = \frac{1}{1-\gamma} > \alpha_i, \forall i$. Since the width of the interval is larger than $\alpha_i, \forall i$, there must exist a smallest (random) iteration $k^\dagger$ such that $\frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma} \leq \sum_{i=1}^{k^\dagger}\alpha_i \leq \frac{\log\frac{V_{max}}{\epsilon}+1}{1-\gamma}$ (i.e. it is not possible that the smallest (random) iteration $k^\dagger$ such that $\frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma} \leq \sum_{i=1}^{k^\dagger}\alpha_i$ also satisfies $\frac{\log\frac{V_{max}}{\epsilon}+1}{1-\gamma} \leq \sum_{i=1}^{k^\dagger}\alpha_i$). This proves the first part of the corollary.

The second part of the corollary immediately follows from **Theorem 1**. $\mu(v_{\pi_*} - v_{\pi_k})$
$\leq \frac{C^{(1,0)}}{(1-\gamma)^2}\sum_{i=1}^{k}\alpha_i\epsilon_i + e^{\{-(1-\gamma)\sum_{i=1}^{k}\alpha_i\}}V_{max}$
$\leq \epsilon\frac{C^{(1,0)}}{(1-\gamma)^2}\sum_{i=1}^{k}\alpha_i + e^{\{-(1-\gamma)\frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma}\}}V_{max}$
$\leq \epsilon\frac{C^{(1,0)}}{(1-\gamma)^2}(\frac{\log\frac{V_{max}}{\epsilon}+1}{1-\gamma}) + e^{\{-(1-\gamma)\frac{\log\frac{V_{max}}{\epsilon}}{1-\gamma}\}}V_{max}$
$\leq \epsilon\frac{C^{(1,0)}(\log\frac{V_{max}}{\epsilon}+1)}{(1-\gamma)^3} + \epsilon$
$\leq (\frac{C^{(1,0)}(\log\frac{V_{max}}{\epsilon}+1)}{(1-\gamma)^3} + 1)\epsilon$
Thereby, we obtain the key results of **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.

Scherrer notes that the analysis of CPI's performance bound can be done with respect to another concentrability constant C_{π_*}, which is the smallest coefficient in $[1,\infty)\cup\{\infty\}$ such that $d_{\pi_*,\mu} = (1-\gamma)\mu(I-\gamma P_{\pi_*})^{-1} \leq C_{\pi_*}\nu$. (Unfortunately, we do not have space for this proof.) This implies that CPI's performance guarantee can be arbitrarily better than that of API, but results in an exponential increase of time complexity since CPI will then require a number of iterations that scales in $O(\frac{1}{\epsilon^2})$. However, for any MDP and any distribution μ, it is possible to find an input distribution ν for CPI such that C_{π_*} is finite, but this is not the case for $C^{(1,0)}$.

3.3 A Convergent Form of Approximate 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 ϵ-soft and Lipschitz continuous in action values. 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, Γ, that maps this function to a stochastic policy. The central theorem is as follows:

Theorem 1 *For any infinite-horizon Markov decision process satisfying Assumption 1 (below), 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 converges to a unique limiting policy $\pi \in \Pi_\epsilon$, regardless of initial policy π_0.*

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.

We will now set out to establish a series of lemmas to serve as a basis for a contraction mapping argument. We first begin by defining that which we hope to show. A policy π is called *ϵ-soft* if $\pi(s,a) \geq \epsilon$ for all s and a. For any $\epsilon > 0$ let Π_ϵ denote the set of ϵ-soft policies. We observe that Π_ϵ can be thought of as a compact subset of $\mathbb{R}^{mn}$. We now assume:

Assumption 1 *Under any policy π, the Markov decision process behaves as an irreducible, aperiodic Markov chain over the states S.*

Assumption 2 *The columns of feature vector Φ are linearly independent.*

A generic policy improvement operator Γ, which maps every $Q \in \mathbb{R}^{mn}$ to a stochastic policy. Γ is *Lipschitz continuous with constant c* if $\forall$Q$_1, Q_2 \in \mathbb{R}^{mn}$, $\|\Gamma(Q_1) - \Gamma(Q_2)\| \leq c\,\|(Q_1) - (Q_2)\|$. Γ is ϵ-soft if, $\forall$Q $\in \mathbb{R}^{mn}$, $\Gamma(Q)$ is ϵ-soft.

Consider P^π the mn square transition matrix.

Lemma 1 *There exists c_P such that for all π_1, π_2, $\|P_1^\pi - P_2^\pi\| \leq c_P\|\pi_1 - \pi_2\|$.*

Fix π_1 and π_2, let $i = (s,a)$ and $j = (s',a')$.
$\|P_{i,j}^{\pi_1} - P_{i,j}^{\pi_2}\| = \|p_{s,s'}^a(\pi_1(s',a') - \pi_2(s',a'))\|$
$\leq \|\pi_1(s',a') - \pi_2(s',a')\|$
$\leq max_{s',a'}|\pi_1(s',a') - \pi_2(s',a')\| = \|\pi_1 - \pi_2\|_\infty$
$\leq \|\pi_1 - \pi_2\|$
$\implies \|P_1^\pi - P_2^\pi\| \leq \sqrt{mn}\|\pi_1 - \pi_2\|$

We now define μ^π to be the length mn vector whose $(s,a)^{th}$ element is $p^\pi(s)\pi(s,a)$

Lemma 2 *For any $\epsilon > 0$, there exists c_μ such that for all $\pi_1, \pi_2 \in \Pi_\epsilon$, $\|\mu_1^\pi - \mu_2^\pi\| \leq c_\mu\|\pi_1 - \pi_2\|$.*

$\forall \pi \in \Pi_\epsilon$, let λ^π be the largest eigenvalue of P^π with modulus less than 1. λ^π well defined by [7] – furthermore, applying continuity of eigenvalues in the elements of a matrix [8] and compactness of $\Pi_\epsilon, \exists \lambda^m ax = max_{\pi \in \Pi_\epsilon \lambda\pi} = \lambda^{\pi_m ax} < 1$ for some $pi_m ax \in \Pi\epsilon$. Let $\pi_1, \pi_2 \in \Pi_\epsilon$. $\|\mu^{\pi_1} - \mu^{\pi_2}\| \leq \|\mu^{\pi_1} - \mu^{\pi_2}\|_1 \leq \frac{mn}{|1-\lambda^{\pi_1}|}\|P^{\pi_1} - P^{\pi_2}\|_\infty \leq \frac{mn}{|1-\lambda^m ax|}\|P^{\pi_1} - P^{\pi_2}\| \leq \frac{mn}{|1-\lambda^m ax|}c_P\|\pi_1\pi_2\|$. Lastly define D^π to be the matrix with diagonal $\mu^\pi \implies \|\mu_1^\pi - \mu_2^\pi\| \leq c_\mu\|\pi_1 - \pi_2\|$

[6] posits that weights converge to unique solution to equation: $\Phi' D^\pi(I - \gamma P^\pi)\mathbf{\Phi w} = \Phi^\pi \mathbf{r}$.
Let $A^\pi = \Phi' D^\pi(I - \gamma P^\pi)\Phi$ and $b^\pi = \Phi' D^\pi \mathbf{r}$

Lemma 3 *There exists c_b and c_A such that for all π_1, π_2, $\|b_1^\pi - b_2^\pi\| \leq c_b\,\|\pi_1 - \pi_2\|$ and $\|A_1^\pi - A_2^\pi\| \leq c_A\,\|\pi_1 - \pi_2\|$.*

$\|b^{pi_1} - b^{pi_2}\| = \|\Phi'(D^{\pi_1} - D^{\pi_2})\mathbf{r}\| \leq \|\Phi'\|\|D^{\pi_1} - D^{\pi_2}\|\|r\| \leq c_\mu\|\Phi'\|\|\mathbf{r}\|\|\pi_1 - \pi_2\|$
$\|A^{pi_1} - A^{pi_2}\| = \|\Phi'[D_1^\pi(I - \gamma P_1^\pi) - D_2^\pi(I - \gamma P_2^\pi)]\Phi\|$
$\leq \|\Phi'\|\|D_1^\pi(I - \gamma P_1^\pi) - D_2^\pi(I - \gamma P_2^\pi)\|\|\Phi\|$
$= \|\Phi'\|\|D_1^\pi - D_2^\pi - \gamma D^{\pi_1}P^{\pi_1} + \gamma D^{\pi_2}P^{pi_2}\|\|\Phi\|$
$= \|\Phi'\|\|D_1^\pi - D_2^\pi - \gamma D^{\pi_1}(P^{\pi_1} - P^{\pi_2} + P^{\pi_2}) + \gamma D^{\pi_2}P^{pi_2}\|\|\Phi\|$
$= \|\Phi'\|\|D_1^\pi - D_2^\pi - \gamma D^{\pi_1}(P^{\pi_1} - P^{\pi_2}) - \gamma(D^{\pi_1} - D^{\pi_2})P^{pi_2}\|\|\Phi\|$
$\leq \|\Phi'\|(\|D_1^\pi - D_2^\pi\| - \gamma\|D^{\pi_1}\|\|P^{\pi_1} - P^{\pi_2}\| - \gamma\|D^{\pi_1} - D^{\pi_2}\|\|P^{\pi_2}\|)\|\Phi\|$
$\leq ((1+\gamma)c_\mu + \gamma c_P)\|\Phi'\|\|\Phi\|\|\pi_2 - \pi_1\|$, by Lemma 1 and 2 and since $\|D^\pi\| \leq 1$ and $\|P^\pi\| = 1$

Lemma 4 $\forall\epsilon > 0$*, there exists* c_w *such that* $\|\mathbf{w}^\pi\| \leq c_w, \forall\pi \in \Pi_\epsilon$.

From Lemma 1 and 2 and [4], $\mathbf{w}^\pi$ continuous $\implies$ $\|\mathbf{w}^\pi\|$ continuous. Since Π_ϵ compact set, boundedness follows.

Lemma 5 *For any* $\epsilon > 0$*, there exists* $c_g > 0$ *such that for all* $\pi \in \Pi_\epsilon, g(A^\pi) \geq c_g$.

From Lemma 7, g is postive, continuous mapping for any non-singular matrix. For all $\pi \in \Pi_\epsilon, inf_{\pi_1 \in \Pi_\epsilon} g(A^{\pi_1}).\exists\pi_{inf} \in \Pi_\epsilon \implies g(A^\pi) \geq g(A^{\pi_{inf}}) > 0$ since $A^{\pi_{inf}}$ non-singular.

Lemma 6 *For any* $\epsilon > 0$*, there exists* c_{w2} *such that for all* $\pi_1, \pi_2 \in \Pi_\epsilon, \|\mathbf{w^{\pi_1}} - \mathbf{w^{\pi_2}} \leq c_{w2}\|\pi_1 - \pi_2\|$

Consider arbitrary $\pi_1, \pi_2 \in \Pi_\epsilon$, from 3.3, $A^{\pi_1}\mathbf{w}^{\pi_1} = b^{\pi_1}$ and $A^{\pi_2}\mathbf{w}^{\pi_2} = b^{\pi_2}$.
$A^{\pi_1}\mathbf{w}^{\pi_1} - A^{\pi_2}\mathbf{w}^{\pi_2} = b^{\pi_1} - b^{\pi_2} \implies A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2} + \mathbf{w}^{\pi_2}) - A^{\pi_2}\mathbf{w}^{\pi_2} = b^{\pi_1} - b^{\pi_2}$
$\implies A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}) + (A^{\pi_1} - A^{\pi_2})\mathbf{w}^{\pi_2} = b^{\pi_1} - b^{\pi_2}$
$\implies \|A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2})\|$
$\leq \|b^{\pi_1} - b^{\pi_2}\| + \|A^{\pi_1} - A^{\pi_2})\mathbf{w}^{\pi_2}\|$
$\implies c_g\|\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}\| \leq c_b\|\pi_1 - \pi_2\| + c_w c_A\|\pi_1 - \pi_2\|$
$\implies |\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}\| \leq c^-1_g(c_b + c_w c_A)\|\pi_1 - \pi_2\|$

Lemma 7 *For 1-1 matrix M, let g(M)=* $min_{\|x\|=1}\|Mx\|$
Then:

$$g(M) \geq 0\forall M, \tag{13}$$

$$\textit{g(M) > 0 when M non-singular} \tag{14}$$

$$\forall x \in \mathbb{R}, \|Mx\| \geq g(M)\|x\| \tag{15}$$

$$\textit{g is continuous} \tag{16}$$

Contraction Argument Consider infinite horizon discounted Markov decision problem with $\epsilon > 0\Gamma$ fixed and Lipschitz continuous with constant c. Let $\pi_1, \pi_2 \in \Pi_\epsilon$ be arbitrary. Observe that
$\|\Gamma(\hat{Q^{\pi_1}}) - \Gamma(\hat{Q^{\pi_2}})\| \leq c\|\hat{Q^{\pi_1}} - \hat{Q^{\pi_2}}\| = c\|\Phi(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2})\| \leq c\|\Phi\|c_{w2}\|\pi_1 - \pi_2\|$.
If $c < \|\Phi\|^{-1}c_{w2}^{-}1, \exists\beta \in [0,1)$ such that $\|\Gamma(\hat{Q^{\pi_1}}) - \Gamma(\hat{Q^{\pi_2}})\| \leq \beta\|\pi_1 - \pi_2\|$.
Since each iteration of the proposed policy iteration is a contraction, by the Contraction Mapping Theorem, there exists a subsequential limit of the policies generated from Γ operator mapping to a fixed limit.

4 Conclusions and Future Directions

Approximate policy iteration seems to be extremely promising, especially for the computational 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 ϵ-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 ϵ-approximate greedy operator. One potential future direction of research is further unpacking the ϵ-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 Γ. However, it remarkably still guarantees convergence of Approximate Policy Iteration if Γ is both ϵ-soft and Lipschitz continuous. The authors hope this type of guarantee can be extended to similar algorithms such as online SARSA or SARSA(λ), 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 paper would provide a comprehensive overview of current approximate policy iteration schemes, and serve as a valuable resource for future research in this area.

References

[1] Scherrer, B. (2014, January). Approximate policy iteration schemes: a comparison. In International Conference on Machine Learning (pp. 1314-1322).

[2] Perkins, T. J., & Precup, D. (2003). A convergent form of approximate policy iteration. In Advances in neural information processing systems (pp. 1627-1634).

[3] Smirnova, E., & Dohmatob, E. (2019). On the Convergence of Approximate and Regularized Policy Iteration Schemes. arXiv preprint arXiv:1909.09621.

[4] Puterman, M. L. (2014). Markov decision processes: discrete stochastic dynamic programming. John Wiley Sons.

[5] Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., & Geist, M. (2015). Approximate modified policy iteration and its application to the game of Tetris. Journal of Machine Learning Research, 16(49), 1629-1676.

[6] Tsitsiklis, J. N., & Van Roy, B. (1997). Analysis of temporal-difference learning with function approximation. In Advances in neural information processing systems (pp. 1075-1081).

[7] Gordon, G. J. (1999). Approximate Solutions to Markov Decision Processes. PhD thesis, Carnegie Mellon University.

[8] Gordon, G. J. (2001). Reinforcement learning with function approximation converges to a region. In Advances in neural information processing systems (pp. 1040-1046).

[9] Bertsekas, D. P. (2011). Approximate policy iteration: A survey and some new methods. Journal of Control Theory and Applications, 9(3), 310-335.

[10] Farahmand, A. M., Szepesvári, C., & Munos, R. (2010). Error propagation for approximate policy and value iteration. In Advances in Neural Information Processing Systems (pp. 568-576).