On the Convergence of Approximate and Regularized Policy Iteration Schemes

Elena Smirnova
Criteo AI Lab
e.smirnova@criteo.com

Elvis Dohmatob
Criteo AI Lab
e.dohmatob@criteo.com

Abstract

Algorithms based on the entropy regularized framework, such as Soft Q-learning and Soft Actor-Critic, recently showed state-of-the-art performance on a number of challenging reinforcement learning (RL) tasks. The regularized formulation modifies the standard RL objective and thus, generally, converges to a policy different from the optimal greedy policy of the original RL problem. Practically, it is important to control the suboptimality of the regularized optimal policy. In this paper, we propose the optimality-preserving regularized modified policy iteration (MPI) scheme that simultaneously (a) provides desirable properties to intermediate policies such as targeted exploration, and (b) guarantees convergence to the optimal policy with explicit rates depending on the decrease rate of the regularization parameter. This result is based on two more general results. First, we show that the approximate MPI scheme converges as fast as the exact MPI if the decrease rate of error sequence is sufficiently fast; otherwise, its rate of convergence slows down to the errors decrease rate. Second, we show the regularized MPI is an instance of the approximate MPI where regularization plays the role of errors. In a special case of negative entropy regularizer (leading to a popular Soft Q-learning algorithm), our result explicitly links the convergence rate of policy / value iterates to exploration.

1 Introduction

The main principle of the entropy regularized approach to RL [1–4] is to modify the standard RL objective to additionally maximize the (relative) entropy of a policy at each visited state. The resulted regularized objective has shown substantially improved performance and exploration targeted at high-value actions [5]. The regularization parameter, referred to as temperature, controls the relative importance of the exploration versus the reward. In this paper, we view different entropy regularized algorithms from a unified perspective using the recently proposed regularized MPI framework [6].

Despite the empirical success, the algorithms under regularized MPI framework would not generally converge to the optimal policy/value pair. A natural way of controlling the optimality gap is through the regularization parameter that, set at zero, recovers the unregularized objective. Thus, a common idea is to gradually decrease the regularization weight to eventually converge into this regime.

Prior works considered decaying temperature during learning in the context of specific algorithms. [7] proved asymptotic convergence of SARSA with Boltzmann policy and decaying temperature. Experimentally, a linear schedule of the inverse temperature over iterations was used with Soft Q-learning [3] and Dual Averaging algorithms [8]. The authors in [6, D.1] suggest time-varying values of the regularizer weight analogous to the learning rate in the gradient descent approach.

In this work, we show that the regularized MPI scheme with a decreasing schedule for the regularization parameter converges to the optimal policy/value pair of the original RL problem. Furthermore, we establish a relation between the decrease rate of regularization weight and the convergence rate of

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

the regularized MPI to optimality. In the entropy regularized case, this relation corresponds to the exploration-exploitation trade-off since the regularization weight matches the temperature parameter that controls the amount of exploration performed by the policy.

This result is a consequence of two more general contributions:

(1) We derive the convergence rate of the approximate MPI [9] to optimality in terms of the decrease rate of error sequence (Theorem 1),

(2) We derive the convergence rate of the regularized MPI to the optimal solution of the non-regularized RL problem through the reduction to the approximate MPI (Theorem 2).

One consequence of the result (1) is that in the approximate MPI scheme the errors decreasing faster than the discount factor (asymptotically) result in the same convergence rate as the exact MPI. From the result (2), it can be explicitly shown that optimality-preserving temperature schedule of the regularized MPI trades the speed of convergence for exploration.

Paper organisation. In Section 2 we detail the notations and introduce MPI-based algorithmic schemes. In Section 3 we present our results (1) and (2). We discuss the related works in Section 4.

2 Preliminaries

2.1 Notations and terminology

Δ_X will denote the set of probability distributions over finite set (or general measurable space) X and Y^X is a set of mappings from set X to set Y. We consider a Markov decision process (MDP) is a tuple $M := (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ where $\mathcal{S}$ is a state space, $\mathcal{A}$ is a finite action space, $P \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ is the transition kernel so that the probability of the environment moving to state s' after the agent takes action a in state s is $P(s'|s,a)$, accompanied by a reward $r(s,a)$ (assumed to be bounded). We define a stochastic stationary policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. We consider the discounted setting with discount factor $\gamma \in [0,1)$. We define the Bellman operator $\mathcal{T}^\pi$ for any function $V \in \mathbb{R}^{\mathcal{S}}$, $\forall s \in \mathcal{S}$ as follows:

$$[\mathcal{T}^\pi V](s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]\right] = r^\pi(s) + \gamma P^\pi(\cdot|s)V, \tag{1}$$

where $r^\pi \in \mathbb{R}^{\mathcal{S}}$ and $P^\pi \in \Delta_{\mathcal{S}}^{\mathcal{S}}$ are defined by $r^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]$ and $P^\pi(s'|s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[P(s'|s,a)]$. $\mathcal{T}^\pi$ is a γ-contraction in ℓ_∞ norm and its unique fixed-point is V^π:= $\lim_{k\to\infty}(\mathcal{T}^\pi)^k V = V^\pi$, where equality holds component-wise. By denoting $Q_V(s,a) := r(s,a) + \gamma\mathbb{E}_{s'\sim p(s'|s,a)}[V(s')]$, Eq. (1) can be re-written as an inner-product $[\mathcal{T}^\pi V](s = \langle \pi(\cdot|s), Q_V(s,\cdot)\rangle$.

Finally, we define the Bellman max-operator as follows (the max is point-wise)

$$\mathcal{T}^\star V := \max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} \mathcal{T}^\pi V, \tag{2}$$

which again is a γ-contraction in ℓ_∞ norm and its unique fixed-point is the optimal value function $V^\star$. We denote by $\mathcal{G}(V)$ the set of optimal policies that achieve the maximum of Eq. (2) state-wise

$$\mathcal{G}(V) := \arg\max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}} \mathcal{T}^\pi V \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}.$$

Equivalently, this set coincides with the set of optimal policies: $\mathcal{G}(V) = \{\pi : \mathcal{T}^\pi V = \mathcal{T}^\star V\}$.

2.2 Modified policy-iteration schemes

Modified Policy Iteration (MPI) [10]. MPI is a classical dynamic programming algorithm that alternates between policy improvement and (partial) policy evaluation steps. For $m \geq 1$, the MPI algorithm is defined as follows

$$\begin{cases} \pi_{t+1} \in \mathcal{G}(V_t) \\ V_{t+1} = (\mathcal{T}^{\pi_{t+1}})^m V_t, \end{cases} \tag{3}$$

where $m = 1$ corresponds to Value Iteration and $m = \infty$ corresponds to Policy Iteration. Here V_t denotes an approximation of V^{π_t}.

Convex conjugate functions. Following [6], we introduce the regularized RL framework through convex conjugate functions, see e.g. Section 3.3.1 in [11]. For a strongly convex function $\Omega : \Delta_{\mathcal{A}} \to \mathbb{R}$ its convex conjugate $\Omega^\star : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ is given by

$$\Omega^\star(q) = \max_{\pi_s \in \Delta_{\mathcal{A}}} \langle \pi_s, q \rangle - \Omega(\pi_s), \ \forall q \in \mathbb{R}^{\mathcal{A}}, \tag{4}$$

where $\langle \pi_s, q \rangle := \mathbb{E}_{a \sim \pi_s}[q(a)]$, where $\pi_s := \pi(\cdot|s)$. $\Omega(\pi)$ will be used as a shorthand for the vector $(\Omega(\pi_s))_{s \in \mathcal{S}}$. Further, we make use of the weighted regularizer $\Omega_\alpha(\pi) := \alpha\Omega(\pi)$ that, by properties of the convex conjugate, results in $\Omega^*_\alpha(q) := \alpha\Omega^*(q/\alpha)$. Another property of the convex conjugate (*Danskin's Theorem*) is that the maximizer of (4) is given by the gradient of the dual function

$$\nabla\Omega^\star(q) = \arg\max_{\pi_s \in \Delta_{\mathcal{A}}} \langle \pi_s, q \rangle - \Omega(\pi_s). \tag{5}$$

Approximate Modified Policy Iteration (AMPI) [9]. AMPI is an approximate counterpart of (3) that can be seen as a generalization of MPI that allows errors in the policy improvement (ϵ'_t) and policy evaluation (ϵ_t) steps

$$\begin{cases} \pi_{t+1} \in \mathcal{G}_{\epsilon'_{t+1}}(V_t) \\ V_{t+1} = (\mathcal{T}^{\pi_{t+1}} V_t)^m + \epsilon_{t+1}, \end{cases} \tag{6}$$

where $\epsilon_t, \epsilon'_t \in \mathbb{R}^{\mathcal{S}}$ are respectively the evaluation step and the policy improvement step error vectors (one component per state) and $\pi \in \mathcal{G}_{\epsilon'}(V) \iff \forall \pi' \ \mathcal{T}^{\pi'} V \leq \mathcal{T}^{\pi} V + \epsilon'$. AMPI naturally arises from MPI in practical settings with large state and / or action spaces.

Regularized Modified Policy Iteration (reg-MPI) [6]. Similarly to standard Bellman operators (1), we define the regularized Bellman operator [6] as follows

$$\mathcal{T}^\pi_\Omega V := \mathcal{T}^\pi V - \Omega(\pi), \tag{7}$$

and by virtue of (5) the corresponding optimal policy $\mathcal{G}_\Omega(V) \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ is given by

$$\mathcal{G}_\Omega(V) := \arg\max_{\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}} \mathcal{T}^\pi_\Omega V = (\nabla\Omega^\star(Q_V(s, \cdot))_{s \in \mathcal{S}}.$$

Reg-MPI is a formulation of MPI that underlies several state-of-the-art RL algorithms [2, 4, 12]

$$\begin{cases} \pi_{t+1} \leftarrow \mathcal{G}_{\Omega_t}(V_t) \\ V_{t+1} \leftarrow (\mathcal{T}^{\pi_{t+1}}_{\Omega_t})^m V_t. \end{cases} \tag{8}$$

Negative entropy regularizer. A practically important instance of the reg-MPI scheme corresponds to the negative entropy regularizer $\Omega_t(\pi(\cdot|s)) = \lambda_t \Omega_{\text{Ent}}(\pi(\cdot|s))$ for a time-varying temperature parameter $\lambda_t > 0$, with $\Omega_{\text{Ent}}(\pi(\cdot|s)) = \sum_a \pi(a|s) \log \pi(a|s)$. Its convex conjugate is the smoothed maximum $\Omega^\star_t(Q_V(s, \cdot)) = \lambda_t \log \sum_a \exp(Q_V(s, a)/\lambda_t)$ and the maximizing policy is given by the Boltzmann policy $\pi_{t+1}(\cdot|s) = \nabla\Omega^\star_t(Q_{V_t}(s, \cdot))$ given by

$$\pi_{t+1}(a|s) = \frac{\exp(Q_{V_t}(s, a)/\lambda_t)}{\sum_{a'} \exp(Q_{V_t}(s, a')/\lambda_t)}. \tag{9}$$

With this regularization and $m = 1$, the reg-MPI scheme (8) describes the core principle of the Soft Q-learning algorithm [3, 2].

3 Contributions: Error analysis and convergence rates of AMPI algorithms

We now present the main contributions of this work, namely a fine-grained error analysis of AMPI-type algorithms, including sufficient conditions for convergence, with explicit convergence rates.

3.1 General AMPI algorithms

The error propagation analysis links the error sequence that occurred at previous iterations to the distance to optimality of the current value iterate. In the following Lemma, we restate the error propagation bounds of AMPI established in [9, Theorem 7].

Lemma 1 (AMPI error propagation [9]). *For any initial value function V_0 and $m \geq 1$, consider the AMPI scheme* (6)*. Then, one has*

$$\|V_N - V^*\|_\infty \leq \frac{2}{1-\gamma}\left(E_N + \gamma^N \|V_0 - V^*\|_\infty\right), \tag{10}$$

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

Thus, convergence of the AMPI algorithm entirely depends on controlling the cumulative error term E_N in the above Lemma. In the next theorem, we show the general convergence of AMPI if the sequence of sums of evaluation step and improvement step errors $\|\epsilon_N\|_\infty + \|\epsilon'_N\|_\infty$ converge to zero. By analysing the decrease rate of the error sequence, we provide explicit rates of convergence of AMPI value iterates to the optimal value function.

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* $r_N \longrightarrow 0$*. Then, the AMPI scheme* (6) *converges to the optimal greedy policy of the exact MPI* (3)*.*

Furthermore, define $\underline{\rho} := \liminf_{N\longrightarrow\infty} r_N/r_{N-1}$ *and* $\overline{\rho} := \limsup_{N\longrightarrow\infty} r_N/r_{N-1}$.

(A) If $\underline{\rho} > \gamma$*, then* $\|V_N - V^*\|_\infty = \mathcal{O}(r_N)$.

(B) If $\overline{\rho} \leq \gamma$*, then* $\|V_N - V^*\|_\infty = \begin{cases} \mathcal{O}(\gamma^N), & \text{if } \overline{\rho} < \gamma, \\ \mathcal{O}(N\gamma^N), & \text{if } \overline{\rho} = \gamma. \end{cases}$

We note that the conditions in the theorem are not restrictive. For example, the maximum error can decrease as slow as inverse logarithmically in the number of iterations and still eventually yield an optimal policy at the rate given in Theorem 1(A). A similar story holds for estimates of the form $r_N \propto 1/N$; $r_N \propto 1/\sqrt{N}$; $r_N \propto 1/\log N$; $r_N \propto \log N/N$; etc. where $\underline{\rho} = \overline{\rho} = 1 > \gamma$. On the other hand, if the error sequences decrease at at rate which is (asymptotically) less than the discount factor γ, then the AMPI converges at the same linear rate as the exact MPI! Fig. 1 illustrates these bounds.

Proof of Theorem 1. The proof is based on basic properties of convergent sequences and series.

General convergence. Since $r_t \longrightarrow 0$, it follows that for any $\delta > 0$, $r_t \lesssim \delta$ (where the symbol "$a_t \lesssim b_t$" means that $a_t \leq b_t$ for sufficiently large t). Thus for sufficiently large N, one has

$$E_N := \sum_{t=1}^{N-1} \gamma^{N-t}(\|\epsilon_t\|_\infty + \|\epsilon'_t\|_\infty) \leq C \sum_{t=1}^{N-1} \gamma^{N-t} r_t \lesssim C\delta \sum_{t=1}^{N-1} \gamma^{N-t} \leq C\frac{\gamma}{1-\gamma}\delta.$$

Thus $E_N \longrightarrow 0$ in the limit $N \to \infty$, and by virtue of the bound (10) of Lemma 1 the algorithm converges to the optimal value function V^* as claimed.

Convergence with explicit rates. We now establish the explicit rates of convergence claimed in the theorem under corresponding additional assumptions.

(A) Suppose $\underline{\rho} := \liminf_{N\longrightarrow\infty} r_N/r_{N-1} > \gamma$. For sufficiently large $t \leq N$, we have $r_N \geq \underline{\rho} r_{N-1} \geq \ldots \geq \underline{\rho}^{N-t} r_t$ and so $r_t \leq r_N \underline{\rho}^{-(N-t)}$. So, for large N, one computes

$$\begin{aligned} E_N &:= \textstyle\sum_{t=1}^{N-1} \gamma^{N-t}(\|\epsilon_t\|_\infty + \|\epsilon'_t\|_\infty) \leq C \sum_{t=1}^{N-1} \gamma^{N-t} r_t \lesssim C \sum_{t=1}^{N-1} \gamma^{N-t} r_N \underline{\rho}^{-(N-t)} \\ &= C r_N \textstyle\sum_{t=1}^{N-1} (\gamma/\underline{\rho})^{N-t} = C r_N (\gamma/\underline{\rho}) \frac{1-(\gamma/\underline{\rho})^N}{1-\gamma/\underline{\rho}} \lesssim \frac{C\gamma}{\underline{\rho}-\gamma} r_N = \mathcal{O}(r_N). \end{aligned}$$

(B) Suppose $\overline{\rho} := \limsup_{N\longrightarrow\infty} r_N/r_{N-1} < \gamma$. Then for sufficiently large $t \leq N$, it holds that $r_t \leq \overline{\rho} r_{t-1} \leq \ldots \leq \overline{\rho}^{t-1} r_1$. Thus for sufficiently large N, one has

$$\begin{aligned} E_N &\leq C \textstyle\sum_{t=1}^{N-1} \gamma^{N-t} r_t \lesssim C \sum_{t=1}^{N-1} \gamma^{N-t} r_1 \overline{\rho}^{t-1} = C r_1 \gamma^{N-1} \sum_{t=0}^{N-2} (\overline{\rho}/\gamma)^t \\ &= C r_1 \gamma^{N-1} \tfrac{1-(\overline{\rho}/\gamma)^{N-1}}{1-\overline{\rho}/\gamma} \lesssim \tfrac{C\gamma}{\gamma-\overline{\rho}} r_1 \gamma^{N-1} = \mathcal{O}(\gamma^N). \end{aligned}$$

Finally, if $\overline{\rho} = \gamma$, then similar arguments yield $E_N \lesssim C r_1 \gamma^{N-1} \sum_{t=0}^{N-2} 1 = \mathcal{O}(N\gamma^N)$. □

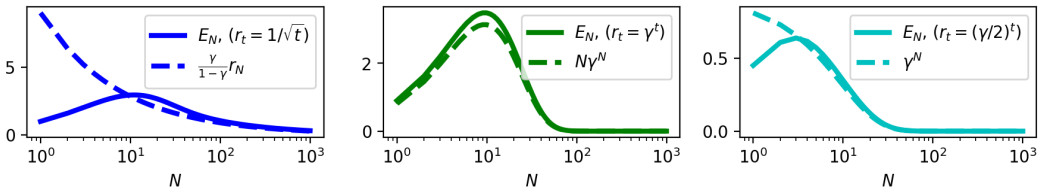

Figure 1: Illustration of the bounds established in Theorem 1, for different regimes of the per-iteration error bounds r_t. In these illustrations, we plugged $\gamma = 0.9$. We see that our proposed upper bounds are quite tight.

3.2 Regularized MPI algorithms

We first show that the reg-MPI (8) is an instance of the AMPI (6). Then, using Theorem 1, we bound the distance between the value iterates of the reg-MPI and the optimal solution of the exact MPI (3). Without loss of generality, we will consider the reg-MPI scheme with weighted regularizer.

Theorem 2 (Reg-MPI convergence)**.** *Consider the reg-MPI algorithm* (8) *with time-varying regularization functions Ω_t, and let the sequence $(\alpha_t)_t$ which uniformly boundes Ω_t, that is*

$$\sup_{\pi} \|\Omega_t(\pi)\|_\infty := \sup_{\pi,\, s} |\Omega(\pi(\cdot|s))| \leq \alpha_t. \tag{11}$$

Then it holds that

$$\|V_{N,\Omega} - V^*\|_\infty \leq \frac{2}{1-\gamma}\left(A_N + \gamma^N \|V_{0,\Omega} - V^*\|_\infty\right), \tag{12}$$

where $A_N := (1 + \frac{1-\gamma^m}{1-\gamma}) \sum_{t=1}^{N-1} \gamma^{N-t}\alpha_t$. Moreover, if $\alpha_t \longrightarrow 0$, then the algorithm converges to the optimal value function V^*.

Furthermore, let $\underline{\rho} := \liminf_{N\longrightarrow\infty} \alpha_N/\alpha_{N-1}$ and $\overline{\rho} := \limsup_{N\longrightarrow\infty} \alpha_N/\alpha_{N-1}$.

- *(A) If $\underline{\rho} > \gamma$, then the algorithm converges to the optimal value function V^* with the same rate as the step-sizes: $\|V_{N,\Omega} - V^*\|_\infty = \mathcal{O}(\alpha_N)$.*
- *(B) If $\overline{\rho} \leq \gamma$, then the algorithm converges to the optimal value function V^* at same rate as the exact MPI* (3) *(i.e linear rate of convergence). More precisely,*

$$\|V_{N,\Omega} - V^*\|_\infty = \begin{cases} \mathcal{O}(\gamma^N), & \text{if } \overline{\rho} < \gamma, \\ \mathcal{O}(N\gamma^N), & \text{if } \overline{\rho} = \gamma. \end{cases}$$

It should be noted that the condition (11) is satisfied by time-varying rgularizers of the form $\Omega_t = \alpha_t\Omega$, for some uniformly bounded Ω. These include the negative entropy regularizer as a special case.

Proof of Theorem 2. We proceed by bounding the policy evaluation and policy improvement step errors of the reg-MPI (8) with respect to the exact MPI (3) defined as in the AMPI (6).

Step 1: bound evaluation step error $\|\epsilon_t\|_\infty$. To begin, it is easy to prove by induction on m (see Appendix A) that for every policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and value function $V \in \mathbb{R}^{\mathcal{S}}$ and one has the formula

$$(\mathcal{T}_\Omega^\pi)^m V = (\mathcal{T}^\pi)^m V - \sum_{j=0}^{m-1} \gamma^j (P^\pi)^j \Omega(\pi), \tag{13}$$

where $(P^\pi)^j$ is the jth power of the matrix P^π. Thus one has

$$\begin{aligned} \|\epsilon_t\|_\infty &= \|V_{t,\Omega} - (\mathcal{T}^{\pi_t})^m V_t\|_\infty = \|(\mathcal{T}_{\Omega_t}^{\pi_t})^m V_t - (\mathcal{T}^{\pi_t})^m V_t\|_\infty = \left\| \sum_{j=0}^{m-1} \gamma^j (P^{\pi_t})^j \Omega_t(\pi_t) \right\|_\infty \\ &\leq \sum_{j=0}^{m-1} \gamma^j \|(P^{\pi_t})^j \Omega_t(\pi_t)\|_\infty \leq \sum_{j=0}^{m-1} \gamma^j \|\Omega_t(\pi_t)\|_\infty = \frac{1-\gamma^m}{1-\gamma} \|\Omega_t(\pi_t)\|_\infty \leq \frac{1-\gamma^m}{1-\gamma} \alpha_t, \end{aligned} \tag{14}$$

where the last inequality follows from the Cauchy-Schwartz inequality

$$\|(P^{\pi})^j \Omega_t(\pi_t)\|_\infty = \max_s |(P^{\pi})^j(\cdot|s)\Omega_t(\pi_t)| \leq \max_s \|(P^{\pi_t})^j(\cdot|s)\|_1 \|\Omega_t(\pi_t)\|_\infty = \|\Omega_t(\pi_t)\|_\infty,$$

since $\|(P^{\pi_t})^j(\cdot|s)\|_1 = 1$ because $(P^{\pi_t})^j(\cdot|s)$ is a probability distribution (over next states).

Step 2: bound policy improvement step error $\|\epsilon'_t\|_\infty$. Using elementary properties of the max operator and definition of the regularized operator $\mathcal{T}^{\pi}_{\Omega}$, one has

$$\|\epsilon'_t\|_\infty = \|\max_\pi \mathcal{T}^{\pi} V_t - \max_\pi \mathcal{T}^{\pi}_{\Omega_t} V_t\|_\infty \leq \max_\pi \|\mathcal{T}^{\pi} V_t - \mathcal{T}^{\pi}_{\Omega_t} V_t\|_\infty = \max_\pi \|\Omega_t(\pi)\|_\infty \leq \alpha_t. \quad (15)$$

By combining per-iteration error bounds (14) and (15) and using Lemma 1, one obtains (12). From this bound and Theorem 1 invoked with $r_t := \alpha_t$ and $C = 1 + \frac{1-\gamma^m}{1-\gamma}$, we get that the algorithm reg-MPI (8) converges to the optimal value function V^*, with the claimed rates of convergence. □

We note that in the case of negative entropy regularizer, the step-sizes α_t correspond to a dynamic temperature parameter and the temperature controls the amount of exploration performed by the policy. From (12) it is apparent that exploration is performed in exchange for slower convergence if the decrease rate is greater than the discount factor γ. On the other hand, if the decrease rate is fast enough (smaller than the discount factor asymptotically), then exploration does not impact the convergence rate that matches the rate of the exact MPI. To the best of our knowledge, this result is the first to relate non-asymptotic performance of MPI to exploration.

One limitation of Theorem 2 is that it does not provide a specific weight schedule to a problem at hand. Indeed, the amount of necessary exploration depends on the MDP structure. Too fast temperature decay implies no regularization and leads to insufficient exploration. On the other hand, too slow temperature decay results in too strong regularization and unnecessary slow convergence. This trade-off has also been shown empirically on a class of entropy regularized algorithms in [8].

4 Related works

We first discuss closely related work [7, 13]. [7] proves convergence to optimality of the SARSA algorithm with GLIE policies ("greedy in the limit with infinite exploration") that include a class of Boltzmann policies with decaying temperature. Another close work [13] studies convergence to optimality of value iteration algorithm with dynamic Boltzmann operator that represents an instance of the reg-MPI scheme with $m = 1$, negative entropy regularizer and decreasing temperature schedule. Our work is different from the above-cited work since (1) we consider MPI-based algorithms, (2) our result on convergence rate holds over a class of the approximate MPI and the regularized MPI algorithms, and (3) we link the exploration in the regularized MPI to its convergence rate through the schedule of regularization parameter.

The optimization perspective on the regularized MDP framework proposed by [8, 6] allows the learning rate interpretation of the regularization weight. In [6, D.1] the regret of the weighted regularized MPI scheme is analysed when it is subject to approximations. Our work is different in that we consider the regularization itself as errors in the approximate MPI scheme.

The temperature schedules obtained in Section 3.1 have similarities with the decrease factors of the Boltzmann exploration in the multi-arm bandit setting, e.g. $\mathcal{O}(1/N)$ and $\mathcal{O}(\log N/N)$ are frequently used [14, 2.2]. Recently, it was shown that temperature schedules of the form $\mathcal{O}(1/\sqrt{N})$ induce near-optimal performance [15]. Despite these similarities, exploration in the RL setup is not as well understood as in bandits setting; our work contributes by providing a link to the convergence rate.

5 Conclusion

Following the success of entropy-regularized methods in RL, we study the convergence to optimality of a class of dynamic programming algorithms unified under the regularized MPI scheme. By the means of reduction to the approximate MPI, we showed the general convergence of this scheme to the solution of the original RL problem under decreasing schedule of the regularization parameter over iterations. Moreover, our analysis showed that the convergence of the regularized MPI is as fast as the exact MPI, if the regularization decay rate is large enough, but otherwise slows down to the decay rate of regularization parameter.

References

[1] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[2] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1352–1361, 2017.

[3] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 202–211, 2016.

[4] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2775–2785, 2017.

[5] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[6] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2160–2169, 2019.

[7] Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308, 2000.

[8] Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

[9] Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *Journal of Machine Learning Research*, 16:1629–1676, 2015.

[10] Martin L Puterman. *Markov decision processes*. Wiley, New York, 1994.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[12] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

[13] Ling Pan, Qingpeng Cai, Qi Meng, Longbo Huang, and Tie-Yan Liu. Reinforcement learning with dynamic boltzmann softmax updates. *arXiv preprint arXiv:1903.05926*, 2019.

[14] Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer, 2005.

[15] Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems*, pages 6284–6293, 2017.

A Supplement to proofs

Proof of formula (13). Let π be a policy and V be a value function. By (7), one has $\mathcal{T}_\Omega^\pi V = \mathcal{T}^\pi V - \Omega(\pi) = \mathcal{T}^\pi V - \gamma^0 (P^\pi)^0 \Omega(\pi)$, and so the formula is valid for $m = 1$ step. Now suppose the formula (13) is valid for m steps. Then

$$\begin{aligned}
(\mathcal{T}_\Omega^\pi)^{m+1} V &= \mathcal{T}_\Omega^\pi((\mathcal{T}_\Omega^\pi)^m V) = \mathcal{T}^\pi((\mathcal{T}_\Omega^\pi)^m V) - \Omega(\pi) = r^\pi + \gamma P^\pi (\mathcal{T}_\Omega^\pi)^m V - \Omega(\pi) \\
&= r^\pi + \gamma P^\pi \left((\mathcal{T}^\pi)^m V - \sum_{j=0}^{m-1} \gamma^j (P^\pi)^j \Omega(\pi) \right) - \Omega(\pi) \\
&= r^\pi + \gamma P^\pi (\mathcal{T}^\pi)^m V - \gamma P^\pi \sum_{j=0}^{m-1} \gamma^j (P^\pi)^j \Omega(\pi) - \Omega(\pi) \\
&= \mathcal{T}^\pi((\mathcal{T}^\pi)^m V) - \sum_{j=0}^{m} \gamma^j (P^\pi)^j \Omega(\pi) = (\mathcal{T}^\pi)^{m+1} V - \sum_{j=0}^{m} \gamma^j (P^\pi)^j \Omega(\pi),
\end{aligned}$$

which is the formula (13) for $m + 1$ steps. □