Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

DATA-DRIVEN SCHEMES FOR RESOLVING MISSPECIFIED MDPS: ASYMPTOTICS AND ERROR ANALYSIS

Hao Jiang

Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
104 S. Mathews Ave.
Urbana, IL 61801, USA

Uday V. Shanbhag

Industrial and Manufacturing Engineering
Pennsylvania State University
310 Leonhard Building
University Park, PA 16802, USA

ABSTRACT

We consider the solution of a finite-state infinite horizon Markov Decision Process (MDP) in which both the transition matrix and the cost function are misspecified, the latter in a parametric sense. We consider a data-driven regime in which the learning problem is a stochastic convex optimization problem that resolves misspecification. Via such a framework, we make the following contributions: (1) We first show that a *misspecified* value iteration scheme converges almost surely to its true counterpart and the mean-squared error after K iterations is $\mathcal{O}(1/K^{1/2-\alpha})$ with $0 < \alpha < 1/2$; (2) An analogous asymptotic almost-sure convergence statement is provided for *misspecified* policy iteration; and (3) Finally, we present a constant steplength *misspecified* Q-learning scheme and show that a suitable error metric is $\mathcal{O}(1/K^{1/2-\alpha}) + \mathcal{O}(\sqrt{\delta})$ with $0 < \alpha < 1/2$ after K iterations where δ is a bound on the steplength.

1 INTRODUCTION

Markov decision processes (MDPs) are an important class of models for analyzing dynamic decision making problems. First examined by Bellman (1957), such models have been used in a number of domains including robotics, control-theory, economics, healthcare, and manufacturing . Specially, a Markov decision process is a discrete time stochastic control process. At each time step, the process is in some state s, and the decision maker may choose an action a that is available in state s. The process responds at the next time step by moving to a new state s', and giving the decision maker a corresponding reward $R_a(s,s')$ or cost $C_a(s,s')$. The next state s' depends on the current state s and the decision maker's action a, but given s and a, it is conditionally independent of all previous states and actions; in other words, the state transitions of an MDP have the Markov property. In an MDP with a discrete state space, the state transition probabilities from time t to $t+1$ are specified by an action U_t at time t, i.e., $\mathbb{P}(s' \mid s,a) \triangleq \mathbb{P}(X_{t+1} = s' \mid X_t = s, U_t = a)$, where at time t, X_t and U_t denotes the state of the process and the transition matrix, respectively.

Suppose $\mathscr{A}$ and $\mathscr{S}$ denote the set of actions and states. Suppose $C(a,s;\psi^*)$ denotes the correctly specified cost of taking action a at state s where $\gamma \in [0,1)$ denotes the discount factor. The probability of the system transitioning from state s' to s'' based on action a is specified by $\mathbb{P}^*(s = s'' \mid s = s', a)$. Furthermore, we define a policy map as $\pi : \mathscr{S} \to \mathscr{A}$ while the value function of a policy π is denoted by $V^\pi : \mathscr{S} \to \mathbb{R}$ and $V^\pi(s)$ denotes the expected discounted cost of policy π when starting at state s. The objective lies in determining a policy π that minimizes the discounted expected sum over an infinite horizon, given by $\sum_{k=0}^{\infty} \gamma^k C(s_k, a_k; \psi^*)$, where $a_{k+1} = \pi(s_k)$.

This paper considers the resolution of such problems in regimes where the transition matrix $\mathbb{P}^*$ and the parametrization of the cost function ψ^* are unavailable a priori. Estimation of transition matrices has been studied extensively in the literature (Anderson and Goodman 1957, Ljung 1987, Han and Liu 2013) while robust optimization approaches have also been employed (cf. (Nilim and El Ghaoui 2005, Delage and

Mannor 2010)). A rather distinct approach in contending with the absence of information is embodied by the Q-learning algorithm presented in (Watkins and Dayan 1992). This is a simulation-based technique for computing estimates to the value function and has a similar structure to stochastic approximation algorithms (Tsitsiklis and Sutton 1994). Simulation-based approaches have also been reviewed by Chang, Hu, Fu, and Marcus (2013), particularly notable being the upper confidence bound (UCB) sampling algorithm (cf. (Brafman and Tennenholtz 2003, Bartlett and Tewari 2009, Auer, Jaksch, and Ortner 2009)).

Given an MDP$(\mathbb{P}^*, \psi^*)$ where $\mathbb{P}^*$ and ψ^* are unavailable, a standard approach is the following:

$$\text{(1) Learn } \mathbb{P}^* \text{ and } \psi^*; \qquad \text{(2) Solve MDP}(\mathbb{P}^*, \psi^*).$$

This technique is afflicted by several challenges, a subset of which we describe next:
(i) *Inability to accommodate streaming data:* Increasingly, MDP-based models have to be constantly updated with new, and possibly, streaming data. Yet the traditionally developed asymptotics and error analysis for resolving MDPs cannot accommodate streaming data.
(ii) *Lack of asymptotics:* Step (1) often requires solving stochastic and/or large-scale learning problems whose solutions are obtained in an asymptotic sense. Any practical implementation of this scheme necessitates that Step (1) terminate finitely; however, premature termination of (1) leads to estimators afflicted by error and may result in significant error in the computed value function. In effect, asymptotic convergence of this scheme cannot be claimed.
(iii) *Practical implementations:* Step (1) may take a significant amount of real-time, particularly since it requires solving stochastic optimization problems and during this period, no estimate of the optimal value function is available. In effect, error bounds can only be prescribed after step (1) is complete.

A simultaneous scheme for learning and computation: We consider an avenue that has found recent application for resolving misspecified optimization and variational problems in stochastic regimes (Jiang and Shanbhag 2014, Jiang and Shanbhag 2013). This necessitates a *simultaneous* approach in which the learning problems for $\mathbb{P}^*$ and ψ^* are resolved simultaneously with the original MDP. In effect, we consider the estimators from the coupled dynamics and examine both the asymptotics and error bounds for a variety of computational schemes. Our scheme relies on the prescription of learning problems.
(i) *Learning of transition matrices:* We consider the following learning problem for transition matrices based on using observational data:

$$\mathbb{P}^* \in \operatorname*{argmin}_{\mathbb{P}\in\mathscr{P}} \mathbb{E}[g(\mathbb{P};\eta)], \tag{$\mathscr{L}^{\mathbb{P}}$}$$

where $\mathscr{P}$ denotes the space of stochastic matrices, i.e. nonnegative matrices with row sums equal to unity.
(ii) *Misspecification of cost functions:* The cost functions are parameterized by a vector ψ^*, representing a set of parameters idiosyncratic to the machine of interest. For instance, it may pertain to the efficiency of the machine, the start-up/shut-down times, the skill of the workers in question etc. All of these parameters may require learning, often via an online approach that incorporates the use of observations, possibly corrupted by noise. Such a problem can be cast as a stochastic optimization problem, defined as follows:

$$\psi^* \in \operatorname*{argmin}_{\psi\in\Psi} \mathbb{E}[R(\psi;\xi)], \tag{$\mathscr{R}^{\Psi}$}$$

where ξ a random variable and Ψ denotes the feasibility set for ψ. By using stochastic approximation, we may generate sequences $\{\mathbb{P}_k\}$ and $\{\psi_k\}$ such that $\mathbb{P}_k \to \mathbb{P}^*$ and $\psi_k \to \psi^*$ as $k \to \infty$ in an a.s. sense.

We provide an illustration of the approach by using the well-studied value iteration scheme as a basis (Howard 1960). In its original form, value iteration maintains an estimate of the value function and updates this belief based on solving a suitable problem. When the change in the value functions falls within a suitably defined threshold in a norm-sense, the scheme terminates. We now provide a relatively quick overview of this scheme (cf. (Powell 2007)). Let $\mathscr{V}$ denote the space of value functions and $\mathscr{M} : \mathscr{V} \to \mathscr{V}$

be a mapping such that for each $s \in \mathscr{S}$, $\mathscr{M}$ is defined as follows:

$$\mathscr{M}v(s) = \max_{a \in \mathscr{A}} \left\{ C(s,a;\psi^*) + \gamma \sum_{s' \in \mathscr{S}} \mathbb{P}^*(s' \mid s,a) v(s') \right\}. \qquad (\text{MDP}(\mathbb{P}^*, \Psi^*))$$

Given a v^0, the value iteration scheme is defined as follows:

$$v^{k+1} := \mathscr{M}v^k, \qquad \text{for } k \geq 0. \qquad \text{(Value Iteration)}$$

Since $\mathscr{M}$ is a contraction mapping on $\mathscr{V}$ if $0 \leq \gamma < 1$ (cf. Proposition 3.10.2 in Powell (2007)), convergence of the scheme can be shown with in a reasonably straightforward fashion. However, one of the challenges lies in the availability of $C(s,a;\psi^*)$ and $\mathbb{P}^*$, motivating the development of a misspecified variant. We assume the cost and matrix to be given by $C(s,a;\widetilde{\psi})$ and $\widetilde{\mathbb{P}}(s'|s,a)$. Then, we may define a *misspecified* operator $\widetilde{\mathscr{M}_k} : \mathscr{V} \to \mathscr{V}$ by utilizing estimates $\widetilde{\mathbb{P}}_k$ and $\widetilde{\psi}_k$:

$$\widetilde{\mathscr{M}_k}v(s) = \max_{a \in \mathscr{A}} \left\{ C(s,a;\widetilde{\psi}_k) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}_k(s'|s,a) v(s') \right\}. \qquad (1)$$

We now present our main research questions and provide an outline of this paper:
(i) Misspecified value iteration: In Section 2, we present a misspecified value iteration scheme for addressing MDPs in which the cost function and transition matrices are misspecified. We examine the asymptotics of the resulting scheme and providing a quantification of the degradation of the rate of convergence based on the presence of learning.
(ii) Misspecified policy iteration: In Section 3, we consider an analogous set of questions in the regime of policy iteration where we provide almost sure convergence statements.
(iii) Misspecified Q-learning: Finally, in Section 4, we consider Q-learning approaches for solving MDPs with misspecified cost functions and present constant steplength error bounds for extensions that resolve the misspecification while solving the original MDP.

2 MISSPECIFIED VALUE ITERATION

Value iteration (Bellman 1957) represents amongst the oldest of schemes for solving an MDP. We begin by presenting a *misspecified* value iteration scheme for resolving MDP$(\mathbb{P}^*, \psi^*)$ and subsequently present asymptotic convergence and error analysis.

We define $\mathscr{P}$ to be set of all transition matrices, vec$(\mathbb{P})$ to be the vector drawn from the entries of $\mathbb{P}$ for all $\mathbb{P} \in \mathscr{P}$, and vec$(\mathscr{P}) \triangleq \{\text{vec}(\mathbb{P}), \mathbb{P} \in \mathscr{P}\}$. Estimating $\mathbb{P}^*$ often requires the resolution of a suitably defined learning problem, given by a stochastic optimization problem $(\mathscr{L}^{\mathbb{P}})$, where vec$(\mathscr{P})$ is a closed and convex set, $\eta : \Lambda \to \mathbb{R}^p$ is a random variable defined on a probability space $(\Lambda, \mathscr{F}_\eta, \mathbb{P}_\eta)$, and $g : \mathbf{P} \times \Lambda \to \mathbb{R}$ is a real-valued function. We may specify our joint scheme for learning and computation as follows:

Algorithm I: Misspecified Value Iteration.
Let $\tilde{v}^0 : \mathscr{S} \to \mathbb{R}$, vec$(\widetilde{\mathbb{P}}_0) \in$ vec$(\mathbf{P})$, $\widetilde{\psi}_0 \in \Psi$, $\alpha_0 > 0$, $\beta_0 > 0$ and $k = 0$.
Step 1: For all $k \geq 0$,

$$\tilde{v}^{k+1} := \widetilde{\mathscr{M}_k}\tilde{v}^k, \qquad \text{(Computation)}$$

$$\text{vec}(\widetilde{\mathbb{P}}_{k+1}) := \Pi_{\text{vec}(\mathbf{P})}\left(\text{vec}(\widetilde{\mathbb{P}}_k) - \alpha_k(\nabla g(\widetilde{\mathbb{P}}_k) + w_k)\right), \qquad \text{(Learning}-\mathbb{P})$$

$$\widetilde{\psi}_{k+1} := \Pi_\Psi\left(\widetilde{\psi}_k - \beta_k(\nabla R(\widetilde{\psi}_k) + u_k)\right) \qquad \text{(Learning}-\psi)$$

where $w_k \triangleq \nabla g(\widetilde{\mathbb{P}}_k;\eta_k) - \nabla g(\widetilde{\mathbb{P}}_k)$, $g(\mathbb{P}) \triangleq \mathbb{E}[g(\mathbb{P};\eta)]$, $u_k = \nabla R(\widetilde{\psi}_k;\xi_k) - \nabla R(\widetilde{\psi}_k)$, $R(\psi) \triangleq \mathbb{E}[R(\psi;\xi)]$, $\widetilde{\mathscr{M}_k}v(s) := \max_{a \in \mathscr{A}}(C(s,a;\widetilde{\psi}_k) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}_n(s'|s,a)v(s'))$, and α_k and β_k are chosen according to Proposition 1.
Step 2: If $k > K$, stop; else $k := k+1$ and go to Step 1.

We begin by showing that the misspecified operator $\widetilde{\mathscr{M}}_k$ is a contraction mapping for any k. We suppress the subscript k in this proof for purposes of clarity.

Lemma 1 (Contractive property of $\widetilde{\mathscr{M}}$) If $0 \leq \gamma < 1$, then $\widetilde{\mathscr{M}}$ is a contraction mapping on $\mathscr{V}$.

Proof. Let $u, v \in \mathscr{V}$ and assume that $\widetilde{\mathscr{M}}v(s) \geq \widetilde{\mathscr{M}}u(s)$ without loss of generality for any state s. For any state s, let $\tilde{a}_s^*(v)$ be defined as follows:

$$\tilde{a}_s^*(v) = \operatorname*{argmax}_{a \in \mathscr{A}} \left\{ C(s,a;\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,a) v(s') \right\}.$$

Then, we have the following sequence of inequalities:

$$0 \leq \widetilde{\mathscr{M}}v(s) - \widetilde{\mathscr{M}}u(s) = C(s,\tilde{a}_s^*(v);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) v(s') - \left(C(s,\tilde{a}_s^*(u);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(u)) u(s') \right)$$

$$\leq \underbrace{C(s,\tilde{a}_s^*(v);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) v(s') - \left(C(s,\tilde{a}_s^*(v);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) u(s') \right)}_{\text{Term (a)}},$$

where the second inequality is a consequence of noting that for all s, we have the following:

$$\widetilde{\mathscr{M}}u(s) = \max_{a \in \mathscr{A}} \left\{ C(s,a;\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,a) u(s') \right\} = \left(C(s,\tilde{a}_s^*(u);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(u)) u(s') \right)$$

$$\geq \left(C(s,\tilde{a}_s^*(v);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) u(s') \right).$$

It follows that Term (a) can be bounded as follows:

$$C(s,\tilde{a}_s^*(v);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) v(s') - \left(C(s,\tilde{a}_s^*(v);\widetilde{\psi}) + \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) u(s') \right)$$

$$= \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) \left(v(s') - u(s') \right) \leq \gamma \sum_{s' \in \mathscr{S}} \widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v)) \|v-u\|_\infty = \gamma \|v-u\|_\infty,$$

Consequently, $\|\widetilde{\mathscr{M}}v - \widetilde{\mathscr{M}}u\|_\infty = \sup_{s \in \mathscr{S}} |\widetilde{\mathscr{M}}v(s) - \widetilde{\mathscr{M}}u(s)| \leq \gamma \|v-u\|_\infty$, implying that $\widetilde{\mathscr{M}}$ is contractive. ∎

Our next proposition shows that when the estimated transition matrix is within some bound of its true counterpart, under a suitable Lipschitzian requirement of $C(s,a,\psi)$ in ψ, we obtain the following relationship between the true operator and its misspecified counterpart. This lemma subsequently finds application in the main convergence result.

Lemma 2 Suppose $\sum_{s' \in \mathscr{S}} |\mathbb{P}^*(s'|s,a) - \widetilde{\mathbb{P}}(s'|s,a)| \leq \delta$ for all s and a. Suppose $C(s,a;\psi)$ is Lipschitz continuous in ψ with constant L_C uniformly in s and a. Then the following holds for all $u, v \in \mathscr{V}$:

$$\|\mathscr{M}v - \widetilde{\mathscr{M}}u\| \leq L_C \|\psi^* - \widetilde{\psi}\| + \gamma\delta(\|v\| + \|u\|) + \gamma\|v-u\|.$$

Proof. Let $u, v \in \mathscr{V}$ and assume without loss of generality that $\mathscr{M}v(s) \geq \widetilde{\mathscr{M}}u(s)$. For a state s, we may define $a_s^*(v)$ and $\tilde{a}_s^*(v)$ as follows:

$$a_s^*(v) \triangleq \underset{a\in\mathscr{A}}{\operatorname{argmax}}(C(s,a;\psi^*)+\gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a)v(s')) \text{ and } \tilde{a}_s^*(v) \triangleq \underset{a\in\mathscr{A}}{\operatorname{argmax}}(C(s,a;\widetilde{\psi})+\gamma\sum_{s'\in\mathscr{S}}\widetilde{\mathbb{P}}(s'|s,a)v(s')).$$

Then, we have the following set of relations:

$$\begin{aligned}
0 &\leq \mathscr{M}v(s)-\widetilde{\mathscr{M}}u(s) = \mathscr{M}v(s)-\widetilde{\mathscr{M}}v(s)+\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s)\\
&= C(s,a_s^*(v);\psi^*)+\gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a_s^*(v))v(s') - \left(C(s,\tilde{a}_s^*(v);\widetilde{\psi})+\gamma\sum_{s'\in\mathscr{S}}\widetilde{\mathbb{P}}(s'|s,\tilde{a}_s^*(v))v(s')\right)\\
&\quad +\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s)\\
&\leq C(s,a_s^*(v);\psi^*)+\gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a_s^*(v))v(s') - \left(C(s,a_s^*(v);\widetilde{\psi})+\gamma\sum_{s'\in\mathscr{S}}\widetilde{\mathbb{P}}(s'|s,a_s^*(v))v(s')\right)\\
&\quad +\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s),
\end{aligned}$$

where the second inequality follows from suboptimality of $a_s^*(v)$ with respect to $a^*(v)$. It follows that

$$\begin{aligned}
&C(s,a_s^*(v);\psi^*)+\gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a_s^*(v))v(s') - \left(C(s,a_s^*(v);\widetilde{\psi})+\gamma\sum_{s'\in\mathscr{S}}\widetilde{\mathbb{P}}(s'|s,a_s^*(v))v(s')\right)\\
&+\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s)\\
&\leq L_C\|\psi^*-\widetilde{\psi}\|+\gamma\sum_{s'\in\mathscr{S}}\left(\mathbb{P}^*(s'|s,a_s^*(v))-\widetilde{\mathbb{P}}(s'|s,a_s^*(v))\right)v(s')+\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s)\\
&\leq L_C\|\psi^*-\widetilde{\psi}\|+\gamma\sum_{s'\in\mathscr{S}}|\mathbb{P}^*(s'|s,a_s^*(v))-\widetilde{\mathbb{P}}(s'|s,a_s^*(v))|\|v\|+\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s)\\
&\leq L_C\|\psi^*-\widetilde{\psi}\|+\gamma\delta\|v\|+\widetilde{\mathscr{M}}v(s)-\widetilde{\mathscr{M}}u(s)\\
&\leq L_C\|\psi^*-\widetilde{\psi}\|+\gamma\delta\|v\|+\gamma\|v-u\|\\
&\leq L_C\|\psi^*-\widetilde{\psi}\|+\gamma\delta(\|v\|+\|u\|)+\gamma\|v-u\|,
\end{aligned}$$

where the first inequality follows from the Lipschitz continuity of $C(s,a;\psi)$ in ψ, the second inequality is a consequence of the Cauchy-Schwartz inequality and the last inequality is a consequence of invoking the contractive property of $\widetilde{\mathscr{M}}$. ∎

We are now ready to prove our main convergence statement.

Proposition 1 (Convergence of misspecified value iteration scheme) Suppose $\{\tilde{v}^k\}$, $\{\widetilde{\mathbb{P}}_k\}$ and $\{\widetilde{\psi}_k\}$ are generated from Algorithm I. Suppose the learning function $g(\cdot)$ is strongly convex in $\text{vec}(\mathbf{P})$, and the learning function $R(\cdot)$ is strongly convex in Ψ. Suppose $\alpha_k = \theta_1/k$ and $\beta_k = \theta_2/k$ with $\theta_1 > 1/2(\mu_g)$, $\theta_2 > 1/(2\mu_R)$, $0 < \alpha < 1/2$, μ_g is the strong convexity constant of g and μ_R is the strong convexity constant of R. Suppose $C(s,a;\psi)$ is Lipschitz continuous in ψ with constant L_C for all s and a. Then, there exists a constant λ such that the following hold:

(i) $\|\tilde{v}^k - v^*\| \to 0$, $\widetilde{\mathbb{P}}_k \to \mathbb{P}^*$ and $\widetilde{\psi}_k \to \psi^*$ *a.s.* as $k \to \infty$.

(ii) For any k, we have that the following holds:

$$\mathbb{E}\left[\|\tilde{v}^{k+1}-v^*\|\right] \leq \gamma^k\mathbb{E}[\|\tilde{v}^0-v^*\|]+\sum_{j=1}^{k}\frac{\gamma^{j-1}\lambda}{\sqrt{k-j-1}}\|v^*\| = \mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}}\right),$$

Proof. **(i)** First, we have that the following holds almost surely:

$$\begin{aligned}\|\tilde{v}^{k+1}-v^*\| &= \|\widetilde{\mathscr{M}}_k\tilde{v}^k-\mathscr{M}v^*\| \\ &= \|\widetilde{\mathscr{M}}_kv^k-\widetilde{\mathscr{M}}_kv^*+\widetilde{\mathscr{M}}_kv^*-\mathscr{M}v^*\| \le \|\widetilde{\mathscr{M}}_k\tilde{v}^k-\widetilde{\mathscr{M}}_kv^*\|+\|\widetilde{\mathscr{M}}_kv^*-\mathscr{M}v^*\| \\ &\le \gamma\|\tilde{v}^k-v^*\|+L_C\|\widetilde{\psi}_k-\psi^*\|+\gamma\|\mathrm{vec}(\widetilde{\mathbb{P}}_k)-\mathrm{vec}(\mathbb{P}^*)\|\|v^*\|, \qquad (2)\end{aligned}$$

where the last inequality follows from invoking Lemmas 1 and 2. Let $a_k = L_C\|\widetilde{\psi}_k-\psi^*\|+\gamma\|\mathrm{vec}(\widetilde{\mathbb{P}}_k)-\mathrm{vec}(\mathbb{P}^*)\|\|v^*\|$. Then, we have

$$\begin{aligned}\|\tilde{v}^{k+1}-v^*\| &\le \gamma\|\tilde{v}^k-v^*\|+a_k \le \gamma(\gamma\|\tilde{v}^{k-1}-v^*\|+a_{k-1})+a_k \\ &= \gamma^2\|\tilde{v}^{k-1}-v^*\|+\gamma a_{k-1}+a_k \le \cdots \le \gamma^{k+1}\|\tilde{v}^0-v^*\|+\sum_{i=0}^{k}\gamma^i a_{k-i}.\end{aligned}$$

Since $\gamma^{k+1}\to 0$, it suffices to show that $\sum_{i=0}^k\gamma^i a_{k-i}\to 0$ as $k\to\infty$ in an a.s. sense. Since the learning problems for ψ^* and $\mathbb{P}^*$ are both strongly convex, we have that $a_k\to 0$ *a.s.* as $k\to\infty$. Then, for almost all $\omega\in\Omega$, given $\varepsilon>0$, there exists $N_1(\omega)$ such that $a_k\le\varepsilon$ for all $k\ge N_1(\omega)$. Also, for almost every $\omega\in\Omega$, $a_k\le L(\omega)$ for all k and some constant $L(\omega)>0$. Thus, for $k\ge N_1(\omega)$,

$$\begin{aligned}\sum_{i=0}^{k}\gamma^i a_{k-i} &= \gamma^n a_0+\ldots+\gamma^{k-N_1(\omega)}a_{N_1(\omega)}+\gamma^{k-N_1(\omega)-1}a_{N_1(\omega)+1}+\ldots+\gamma^0 a_k \\ &\le (\gamma^k+\ldots+\gamma^{k-N_1(\omega)})L(\omega)+\frac{\varepsilon}{1-\gamma}.\end{aligned}$$

Since $\gamma^k\to 0$, there exists $N_2(\omega)$ such that $\gamma^k\le\frac{\varepsilon}{N_1(\omega)+1},\ldots,\gamma^{k-N_1(\omega)}\le\frac{\varepsilon}{N_1(\omega)+1}$ for all $k\ge N_2(\omega)$. So, when $k\ge N(\omega)\triangleq\max\{N_1(\omega),N_2(\omega)\}$, we have that

$$\sum_{i=0}^{k}\gamma^i a_{k-i}\le L(\omega)\varepsilon+\frac{\varepsilon}{1-\gamma}=\left(L(\omega)+\frac{1}{1-\gamma}\right)\varepsilon.$$

Since $L(\omega)$ is finite in an a.s. sense and ε is arbitrarily chosen, proving that $\sum_{i=0}^k\gamma^i a_{n-i}\to 0$ *a.s.*. We may then conclude that $\|\tilde{v}^{k+1}-v^*\|\to 0$ in an a.s. sense as $k\to\infty$.

(ii) By taking expectations on both sides of (2), we have the following:

$$\mathbb{E}[\|\tilde{v}^{k+1}-v^*\|]\le\gamma\mathbb{E}[\|\tilde{v}^k-v^*\|]+L_C\mathbb{E}[\|\widetilde{\psi}_k-\psi^*\|]+\gamma\mathbb{E}[\|\mathrm{vec}(\widetilde{\mathbb{P}}_k)-\mathrm{vec}(\mathbb{P}^*)\|]\|v^*\|.$$

Recall that the learning problem for ψ^* and $\mathbb{P}^*$ are both strongly convex. Then, we can use the standard rate estimate (see (5.292) in Shapiro et al. (2009)) to get the following for suitably chosen λ_1 and λ_2:

$$\mathbb{E}[\|\widetilde{\psi}_k-\psi^*\|]\le\frac{\lambda_1}{\sqrt{k}}\text{ and }\mathbb{E}[\|\mathrm{vec}(\widetilde{\mathbb{P}}_k)-\mathrm{vec}(\mathbb{P}^*)\|]\le\frac{\lambda_2}{\sqrt{k}}.$$

Consequently, we obtain the following:

$$\mathbb{E}[\|\tilde{v}^{k+1}-v^*\|]\le\gamma\mathbb{E}[\|\tilde{v}^k-v^*\|]+\frac{L_C\lambda_1+\gamma\lambda_2}{\sqrt{k}}\|v^*\|.$$

Let $\lambda=L_C\lambda_1+\gamma\lambda_2$. Then, we have

$$\begin{aligned}\mathbb{E}[\|\tilde{v}^{k+1}-v^*\|] &\le \gamma\mathbb{E}[\|\tilde{v}^k-v^*\|]+\frac{\lambda}{\sqrt{k}}\|v^*\| \le \gamma^2\mathbb{E}[\|\tilde{v}^{k-1}-v^*\|]+\frac{\gamma\lambda}{\sqrt{k-1}}\|v^*\|+\frac{\lambda}{\sqrt{k}}\|v^*\| \\ &\le \gamma^k\mathbb{E}[\|\tilde{v}^0-v^*\|]+\sum_{j=1}^{k}\frac{\gamma^{j-1}\lambda}{\sqrt{k-j+1}}\|v^*\|.\end{aligned}$$

If we fix K, then the second term in the above inequality becomes

$$\sum_{j=1}^{k} \frac{\gamma^{j-1}\lambda}{\sqrt{k-j+1}}\|v^*\| = \lambda\|v^*\| \sum_{j=1}^{k} \frac{\gamma^{j-1}}{\sqrt{k-j+1}} = \lambda\|v^*\| \left(\sum_{j=1}^{K} \frac{\gamma^{j-1}}{\sqrt{k-j+1}} + \sum_{j=K+1}^{k} \frac{\gamma^{j-1}}{\sqrt{k-j+1}} \right)$$
$$\leq \lambda\|v^*\| \left(\frac{K}{\sqrt{k-K+1}} + \sum_{j=K+1}^{k} \gamma^{j-1} \right) \leq \lambda\|v^*\| \left(\frac{K}{\sqrt{k-K+1}} + \frac{\gamma^K}{1-\gamma} \right),$$

where the second inequality follows from noting that $1/\sqrt{k-j+1} \leq 1$ for $K+1 \leq j \leq k$. If K is chosen to be k^α with $0 < \alpha < 1/2$, then we have

$$\sum_{j=1}^{k} \frac{\gamma^{j-1}\lambda}{\sqrt{k-j+1}}\|v^*\| \leq \lambda\|v^*\| \left(\frac{k^\alpha}{\sqrt{k-k^\alpha+1}} + \frac{\gamma^{k^\alpha}}{1-\gamma} \right) = \lambda\|v^*\| \left(\frac{k^\alpha}{\sqrt{k}\sqrt{1-k^{-(1-\alpha)}+1/k}} + \frac{\gamma^{k^\alpha}}{1-\gamma} \right)$$
$$= \lambda\|v^*\| \left(\frac{1}{k^{1/2-\alpha}\sqrt{1-k^{-(1-\alpha)}+1/k}} + \frac{\gamma^{k^\alpha}}{1-\gamma} \right) \leq \lambda\|v^*\| \left(\frac{1}{k^{1/2-\alpha}} + \frac{\gamma^{k^\alpha}}{1-\gamma} \right) \leq \mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}} \right),$$

since the second term diminishes to zero at a faster rate than $1/(k^{1/2-\alpha})$. ∎

3 MISSPECIFIED POLICY ITERATION

In this section, we consider a policy iteration scheme for the resolution of misspecified MDPs. We initiate our discussion with a formal statement of the misspecified policy iteration scheme and subsequently prove its asymptotic convergence. If $c^\pi(\cdot) \triangleq C(\cdot, \pi(\cdot); \psi^*)$ and $\tilde{c}^{\pi_k}(\cdot) \triangleq C(\cdot, \pi_k(\cdot); \tilde{\psi}_k)$, then the operators $\mathscr{M}^\pi$ and $\widetilde{\mathscr{M}}_k^{\pi_k}$ may be defined as follows for policies π and π_k, respectively:

$$\mathscr{M}^\pi v \triangleq c^\pi + \gamma(\mathbb{P}^*)^\pi v \text{ and } \widetilde{\mathscr{M}}_k^{\pi_k} v \triangleq \tilde{c}^{\pi_k} + \gamma\widetilde{\mathbb{P}}_k^{\pi_k} v,$$

Next, we define the misspecified policy iteration scheme.

Algorithm II: Misspecified policy iteration.
Let $\tilde{v}^0 : \mathscr{S} \to \mathbb{R}$, $\text{vec}(\widetilde{\mathbb{P}}_0) \in \text{vec}(\mathbf{P})$, $\alpha_k > 0$, $\tilde{\psi}_0 \in \Psi$, $\alpha_0 > 0$, $\beta_0 > 0$ and $k = 0$.
Step 1: For all $k \geq 0$,

$$a_{k+1}(s) := \underset{a \in \mathscr{A}}{\text{argmax}} (C(s,a;\tilde{\psi}_k) + \gamma\widetilde{\mathbb{P}}_k^{\pi_{k+1}} \tilde{v}^{k+1}), \qquad \text{(Computation)}$$
$$\text{vec}(\widetilde{\mathbb{P}}_{k+1}) := \Pi_{\text{vec}(\mathbf{P})} \left(\text{vec}(\widetilde{\mathbb{P}}_k) - \alpha_k(\nabla g(\widetilde{\mathbb{P}}_k) + w_k) \right), \qquad \text{(Learning}-\mathbb{P}\text{)}$$
$$\tilde{\psi}_{k+1} := \Pi_\Psi \left(\tilde{\psi}_k - \beta_k(\nabla R(\tilde{\psi}_k) + u_k) \right) \qquad \text{(Learning}-\Psi\text{)}$$

where $w_k \triangleq \nabla g(\widetilde{\mathbb{P}}_k; \eta_k) - \nabla g(\widetilde{\mathbb{P}}_k)$ with $g(\mathbb{P}) \triangleq \mathbb{E}[g(\mathbb{P}; \eta)]$, $u_k = \nabla R(\tilde{\psi}_k; \xi_k) - \nabla R(\tilde{\psi}_k)$, $R(\psi) \triangleq \mathbb{E}[R(\psi; \xi)]$ and $(I - \gamma\widetilde{\mathbb{P}}_k^{\pi_k})\tilde{v}^{k+1} = \tilde{c}^{\pi_k}$.
Step 2: If $k > K$, stop; else $k := k+1$ and go to Step 1.

Analogous to Proposition 1 for the value iteration, we can get the following convergence statement where $\|\bullet\|$ denotes the infinity norm for both matrices and vectors.

Proposition 2 (Convergence of misspecified policy iteration) Suppose $\{\tilde{v}^n\}$, $\{\widetilde{\mathbb{P}}_k\}$ and $\{\tilde{\psi}_k\}$ are generated by Algorithm II and the learning functions $g(\cdot)$ and $R(\cdot)$ are strongly convex. Finally, suppose $C(s,a;\psi)$ is Lipschitz continuous in ψ with constant L_C for all s and a and $\|\tilde{v}^k\|$ is bounded for all k. Then $\|\tilde{v}^k - v^*\| \to 0 \quad a.s. \quad \text{as } k \to \infty$.

Proof. We proceed to show that $\|v_k - \tilde{v}^k\| \to 0$ as $k \to \infty$ whereby the result follows by recalling that by the convergence of policy iteration, $\|v^k - v^*\| \to 0$ as $k \to \infty$. From Algorithm II, we have

$$\begin{aligned}\|v^{k+1} - \tilde{v}^{k+1}\| &= \|c^{\pi_k} + \gamma(\mathbb{P}^*)^{\pi_k} v^{k+1} - (\tilde{c}^{\pi_k} + \gamma\widetilde{\mathbb{P}}_k^{\pi_k}\tilde{v}^{k+1})\| \\ &= \|c^{\pi_k} - \tilde{c}^{\pi_k} + \gamma(\mathbb{P}^*)^{\pi_k}(v^{k+1} - \tilde{v}^{k+1}) + \gamma((\mathbb{P}^*)^{\pi_k} - \widetilde{\mathbb{P}}_k^{\pi_k})\tilde{v}^{k+1}\| \\ &\leq L_C N\|\psi^* - \widetilde{\psi}_k\| + \gamma\|(\mathbb{P}^*)^{\pi_k}\|\|v^{k+1} - \tilde{v}^{k+1}\| + \gamma\|(\mathbb{P}^*)^{\pi_k} - \widetilde{\mathbb{P}}_k^{\pi_k}\|\|\tilde{v}^{k+1}\|.\end{aligned}$$

It follows that

$$\|v^{k+1} - \tilde{v}^{k+1}\| \leq \frac{L_C N\|\psi^* - \widetilde{\psi}_k\| + \gamma\|(\mathbb{P}^*)^{\pi_k} - \widetilde{\mathbb{P}}_k^{\pi_k}\|\|\tilde{v}^{k+1}\|}{1 - \gamma\|(\mathbb{P}^*)^{\pi_k}\|} = \frac{L_C N\|\psi^* - \widetilde{\psi}_k\| + \gamma\|(\mathbb{P}^*)^{\pi_k} - \widetilde{\mathbb{P}}_k^{\pi_k}\|\|\tilde{v}^{k+1}\|}{1-\gamma}.$$

Recall that the learning problem for ψ^* and $\mathbb{P}^*$ are both strongly convex, implying that $\widetilde{\psi}_k \to \psi^*$ and $\text{vec}(\widetilde{\mathbb{P}}_k) \to \text{vec}(\mathbb{P}^*)$ *a.s.* as $k \to \infty$. Thus, by the boundedness of $\tilde{v}^k$ and by invoking the property that $\|v^k - v^*\| \to 0$ as $k \to \infty$, we have that $\|v^k - \tilde{v}^k\| \to 0$ *a.s.* as $k \to \infty$. Therefore, $\|\tilde{v}^k - v^*\| \to 0$ *a.s.* as $k \to \infty$. ∎

4 MISSPECIFIED Q-LEARNING

When transition matrices are unavailable, a commonly adopted approach is Q-learning (Watkins and Dayan 1992). We consider a misspecified variant of Q-learning that incorporates learning of the misspecified cost and examines the resulting sequence of estimators. We begin by defining the Q-function as $Q(s,a) \triangleq C(s,a;\psi^*) + \gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a)v(s')$, which allows for restating as follows:

$$Q(s,a) \triangleq C(s,a;\psi^*) + \gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a)\max_{b\in\mathscr{A}}Q(s',b). \tag{3}$$

We define the operator $\mathscr{T}$ as

$$\mathscr{T}[Q(s,a)] \triangleq C(s,a;\psi^*) + \gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a)\max_{b\in\mathscr{A}}Q(s',b).$$

Then the Q-function is the fixed point of the operator $\mathscr{T}$; i.e. $Q = \mathscr{T}[Q]$. Given the vector $\widetilde{\psi}_k$ in the cost at iteration n, we may define the misspecified operator $\widetilde{\mathscr{T}_k}$ at iteration n as

$$\widetilde{\mathscr{T}_k}Q(s,a) \triangleq C(s,a;\widetilde{\psi}_k) + \gamma\sum_{s'\in\mathscr{S}}\mathbb{P}^*(s'|s,a)\max_{b\in\mathscr{A}}Q(s',b).$$

As in previous sections, we may specify our misspecified Q-learning scheme as follows:

Algorithm III: Misspecified Q-learning.
Let $\widetilde{Q}_0(s,a) \in \mathbb{R}$, $\widetilde{\psi}_0 \in \Psi$, $\beta_0 > 0$ and $k = 0$.
Step 1: For all $n \geq 0$,

$$\widetilde{Q}_{k+1}(s,a) := (1-\delta)\widetilde{Q}_k(s,a) + \delta\left[C(s,a;\widetilde{\psi}_k) + \gamma\max_{b\in\mathscr{A}}\widetilde{Q}_k(s',b)\right], \qquad (Q\text{−update})$$

$$\widetilde{\psi}_{k+1} := \Pi_\Psi\left(\widetilde{\psi}_k - \beta_k(\nabla R(\widetilde{\psi}_k) + u_k)\right), \qquad (\text{Learning-}\psi)$$

where $\delta \in (0,1)$, s' is the random next state reached when the current state is s and action is a, and $u_k = \nabla R(\widetilde{\psi}_k;\xi_k) - \nabla R(\widetilde{\psi}_k)$ with $R(\psi) \triangleq \mathbb{E}[R(\psi;\xi)]$.
Step 2: If $n > K$, stop; else $k := k+1$ and go to Step 1.

Our convergence analysis begins with a reproduction of two classical results regarding the operator $\widetilde{\mathscr{T}}$, which may be directly applied to the misspecified operator $\widetilde{\mathscr{T}_k}$. First, $\widetilde{\mathscr{T}_k}$ is a contraction mapping.

Proposition 3 (**Contractive property of** $\widetilde{\mathscr{T}_k}$ (Tsitsiklis and Sutton 1994)) If $0 \leq \gamma < 1$, then $\|\widetilde{\mathscr{T}_k}[Q_1] - \widetilde{\mathscr{T}_k}[Q_2]\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$ for any two vectors Q_1 and Q_2. ∎

Second, the estimated Q-function stays bounded.

Proposition 4 (**Boundedness of** Q **function** (Gosavi 2006)) There exists $\hat{Q}_{\max}$ such that $\|\hat{Q}_k\|_\infty \leq \hat{Q}_{\max}$ for any k. ∎

We now provide an intermediate lemma that provides a constant steplength error bound on a suitably defined metric D.

Lemma 3 For any state-action pair (s,a), suppose $D_k(s,a) = \bar{Q}(s,a) - z_k$ and $z_k(s,a)$ be defined as follows:

$$z_{k+1}(s,a) = (1-\delta) z_k(s,a) + \delta\gamma w_k(s,a), \quad z_0(s,a) = 0. \tag{4}$$

Then for any k, we have that $\mathbb{E}[\|D_k\|_\infty] \leq \left(\mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}}\right) + \frac{\gamma^2}{1-\gamma}\sqrt{\frac{\delta W_{\max}^2}{2-\delta}} \right)$, where $0 < \alpha < 1/2$.

Proof. We utilize an approach employed by Beck and Srikant (2012) and begin by defining the error $\bar{Q}_k(s,a)$ as $\overline{Q}_k(s,a) \triangleq \widetilde{Q}_k(s,a) - Q(s,a)$. Using (3) and ($Q-$update), the error can be written as

$$\begin{aligned}
\overline{Q}_{k+1}(s,a) &= (1-\delta)\overline{Q}_k(s,a) + \delta\left[C(s,a;\widetilde{\psi}_k) + \gamma \max_{b\in\mathscr{A}} \widetilde{Q}_k(s',b) - Q(s,a)\right] \\
&= (1-\delta)\overline{Q}_k(s,a) + \delta\left[C(s,a;\widetilde{\psi}_k) - C(s,a;\psi^*) + \gamma \max_{b\in\mathscr{A}} \widetilde{Q}_k(s',b) - \gamma \sum_{s'\in\mathscr{S}} \mathbb{P}^*(s'|s,a) \max_{b\in\mathscr{A}} Q(s',b)\right] \\
&= (1-\delta)\overline{Q}_k(s,a) + \delta\left(C(s,a;\widetilde{\psi}_k) - C(s,a;\psi^*)\right) \\
&\quad + \delta\gamma \sum_{s'\in\mathscr{S}} \mathbb{P}^*(s'|s,a)\left(\max_{b\in\mathscr{A}} \widetilde{Q}_k(s',b) - \max_{b\in\mathscr{A}} Q(s',b)\right) + \delta\gamma w_k(s,a) \\
&= (1-\delta)\overline{Q}_k(s,a) + \delta\left(C(s,a;\widetilde{\psi}_k) - C(s,a;\psi^*)\right) + \delta(T\widetilde{Q}_k(s,a) - TQ(s,a)) + \delta\gamma w_n(s,a),
\end{aligned}$$

where $w_k(s,a) = \max_{b\in\mathscr{A}} \widetilde{Q}_k(s',b) - \sum_{s'\in\mathscr{S}} \mathbb{P}^*(s'|s,a) \max_{b\in\mathscr{A}} \widetilde{Q}_n(s',b)$. If z_k is defined by (4) (as done by Beck and Srikant (2012)), then the following holds for the second moment:

$$\mathbb{E}[\|z_k\|_2] \leq \sqrt{\frac{\gamma^2 \delta W_{\max}^2}{2-\delta}}, \tag{5}$$

where $W_{\max}^2 = |\mathscr{S}\times\mathscr{A}| 4\widehat{Q}_{\max}^2$ with $|\mathscr{S}|$ being the cardinality of the set of states and $|\mathscr{A}|$ being the cardinality of the set of possible actions. By defining the sequence $D_k \triangleq \overline{Q}_k - z_k$, we may bound it as follows:

$$\begin{aligned}
& D_{k+1}(s,a) = (1-\delta)D_k(s,a) + \delta\left(C(s,a;\widetilde{\psi}_k) - C(s,a;\psi^*)\right) + \delta(T\widetilde{Q}_k(s,a) - TQ(s,a)) \\
\implies \quad & |D_{k+1}(s,a)| \leq (1-\delta)|D_k(s,a)| + \delta L_C \|\widetilde{\psi}_k - \psi^*\| + \delta \|T\widetilde{Q}_k(s,a) - TQ(s,a)\|_\infty \\
& \leq (1-\delta)|D_k(s,a)| + \delta L_C \|\widetilde{\psi}_k - \psi^*\| + \delta\gamma \|\widetilde{Q}_k - Q\|_\infty \leq (1-\delta)\|D_k\|_\infty + \delta L_C \|\widetilde{\psi}_k - \psi^*\| + \delta\gamma \|\overline{Q}_k\|_\infty,
\end{aligned}$$

where the first inequality follows from the Lipschitz continuity of the cost function and the second inequality follows from Proposition 3. Therefore,

$$\begin{aligned}
\|D_{k+1}\|_\infty &\leq (1-\delta)\|D_k\|_\infty + \delta L_C \|\widetilde{\psi}_k - \psi^*\| + \delta\gamma \|\overline{Q}_k\|_\infty \\
&\leq (1-\delta)\|D_k\|_\infty + \delta L_C \|\widetilde{\psi}_k - \psi^*\| + \delta\gamma(\|D_k\|_\infty + \|z_n\|_\infty) \\
&= (1-\delta(1-\gamma))\|D_k\|_\infty + \delta L_C \|\widetilde{\psi}_k - \psi^*\| + \delta\gamma \|z_k\|_\infty.
\end{aligned}$$

We may then derive a bound for D_k:

$$\begin{aligned}\|D_k\|_\infty &\leq (1-\delta(1-\gamma))\|D_{k-1}\|_\infty + \delta L_C\|\widetilde{\psi}_{k-1}-\psi^*\| + \delta\gamma\|z_{k-1}\|_\infty \\ &\leq (1-\delta(1-\gamma))^2\|D_{k-2}\|_\infty + (1-\delta(1-\gamma))\delta L_C\|\widetilde{\psi}_{k-2}-\psi^*\| + \delta L_C\|\widetilde{\psi}_{k-1}-\psi^*\| \\ &\quad + (1-\delta(1-\gamma))\delta\gamma\|z_{k-2}\|_\infty + \delta\gamma\|z_{k-1}\|_\infty \\ &\leq \quad \vdots \\ &\leq (1-\delta(1-\gamma))^k\|D_0\|_\infty + \delta L_C\sum_{l=0}^{k-1}(1-\delta(1-\gamma))^l\|\widetilde{\psi}_{k-1-l}-\psi^*\| + \delta\gamma\sum_{l=0}^{k-1}(1-\delta(1-\gamma))^l\|z_{k-1-l}\|_\infty.\end{aligned}$$

Recall that the learning problem for ψ^* is strongly convex implying that for some λ and for all k, we have $\mathbb{E}[\|\widetilde{\psi}_k-\psi^*\|] \leq \frac{\lambda}{\sqrt{k}}$. Therefore, for $0<\alpha<1/2$

$$\begin{aligned}\mathbb{E}[\|D_k\|_\infty] &\leq (1-\delta(1-\gamma))^k\|\overline{Q}_0\|_\infty + \delta L_C\sum_{l=0}^{k-1}\frac{(1-\delta(1-\gamma))^l\lambda}{\sqrt{k-1-l}} + \delta\gamma\sum_{l=0}^{k-1}(1-\delta(1-\gamma))^l\|z_{k-1-l}\|_\infty \\ &\leq (1-\delta(1-\gamma))^k\|\overline{Q}_0\|_\infty + \delta L_C\sum_{l=0}^{k-1}\frac{(1-\delta(1-\gamma))^l\lambda}{\sqrt{k-1-l}} + \frac{\delta\gamma}{\delta(1-\gamma)}\sqrt{\frac{\gamma^2\delta W_{\max}^2}{2-\delta}} \\ &= \mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}}\right) + \frac{\gamma^2}{1-\gamma}\sqrt{\frac{\delta W_{\max}^2}{2-\delta}},\end{aligned}$$

where the second inequality utilizes $\mathbb{E}[\|z_k\|_\infty] \leq \mathbb{E}[\|z_k\|_2]$ together with the bound (5) and the last equality utilizes a proof technique similar to that adopted in Prop. 1. ∎

Proposition 5 (Constant steplength error bound for misspecified Q-learning) Suppose $\{\widetilde{Q}_k\}$, and $\{\widetilde{\psi}_k\}$ are generated from Algorithm III. Suppose the learning function $R(\cdot)$ is strongly convex in Ψ and $C(s,a;\psi)$ is Lipschitz continuous in ψ with constant L_C for all s and a. Then, the following holds for any k, $\delta<1$, and $0<\alpha<1/2$:

$$\mathbb{E}\left[\|\overline{Q}_k\|_\infty\right] \leq \mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}}\right) + \frac{\gamma}{1-\gamma}\sqrt{\frac{\delta W_{\max}^2}{2-\delta}}.$$

Proof. The result follows directly from Lemma 3, expression (5), and $\delta<1$:

$$\mathbb{E}\left[\|\overline{Q}_k\|_\infty\right] \leq \mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}}\right) + \frac{\gamma^2}{1-\gamma}\sqrt{\frac{\delta W_{\max}^2}{2-\delta}} + \sqrt{\frac{\gamma^2\delta W_{\max}^2}{2-\delta}} = \mathscr{O}\left(\frac{1}{k^{\frac{1}{2}-\alpha}}\right) + \mathscr{O}\left(\sqrt{\delta}\right).$$

∎

5 FINAL COMMENTS AND CONCLUDING REMARKS

Motivated by the increasing role of streaming data and misspecification in decision-making problems, we consider the resolution of MDPs in which transition matrices are unknown and the cost functions are misspecified. We develop extensions to value iteration, policy iteration and Q-learning through which both misspecification is resolved while solving the original MDP in an asymptotic sense. A precise characterization of the impact of learning on the resulting error bounds is provided in the context of value iteration and Q-learning.

We conclude with a short commentary on the nature of the error bounds. First, we assume that the learning problems are strongly convex since deriving overall rate statements requires bounds on the expected error in parameter estimates. In fact, the knowledge of the convexity constant in the learning problem is crucial in the development of bounds. It is worth emphasizing that if mere convexity assumptions are imposed on the learning problems, the currently adopted avenue cannot be utilized since error bounds are only available in a functional value sense. Furthermore, while averaging-based techniques may be utilized to resolve merely convex learning problems, such approaches provide bounds on the averaged iterates in a functional sense but not on the solution iterates; in the absence of bounds on the solution iterates, one cannot derive rate statements. Second, in the context of Q-learning, we develop a misspecified variant of the constant steplength scheme. Naturally, diminishing steplength versions can also be developed which will be the subject of future work. Third, throughout the paper, we assume that the learning problems are static and consequently, rather than regret-based bounds, we derive error bounds on the optimal functional value or solution.

REFERENCES

Anderson, T. W., and L. A. Goodman. 1957. "Statistical Inference about Markov Chains". *Ann. Math. Statist.* 28:89–110.

Auer, P., T. Jaksch, and R. Ortner. 2009. "Near-Optimal Regret Bounds for Reinforcement Learning". In *Advances in Neural Information Processing Systems 21*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, 89–96. Curran Associates, Inc.

Bartlett, P. L., and A. Tewari. 2009. "REGAL: A Regularization Based Algorithm for Reinforcement Learning in Weakly Communicating MDPs". In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, 35–42. Arlington, Virginia, United States: AUAI Press.

Beck, C. L., and R. Srikant. 2012. "Error Bounds for Constant Step-Size Q-Learning". *Systems & Control Letters* 61 (12): 1203 – 1208.

Bellman, R. 1957. *Dynamic Programming*. 1st ed. Princeton, NJ, USA: Princeton University Press.

Brafman, R. I., and M. Tennenholtz. 2003, March. "R-MAX - A General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning". *J. Mach. Learn. Res.* 3:213–231.

Chang, H., J. Hu, M. Fu, and S. Marcus. 2013. *Simulation-Based Algorithms for Markov Decision Processes*. Communications and Control Engineering. Springer London.

Delage, E., and S. Mannor. 2010. "Percentile Optimization for Markov Decision Processes with Parameter Uncertainty". *Operations Research* 58 (1): 203–213.

Gosavi, A. 2006. "Boundedness of Iterates in Q-Learning". *Systems & Control Letters* 55 (4): 347–349.

Han, F., and H. Liu. 2013, May. "Transition Matrix Estimation in High Dimensional Time Series". In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, edited by S. Dasgupta and D. Mcallester, Volume 28, 172–180: JMLR Workshop and Conference Proceedings.

Howard, R. 1960. *Dynamic Programming and Markov Processes*. The MIT press, New York London.

Jiang, H., and U. V. Shanbhag. 2013. "On the Solution of Stochastic Optimization Problems in Imperfect Information Regimes". In *Winter Simulation Conference*, 821–832.

Jiang, H., and U. V. Shanbhag. 2014. "On the Solution of Stochastic Optimization and Variational Problems in Imperfect Information Regimes". *http://arxiv.org/abs/1402.1457*.

Ljung, L. 1987. "System Identification: Theory for the User". *Prentice Hall Inf and System Sciencess Series, New Jersey* 7632.

Nilim, A., and L. El Ghaoui. 2005, September-October. "Robust Control of Markov Decision Processes with Uncertain Transition Matrices". *Operations Research* 53 (5): 780–798.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2009. *Lectures on Stochastic Programming*, Volume 9 of *MPS/SIAM Series on Optimization*. Philadelphia, PA: SIAM. Modeling and theory.

Tsitsiklis, J. N., and R. Sutton. 1994. "Asynchronous Stochastic Approximation and Q-Learning". In *Machine Learning*, 185–202.

Watkins, C. J. C. H., and P. Dayan. 1992. "Technical Note Q-Learning". *Machine Learning* 8:279–292.

AUTHOR BIOGRAPHIES

HAO JIANG is a Ph.D. student in the Department of Industrial and Enterprise Systems Engineering at University of Illinois at Urbana-Champaign. He received both his bachelor and master degrees in mathematics from Nanjing University in Nanjing, China. His research interests include learning, variational inequalities and game theory. His email address is jiang23@illinois.edu.

UDAY V. SHANBHAG has been an Associate Professor in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Penn. State University since 2012. From 2006, he was an assistant and subsequently a tenured associate professor at the department of Industrial and Enterprise Systems Engineering at the University of Illinois. He received his Ph.D. degree in operations research from the Department of Management Science and Engineering, Stanford University, Stanford, CA, in 2006. His interests lies in the development of analytical and algorithmic tools in the context of optimization and variational problems, in regimes complicated by uncertainty, dynamics and nonsmoothness. Dr. Shanbhag received the triennial A.W. Tucker Prize for his dissertation from the mathematical programming society (MPS) in 2006, the Computational Optimization and Applications (COAP) Best Paper Award in 2007 (with W. Murray), and the best theory paper award at the Winter Simulation Conference (2013) (with F. Yousefian and A. Nedić). His email address is udaybag@psu.edu and his web page is http://www2.ie.psu.edu/shanbhag/personal/index.htm.