Zap Q-Learning for Optimal Stopping Time Problems

Shuhang Chen*   Adithya M. Devraj†   Ana Bušić‡   Sean Meyn§

Abstract—We propose a novel reinforcement learning algorithm that approximates solutions to the problem of discounted optimal stopping in an irreducible, uniformly ergodic Markov chain evolving on a compact subset of $\mathbb{R}^n$. A dynamic programming approach has been taken by Tsitsiklis and Van Roy to solve this problem, wherein they propose a Q-learning algorithm to estimate the value function, in a linear function approximation setting. The Zap-Q learning algorithm proposed in this work is the first algorithm that is designed to achieve optimal asymptotic variance. We prove convergence of the algorithm using ODE analysis, and the optimal asymptotic variance property is reflected via fast convergence in a finance example.

I. INTRODUCTION

Consider a discrete-time Markov chain $X = \{X_n : n \geq 0\}$ evolving on a general state-space $X$. The goal in optimal stopping time problems is to minimize over all stopping times $\tau$, the associated expected cost:

$$E \left[ \sum_{n=0}^{\tau-1} \beta^n c(X_n) + \beta^\tau c_s(X_\tau) \right]$$

(1)

where $c : X \to \mathbb{R}$ denotes the per-stage cost, $c_s : X \to \mathbb{R}$ the terminal cost, and $\beta \in (0, 1)$ is the discount factor. Examples of such a problem arise mostly in financial applications such as derivatives analysis (see Section V), timing of a purchase or sale of an asset, and more generally in sequential analysis problems.

In this work, the optimal decision rule is approximated using reinforcement learning techniques. We propose and analyze an optimal variance algorithm to approximate the value function associated with the optimal stopping rule.

A. Problem Setup

It is assumed that $X \subset \mathbb{R}^m$ is compact, and we let $B$ denote the associated Borel $\sigma$-algebra. The time-homogeneous Markov chain $X$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and its distribution is determined by an initial distribution and a transition kernel $P$, defined for each $x \in X$ and $A \in B$ by

$$P(x, A) = \Pr(X_{n+1} \in A \mid X_n = x)$$

It is assumed that $X$ is uniformly ergodic, so in particular it has a unique invariant probability measure denoted $\pi$ [9].

Denote by $\{\mathcal{F}_n : n \geq 0\}$ the filtration associated with $X$. The Markov property asserts that for bounded measurable functions $h : X \to \mathbb{R}$,

$$E[h(X_{n+1}) \mid \mathcal{F}_n] = \int P(x, dy)h(y) \quad \text{with } x = X_n$$

In this paper a stopping time $\tau$ is a random variable taking on values in the non-negative integers, with the defining property $\{\omega : \tau(\omega) \leq n\} \in \mathcal{F}_n$ for each $n$. A stationary policy is a measurable function $\phi : X \to \{0, 1\}$ that defines a stopping time as follows:

$$\tau^\phi = \inf\{n \geq 0 : \phi(X_n) = 1\}$$

(2)

The (optimal) value function is defined as the infimum of (1) over all stopping times: for any $x \in X$,

$$h^* (x) := \inf_{\tau} \left\{ E \left[ \sum_{n=0}^{\tau-1} \beta^n c(X_n) + \beta^\tau c_s(X_\tau) \mid X_0 = x \right] \right\}$$

(3)

The associated $Q$-function is defined by $Q^* (x) := c(x) + \beta E[h^*(X_1) \mid X_0 = x]$. The optimal stopping problem is a special cast of discounted-cost optimal control [1]. It follows that $Q^*$ solves the following Bellman equation: for each $x \in X$,

$$Q^*(x) = c(x) + \beta E[\min(c_s(X_1), Q^*(X_1)) \mid X_0 = x]$$

(4)

Moreover, the optimal stopping rule is defined by the stationary policy,

$$\phi^*(x) = \mathbb{1}\{c_s(x) \leq Q^*(x)\}$$

(5)

An optimal stopping time is $\tau^* = \tau^{\phi^*}$, using the general definition (2).

The Bellman equation (4) can be expressed as the functional fixed point equation:

$$Q^* = FQ^*$$

(6)

where $F$ denotes the dynamic programming operator: for any function $Q : X \to \mathbb{R}$, and $x \in X$,

$$FQ(x) := c(x) + \beta E[\min(c_s(X_1), Q(X_1)) \mid X_0 = x]$$

Acknowledgements: Financial support from ARO grant W911NF1810334 is gratefully acknowledged. Additional support from EPCN 1609131 & CPS 1646229, and French National Research Agency grant ANR-16-CE05-0008.
Analysis is framed in the usual Hilbert space $L_2(\pi)$ of real-valued measurable functions on $X$ with inner product:

$$\langle f, g \rangle_\pi = E[f(X)g(X)]$$

(7)

and norm:

$$\|f\|_\pi = \sqrt{\langle f, f \rangle_\pi}$$

(8)

where the expectation in (7) is with respect to the steady state distribution $\pi$. It is assumed throughout that the cost functions $c$ and $c_s$ are in $L_2(\pi)$.

B. Objective

The objective is to approximate $Q^*$ using a parameterized family of functions $\{Q^\theta\}$, where $\theta \in \mathbb{R}^d$ denotes the parameter vector. We restrict to linear parameterization throughout, so that:

$$Q^\theta(x) := \theta^T \psi(x), \quad x \in X$$

(9)

where $\psi := [\psi_1, \ldots, \psi_d]^T$ with $\psi_i : X \to \mathbb{R}$, $\psi_i \in L_2(\pi)$, $1 \leq i \leq d$, denotes the basis functions. For any parameter vector $\theta \in \mathbb{R}^d$, we denote the Bellman error

$$B^\theta_c = FQ^\theta - Q^\theta.$$ 

It is assumed that the basis functions are linearly independent: The $d \times d$ covariance matrix $\Sigma_\psi$ is full rank, where

$$\Sigma_\psi(i, j) = \langle \psi_i, \psi_j \rangle_\pi, \quad 1 \leq i, j \leq d$$

(10)

In a finite state-space setting, it is possible to construct a consistent algorithm that computes the Q-function exactly [8]. The Q-learning algorithm of Watkins [18], [19] can be used in this case (see [21] for a discussion).

In a function approximation setting, we need to relax the goal of solving (4). As in previous works [17], [6], [21], the goal in this paper is to obtain the solution to a *Galerkin relaxation* of (4): Find $\theta^*$ such that,

$$E[B^\theta_c(X_n)\zeta_n(i)] = 0, \quad 1 \leq i \leq d,$$

(11)

where $\{\zeta_n\}$ is adapted to $\mathcal{F}_n$, and the expectation is in steady state. For a given constant $\lambda \in [0, 1)$, the Zap-$Q(\lambda)$ algorithm that is proposed in this work intends to solve (11) with

$$\zeta_n = \sum_{k=0}^{\infty} \lambda^{n-k} \psi(X_k)$$

in which $X$ is a stationary realization. This is similar to what is considered in the TD($\lambda$) algorithm [12], [16].

In much of this paper we fix $\lambda = 0$ for simplicity. In this special case, we have: $\zeta_n \equiv \psi(X_n)$, for each $n \geq 0$, and the goal is to find $\theta^*$ such that, for each $1 \leq i \leq d$,

$$\langle FQ^{\theta^*} - Q^{\theta^*}, \psi_i \rangle_\pi = 0$$

(12)

C. Literature Survey

Obtaining an approximate solution to the original problem (4) using a modified objective (12) was first considered in [17]. The authors propose an extension of the TD(0) algorithm of [13], and prove convergence under general conditions. Though it is not obvious at first sight, the algorithm in [17] is more closely connected to Watkins’ Q-learning algorithm [18], [19], than the TD(0) algorithm. This is specifically due to a minimum term that appears in (12) (see definition of $F$ in (6)), similar to what appears in Q-learning. This is important to note, because Q-learning algorithms are not known to converge under function approximation settings, and this is due to the fact that the dynamic programming operator may not be a contraction in general [1]. The operator $F$ defined in (6) is quite special in this sense: it can be shown that it is a contraction with respect to the $\pi$-norm defined in (8) [15]:

$$\|FQ - FQ'\| \leq \beta \|Q - Q'\|, \quad \text{for all } Q, Q' \in L_2(\pi)$$

Since [15], many other algorithms have been proposed to improve the convergence rate. In [6] the authors propose a matrix gain variant of the algorithm presented in [15], improving the rate of convergence in numerical studies [7]. In [21], the authors take a least squares approach to solve the problem, and propose the *least squares Q-learning algorithm*, that has close resemblance to the least squares policy evaluation algorithm (LSPE (0) of [10]). The authors recognize the high computational complexity of the algorithm, and propose variants in [20]. In prior works [6] and [20], though a function-approximation setting is considered, the state-space is assumed finite.

More recently, in [8], [7], the authors propose the Zap Q-learning algorithm to solve for a solution to a fixed point equation similar to (but more general than) (4). The proof of convergence is provided only for the tabular case (wherein the $\psi_i$’s span all possible functions), and when the state-action space is finite.

The remainder of the paper is organized as follows: Section II contains the approximation architecture, and introduces the Zap Q-learning algorithm. The assumptions and main results are contained in Section III. Section IV provides a high-level proof of the results, numerical results are collected together in Section V, and conclusions in Section VI. Full proofs are available in an extended version of this paper, available on arXiv [4].

II. Q-LEARNING FOR OPTIMAL STOPPING

A. Notation

The following notation will be used throughout. Define for each $\theta \in \mathbb{R}^d$, the corresponding policy $\phi^\theta : X \to \mathbb{R}$.
\[ \{0, 1\}, \quad \phi^\theta(x) := \mathbb{I}\{c_s(x) \leq Q^\theta(x)\}, \quad \text{(13)} \]

For any function \( f \) with domain \( X \), two operators are defined as the simple products,
\[
S_b f(x) := \mathbb{I}(Q^\theta(x) < c_s(x)) f(x) \quad \text{(14a)}
\]
\[
S_b^0 f(x) := \mathbb{I}(c_s(x) \leq Q^\theta(x)) f(x) \quad \text{(14b)}
\]

Observe that \( S_b f = (1 - \phi^\theta) f \).

We then denote a \( d \times d \) matrix, two \( d \)-dimensional vectors as follows:
\[
A(\theta) := \mathbb{E}[\psi(X_n)\beta S_b\psi^\prime(X_{n+1}) - \psi(X_n)\psi^\prime(X_n)] \quad \text{(15)}
\]
\[
b^* := \mathbb{E}[\psi(X_n)c_s(X_n)] \quad \text{(16)}
\]
\[
\tau_s(\theta) := \mathbb{E}[\psi(X_n)S_b^0 c_s(X_{n+1})] \quad \text{(17)}
\]

The objective (12) can be expressed:
\[
A(\theta^*)\theta^* + \beta\tau_s(\theta^*) + b^* = 0 \quad \text{(18)}
\]

**B. Zap Q-Learning**

It is useful to first look at a more general class of “matrix gain Q-learning” algorithms. Given a \( d \times d \) matrix gain sequence \( \{G_n : n \geq 0\} \) with each \( G_n \) invertible, and a scalar step-size sequence \( \{\alpha_n : n \geq 0\} \), the corresponding matrix gain Q-learning algorithm for optimal stopping is given by the following recursion:
\[
\theta_{n+1} = \theta_n + \alpha_n G_{n+1}^{-1}\psi(X_{n+1})d_{n+1} \quad \text{(19)}
\]

with \( \{d_n\} \) denoting the “temporal difference” sequence:
\[
d_{n+1} = c(X_n) + \beta \min(c_s(X_{n+1}), Q^\theta(X_{n+1})) - Q^\theta(X_n). \quad \text{(20)}
\]

The algorithm proposed in [17] is (19), with \( G_n \equiv I \) (the \( d \times d \) identity matrix). This is similar to the TD(0) algorithm [16], [13]. The fixed point Kalman filter algorithm of [6] can also be written as a special case of (19): Each \( G_n \) is an estimate of the matrix \( \Sigma_n \) defined in (10), which can be obtained recursively:
\[
G_{n+1} = G_n + \alpha_n G_{n+1}^{-1}\left[\psi(X_n)\psi^\prime(X_{n+1}) - G_n\right] \quad \text{(20)}
\]

In the Zap Q(\( \lambda \)) algorithm, the matrix gain sequence \( \{G_n\} \) is designed so that the asymptotic covariance of the resulting algorithm is minimized (see Section III for details). Similar to the Zap-Q algorithm of [8], it uses matrix gain \( A_{n+1}^b \) (the projected pseudo-inverse of \( A_n \)); an estimate of \( A^{-1}(\theta_n) \), with \( A(\theta) \) defined in (15).

The term inside the expectation in (15), following the substitution \( \theta = \theta_n \), is denoted
\[
A_{n+1} := \psi(X_n)\left[\beta S_b\psi^\prime(X_{n+1}) - \psi(X_n)\right]^T \quad \text{(21)}
\]

Using (21), the matrix \( A(\theta_n) \) is recursively estimated using Monte Carlo in the Zap Q(\( \lambda \)) algorithm. For simplicity we give details only for \( \lambda = 0 \):

**Algorithm 1 Zap Q(0) Algorithm for Optimal Stopping**

**Input:** Initial \( \theta_0 \in \mathbb{R}^d \), \( \tilde{A}_n \); \( d \times d \), negative definite; step-size sequences \( \{\alpha_n\} \) and \( \{\gamma_n\} \) and \( n = 0 \)

1: \textbf{repeat}
2: \text{Obtain the Temporal Difference term:}
\[
d_{n+1} = c(X_n) + \beta \min(c_s(X_{n+1}), Q^\theta(X_{n+1})) - Q^\theta(X_n) \quad \text{(22)}
\]
3: \text{Update the matrix gain estimate} \( \tilde{A}_n \) of \( A(\theta_n) \), with \( A_{n+1} \) defined in (21):
\[
\tilde{A}_{n+1} = \tilde{A}_n + \gamma_{n+1}\left[A_{n+1} - \tilde{A}_n\right] \quad \text{(23)}
\]
4: \text{Update the parameter vector:}
\[
\theta_{n+1} = \theta_n - \alpha_{n+1}\tilde{A}_{n+1}^\dagger\psi(X_{n+1})d_{n+1} \quad \text{(24)}
\]
5: \( n = n + 1 \)
6: \textbf{until} \( n \geq N \)

**Output:** \( \theta = \theta_N \)

The gain sequences \( \{\alpha_n\} \) and \( \{\gamma_n\} \) in the above algorithm are chosen as: for some \( \rho \in (0.5, 1) \),
\[
\alpha_n = 1/n, \quad \gamma_n = 1/n^\rho. \quad \text{(24)}
\]

For each \( \theta \in \mathbb{R}^d \), consider the following terms:
\[
b(\theta) = -A(\theta)\theta - \beta\tau_s(\theta) \quad \text{(25a)}
\]
\[
c^\theta(x) = Q^\theta(x) \quad \text{(25b)}
\]
\[
- \mathbb{E}[\beta \min(c_s(X_n), Q^\theta(X_n))] | X_{n+1} = x \quad \text{(26)}
\]

The vector \( b(\theta) \) is analogous to \( b^* \) in (18), and (25b) recalls the Bellman equation (4). Prop. 2.1 follows from these definitions. It shows that \( b(\theta) \) is the “projection” of the cost function \( c^\theta \), similar to how \( b^* \) is related to \( c \) through (16).

**Proposition 2.1:** For each \( \theta \in \mathbb{R}^d \), we have:
\[
b(\theta) = \mathbb{E}[c^\theta(X_n)\psi(X_n)] \quad \text{(26)}
\]

where the expectation is in steady state. In particular, \( b^* = b(\theta^*) \)

\( \square \)

**III. ASSUMPTIONS AND MAIN RESULTS**

**A. Preliminaries**

Preliminary results are summarized here that will be useful in establishing the main results. We start with the contraction property of the dynamic programming operator \( F \) defined in (6). This is a result directly obtained from [17] (Lemma 4 on p. 1844).

**Lemma 3.1:** The dynamic programming operator \( F \) defined in (6) satisfies:
\[
\|FQ - FQ'\| \leq \beta\|Q - Q'\|, \quad Q, Q' \in L_2(\pi). \quad \text{(27)}
\]

Furthermore, \( Q^* \) is the unique fixed point of \( F \) in \( L_2(\pi) \).

\( \square \)
Recall that \( Q^\theta : X \to \mathbb{R} \) is defined in (9). Similar to the operator \( F \), for each \( \theta \in \mathbb{R}^d \) we define operators \( H^\theta \) and \( F^\theta \) that operate on functions \( Q : X \to \mathbb{R} \) as follows:

\[
H^\theta Q(x) = \begin{cases} Q(x), & \text{if } Q^\theta(x) < c_s(x) \\ c_s(x), & \text{otherwise} \end{cases}
\]  

(27)

\[
F^\theta Q = c + \beta PH^\theta Q.
\]  

(28)

The following Lemma is a slight extension of Lemma 3.1.

**Lemma 3.2:** For each \( \theta \in \mathbb{R}^d \), the operator \( F^\theta \) satisfies:

\[
\| F^\theta Q - F^\theta Q' \| \leq \beta \| Q - Q' \|, \quad \forall Q, Q' \in L_2(\pi).
\]

**Proof:** Based on the definition (28), we have:

\[
\| F^\theta Q - F^\theta Q' \| = \beta \| PH^\theta Q - PH^\theta Q' \|
\]

\[
\leq \beta \| H^\theta Q - H^\theta Q' \|
\]

\[
\leq \beta \| Q - Q' \|,
\]

where the first inequality follows from the fact that \( \| P \| \leq 1 \) (with \( \| P \| \) the induced operator norm in \( L_2(\pi) \)). The last inequality is true because:

\[
H^\theta Q(x) - H^\theta Q'(x) = S_\theta (Q - Q')(x), \quad x \in X
\]

The next result is a direct consequence of Lemma 3.2.

**Lemma 3.3:** For each \( \theta \in \mathbb{R}^d \),

(i) The \( d \times d \) matrix \( A(\theta) \) defined in (15) satisfies:

\[
- v^T A(\theta) v \geq (1 - \beta) v^T \Sigma_\psi v,
\]

(29)

for each \( v \in \mathbb{R}^d \), with \( \Sigma_\psi \) defined in (10).

(ii) Eigenvalues of \( A(\theta) \) are strictly bounded away from 0, and \( \{ A^{-1}(\theta) : \theta \in \mathbb{R}^d \} \) are uniformly bounded.

Prop. 2.1 implies a Lipschitz bound on the function \( b \) defined in (25a):

**Lemma 3.4:** The mapping \( b \) is Lipschitz: For some \( \ell_1 > 0 \), and each \( \theta^1, \theta^2 \in \mathbb{R}^d \),

\[
\| b(\theta^1) - b(\theta^2) \| \leq \ell_1 \| \theta^1 - \theta^2 \|
\]

**Assumption A3:** The conditional distribution of \( \psi(X_{n+1}) \) given \( X_n = x \) has a density, \( f_{\psi|\psi}(z) \). This density is also assumed to have uniformly bounded likelihood ratio \( \ell(z \mid x) \) with respect to the Gaussian density \( \mathcal{N}(\psi(x), I) \). Consequently, \( \Sigma_\psi > 0 \).

It is assumed moreover that the function \( c_s \) is in the span of \( \{ \psi_i \} \).

**Assumption A4:** The parameter sequence \( \{ \theta_n : n \geq 1 \} \) is bounded a.s.

Assumption (A3) consists of technical conditions in the proof. The density assumption is imposed to ensure that the conditional expectation given \( X_n \) of functions such as \( S_\theta \psi^i(X_{n+1}) \) are smooth as a function of \( \theta \).

As for (A4), it is highly likely that boundedness can be established via an extension of the “Borkar & Meyn Theorem” of [3], [2]. The “ODE at infinity” posed there is stable as required, but the two-time scale algorithm presents a challenge with application.

**C. Main Result**

The main result of this paper establishes convergence of iterates \( \{ \theta_n \} \) obtained using Algorithm 1:

**Theorem 3.5:** Suppose that Assumptions A1-A4 hold. Then,

(i) The parameter sequence \( \{ \theta_n \} \) obtained using the Zap-Q(0) algorithm converges to \( \theta^* \) a.s., where \( \theta^* \) satisfies (12).

(ii) An ODE approximation holds for the sequences \( \{ \theta_n, b(\theta_n) \} \) by continuous time functions \( (w, b) \) satisfying

\[
\frac{d}{dt} b(t) = -b(t) + b
\]

\[
b(t) = -A(w(t))w(t) - \beta \Sigma_\psi(w(t))
\]

(30)

The term **ODE approximation** is standard in the SA literature: For \( t \geq s \), let \( w^s(t) \) denote the solution to:

\[
\frac{d}{dt} w^s(t) = \xi(w^s(t)), \quad w^s(s) = \bar{w}(s)
\]

for some \( \xi : \mathbb{R}^d \to \mathbb{R}^d \). We say that the ODE approximation:

\[
\frac{d}{dt} w(t) = \xi(w(t))
\]

holds for the sequence \( \{ \theta_n : n \geq 1 \} \) if the following is true for any \( T > 0 \):

\[
\lim_{s \to -\infty} \sup_{t \in [s, s+T]} \| \bar{w}(t) - w^s(t) \| = 0, \text{ a.s.}
\]

where \( \bar{w}(t) \) denotes the continuous time process constructed from the sequence \( \{ \theta_n : n \geq 0 \} \), whose meaning will be made precise in Sec IV-B. The optimality of the algorithm in terms of the asymptotic variance is discussed next.
D. Asymptotic Variance

The asymptotic covariance $\Sigma_\theta^*$ of any algorithm is defined to be the following limit:

$$\Sigma_\theta^* := \lim_{n \to \infty} nE\left[(\theta_n - \theta^*)(\theta_n - \theta^*)^T\right]$$  \hfill (32)

Consider the matrix gain Q-learning algorithm (19), and suppose the matrix sequence $\{G_n: n \geq 0\}$ is constant: $G_n \equiv G$. Also, suppose that all eigenvalues of $GA(\theta^*)$ satisfy $\lambda(GA(\theta^*)) < -\frac{1}{2}$. Following standard analysis (see Section 2.2 of [8] and references therein), it can be shown that, under general assumptions, the asymptotic covariance of the algorithm (19) can be obtained as a solution to the Lyapunov equation:

$$(GA(\theta^*) + \frac{1}{2}I)\Sigma_\theta + \Sigma_\theta (GA(\theta^*) + \frac{1}{2}I)^T + G\Sigma_\theta G^T = 0$$  \hfill (33)

where $\Sigma_\theta$ is the “noise covariance matrix”, which is defined as follows.

A “noise sequence” $\{\varepsilon_n\}$ is defined as

$$\varepsilon_n := A_{n+1} \theta^* + \tilde{b}_{n+1} + \tilde{A}_{n+1} \tilde{\theta}_n$$  \hfill (34)

where $\tilde{A}_{n+1} := A_{n+1} - A(\theta^*)$, $\tilde{b}_{n+1} := b_{n+1} - b^*$, $\tilde{\theta}_n := \theta_n - \theta^*$, with $A_{n+1}$ defined in (21). $\theta^*$ defined in (15).

$$b_{n+1} := \psi(X_n)[c(X_n) + S_{\theta_n} c_s(X_{n+1})]$$  \hfill (35)

and $b^*$ defined in (16). For the matrix gain algorithm with $G_n \equiv I$, the algorithm would be deterministic in the ideal case $\varepsilon_n \equiv 0$.

The noise covariance matrix is defined as the limit

$$\Sigma_\varepsilon = \lim_{T \to \infty} \frac{1}{T} E[\varepsilon_n^T \varepsilon_1]$$  \hfill (36)

in which $S_T = \sum_{n=1}^T \varepsilon_n$, and the expectation is in steady state.

Optimality of the asymptotic covariance: The asymptotic covariance $\Sigma_\theta^*$ can be obtained as a solution to (33) only when all eigenvalues satisfy $\lambda(GA(\theta^*)) < -\frac{1}{2}$. If there exists at least one eigenvalue such that $\lambda(GA(\theta^*)) \geq -\frac{1}{2}$, then, under general conditions, it can be shown that the asymptotic covariance is not finite. This implies that the rate of convergence of $\theta_n$ is slower than $O(1/\sqrt{n})$.

It is possible to optimize the covariance $\Sigma_\theta^*$ over all matrix gains $G$ using (33). Specifically, it can be shown that letting $G^* = \tilde{A}(\theta^*)^{-1}$ will result in the minimum asymptotic covariance $\Sigma^*$, where

$$\Sigma^* = \tilde{A}(\theta^*)^{-1} \Sigma_\varepsilon \tilde{A}(\theta^*)^{-1}^T$$  \hfill (37)

That is, for any other gain $G_n \equiv G$, denoting $\Sigma_\theta^G$ to be the asymptotic covariance of the algorithm (19), obtained as a solution to the Lyapunov equation (33), the difference $\Sigma_\theta^G - \Sigma^*$ is positive semi-definite.

The Zap Q(0) algorithm is specifically designed to achieve the optimal asymptotic covariance. A full proof of optimality will require extra effort. Thm. 3.5 tells us that we have the required convergence $G_n \to G^*$ for this algorithm. Provided we can obtain additional tightness bounds for the scaled error $\sqrt{n}\tilde{\theta}_n$, we obtain a functional Central Limit Theorem with optimal covariance $\Sigma^*$ [2]. Minor additional bounds ensure convergence of (32) to the optimal covariance $\Sigma^*$.

The next section is dedicated to the proof of Thm. 3.5.

IV. PROOF OF THEOREM 3.5

A. Overview of the Proof

Unlike martingale difference assumptions in standard stochastic approximation [2], the noise in our algorithm is Markovian. The first part of this section establishes that our noise sequence satisfies the so called ODE friendly property [14], such that their asymptotic effect over the parameter update is zero. This enables the argument that the gain matrices are close to their equilibrium over the fast time scale defined by $\{\gamma_n\}$. We then go back to the slow time scale defined by $\{\alpha_n\}$, and obtain the ODE approximations for $\{\theta_n\}$ and the expected projected cost $b(\theta_n)$.

B. ODE Analysis

The remainder of this section is devoted to the proof of the ODE approximation (30). The construction of an approximating ODE involves first defining a continuous time process $\overline{w}$. Denote

$$t_n = \sum_{i=1}^n \alpha_i, \quad n \geq 1, \quad t_0 = 0,$$  \hfill (38)

and define $\overline{w}(t_n) = \tilde{\theta}_n$ at these time points, with the definition extended to $\mathbb{R}_+$ via linear interpolation.

Along with the piecewise linear continuous-time process $\{\overline{w}_t: t \geq 0\}$, denote by $\{\overline{A}_n: t \geq 0\}$ the piecewise linear continuous-time process defined similarly, with $\overline{A}_n = \tilde{A}_n$, $n \geq 1$. Furthermore, for each $t \geq 0$, denote

$$\tilde{b}_t \equiv b(\overline{w}_t) := -\tilde{A}(\overline{w}_t)\overline{w}_t - \beta \overline{c}_s(\overline{w}_t)$$

To construct an ODE, it is convenient first to obtain an alternative and suggestive representation for the pair of equations (22,23).

The following definition is needed in Lemma 4.1:

ODE-friendly sequence: A vector-valued sequence of random variables $\{\varepsilon_k\}$ will be called ODE-friendly if it admits the decomposition,

$$\varepsilon_k = \Delta_k + \overline{T}_k - \overline{T}_{k-1} + \varepsilon_k, \quad k \geq 1$$  \hfill (39)

in which:

(i) $\{\Delta_k: k \geq 1\}$ is a martingale-difference sequence satisfying $E[\|\Delta_k\|^2 | F_k] \leq \sigma^2 < \infty$ a.s. for all $k$
(ii) \( \{ T_k : k \geq 1 \} \) is a bounded sequence.

(iii) The final sequence \( \{ \varepsilon_k \} \) is bounded and satisfies:
\[
\sum_{k=1}^{\infty} \gamma_k \| \varepsilon_k \| < \infty \quad \text{a.s.} \tag{40}
\]

Lemma 4.1 establishes that the error sequences that appear in the updates for \( \{ \theta_n \} \) and \( \{ \hat{A}_n \} \) are ODE-friendly.

**Lemma 4.1:** The pair of equations (22, 23) can be expressed:
\[
\theta_{n+1} = \theta_n - \alpha_{n+1} \hat{A}_{n+1}^1 \left[ A(\theta_n) \theta_n + \beta \sigma(\theta_n) + b^* \right]
\]
\[
\hat{A}_{n+1} = \hat{A}_n + \gamma_{n+1} \left[ A(\theta_n) - \hat{A}_n + \varepsilon_{n+1}^A \right]
\]
where \( A(\theta_n) \) is a linear operator defined in Section 20.1 of [9].

The proof argument is decomposed noise sequences into tractable terms, each of which is shown to be ODE-friendly by standard arguments based on solutions to Poisson’s equation [11]. We only present the treatment for \( \varepsilon_{n+1}^A \) here through Lemma 4.2 to 4.5 since same technique can be applied to \( \varepsilon_{n+1}^A \).

Denote for \( n \geq 0, \)
\[
M_{\psi, \theta}(n+1) = \psi(X_n) S_{\psi, \theta} \psi(X_{n+1}) - \psi(X_n)
\]
For noise sequence \( \varepsilon_{n+1}^A \) in (41) there are \( A_{n+1} - A(\theta_n) \) to Poisson’s equation with forcing \( A_{n+1} - A(\theta_n) \). The argument proceeds by decomposing noise sequences into tractable terms, each of which is shown to be ODE-friendly by standard arguments based on solutions to Poisson’s equation [11]. We only present the treatment for \( A_{n+1} - A(\theta_n) \) here through Lemma 4.2 to 4.5 since same technique can be applied to \( A_{n+1} - A(\theta_n) \).

**Lemma 4.2:** Suppose that \( f(\theta) \) is a bounded function on \( X \) with zero mean. Then the sequence \( \{ f(X_n) \} \) is ODE-friendly with \( \{ \varepsilon_k \} \) equal to zero, and \( T_k = f(X_k) \), with \( f \) the solution to Poisson’s equation with forcing function \( f \).

The sequence \( \{ \hat{A}_{n+1}^1 \} \) is then ODE-friendly since it is a bounded over state space \( X \) with zero mean. Before we show the same for sequences \( \{ \hat{A}_{n+1}^2, \hat{A}_{n+1}^3 \} \), we need two preliminary results: the Lipschitz continuity of \( E[M_{\psi, \theta}(n+1) | F_n] \), and that this continuity is preserved in associated Poisson equation solutions.

**Lemma 4.3:** There is a deterministic constant \( \ell_M < \infty \) such that, with probability one,
\[
\| E[M_{\psi, \theta}(n+1) - M_{\psi, \theta}(n+1) | F_n] \| \leq \ell_M \theta^1 - \bar{\theta}^2
\]

The second result is similar to Lemma 4.2. Both are consequences of the fact that the fundamental kernel defined in Section 20.1 of [9] is a bounded linear operator on \( L_\infty \) when the Markov chain is uniformly ergodic.

**Lemma 4.4:** There is a constant \( B_Z \) such that the following holds: For any family of zero mean functions \( \{ f_0 \} \) satisfying for some constants \( B_F, \ell_F, \)
\[
\sup_x \| f_0(x) \| \leq B_F, \quad \sup_x \| f_{\theta_1}(x) - f_{\theta_2}(x) \| \leq \ell_F \theta^1 - \bar{\theta}^2
\]
for all \( \theta, \theta^1, \theta^2 \), then there are solutions to corresponding Poisson’s equation, \( \{ \hat{f}_0 \} \), with zero mean, and satisfying
\[
\sup_x \| \hat{f}_{\theta_1}(x) - \hat{f}_{\theta_2}(x) \| \leq B_Z \ell_F \theta^1 - \bar{\theta}^2
\]

Now we are ready to claim:

**Lemma 4.5:** The sequences \( \{ \hat{A}_{n+1}^2, \hat{A}_{n+1}^3 \} \) are ODE-friendly.

The following Lemma shows that the matrix gain \( \hat{A}_n \), recursively obtained by (22), approximates the mean \( A(\theta) \), with \( \theta = \theta_n \).

**Lemma 4.6:** Suppose the sequence \( \{ \varepsilon_{n+1}^A : n \geq 1 \} \) is ODE-friendly. Then,

(i) \( \lim_{n \to \infty} \| \hat{A}_n - A(\theta_n) \| = 0, \ a.s. \)

(ii) Consequently, \( \hat{A}_n^1 \neq \hat{A}_n^{-1} \) only finitely often, and \( \lim_{n \to \infty} \| \hat{A}_n^1 - A^{-1}(\theta_n) \| = 0, \ a.s. \)

With the definition of ODE approximation below (31), we have:

**Lemma 4.7:** The ODE approximation for \( \{ \theta_n \} \) holds with probability one, \( \bar{w}(t) \) asymptotically tracks the ODE:
\[
\frac{d}{dt} w(t) = -A^{-1}(w(t)) \left[ b^* - b(w(t)) \right]
\]

For a fixed but arbitrary time horizon \( T > 0 \), we define two families of uniformly bounded and uniformly Lipschitz continuous functions:
\[
\{ \bar{w}(s + t), t \in [0, T] \}_{s \geq 0} \quad \text{and} \quad \{ \bar{b}(s + t), t \in [0, T] \}_{s \geq 0}.
\]
Sub-sequential limits of \( \{ \bar{w}(s + t), t \in [0, T] \}_{s \geq 0} \) and \( \{ \bar{b}(s + t), t \in [0, T] \}_{s \geq 0} \) are denoted \( w_t \) and \( b_t \) respectively.

We recast the ODE limit of the projected cost as follows:

**Lemma 4.8:** For any sub-sequential limits \( \{ w_t, b_t \} \),

(i) They satisfy \( b_t = b(w_t) \).

(ii) For a.e. \( t \in [0, T], \)
\[
\frac{d}{dt} b_t = -A(w_t) \frac{d}{dt} w_t = -b_t + b^*
\]
Proof of Thm. 3.5: Boundedness of sequences \{\hat{A}_n : n \geq 0\} and \{\hat{A}_n^{-1} : n \geq 0\} is established in Lemma 4.6. Together with boundedness assumption of \{\theta_n : n \geq 0\}, the ODE approximation is established in Lemma 4.8. Result (i) then follows from those two results using standard arguments from [2]. ■

V. NUMERICAL RESULTS

In this section we illustrate the performance of the Zap Q-learning algorithm in comparison with existing techniques, on a finance problem that has been studied in prior work [6], [17]. We observe that the Zap algorithm performs very well, despite the fact that some of the technical assumptions made in Section III do not hold.

A. Finance model

The following finance example is used in [6], [17] to evaluate the performance of their algorithms for optimal stopping. The reader is referred to these references for complete details of the problem set-up.

The Markovian state process considered is the vector of ratios: \(X_n = (\hat{p}_{n-99}, \hat{p}_{n-98}, \ldots, \hat{p}_{n-100})/\hat{p}_{n-100}, n \geq 0\), in which \{\hat{p}_t : t \in \mathbb{R}\} is a geometric Brownian motion (derived from an exogenous price-process). This uncontrolled Markov chain is positive Harris recurrent on the state space \(\mathbb{X} \equiv \mathbb{R}^{100}\), so \(X\) is not compact. The Markov chain is uniformly ergodic.

The “time to exercise” is modeled as a stopping time \(\tau \in \mathbb{Z}^+\). The associated expected reward is defined as \(E[\beta^\tau r(X_\tau)]\), with \(r(X_n) := X_n(100) = \hat{p}_n/\hat{p}_{n-100}\) and \(\beta \in (0, 1)\) fixed. The objective of finding a policy that maximizes the expected reward is modeled as an optimal stopping time problem.

The value function is defined to be the infimum (3), with \(c \equiv 0\) and \(c_{\tau} \equiv -r\) (the objective in Section I is to minimize the expected cost, while here, the objective is to maximize the expected reward). The associated Q-function is defined using (4), and the associated optimal policy using (5): \(\phi^*(x) := \mathbb{I}\{r(x) \geq Q^*(x)\}\).

When the Q-function is linearly approximated using (9), for a fixed parameter vector \(\theta\), the associated value function can be expressed:

\[
\hat{h}_{\phi^*}(x) := E[\beta^\tau r(X_\tau) \mid x_0 = x],
\]

where

\[
\tau_0 := \min\{n: \phi^*(X_n) = 1\}
\]

and

\[
\phi^*(x) := \mathbb{I}\{r(x) \geq Q^*(x)\}.
\]

Given a parameter estimate \(\hat{\theta}\) and the initial state \(X(0) = x\), the corresponding average reward \(\hat{h}_{\phi^*}(x)\) was estimated using Monte-Carlo in the numerical experiments that follow.

B. Approximation & Algorithms

Along with Zap Q-learning algorithm we also implement the fixed point Kalman filter algorithm of [6] to estimate \(\theta^*\). This algorithm is given by the update equations (19) and (20). The computational as well as storage complexities of the least squares Q-learning algorithm (and its variants) [20] are too high for implementation.

C. Implementation Details

The experimental setting of [6], [17] is used to define the set of basis functions and other parameters. We choose the dimension of the parameter vector \(d = 10\), with the basis functions defined in [6]. The objective here is to compare the performances of the fixed point Kalman filter algorithm with the Zap-Q learning algorithm in terms of the resulting average reward (45).

Recall that the step-size for the Zap Q-learning algorithm is given in (24). We set \(\gamma_n = n^{-0.85}\) for all implementations of the Zap algorithm, but similar to what is done in [6], we experiment with different choices for \(\alpha_n\). Specifically, in addition to \(\alpha_n = n^{-1}\), we let:

\[
\alpha_n = \frac{b}{b+n}
\]

with \(b = 1e4\) and experiment with \(g = 2, 5,\) and 10. In addition, we also implement Zap with \(\alpha_n = 0.1/n\) and \(\gamma_n = 1/n^{0.85}\). Based on the discussion in Section III-D, we expect this choice of step-size sequences to result in infinite asymptotic variance.

In the implementation of the fixed point Kalman filter algorithm, as suggested by the authors, we choose step-size \(\alpha_n = 1/n\) for the matrix gain update rule in (20), and step-size of the form (47) for the parameter update in (19). Specifically, we let \(b = 1e4\), and \(g = 100\) and 200.

The number of iterations for each of the algorithm is fixed to be \(N = 2e6\).

D. Experimental Results

The average reward histogram was obtained by the following steps: We simulate 500 parallel simulations of each of the algorithms to obtain as many estimates of \(\theta^*\). Each of these estimates defines a policy \(\phi^*\) defined in (46). We then estimate the corresponding average reward \(\hat{h}_{\phi^*}(x)\) defined in (45), with \(X(0) = x = 1\).

Along with the average discounted rewards, we also plot the histograms to visualize the asymptotic variance (32), for each of the algorithms. The theoretical values of the covariance matrices \(\Sigma^*\) and \(\Sigma^*_G\) were estimated through the following steps: The matrices \(A(\theta^*)\) and \(\Sigma_\phi\) (the limit of the matrix gain used in [6]) were estimated via Monte-Carlo. Estimation of \(A(\theta^*)\) requires an estimate of \(\theta^*\); this was taken to be \(\hat{\theta}_N\) obtained using the Zap-Q algorithm with \(\alpha_n = n^{-1}\) and \(\gamma_n = n^{-0.85}\). This estimate of \(\theta^*\) was also used to estimate the covariance matrix \(\Sigma_\Delta\) defined in (36) using the batch means method. The matrices \(\Sigma^*_G\) and \(\Sigma^*\) were then obtained using (33) and (37), respectively.

Fig. 1 contains the histograms of the average rewards obtained using the above algorithms. Fig. 2 contains the
histograms of $\sqrt{N}(\theta_N(8) - \theta^*(8))$ along with a plot of the theoretical prediction.

It was observed that the eigenvalues of the matrix $A(\theta^*)$ have a wide spread: The condition-number is of the order $10^4$. Despite a badly conditioned matrix gain, it is observed in Fig. 1, that the average rewards of the Zap-Q algorithms are better than its competitors. It is also observed that the algorithm is robust to the choice of step-sizes. In Fig. 2 we observe that the asymptotic behavior of the algorithms are a close match to the theoretical prediction. Specifically, large variance of Zap-Q with step-size $\alpha_n = 0.1/n$ confirms that the asymptotic variance is very large (ideally, infinity), if the eigenvalues of the matrix $GA(\theta^*) > -\frac{1}{2}$.

VI. Conclusion

In this paper, we extend the theory for the Zap Q-learning algorithm to a linear function approximation setting, with application to optimal stopping. We prove convergence of the algorithm using ODE analysis, and also observe that it achieves optimal asymptotic variance.

REFERENCES

[1] D. Bertsekas and J. N. Tsitsiklis, Neuro-Dynamic Programming. Athena Scientific, Cambridge, Mass, 1996.
[2] V. S. Borkar. Stochastic Approximation: A Dynamical Systems Viewpoint. Hindustan Book Agency and Cambridge University Press (jointly), Delhi, India and Cambridge, UK, 2008.
[3] V. S. Borkar and S. P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. SIAM J. Control Optim., 38(2):447–469, 2000. (also presented at the IEEE CDC, December, 1998).
[4] S. Chen, A. M. Devraj, A. Bušić, and S. P. Meyn. Zap q-learning for optimal stopping time problems. arXiv preprint arXiv, 2019.
[5] G. Cheng. Note on some upper bounds for the condition number. JOURNAL OF MATHEMATICAL INEQUALITIES, 8(2):369–374, 2014.
[6] D. Choi and B. Van Roy. A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning. Discrete Event Dynamic Systems: Theory and Applications, 16(2):207–239, 2006.
[7] A. M. Devraj and S. Meyn. Zap Q-learning. In Advances in Neural Information Processing Systems, pages 2235–2244, 2017.
[8] A. M. Devraj and S. P. Meyn. Fastest convergence for Q-learning. ArXiv e-prints, July 2017.
[9] S. P. Meyn and R. L. Tweedie. Markov chains and stochastic stability. Cambridge University Press, Cambridge, second edition, 2009. Published in the Cambridge Mathematical Library. 1993 edition online.
[10] A. Nedic and D. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. Discrete Event Dynamic Systems: Theory and Applications, 13(1-2):79–110, 2003.
[11] A. Shwartz and A. Makowski. On the Poisson equation for Markov chains: existence of solutions and parameter dependence. Technical Report, Technion—Israel Institute of Technology, Haifa 32000, Israel., 1991.
[12] R. Sutton and A. Barto. Reinforcement Learning: An Introduction. MIT Press. On-line edition at http://www.cs.ualberta.ca/~sutton/book/the-book.html, Cambridge, MA, 1998.
[13] R. S. Sutton. Learning to predict by the methods of temporal differences. Mach. Learn., 3(1):9–44, 1988.
[14] V. B. Tadic and S. P. Meyn. Asymptotic properties of two time-scale stochastic approximation algorithms with constant step sizes. In Proceedings of the 2003 American Control Conference, 2003., volume 5, pages 4426–4431. IEEE, 2003.
[15] J. N. Tsitsiklis and B. V. Roy. Average cost temporal-difference learning. Automatica, 35(11):1799 – 1808, 1999.
[16] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. IEEE Trans. Automat. Control, 42(5):674–690, 1997.
[17] J. N. Tsitsiklis and B. Van Roy. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. IEEE Trans. Automat. Control, 44(10):1840–1851, 1999.
[18] C. J. C. H. Watkins. Learning from Delayed Rewards. PhD thesis, King’s College, Cambridge, Cambridge, UK, 1989.
[19] C. J. C. H. Watkins and P. Dayan. Q-learning. Machine Learning, 8(3-4):279–292, 1992.
[20] H. Yu and D. P. Bertsekas. A least squares q-learning algorithm for optimal stopping problems. 2006.
[21] H. Yu and D. P. Bertsekas. Q-learning algorithms for optimal stopping based on least squares. In 2007 European Control Conference (ECC), pages 2368–2375. IEEE, 2007.
**Appendix**

*Proof of Lemma 3.3:* To show result (i), we rewrite \( A(\theta) \) as the difference of two matrices, \( A(\theta) = A_{NL}(\theta) - A_L \), denoting \( A_{NL}(\theta) \) to be the part of the matrix that depends on \( \theta \) and \( A_L \) to be the one that is independent of \( \theta \):

\[
A_{NL}(\theta) := E[\psi(X_n)\beta S_\theta \psi'(X_{n+1})] \\
A_L := E[\psi(X_n)\psi'(X_n)]
\]

Proving (29) is equivalent to proving:

\[
v^T A_{NL}(\theta)v - v^T A_L v \leq (\beta - 1)v^T \Psi v, \quad v \in \mathbb{R}^d.
\]

The proof is easier to follow if we suppose that the vector \( v \) is a difference of two parameter vectors, \( v = \theta^1 - \theta^2 \). Expanding the left hand side of the above inequality:

\[
v^T A_{NL}(\theta)v = (\theta^1 - \theta^2)^T E[\psi(X_n)\beta PH^\theta (Q^{\theta^1}(X_n) - Q^{\theta^2}(X_n))] \\
= \beta E[(Q^{\theta^1}(X_n) - Q^{\theta^2}(X_n)) PH^\theta (Q^{\theta^1}(X_n) - Q^{\theta^2}(X_n))].
\]

Next, using Cauchy-Schwartz and the fact that \( \|P\|_{\pi} \leq 1 \),

\[
v^T A_{NL}(\theta)v \leq \beta\|Q^{\theta^1} - Q^{\theta^2}\| \|PH^\theta (Q^{\theta^1}(X_n) - Q^{\theta^2}(X_n))\| \\
\leq \beta\|Q^{\theta^1} - Q^{\theta^2}\|^2 \\
= \beta v^T A_L v. \\
= (\beta - 1)v^T A_L v + v^T A_L v.
\]

Rearranging the terms, and noting that \( \Psi = A_L \), the statement of the Lemma follows:

\[
v^T A(\theta)v = v^T A_{NL}(\theta)v - v^T A_L v \leq (\beta - 1)v^T \Psi v \tag{48}
\]

Next, for fixed matrix \( A(\theta) \) with eigenvalue-eigenvector pair \( \lambda_a \in \mathbb{C}, v = a + bi \in \mathbb{C}^d \), we consider

\[
v^* A(\theta)v = (a^T - b^T i)A(\theta)(a + bi) = a^T A(\theta)a + b^T A(\theta)b + [a^T A(\theta)b - b^T A(\theta)a]i
\]

where \( v^* \) denotes the conjugate transpose of \( v \). With \( v^* A(\theta)v = \lambda_A v^* v \), it follows that

\[
\text{Re}\{\lambda_A\} v^* v = \text{Re}\{v^* A(\theta)v\} = a^T A(\theta)a + b^T A(\theta)b
\]

Let \( \lambda_\psi > 0 \) be the largest eigenvalue of \( \Sigma_\psi \), by the inequality (48), the following relation holds

\[
\text{Re}\{\lambda_A\} v^* v = a^T A(\theta)a + b^T A(\theta)b \\
\leq (\beta - 1)\lambda_\psi[a^T a + b^T b] \\
= (\beta - 1)\lambda_\psi v^* v
\]

Therefore, \( \text{Re}\{\lambda_A\} \) is negative and bounded above by \( (\beta - 1)\lambda_\psi \). For the last part, \( \|A(\theta)^{-1}\|_F \) is bounded using an inequality from [5]

\[
\|A(\theta)^{-1}\|_F \leq \frac{\sqrt{d}}{|\det A(\theta)|} \left( \frac{\|A(\theta)\|_F}{d - 1} \right)^{\frac{d-1}{d}} \tag{49}
\]

where \( d \) is the dimension. Provided the bound over eigenvalues of \( A(\theta) \) and compactness assumption of state space \( X \), there exists some constant \( \ell_A \) such that

\[
\|A(\theta)^{-1}\|_F \leq \frac{\sqrt{d}}{|1 - \beta|\lambda_\psi^d} \left( \frac{\ell_A}{d - 1} \right)^{\frac{d-1}{d}} \tag{50}
\]

The claim of uniform boundedness of \( \{A(\theta)^{-1} : \theta \in \mathbb{R}^d\} \) then follows. □
Proof of Lemma 3.4: For any two parameter vectors \( \theta_1, \theta_2 \in \mathbb{R}^d \), we have:
\[
\| b(\theta_1) - b(\theta_2) \| = \| \psi(-FQ^{\theta_1} + \hat{F}Q^{\theta_2} + Q^{\theta_1} - Q^{\theta_2}) \|
\leq \| \psi(FQ^{\theta_1} - \hat{F}Q^{\theta_2}) \| + \| \psi(Q^{\theta_1} - Q^{\theta_2}) \|
\leq \beta \| \psi(Q^{\theta_1} - Q^{\theta_2}) \| + \| \psi(Q^{\theta_1} - Q^{\theta_2}) \|
\leq (1 + \beta) \| \psi \| \| \theta_1 - \theta_2 \|.
\]

Proof of Lemma 4.2: Let \( \hat{f} : X \to \mathbb{R}^d \) solve Poisson’s equation:
\[
\mathbb{E} [\hat{f}(X_{n+1}) - f(X_n) | F_n] = f(X_n)
\]
with the forcing function \( f(X_n) \) in Lemma 4.2. The following decomposition is obtained:
\[
f(X_n) = \mathbb{E}[\hat{f}(X_{n+1}) - \hat{f}(X_n) | F_n]
= \mathbb{E}[\hat{f}(X_{n+1}) | F_n] - \hat{f}(X_{n+1}) + \hat{f}(X_{n+1}) - \hat{f}(X_n)
\]
Martingale difference

where \( \hat{f}(X_n) \) is bounded since the forcing function \( f(X_n) \) is bounded [9].

Proof of Lemma 4.3: For \( \theta_1, \theta_2 \), by Markov property with \( X_n = x \), we rewrite the term as
\[
\mathbb{E} [\psi(x)S_{\theta_1}\psi(X_{n+1}) | F_n] - \mathbb{E} [\psi(x)S_{\theta_2}\psi(X_{n+1}) | F_n]
= \mathbb{E} [\psi(x)S_{\theta_1}\psi(X_{n+1}) | X_n = x] - \mathbb{E} [\psi(x)S_{\theta_2}\psi(X_{n+1}) | X_n = x]
\]
By Assumption A3, the terminating cost function \( c_s - 1 \) is in the span of basis. There exists some constant \( \theta_{cs} \) such that
\[
c_s(y) \equiv \psi^T(y)\theta_{cs} + 1, \quad \forall y \in X.
\]
Denote \( Z \sim \mathcal{N}(\psi(x), I) \) as the random variable with density \( p_Z(\cdot) \), we have, for some \( \ell_E > 0 \),
\[
\| \mathbb{E} [\psi(x)S_{\theta_1}\psi(X_{n+1}) | X_n = x] - \mathbb{E} [\psi(x)S_{\theta_2}\psi(X_{n+1}) | X_n = x] \|
\leq \mathbb{E} [\| \psi(x)\| \| \psi(X_{n+1}) \| \| \theta_1^2 - \theta_2^2 \| \leq c_s(\theta_{cs})] - \mathbb{E} [\| \psi(x)\| \| \psi(X_{n+1}) \| \| \theta_1^2 - \theta_2^2 \| \leq c_s(\theta_{cs})] | X_n = x
\]
\[
= \int \| \psi(x)z' \| \cdot \ell(z|x) \cdot \| \{ z' (\theta_1^2 - \theta_2^2) \leq 1 \} - \{ z' (\theta_2^2 - \theta_2^2) \leq 1 \} \| p_Z(z) dz
\]
\[
\leq \ell_E \int \| \{ z' (\theta_1^2 - \theta_2^2) \leq 1 \} - \{ z' (\theta_2^2 - \theta_2^2) \leq 1 \} \| p_Z(z) dz
\]
As a result, we only need to show that the expectation in the last line is Lipschitz continuous for a Gaussian random variable, for which a bounded derivative of \( \mathbb{E}[\{ Z' \theta \leq 1 \}] \) w.r.t \( \theta \) will suffice. For each \( \theta, Z' \theta \) follows Gaussian distribution \( \mathcal{N}(\theta^T \psi(x), \theta^T \theta) \). Denote \( \mathbb{E}[\{ Z' \theta \leq 1 \}] \) as \( m(\theta) \)
\[
m(\theta) := \mathbb{E}[\{ Z' \theta \leq 1 \}] = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi\theta^2}} \exp \left( -\frac{1}{2\theta^2} (v - \theta^T \psi(x))^2 \right) dv
\]
Applying the change of variable \( u = \frac{1}{\sqrt{2\pi\theta^2}} (v - \theta^T \psi(x)) \), we have
\[
m(\theta) = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2\theta^2} u^2 \right) du \sqrt{2\pi\theta^2}
\]
\[
= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{1} \exp \left( -u^2 \right) du
\]
Its derivative is
\[
\frac{d}{d\theta} m(\theta) = \left[ \frac{1}{\sqrt{2\pi}} - \psi(x) \frac{||\theta|| - (1 - \psi(x)) \cdot ||\theta||^{-1} \theta}{||\theta||^2} \right] \exp \left( - \frac{(1 - \theta^T \psi(x))^2}{2||\theta||^2} \right)
\]
We observe that
1) If $||\theta|| \to \infty$, the exponential term is bounded by 1, and the coefficient before exponential term goes to zero since
\[
\frac{\psi(x)}{||\theta||} + \frac{(1 - \psi(x))\theta}{||\theta||^2} \to 0
\]
2) $m'(\theta)$ also vanishes as $||\theta|| \to 0$ since
\[
\frac{\psi(x)}{||\theta||} + \frac{(1 - \psi(x))\theta}{||\theta||^2} \exp \left( - \frac{(1 - \theta^T \psi(x))^2}{2||\theta||^2} \right) \to 0
\]
Since the state space $X$ is compact, there exists a deterministic bound over $m'(\theta)$ that does not depend on $x$. As a result, the objective in Lemma 4.3 is Lipschitz continuous with some deterministic Lipschitz constant $\ell_M$.

**Proof of Lemma 4.4:** Let $P$ denote transition kernel of the Markov chain, we define the following fundamental kernel $Z$
\[
Z := [I - P + \mathbb{1} \otimes \pi]^{-1} = \sum_{n=0}^{\infty} [P - \mathbb{1} \otimes \pi]^n = I + \sum_{n=1}^{\infty} [P^n - \mathbb{1} \otimes \pi]
\]
(51)
Note that $\hat{f}_\theta = Z f_\theta$ solves Poisson’s equation:
\[
P \hat{f}_\theta = \hat{f}_\theta - f_\theta
\]
Since $Z : L_\infty \to L_\infty$ is a bounded linear operator, we have that for any $x \in X$,
\[
||h_{\theta_1}(x) - h_{\theta_2}(x)|| \leq ||h_{\theta_1} - h_{\theta_2}||_\infty \leq ||Z||_\infty ||f_{\theta_1} - f_{\theta_2}||_\infty \leq \ell_F ||Z||_\infty ||\theta_1 - \theta_2||
\]
where $||Z||_\infty$ denotes the induced operator norm.

**Proof of Lemma 4.5:** $\{\tilde{A}^2_{n+1}\}$ is ODE friendly as it is martingale difference sequence with bounded moments. For $\tilde{A}^2_{n+1}$, let $\hat{f}_\theta : \Theta \times X \to \mathbb{R}^{d \times d}$ solve Poisson’s equation:
\[
E[f_\theta(X_{n+1}) - f_\theta(X_n)|\mathcal{F}_n] = f_\theta(X_n)
\]
with forcing function $f_\theta(X_n) := \tilde{A}^2_{n+1}$. The following representation is obtained:
\[
f_{\theta_n}(X_n) = E[f_{\theta_n}(X_{n+1}) - f_{\theta_n}(X_n)|\mathcal{F}_n]
\]
\[
= E[f_{\theta_n}(X_{n+1})|\mathcal{F}_n] - f_{\theta_n}(X_{n+1}) + f_{\theta_n}(X_{n+1}) - f_{\theta_n}(X_n)
\]
\[
= E[f_{\theta_n}(X_{n+1})|\mathcal{F}_n] - f_{\theta_n}(X_n) + f_{\theta_{n+1}}(X_{n+1}) - f_{\theta_{n+1}}(X_{n+1})
\]
which admits the form of being ODE-friendly, provided the perturbation term $\varepsilon_n := \hat{f}_{\theta_n}(X_{n+1}) - \hat{f}_{\theta_{n+1}}(X_{n+1})$ satisfies
\[
\sum_{k=1}^{\infty} \gamma_k ||\varepsilon_k|| < \infty, \quad a.s..
\]
Recall that Lemma 4.3 shows $\tilde{A}^2_{n+1}$ is Lipschitz continuous with deterministic Lipschitz constant. Combining this with Lemma 4.4 shows that $\hat{f}_\theta(X_{n+1})$ is uniformly Lipschitz continuous w.r.t $\theta$ with Lipschitz constant $B_Z \ell_F$, which indicates
\[
\sum_{k=1}^{\infty} \gamma_k ||\varepsilon_k|| = \sum_{k=1}^{\infty} \gamma_k ||f_{\theta_k}(X_{k+1}) - f_{\theta_{k+1}}(X_{k+1})|| \leq \sum_{k=1}^{\infty} \gamma_k B_Z \ell_F ||\theta_k - \theta_{k+1}||
\]
\[
\leq B_Z \ell_F \sum_{k=1}^{\infty} \gamma_k \alpha_k ||\tilde{A}^2_k \psi(X_k)(c(X_k) + \beta \min(c_\varepsilon(X_{k+1}), Q^{0*}(X_{k+1})) - Q^{0*}(X_k))||
\]
which is bounded by the boundedness assumption over \{θ_n\} and the fact that we use projected pseudo-inverse of \( \hat{A}_{n+1} \).

Proof of Lemma 4.6: Consider the update rules for θ_n, \( \hat{A}_n \) in (22) and (23). When both update rules are viewed as over the faster time-scale (that is, with step-size sequence \{γ_n\}), they can be rewritten as [2]:

\[
\begin{align*}
θ_{n+1} &= θ_n + γ_{n+1} [o(1)] \\
\hat{A}_{n+1} &= \hat{A}_n + γ_{n+1} [A(θ_n) - \hat{A}_n + \mathcal{E}^A_{n+1}]
\end{align*}
\]

where the \( o(1) \) term is:

\[
o(1) = -\frac{α_{n+1}}{γ_{n+1}} \hat{A}^T_n \psi(X_n) [c(X_n) + β \min(c_s(X_{n+1}), Q^δ_n(X_{n+1}))) - Q^δ(X_n)]
\]

It goes to zero provided stability assumption of \{θ_n\} and boundedness of \( \hat{A}^T \). It follows that \{θ_n, \hat{A}_n\} converges a.s. to the internally chain transitive invariant set of the following ODE [2]

\[
\begin{align*}
\dot{w}(t) &= 0 \\
\hat{A}(t) &= A(w_t) - A(t)
\end{align*}
\]

Which is \( \{(θ, A(θ)) : θ ∈ \mathbb{R}^d\} \). In other words, \( ||\hat{A}_n - A(θ_n)|| \) converges to 0 a.s.. While the invertibility of \( A(θ) \) has been established in Lemma 3.3.

Proof of Lemma 4.7: Within time interval \([s, s + T]\), consider the evolution of \( \bar{w}_{tk} \) over slow time scale defined by \{α_k\}, where \( t_k \) denote the time of \( k \)-th update of \( θ \).

\[
\begin{align*}
\bar{w}_{tk+1} &= \bar{w}_{tk} - α_{k+1} \hat{A}^T_{k+1} \psi(X_k) [c(X_k) \\
&+ β \min(c_s(X_{k+1}), \bar{w}_{tk} \psi(X_{k+1})) - \bar{w}_{tk} \psi(X_k)]
\end{align*}
\]

Write it in standard stochastic approximation form and replace \( \hat{A}^T_{k+1} \) with \( -A(θ_{k+1}) + o(1) \) by Lemma 3.3 and 4.6

\[
\begin{align*}
\bar{w}_{tk+1} &= \bar{w}_{tk} - α_{k+1} A^{-1}(\bar{w}_{tk}) [A(\bar{w}_{tk}) \bar{w}_{tk} + βc_s(\bar{w}_{tk}) + b^* \\
&+ \mathcal{E}_{k+1}^A \bar{w}_{tk} + \mathcal{E}_{k+1}^B] + o(α_{k+1})
\end{align*}
\]

With the noise terms being

\[
\begin{align*}
\mathcal{E}_{k+1}^A \bar{w}_{tk} &= A_{k+1} \bar{w}_{tk} - A(\bar{w}_{tk}) \bar{w}_{tk} \\
\mathcal{E}_{k+1}^B &= c(X_k) \psi(X_k) - b^* + βc_s(X_{k+1}) \mathbb{I}\{Q^δ(X_{k+1}) > c_s(X_{k+1})\} - βc_s(θ_k)
\end{align*}
\]

With them being shown ODE-friendly in Lemma 4.1, the ODE approximation for \( \bar{w}_t \) follows under the stability assumption of \( θ_n \) [14].

Proof of Lemma 4.8: For result in (i), the first relation holds since mapping \( b(\cdot) \) is Lipschitz continuous. The sub-sequential limit \( w_t \) is the solution of the ODE in (43) since the right hand side of (43) is Lipschitz continuous. It is differentiable everywhere within \([0, T]\).

By definition in (25a), \( b_t = b(w_t) \) can be written as follows:

\[
\begin{align*}
b_t &= -A(w_t)w_t - βc_s(w_t) \\
&= -E[\psi(X_n)(β \min(c_s(X_{n+1}), Q^u(X_{n+1}))) - Q^u(X_n))] \\
&= E[\psi(X_n)c^u(X_n)]
\end{align*}
\]

where, for any \( w ∈ \mathbb{R}^d \), \( c^u \) is the cost function that solves the fixed point equation (4) with \( θ^* \) replaced by \( w \):

\[
c^u(x) := -E[β \min(c_s(X_{n+1}), Q^u(X_{n+1}) | X_n = x] + Q^u(x), \quad x ∈ X
\]
The cost function $c^w$ is Lipschitz in $w$. Therefore it is Lipschitz continuous over $[0, T]$ and absolutely continuous over $[0, T]$ that has derivative almost everywhere. Let $t_0$ be a point of differentiability for $c_t = c^w_t$. Both $w_t$ and $c_t(x)$ (for each $x \in X$) are approximated by a line at this time point:

$$w_t = w_{t_0} + (t - t_0)v^w + o(|t - t_0|)$$
$$c_t(x) = c_{t_0}(x) + (t - t_0)v^c(x) + o(|t - t_0|), \quad t \sim t_0$$

where $v^w, v^c(x)$ are the respective derivatives. The assertion

$$v^c(x) = -\mathbb{E}[\beta S_{w_{t_0}} \psi(X_{n+1})^T - \psi(X_n)^T \mid X_n = x]v^w$$

will then imply the statement of the Lemma. Denote $L^c_t = c_{t_0} + (t - t_0)v^c$, $L^w_t = w_{t_0} + (t - t_0)v^w$. For each $t$, we have:

$$L^c_t(x) = -\mathbb{E}[\beta \min(c_{s}(X_{n+1}), Q^{L^c_t}(X_{n+1})) - Q^{L^c_t}(X_n) \mid X_n = x]$$
$$= -\mathbb{E}[\beta S_{L^c_t} Q^{L^c_t}(X_{n+1}) + \beta S^{w_{t_0}} c_{s}(X_{n+1}) - Q^{L^c_t}(X_n) \mid X_n = x]$$
$$\geq -\mathbb{E}[\beta S_{w_{t_0}} Q^{L^c_t}(X_{n+1}) + \beta S^{w_{t_0}} c_{s}(X_{n+1}) - Q^{L^c_t}(X_n) \mid X_n = x]$$
$$= c_{t_0}(x) - (t - t_0)\mathbb{E}[\beta S_{w_{t_0}} \psi(X_{n+1})^T - \psi(X_n)^T \mid X_n = x]v^w$$

The above inequality is true for both $t > t_0$ and $t < t_0$. For $t \sim t_0$, it follows that the inequality becomes an equality, and therefore (54) holds. The ODE (44) for $b_t$ holds by replacing $w_t$ with $b_t$ in (43) using the above derivative.