Mach Learn (2006) 63:107–133
DOI 10.1007/s10994-006-5835-z

Asymptotic analysis of temporal-difference learning algorithms with constant step-sizes

Vladislav B. Tadić

Received: 17 July 2002 / Revised: 28 March 2005 / Accepted: 7 October 2005 /
Published online: 9 March 2006
Springer Science + Business Media, LLC 2006

Abstract The mean-square asymptotic behavior of temporal-difference learning algorithms with constant step-sizes and linear function approximation is analyzed in this paper. The analysis is carried out for the case of discounted cost function associated with a Markov chain with a finite dimensional state-space. Under mild conditions, an upper bound for the asymptotic mean-square error of these algorithms is determined as a function of the step-size. Moreover, under the same assumptions, it is also shown that this bound is linear in the step size. The main results of the paper are illustrated with examples related to $M/G/1$ queues and nonlinear AR models with Markov switching.

Keywords Temporal-difference learning · Neuro-dynamic programming · Reinforcement learning · Stochastic approximation · Markov chains

1. Introduction

The mean-square asymptotic behavior of temporal-difference learning with linear function approximation is the subject of this paper. Temporal-difference learning could be considered as a recursive parametric method for approximating a cost function associated with a Markov chain. The aim of these algorithms is determining the optimal value of the approximator parameters by using only the available observations of the underlying chain. In order to minimize the approximation error, temporal-difference learning algorithms update the approximator parameter whenever a new observation of the underlying chain becomes available.

The prediction and approximation of a cost-to-go function associated with a Markov chain are problems arising in the area of dynamic programming (e.g., the policy evaluation step of

Editor: Robert Schapire

V. B. Tadić (✉)
Department of Automatic Control and Systems Engineering, University of Sheffield,
S1 3JD, Sheffield, United Kingdom
e-mail: v.tadic@sheffield.ac.uk

the policy iteration algorithm is based on the estimation of a cost-to-go function), as well as in areas such as automatic control and time-series analysis. Several methods have been proposed for solving these problems (e.g., Monte Carlo methods in statistics and maximum likelihood methods in automatic control), among which temporal-difference learning is probably the most general. Moreover, it is easy to be implemented and computationally low or moderate complex. Due to their excellent performance, temporal-difference learning algorithms have found a wide range of application (for details see e.g., (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) and references cited therein), while a great number of papers have been devoted to the analysis of their asymptotic behavior (see Dayan, 1992; Dayan & Sejnowski, 1994; Jaakola, Jordan, & Singh, 1994; Konda, 2002; Nedić & Bertsekas, 2003; Sutton, 1988; Tadić, 2000; Tsitsiklis & Van Roy, 1997; see also Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998 and references cited therein). Unfortunately, none of the existing results provide an insight into asymptotic properties of temporal-difference learning algorithms with constant step-sizes. Since temporal-difference learning algorithms (as well as other reinforcement learning algorithms) are usually implemented with constant step sizes, it seems that the asymptotic results obtained for the case of constant step sizes are more important and interesting (at least from the practical point of view) than results on the asymptotic behavior of decreasing step size algorithms.

In this paper, the mean-square asymptotic behavior of temporal-difference learning algorithms with constant step-sizes and linear function approximation is analyzed. The analysis is carried out for the case of discounted cost function associated with a Markov chain with a finite dimensional state-space. Under mild conditions, an upper bound for the asymptotic mean-square error of these algorithms (i.e., for the their asymptotic mean-square deviation from the optimal value of the approximator parameters) is determined as a function of the step-size. Moreover, under the same assumptions, it is also shown that this bound is linear in the step size. The main results of the paper are illustrated with examples related to $M/G/1$ queues and nonlinear autoregressive (AR) models with Markov switching. The results of this paper are an extension of the results of (Tsitsiklis & Van Roy, 1997) and a continuation of the author's work presented in (Tadić, 2000). Moreover, to the best of the author's knowledge, there does not exist a similar result in the available literature on reinforcement learning.

The paper is organized as follows. In Section 2, temporal-difference learning algorithms are formally defined and the assumptions under which their rate of convergence is analyzed are introduced. The statement of the main result is also presented in Section 2, while its proof is given in Section 4. A special case where the underlying Markov chain is geometrically ergodic is considered in Section 5, while the examples related to $M/G/1$ queues and nonlinear AR models with Markov switching are presented in Section 6. In Section 3, the existence and properties of solutions of certain Poisson equations associated with the underlying Markov chain are analyzed. The results presented in this section are a crucial prerequisite for the analysis carried out in Section 4.

2. Main results

Temporal-difference learning algorithms analyzed in this paper are defined by the following equations:

$$\theta_{n+1} = P_Q(\theta_n + \gamma d_{n+1} e_{n+1}), \tag{1}$$

$$d_{n+1} = c(X_n, X_{n+1}) + \alpha\theta_n^T \phi(X_{n+1}) - \theta_n^T \phi(X_n), \tag{2}$$

$$e_{n+1} = \sum_{i=0}^{n} (\alpha\lambda)^{n-i} \phi(X_i), \quad n \geq 0. \tag{3}$$

$\gamma \in (0, \infty), \alpha \in (0, 1)$ and $\lambda \in (0, 1]$ are constants (γ is the algorithm step size). $c : R^{d'} \times R^{d'} \to R$ and $\phi : R^{d'} \to R^d$ are Borel-measurable functions. θ_0 is an R^d-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, while $\{X_n\}_{n\geq 0}$ is an $R^{d'}$-valued homogeneous Markov chain defined on the same probability space. $Q \subset R^d$ is a convex compact set, while $P_Q(\cdot)$ is the projection on Q, i.e.,

$$P_Q(\theta) = \arg\inf_{\theta' \in Q} \|\theta' - \theta\|$$

for $\theta \in R^d(\| \cdot \|)$ is the Euclidean norm in R^d).

Temporal-difference leaning algorithms appearing the literature on reinforcement learning typically do not have projection. However, due to the finite precision of digital computers, any implementation of these algorithms (as well as other reinforcement learning algorithms) implicitly involves projection of the algorithm iterates. Moreover, if the algorithm limit points can a priori be located within a convex compact set (which is a typical situation in practice), the projection to this set usually improves significantly the algorithm asymptotic properties (stability and convergence).

For $x \in R^{d'}$, let

$$J_*(x) = E\left(\sum_{n=0}^{\infty} \alpha^n c(X_n, X_{n+1}) \middle| X_0 = x \right)$$

(provided that $J_*(\cdot)$xs is well-defined). In the context of dynamic programming, $J_*(\cdot)$ is interpreted as a discounted cost function associated with the chain $\{X_n\}_{n\geq 0}$ (for details see e.g., (Bertsekas & Tsitsiklis, 1996). The task of the algorithm (1) – (3) is to approximate the function $J_*(\cdot)$ by $\theta^T \phi(\cdot)$. It determines the optimal value θ_* of the parameter $\theta \in R^d$ such that the $\theta_*^T \phi(\cdot)$ is the best approximator of $J_*(\cdot)$ in the sense explained in [Tsitsiklis & Van Roy, 1997, Section III]. If $\lambda = 1$ and $\{X_n\}_{n\geq 0}$ has a unique invariant probability measure $\pi(\cdot)$, the algorithm (1) – (3) determines $\theta_* \in R^d$ such that $\theta_*^T \phi(\cdot)$ approximates $J_*(\cdot)$ optimally in the $L^2(\pi)$-sense, i.e., it searches for the minimum of the function $J(\theta) = \int (\theta^T \phi(x) - J_*(x))^2 \pi(dx), \theta \in R^d$.

It can easily be noticed from (1) – (3) that temporal-difference learning algorithms belong to the category of stochastic approximation algorithms (for more details on stochastic approximation see e.g., (Benveniste, Metivier, & Priouret, 1990 and Kushner & Yin, 1997). Therefore, the asymptotic analysis of temporal-difference learning is usually based on the methods developed for stochastic approximation (see e.g., Bertsekas & Tsisiklis, 1996; Sutton & Barto, 1998 and references cited therein). The analysis carried out in this paper relies on the Poisson equation based general methodology for the asymptotic analysis of stochastic approximation (for details see Benveniste, Metivier, & Priouret, 1990).

The following notation is used throughout the paper. $\| \cdot \|$ denotes the Euclidean vector norm and the matrix norm induced by the Euclidean vector norm (i.e., $\|A\| = \sup_{\|\theta\|=1} \|A\theta\|$, $A \in R^{d\times d}$, while $\mathcal{B}^{d'}$ is the family of Borel measurable sets from $R^{d'}$. For $x \in R^{d'}$, let $P(x, \cdot)$

and $P^n(x, \cdot)$ be the single and n-th step transition probability kernel of $\{X_n\}_{n\geq 0}$ (respectively), i.e.,

$$P^n(x, B) = \mathcal{P}(X_n \in B \mid X_0 = x) \;\; w.p.1$$

and $P(x, B) = P^1(x, B)$ for all $x \in R^{d'}$, $B \in \mathcal{B}^{d'}$, $n \geq 0$.

In this paper, the asymptotic behavior of temporal-difference learning algorithms with linear function approximation is analyzed under the following conditions.

A1. *$\{X_n\}_{n\geq 0}$ has a (unique) invariant probability measure $\pi(\cdot)$.*
A2. *There exist a constant $K \in [1, \infty)$ and a Borel-measurable function $f : R^{d'} \to [1, \infty)$ such that $\|\phi(x)\| \leq f(x)$ and*

$$\int f^4(x')\pi(dx') < \infty, \tag{4}$$

$$(P^n f^4)(x) \leq K f^4(x), \tag{5}$$

$$\int |c(x, x')|^4 P(x, dx') \leq f^4(x) \tag{6}$$

for all $x \in R^{d'}$, $n \geq 0$.
A3. *There exist a constant $L \in [1, \infty)$ and a Borel-measurable function $g : R^{d'} \to [1, \infty)$ such that*

$$\int g^2(x)\pi(dx) < \infty,$$

$$(P^n g^2)(x) \leq L g^2(x),$$

$$\sum_{m=0}^{\infty} \left\| \int \phi(x')(P^n \phi^T)(x')(P^m - \pi)(x, dx') \right\| \leq g(x), \tag{7}$$

$$\sum_{m=0}^{\infty} \left\| \int \phi(x')(P^n \tilde{c})(x')(P^m - \pi)(x, dx') \right\| \leq g(x) \tag{8}$$

for all $x \in R^{d'}$, $n \geq 0$, where $\tilde{c}(x) = \int c(x, x') P(x, dx')$.
A4. *$\int \phi(x)\phi^T(x)\pi(dx)$ is positive definite.*

Remark: *Using Lemma 1, it can easily be deduced that $\tilde{c}(\cdot)$ and the left-hand sides of (7) and (8) are well-defined and finite.*

Assumption A1 is related to the stationarity properties of $\{X_n\}_{n\geq 0}$. Assumptions of this type are standard for the asymptotic analysis of temporal-difference learning algorithms (see Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998 and references cited therein; see also Tadicć, 2000; and Tsitsiklis & Van Roy, 1997).

Assumption A2 corresponds with the growth rate of $c(\cdot, \cdot)$ and $\phi(\cdot)$. It requires these functions not to grow too fast so that their upper bound $f(\cdot)$ satisfies (4) and (5). The role of A2 is to ensure that $J_*(\cdot)$ and A_*, b_* (defined in (10) and (11)) are well-defined and finite, as well as that solutions of certain Poisson equations associated with the algorithm (1) – (3) (defined in the statement of Lemma 4, Eqs. (31), (32)) have finite second-order moments. A2 is satisfied if $c(\cdot,\cdot)$ and $\phi(\cdot)$ are globally bounded or if $c(\cdot,\cdot)$, and $\phi(\cdot)$ are locally bounded and there exists a constant $M \in [1,\infty)$ such that $\|X_n\| \leq M w.p.1, n \geq 0$. It is also satisfied if $\{X_n\}_{n\geq 0}$ is geometrically ergodic (see Section 5).

A2 and particularly the requirement expressed by (5) are motivated by the necessary and sufficient conditions for the V-uniform ergodicity of homogeneous Markov chains. Namely, an irreducible and aperiodic $R^{d'}$-valued Markov chain with a transition probability kernel $P(x,\cdot), x \in R^{d'}$, is uniformly ergodic with respect to a Borel-measurable function $V : R^{d'} \to (1,\infty)$ if and only if there exists a Borel-measurable function $V_0 : R^{d'} \to (1,\infty)$, constants $\beta \in (0,1)$, $b, c \in (1,\infty)$ and a set $C \in \mathcal{B}^{d'}$ such that $c^{-1}V(x) \leq V_0(x) \leq cV(x)$ and

$$(PV_0)(x) - V_0(x) \leq -\beta V_0(x) + b I_C(x) \tag{9}$$

for all $x \in R^{d'}$ (for details see e.g., (Meyn et. al., 1993, Section 16]; $I_C(\cdot)$ denotes the indicator function of the set C). Iterating (9), it can easily be deduced that

$$(P^n V_0)(x) \leq (1-\beta)^n V_0(x) + b\sum_{i=1}^{n}(1-\beta)^{n-i} \leq V_0(x) + b(1-\beta)^{-1}$$

for all $x \in R^{d'}, n \geq 0$. Consequently,

$$\begin{aligned}(P^n V)(x) &\leq c(V_0(x) + b(1-\beta)^{-1}) \\ &\leq c^2 V(x) + bc(1-\beta)^{-1} \leq (b+c)^2(1-\beta)^{-1}V(x)\end{aligned}$$

for all $x \in R^{d'}, n \geq 0$. On the other hand, if a Borel-measurable function $h : R^{d'} \to [0,\infty)$ does not satisfy $h(x) \leq V(x)$ for all $x \in R^{d'}$, it is possible that $\int h(x)\pi(dx) = \infty$ or $(P^n h)(x) = \infty$ for some $x \in R^{d'}, n \geq 0$. Therefore, in order to ensure that an upper bound of $\|\phi(x)\|$ and

$$\left(\int |c(x,x')|^4 P(x,dx')\right)^{1/4}$$

have finite fourth-order moments with respect to $\pi(\cdot)$ and $P(x,\cdot), n \geq 0$, it is quite reasonable (and natural) to assume that (5) holds (which is equivalent to $f(x) = V^{1/4}(x)$ if $\{X_n\}_{n\geq 0}$ is V-uniformly ergodic).

Assumption A3 is related to the stability of $\{X_n\}_{n\geq 0}$. Basically, A3 requires $\{X_n\}_{n\geq 0}$ to exhibit sufficient "degree of stability" (i.e., $P^n(x,\cdot), x \in R^{d'}$, to converge to $\pi(\cdot)$ sufficiently fast) so that (7) and (8) hold. Its role is to ensure that the Poisson equations associated with the algorithm (1) – (3) have unique solutions (see Lemma 4). A3 is satisfied under geometric ergodicity conditions (see Section 5) and is typical for the asymptotic analysis of temporal-difference learning algorithms (see Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; and references cited therein; see also Tadić, 2000 and Tsitsiklis & Van Roy, 1997).

Assumption A4 is a "persistancy of excitation" condition. These conditions are typical for the areas of system identification, adaptive control and adaptive signal processing (see e.g., Goodwin & Sin, 1984; Solo & Kong, 1995 and references cited therein). A4 requires $\{\phi(X_n)\}_{n\geq 0}$ to be sufficiently "rich" with respect to all directions in R^d at the asymptotic steady-state characterized by $\pi(\cdot)$. Assumptions of this kind are standard for the asymptotic analysis of temporal-difference learning algorithms (see Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; and references cited therein; see also Tadić, 2000 and Tsitsiklis & Van Roy, 1997).

Let

$$A_* = -\int \phi(x)\phi^T(x)\pi(dx) + \alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P^{n+1}\phi^T)(x)\pi(dx), \quad (10)$$

$$b_* = \sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P^n \tilde{c})(x)\pi(dx), \quad (11)$$

while $\theta_* = -A_*^{-1}b_*$. Moreover, let $\lambda_{\min}$ be the minimal eigenvalue of $-A_*$, while

$$\rho_Q = \sup_{\theta,\theta'\in Q} \max\{\|\theta\|, \|\theta-\theta'\|\},$$

$$M = 16(1-\alpha\lambda)^{-2}(K+L)\int f^2(x)\pi(dx),$$

$$K_Q = 6M^2(1+\rho_Q)^2(1+\lambda_{\min})(1+\lambda_{\min}^{-1}).$$

Furthermore, for $x \in R^{d'}$, let

$$h_Q(x) = K_Q(f^4(x) + g^2(x)).$$

The main results of the paper are contained in the next theorem.

Theorem 1. *Let A1 – A4 hold. Suppose that $\theta_* \in Q$ and $\gamma < \lambda_{\min}^{-1}$. Then,*

$$\overline{\lim_{n\to\infty}} E(\|\theta_n - \theta_*\|^2 | X_0 = x) \le h_Q(x)\gamma \quad (12)$$

for all $x \in R^{d'}$

Theorem 1 basically claims that if the step-size γ is less than the constant $\lambda_{\min}^{-1}$ (which depends only on $\pi(\cdot)$, $\phi(\cdot)$, α, λ), the algorithm (1) – (3) is stable in the mean-square sense, and its conditional mean-square error given the chain initial state $X_0 = x$ is asymptotically bounded by a linear function of the step-size γ. Hence, if the step-size γ is sufficiently small, the algorithm iterates $\{\theta_n\}_{n\ge 0}$ fluctuate asymptotically around θ_* with a variance which is linearly bounded by the step-size.

Let

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n + \gamma(A_*\tilde{\theta}_n + \tilde{b}_{n+1}), \quad n \ge 0,$$

where $\tilde{\theta}_0$ and $\{\tilde{b}_n\}_{n\geq 1}$ are vectors from R^d. Then, it is straightforward to demonstrate the following:

(i) $\{\tilde{\theta}_n\}_{n\geq 0}$ is bounded (i.e., Lagrange stable) for any bounded sequence $\{\tilde{b}_n\}_{n\geq 1}$ only if $\gamma < \lambda_{\min}^{-1}$.

(ii) If $\gamma < \lambda_{\min}^{-1}\tilde{\theta}_0$ is deterministic vector and $\{\tilde{b}_n\}_{n\geq 1}$ is an i.i.d. sequence with $\tilde{b}_* = E\,(\tilde{b}_1)$ $\tilde{\sigma}_*^2 = E\|\tilde{b}_1 - \tilde{b}_*\|^2$, then

$$\lim_{n\to\infty} E\|\tilde{\theta}_n - \tilde{\theta}_*\|^2 = \gamma^2 \sum_{n=0}^{\infty} E((\tilde{b}_1 - \tilde{b}_*)^T (I + \gamma A_*)^{2n} (\tilde{b}_1 - \tilde{b}_*))$$
$$\geq \tilde{\sigma}_*^2 \sum_{n=0}^{\infty} (1 - \gamma\lambda_{\max})^{2n} \tilde{\sigma}_*^2$$
$$= \tilde{\sigma}_*^2 \gamma^2 (1 - (1 - \gamma\lambda_{\max})^2)^{-1} \geq 2^{-1}\lambda_{\max}^{-1}\tilde{\sigma}_*^2\gamma$$

where $\lambda_{\max}$ is the maximal eigenvalue of A_* and

$$\tilde{\theta}_* = \lim_{n\to\infty} E(\tilde{\theta}_n) = \sum_{n=0}^{\infty} (I + \gamma A_*)^n \tilde{b}_*.$$

Since the algorithm (1) – (3) can be rewritten as

$$\theta_{n+1} = \theta_n + \gamma(A_{n+1}\theta_n + b_{n+1}), \quad n \geq 0,$$

where $\lim_{n\to\infty} E(A_n) = A_*$ $\lim_{n\to\infty} E(b_n) = b_*$ this is a direct consequence of Lemma 4, Section 3), the results (i) and (ii) on the asymptotic behavior of $\{\tilde{\theta}_n\}_{n\geq 0}$ suggest that the mean-square stability of $\{\theta_n\}_{n\geq 0}$ cannot be guaranteed if$\gamma \geq \lambda_{\min}^{-1}$, as well as that the asymptotic mean-square error of $\{\theta_n\}_{n\geq 0}$ cannot be bounded by a function of the step-size γ which tends to zero at zero at a rate faster than linear. Hence, the results of Theorem 1 are tight regarding the step-size interval $(0, \lambda_{\min}^{-1})$ of the guaranteed stability and the linear dependence of the upper bound $\gamma h_Q(\cdot)$ on the step-size. Unfortunately, the constant K_Q appears to be conservative. However, is seems very hard (if possible at all) to improve K_Q using any existing technique for the asymptotic analysis of stochastic approximation.

It is also important to emphasize that an upper bound for the asymptotic unconditional mean-square error of $\{\theta_n\}_{n\geq 0}$ can be obtained from Theorem 1. Namely, Markov property and Theorem 1 imply that

$$\overline{\lim_{n\to\infty}} E(\|\theta_n - \theta_*\|^2 | X_m) = \overline{\lim_{n\to\infty}} E(\|\theta_{m+n} - \theta_*\|^2 | X_m) \leq h_Q(X_m)\gamma \quad w.p.1$$

for m $\geq$ 0. Then, the Fatou lemma yields that

$$\overline{\lim_{n\to\infty}} E\|\theta_n - \theta_*\|^2 \leq E\left(\overline{\lim_{n\to\infty}} E(\|\theta_n - \theta_*\|^2 | X_m)\right) \leq \gamma E(h_Q(X_m))$$

for m $\geq$ 0. Letting $m \to \infty$ in the previous relation, we get

$$\overline{\lim_{n\to\infty}} E\|\theta_n - \theta_*\|^2 \leq \gamma \lim_{m\to\infty} E(h_Q(X_m)). \tag{13}$$

If $\{X_n\}_{n\geq 0}$ is geometrically ergodic, then

$$\lim_{n\to\infty} E(h_Q(X_n)) = \int h_Q(x)\pi(dx).$$

Asymptotic behavior of temporal-difference learning algorithms has been considered in a large number of papers (see Dayan, 1992, 1994; Jaakola, Jordan, & Singh, 1994; Konda, 2002; Nedić & Bertsekas, 2003; Sutton, 1988; Tadić, 2000; Tsitsiklis & Van Roy, 1997; see also Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998 and references cited therein). Although the existing results provide a good insight into the asymptotic behavior of temporal-difference learning algorithms, not much is known about the asymptotic properties of the temporal-difference learning algorithms with constant step sizes. The strongest existing results on their asymptotic behavior are probably contained in (Tsitsiklis & Van Roy, 1997) recently, the results of (Tsitsiklis & Van Roy, 1997) have been extended in (Tadić, 2000). In comparison with the assumptions adopted in (Tsitsiklis & Van Roy, 1997) A1 – A4 are just slightly more restrictive: the assumptions of (Tsitsiklis & Van Roy, 1997) would be a special case of A1 – A4 if A2 were replaced with the requirement that there exists a constant $K \in [1, \infty\,)$ and a Borel-measurable function $f : R^{d'} \to [1,\, \infty]$ such that $\int f^2(x)\pi(dx) < \infty$, $\|\phi(x)\| \leq f(x)$ and

$$\int |c\,(x,\, x')|^2 P(x, dx') \leq f^2(x),$$
$$(P^n\, f^2)\,(x) \leq K f^2(x)$$

for all $x \in R^{d+2d'}$, $n \geq 0$. However, only the algorithms with decreasing step sizes have been analyzed in (Tsitsiklis & Van Roy, 1997. On the other hand, implementations of temporal-difference learning algorithms are based on constant step sizes. Therefore, the results on the asymptotic behavior of constant step size algorithms seem to be more important and interesting than the results obtained for the case of decreasing step-sizes (at least from the practical point of view). To the best knowledge of the present author, the asymptotic behavior of temporal-difference learning algorithms with constant step sizes has not been considered in the available literature on reinforcement learning.

3. Preliminary results

In this section, we consider the existence of $J_*(\cdot)$, A_*, b_*, θ_*, as well as the existence and properties of solutions of certain Poisson equations associated with the algorithm (1) – (3) (which are defined in the statement of Lemma 4, Eqs. (31) and (32)). The results of this section are a crucial prerequisite for the analysis carried out in the next sections.

Throughout the paper, the following notation is used. For $x, x' \in R^{d'}$, $y \in R^d$, $B \in \mathcal{B}^{d+2d'}$ and $z = (x, x, y)$, let

$$A(z) = y(\alpha\,\phi\,(x') - \phi\,(x))^T,$$
$$b(z) = yc\,(x,\, x'),$$
$$\Pi\,(z,\, B) = \int I_B\,(x',\, x'',\, \alpha\,\lambda\, y + \phi\,(x'))P\,(x',\, dx''),$$

where $I_B(\cdot)$ denotes the indicator function of the set B. Let $Z_{n+1} = (X_n, X_{n+1}, e_{n+1})$, $n \geq 0$. Then, it is straightforward to verify that

$$\theta_{n+1} = \theta_n + \gamma_{n+1}(A(Z_{n+1})\theta_n + b(Z_{n+1})),$$

$$\mathcal{P}(Z_{n+1} \in B \mid Z_1, \dots, Z_n) = \Pi(Z_n, B) \;\; w.p.1$$

for all $B \in \mathcal{B}^{d+2d'}$, $n \geq 0$. Moreover, if $\varphi : R^{d'} \times R^{d'} \times R^d \to R$ is a Borel-measurable function, $\psi(z) = \varphi(x, x', y)$ for all $x, x' \in R^{d'}$, $y \in R^d$, $z = (x, x', y)$, and $\int |\psi(z')|\Pi(z, dz') < \infty$ for all $z \in R^{d+2d'}$, then

$$\int \psi(z')\Pi(z, dz') = \int \varphi(x', x'', \alpha\lambda y + \phi(x'))P(x', dx'') \tag{14}$$

for all $x,\ x' \in R^{d'}$, $y \in R^d$, $z = (x,\ x',\ y)$.

Outline of the Results of Section 3: The most important results of Section 3 are contained in Lemma 4. Lemma 4 is concerned with the existence and properties of solutions of Poisson Eqs. (31), (32) and is of crucial importance for the proofs of Lemma 6 and Theorem 1 (see the outline of the results of Section 4 on page 20). It also provides an explanation for the selection of the constant M in the definition of K_Q, $h_Q(\cdot)$ (page 7): M is selected in such a way that the constant terms in the last inequalities of (35), (36) (i.e., $3(1-\alpha\lambda)^{-1}\int f^2(x'')\pi(dx'')$, $2(1-\alpha\lambda)^{-1}\int f^2(x'')\pi(dx'')$) are not greater than $5^{-1}M$. The proof of Lemma 4 is essentially based on inequalities (35), (36), which themselves are direct consequences of the results of Lemma 3. Lemma 3 itself determines the conditional expectations of $A(\cdot)$, $b(\cdot)$ with respect to the kernels $\Pi^n(z,\cdot)$, $z \in R^{d+2d'}$, $n \geq 1$, while its proof uses only mathematical induction and the results of Lemma 1. On the other hand, Lemma 2 is related to the existence of A_*, b_*, θ_*, while its proof is exclusively based on the Cauchy-Schwartz inequality and Lemma 1. Lemma 1 itself is concerned with the existence of $J_*(\cdot)$, $\tilde{c}(\cdot)$, as well as with the existence and upper bounds of the conditional expectations of $\phi(\cdot)$, $\tilde{c}(\cdot)$ with respect to the kernels $P^n(x,\cdot)$, $x \in R^{d'}$, $n \geq 1$.

Lemma 1. *Let A1 and A2 hold. Then,* $J_*(\cdot)$, $\tilde{c}(\cdot)$, $(P^n\phi)(\cdot)$, $(P^n\tilde{c})(\cdot)$ *are well-defined and finite for all* $n \geq 0$. *Moreover,*

$$\max\{\|(P^n\,\phi)\,(x)\|, |(P^n\tilde{c})(x)|\} \leq Kf\,(x), \tag{15}$$

$$(P^n f^p)(x) \leq Kf^p(x) \tag{16}$$

for all $x \in R^{d'}$, $n \geq 0$.

Proof: Due to the Jensen inequality and A2,

$$((P^n\,f^p)\,(x))^{1/p} \leq ((P^n\,f^4)\,(x))^{1/4} \leq K^{1/4}\,f(x), \tag{17}$$

$$\int |c(x,\ x')|P(x,\ dx') \leq \left(\int |c(x,\ x')|^4 P(x,\ dx')\right)^{1/4} \leq f(x) \tag{18}$$

for all $p \in [1,4], x \in R^{d'}, n \geq 0$. Therefore, $\tilde{c}(\cdot)$ is well-defined, finite and satisfy $|\tilde{c}(x)| \leq f(x)$ for all $x \in R^{d'}$. Then, A2 and (17), (18) imply

$$\max\{\|(P^n\phi)(x)\|, |(P^n\tilde{c})(x)|\} \leq Kf(x),$$

$$\sum_{m=0}^{\infty} \alpha^m \int \int |c(x', x'')| P(x', dx'') P^m(x, dx')$$

$$\leq \sum_{m=0}^{\infty} \alpha^m (P^m f)(x) \leq K(1-\alpha)^{-1} f(x) < \infty$$

for all $x \in R^{d'}$, $n \geq 0$. Hence, $J_*(\cdot)$, $(P^n\phi)(\cdot)$, $(P^n\tilde{c})(\cdot)$ are well-defined and finite for all $n \geq 0$, while (15), (16) hold for all $x \in R^{d'}$, $n \geq 0$. □

Lemma 2. *Let A1, A2 and A4 hold. Then, A_*, b_* and θ_* are well-defined and finite. Moreover, A_* is negative definite and*

$$\max\{\|A_*\|, \|b_*\|\} \leq K(1-\alpha\lambda)^{-1} \int f^2(x)\pi(dx) \leq M^{1/2}. \tag{19}$$

Proof: Due to the Jensen inequality, A2 and Lemma 1,

$$\sum_{n=0}^{\infty} (\alpha\lambda)^n \int \|\phi(x)(P^{n+1}\phi^T)(x)\|\pi(dx) \leq K(1-\alpha\lambda)^{-1} \int f^2(x)\pi(dx) < \infty,$$

$$\sum_{n=0}^{\infty} (\alpha\lambda)^n \int \|\phi(x)(P^n\tilde{c})(x)\|\pi(dx) \leq K(1-\alpha\lambda)^{-1} \int f^2(x)\pi(dx) < \infty. \tag{20}$$

Then, it is obvious that A_* and b_* are well-defined and finite. On the other hand, owing to the Jensen inequality,

$$\int (\theta(P^n\phi)(x))^2 \pi(dx) \leq \int \int (\theta^T \phi(x'))^2 P^n(x, dx') = \int (\theta^T \phi(x))^2 \pi(dx)$$

for all $\theta \in R^d$, $n \geq 0$. Therefore,

$$\left| \int \theta^T \phi(x)(P^n\phi^T)(x)\theta\pi(dx) \right| \leq \left(\int (\theta^T \phi(x))^2 \pi(dx) \right)^{1/2}$$

$$\times \left(\int (\theta^T (P^n\phi)(x))^2 \pi(dx) \right)^{1/2} \leq \int (\theta^T \phi(x))^2 \pi(dx)$$

for all $\theta \in R^d$, $n \geq 0$. Consequently,

$$\theta^T A_* \theta = -\int (\theta^T \phi(x))^2 \pi(dx) + \alpha(1-\lambda) \sum_{n=0}^{\infty} (\alpha\lambda)^n \int \theta^T \phi(x)(P^{n+1}\phi^T)(x)\theta\pi(dx)$$

$$\leq -\left(1-\alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\,\lambda)^{n}\right)\int(\theta^{T}\phi\,(x))^{2}\pi\,(dx)$$

$$= -(1-\alpha)(1-\alpha\lambda)^{-1}\theta^{T}\left(\int\phi(x)\phi^{T}(x)\pi(dx)\right)\theta \tag{21}$$

for all $\theta \in R^{d}$. Then, it is obvious that A_{*} is negative definite, as well as that θ_{*} is well-defined and finite. Due to A2 and (21),

$$\|A_{*}\| \leq (1-\alpha)(1-\alpha\lambda)^{-1}\int\|\phi(x)\phi^{T}(x)\|\pi(dx) \leq (1-\alpha\lambda)^{-1}\int f^{2}(x)\pi(dx). \tag{22}$$

Hence, (19) follows from (20) and (22). □

Lemma 3. *Let A2 and A3 hold. Then,* $(\Pi^{n}\,A)(\cdot)$ *and* $(\Pi^{n}\,b)(\cdot)$ *are well-defined, finite and satisfy the following relations for all* $x, x' \in R^{d'}$, $y \in R^{d}$, $z = (x, x', y)$, $n \geq 1$*:*

$$(\Pi^{n+1}A)(z) = \sum_{i=0}^{n}(\alpha\lambda)^{i}\int\phi(x'')(\alpha(P^{i+1}\phi)(x'') - (P^{i}\phi)(x''))^{T}\cdot P^{n-i}(x', dx'')$$

$$+ (\alpha\lambda)^{n+1}y(\alpha(P^{n+1}\phi)(x'') - (P^{n}\phi)(x''))^{T}, \tag{23}$$

$$(\Pi^{n+1}b)(z) = \sum_{i=0}^{n}(\alpha\lambda)^{i}\int\phi(x'')(P^{i}\tilde{c})(x'')P^{n-i}(x', dx'') + (\alpha\lambda)^{n+1}y(P^{n}\tilde{c})(x''). \tag{24}$$

Proof: The assertion of this lemma is shown by the mathematical induction. Due to A2,

$$\|A(z)\| \leq (\|\phi(x)\| + \|\phi(x')\|)\|y\| \leq (f(x) + f(x'))\|y\|, \tag{25}$$

$$\|b(z)\| \leq |c(x, x')|\|y\| \tag{26}$$

for all $x, x' \in R^{d'}$, $y \in R^{d}$, $z = (x, x', y)$. Owing to Lemma 1 and (14), (25), (26),

$$\int\|A(z')\|\Pi(z, dz') = \int\|(\alpha\lambda y + \phi(x'))(\alpha\phi(x'') + \phi(x'))^{T}\|P(x', dx'')$$

$$\leq \int(f(x') + \|y\|)(f(x') + f(x''))P(x', dx'')$$

$$= (f(x') + \|y\|)(f(x') + (Pf)(x')) < \infty,$$

$$\int\|b(z')\|\Pi(z, dz') = \int\|(\alpha\lambda y + \phi(x'))c(x', x'')\|P(x', dx'')$$

$$\leq \int(f(x') + \|y\|)|c(x', x'')|P(x', dx'')$$

$$= (f(x') + \|y\|)\int|c(x', x'')|P(x', dx'') < \infty$$

for all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Consequently, $(\Pi A)(\cdot)$ and $(\Pi b)(\cdot)$ are well-defined and finite, while (14) implies

$$\begin{aligned}\int A(z')\Pi(z, dz') &= \int (\alpha\lambda y + \phi(x'))(\alpha\phi(x'') - \phi(x'))^T P(x', dx'')\\ &= (\alpha\lambda y + \phi(x'))(\alpha(P\phi)(x') - \phi(x')),\\ \int b(z')\Pi(z, dz') &= \int (\alpha\lambda y + \phi(x'))c(x', x'')P(x', dx'')\\ &= (\alpha\lambda y + \phi(x'))\tilde{c}(x')\end{aligned}$$

for all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Hence, (23) and (24) hold for $n = 0$ and all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Suppose that $(\Pi^{n+1} A)(\cdot)$ and $(\Pi^{n+1} b)(\cdot)$ are well-defined, finite and satisfy (23), (24) for some $n \geq 0$ and all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Then, Lemma 1 implies that

$$\begin{aligned}\|(\Pi^{n+1}A)(z)\| &\leq \sum_{i=0}^{n} \int \|\phi(x'')\|(\|(P^i\phi)(x'')\| + \|(P^{i+1}\phi)(x'')\|) \cdot P^{n-i}(x', dx'')\\ &\quad +(\|(P^n\phi)(x')\| + \|(P^{n+1}\phi)(x')\|)\|y\|\\ &\leq \sum_{i=0}^{n} \int f(x'')((P^i f)(x'') + (P^{i+1} f)(x''))P^{n-i}(x', dx'')\\ &\quad +((P^n f)(x') + (P^{n+1} f)(x'))\|y\|,\\ &\leq 2K \sum_{i=0}^{n} (P^{n-i} f^2)(x') + 2K f(x')\|y\| \end{aligned} \tag{27}$$

$$\begin{aligned}\|(\Pi^{n+1}b)(z)\| &\leq \sum_{i=0}^{n} \int \|\phi(x'')\||(P^i\tilde{c})(x'')|P^{n-i}(x', dx'') + |(P^n\tilde{c})(x')\|y\|\\ &\leq \sum_{i=0}^{n} \int f(x'')(P^i f)(x'')P^{n-i}(x', dx'') + (P^n f)(x')\|y\|\\ &\leq K \sum_{i=0}^{n} (P^{n-i} f^2)(x') + K f(x')\|y\| \end{aligned} \tag{28}$$

for all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Due to Lemma 1 and (14), (27), (28),

$$\begin{aligned}\int \|(\Pi^{n+1}A(z')\|\Pi(z, dz') &\leq 2K \int \left(\sum_{i=0}^{n} (P^{n-i} f^2)(x'') + f(x'')\|\alpha\lambda y + \phi(x')\|\right) P(x', dx'')\\ &\leq 2K \sum_{i=0}^{n} (P^{n-i+1} f^2)(x') + 2K(\|y\| + f(x'))(Pf)(x') < \infty,\end{aligned}$$

$$\int \|(\Pi^{n+1}b(z')\|\Pi(z,dz') \leq K\int\left(\sum_{i=0}^{n}(P^{n-i}f^2)(x'') + f(x'')\|\alpha\lambda y + \phi(x')\|\right)P(x',dx'')$$

$$\leq K\sum_{i=0}^{n}(P^{n-i+1}f^2)(x') + K(\|y\| + f(x'))(Pf)(x') < \infty$$

for all $x, x' \in R^{d'}, y \in R^{d'}, z = (x, x', y)$. Consequently, $(\Pi^{n+2}A)(\cdot)$ and $(\Pi^{n+2}b)(\cdot)$ are well-defined and finite, while (14) implies that

$$\begin{aligned}(\Pi^{n+2}A)(z) &= \sum_{i=0}^{n}(\alpha\lambda)^i\int\int\phi(x''')(\alpha(P^{i+1}\phi)(x''') - (P^i\phi)(x'''))^T \\ &\quad\cdot P^{n-i}(x'',dx''')P(x',dx'') \\ &\quad + (\alpha\lambda)^{n+1}\int(\alpha\lambda y + \phi(x'))(\alpha(P^{n+1}\phi)(x'') - (P^n\phi)(x''))^T\cdot P(x',dx'') \\ &= \sum_{i=0}^{n}(\alpha\lambda)^i\int\phi(x'')(\alpha(P^{i+1}\phi)(x'') - (P^i\phi)(x''))^T\cdot P^{n-i+1}(x',dx'') \\ &\quad + (\alpha\lambda)^{n+1}(\alpha\lambda y + \phi(x'))(\alpha(P^{n+2}\phi)(x') - (P^{n+1}\phi)(x'))^T \\ &= \sum_{i=0}^{n+1}(\alpha\lambda)^i\int\phi(x'')(\alpha(P^{i+1}\phi)(x'') - (P^i\phi)(x''))^T\cdot P^{n-i+1}(x',dx'') \\ &\quad + (\alpha\lambda)^{n+2}y(\alpha(P^{n+2}\phi)(x') - (P^{n+1}\phi)(x'))^T,\end{aligned}$$

$$\begin{aligned}(\Pi^{n+2}b)(z) &= \sum_{i=0}^{n}(\alpha\lambda)^i\int\int\phi(x''')(P^i\tilde{c})(x''')P^{n-i}(x'',dx''')P(x',dx'') \\ &\quad + (\alpha\lambda)^{n+1}\int(\alpha\lambda y + \phi(x'))(P^n\tilde{c})(x'')P(x',dx'') \\ &= \sum_{i=0}^{n}(\alpha\lambda)^i\int\phi(x'')(P^i\tilde{c})(x'')P^{n-i+1}(x',dx'') \\ &\quad + (\alpha\lambda)^{n+1}(\alpha\lambda y + \phi(x'))(P^{n+1}\tilde{c})(x') \\ &= \sum_{i=0}^{n+1}(\alpha\lambda)^i\int\phi(x'')(P^i\tilde{c})(x'')P^{n-i+1}(x',dx'') + (\alpha\lambda)^{n+2}y(P^{n+1}\tilde{c})(x')\end{aligned}$$

for all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Hence, (23) and (24) hold for $n + 2$ and all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. Then, using the mathematical induction, it can easily be deduced that the assertion of this lemma holds. □

Lemma 4. *Let A1 – A3 hold. Then, there exist Borel-measurable functions* $\tilde{A} : R^{d+2d'} \rightarrow R^{d\times d}$ *and* $\tilde{b} : R^{d+2d'} \rightarrow R^d$ *such that*

$$\int\|\tilde{A}(z')\|^2\Pi(z,dz') \leq h^2(z), \tag{29}$$

$$\int \|\tilde{b}(z')\|^2 \Pi(z, dz') \leq h^2(z), \tag{30}$$

$$A(z) - A_* = \tilde{A}(z) - (\Pi\tilde{A})(z), \tag{31}$$

$$b(z) - b_* = \tilde{b}(z) - (\Pi\tilde{b})(z) \tag{32}$$

for all $z \in R^{d+2d'}$, where

$$h(z) = M(f^2(x') + g^2(x') + \|y\|^2)$$

for $x, x' \in R^{d'}$, $y \in R^d$, $z = (x, x', y)$ (M is defined on page 7).

Proof: For $x, x' \in R^{d'}$, $y \in R^d$ and $z = (x, x', y)$, let

$$\hat{a}(z) = 5^{-1}M + f(x)\|y\| + f(x')\|y\| + 2(1-\alpha\lambda)^{-1}g(x'),$$

$$\hat{b}(z) = 5^{-1}M + |c(x, x')|\|y\| + (1-\alpha\lambda)^{-1}g(x').$$

Then, using the Jensen and Minkowski inequality, it can easily be deduced from A2 that

$$\begin{aligned}\left(\int \hat{a}^2(z')\Pi(z, dz')\right)^{1/2} &\leq 5^{-1}M + 2(1-\alpha\lambda)^{-1}((Pg^2)(x'))^{1/2} + f(x')\|\alpha\lambda y + \phi(x')\| \\ &\quad + \left(\int f^2(x'')\|\alpha\lambda y + \phi(x')\|^2 P(x', dx'')\right)^{1/2} \\ &\leq 5^{-1}M + 2(1-\alpha\lambda)^{-1}((Pg^2)(x'))^{1/2} \\ &\quad + (f(x') + ((Pf^4)(x'))^{1/4})(f(x') + \|y\|) \\ &\leq h(z),\end{aligned} \tag{33}$$

$$\begin{aligned}\left(\int \hat{b}^2(z')\Pi(z, dz')\right)^{1/2} &\leq 5^{-1}M + (1-\alpha\lambda)^{-1}((Pg^2)(x'))^{1/2} \\ &\quad + \left(\int |c(x', x'')|^2\|\alpha\lambda y + \phi(x')\|^2 P(x', dx'')\right)^{1/2} \\ &\leq 5^{-1}M + (1-\alpha\lambda)^{-1}((Pg^2)(x'))^{1/2} \\ &\quad + \left(\int |c(x, x')|^4 P(x', dx'')\right)^{1/4}(f(x') + \|y\|) \\ &\leq h(z)\end{aligned} \tag{34}$$

for all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y)$. On the other hand, using Lemma 3, it is straightforward to verify that

$$\begin{aligned}(\Pi^{n+1}A)(z) - A_* &= \sum_{i=0}^{n}(\alpha\lambda)^i \int \phi(x'')(\alpha(P^{i+1}\phi)(x'') - (P^i\phi)(x''))^T \cdot (P^{n-i} - \pi)(x', dx'') \\ &\quad - \sum_{i=n+1}^{\infty}(\alpha\lambda)^i \int \phi(x'')(\alpha(P^{i+1}\phi)(x'') - (P^i\phi)(x''))^T \pi \cdot (dx'') \\ &\quad + (\alpha\lambda)^{n+1} y(\alpha(P^{n+1}\phi)(x') - (P^n\phi)(x'))^T, \end{aligned} \tag{35}$$

$$\begin{aligned}(\Pi^{n+1}b)(z) - b_* &= \sum_{i=0}^{n}(\alpha\lambda)^i \int \phi(x'')(P^i\tilde{c})(x'')(P^{n-i} - \pi)(x', dx'') \\ &\quad - \sum_{i=n+1}^{\infty}(\alpha\lambda)^i \int \phi(x'')(P^i\tilde{c})(x'')\pi(dx'') + (\alpha\lambda)^{n+1} y(P^n\tilde{c})(x') \end{aligned} \tag{36}$$

for all $x, x' \in R^{d'}, y \in R^d, z = (x, x', y), n \geq 0$ (note that the infinite sums in (35) and (36) are well-defined and finite due to Lemma 2). Owing to A3,

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{i=0}^{n}(\alpha\lambda)^i \left\| \int \phi(x')(P^{m+i}\phi^T)(x')(P^{n-i} - \pi)(x, dx') \right\| \\ &\quad \leq \sum_{i=0}^{\infty}(\alpha\lambda)^i \sum_{n=i}^{\infty} \left\| \int \phi(x')(P^{m+i}\phi^T)(x')(P^{n-i} - \pi)(x, dx') \right\| \\ &\quad \leq \sum_{i=0}^{\infty}(\alpha\lambda)^i g(x) = (1 - \alpha\lambda)^{-1} g(x), \end{aligned} \tag{37}$$

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{i=0}^{n}(\alpha\lambda)^i \left\| \int \phi(x')(P^i\tilde{c})(x')(P^{n-i} - \pi)(x, dx') \right\| \\ &\quad \leq \sum_{i=0}^{\infty}(\alpha\lambda)^i \sum_{n=i}^{\infty} \left\| \int \phi(x')(P^i\tilde{c})(x')(P^{n-i} - \pi)(x, dx') \right\| \\ &\quad \leq \sum_{i=0}^{\infty}(\alpha\lambda)^i g(x) = (1 - \alpha\lambda)^{-1} g(x) \end{aligned} \tag{38}$$

for all $x \in R^{d'}, m \geq 0,$ while A2 and Lemma 1 imply that

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{i=n+1}^{\infty}(\alpha\lambda)^i \int \|\phi(x)(P^{m+i}\phi^T)(x)\|\pi(dx) \\ &\quad \leq K\sum_{n=0}^{\infty}\sum_{i=n}^{\infty}(\alpha\lambda)^i \int f^2(x)\pi(dx) = K(1 - \alpha\lambda)^{-2} \int f^2(x)\pi(dx), \end{aligned} \tag{39}$$

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{i=n+1}^{\infty}(\alpha\lambda)^i\int\|\phi(x)(P^{m+i}\tilde{c})(x)\|\pi(dx)\\&\qquad\le K\sum_{n=0}^{\infty}\sum_{i=n}^{\infty}(\alpha\lambda)^i\int f^2(x)\pi(dx)=K(1-\alpha\lambda)^{-2}\int f^2(x)\pi(dx)\end{aligned}\tag{40}$$

for $m \geq 0$. *Since* $\|A(z)\| \leq (f(x) + f(x'))\|y\|$ for all $x, x' \in R^{d'}$, $y \in R^d$, $z = (x,x',y)$ (due to A2), it can easily be deduced from Lemma 2 and (35),(36), (37), (38), (39), (40) that

$$\begin{aligned}\sum_{n=0}^{\infty}\|(\Pi^n A)(z)-A_*\|\le{}&\sum_{n=0}^{\infty}\sum_{i=0}^{n}(\alpha\lambda)^i\left\|\int\phi(x'')(P^{i+1}\phi^T)(x'')(P^{n-i}-\pi)(x',dx'')\right\|\\&+\sum_{n=0}^{\infty}\sum_{i=0}^{n}(\alpha\lambda)^i\left\|\int\phi(x'')(P^{i}\phi^T)(x'')(P^{n-i}-\pi)(x',dx'')\right\|\\&+\sum_{n=0}^{\infty}\sum_{i=n+1}^{\infty}(\alpha\lambda)^i\int\|\phi(x'')(P^{i+1}\phi^T)(x'')\|\pi(dx'')\\&+\sum_{n=0}^{\infty}\sum_{i=n+1}^{\infty}(\alpha\lambda)^i\int\|\phi(x'')(P^{i}\phi^T)(x'')\|\pi(dx'')\\&+(f(x)+f(x'))\|y\|+\|A_*\|\\\le{}&3(1-\alpha\lambda)^{-1}\int f^2(x'')\pi(dx'')+(f(x)+f(x'))\|y\|\\&+2(1-\alpha\lambda)^{-1}g(x)\le\hat{a}(z)<\infty,\end{aligned}\tag{41}$$

$$\begin{aligned}\sum_{n=0}^{\infty}\|(\Pi^n b)(z)-b_*\|\le{}&\sum_{n=0}^{\infty}\sum_{i=0}^{n}(\alpha\lambda)^i\left\|\int\phi(x'')(P^{i}\tilde{c})(x'')(P^{n-i}-\pi)(x',dx'')\right\|\\&+\sum_{n=0}^{\infty}\sum_{i=n+1}^{\infty}(\alpha\lambda)^i\int\|\phi(x'')(P^{i}\tilde{c})(x'')\|\pi(dx'')\\&+|c(x,x')|\|y\|+\|b_*\|\le2(1-\alpha\lambda)^{-1}\int f^2(x'')\pi(dx'')\\&+|c(x,x')|\|y\|+(1-\alpha\lambda)^{-1}g(x)\le\hat{b}(z)<\infty\end{aligned}\tag{42}$$

for all $x, x' \in R^{d'}$ $y \in R^d$, $z = (x\ , x', y)$. *Let* $\tilde{A}(z) = \sum_{n=0}^{\infty}((\Pi^n A)(z) - A_*)$ and $\tilde{b}(z) = \sum_{n=0}^{\infty}((\Pi^n b)(z) - b_*)$, $z \in R^{d+2d'}$. Then, (41) and (42) imply that $\tilde{A}(\cdot)$ and $\tilde{b}(\cdot)$ are well-defined, finite and satisfy (31), (32), while (29) and (30) directly follow from (33), (34), (41) and (42). □

4. Mean-square error analysis

In this section, Theorem 1 is proved. The following notation is used in the section. For $x \in R^{d'}$ $x \in R^{d+2d'}$, let $E_x(\cdot\ ) = E(\cdot|X_0 = x)$ and

$$\xi(z) = A(z)\theta_* + b(z),$$
$$\tilde{\xi}(z) = \tilde{A}(z)\theta_* + \tilde{b}(z),$$

while $\tilde{K}_Q = M^2(1+\rho_Q)$ and $\tilde{L}_Q = 6\tilde{K}_Q(1+\rho_Q)^2(1+\lambda_{\min})$ (ρ_Q, $\lambda_{\min}$, M are defined on page 7). Moreover, for $n \geq 0$, let

$$\theta'_{n+1} = \theta_n + \gamma(A(Z_{n+1})\theta_n + b(Z_{n+1})),$$

while $\vartheta_n = \theta_n - \theta_*$ and $\vartheta'_n = \theta'_n - \theta_*$. Furthermore, for $n \geq 1$, let

$$\begin{aligned}
p_{n+1} &= 2\gamma\vartheta_n^T(\tilde{A}(Z_{n+1}) - (\Pi\tilde{A})(Z_n))\vartheta_n + 2\gamma\vartheta_n^T(\tilde{\xi}(Z_{n+1}) - (\Pi\tilde{\xi})(Z_n)),\\
q_{n+1} &= 2\gamma\left(\vartheta_{n+1}^T(\Pi\tilde{A})(Z_{n+1})\vartheta_{n+1} - \vartheta_n^T(\Pi\tilde{A})(Z_{n+1})\vartheta_n\right) + 2\gamma(\vartheta_{n+1} - \vartheta_n)^T(\Pi\tilde{\xi})(Z_{n+1}),\\
r_{n+1} &= \gamma^2\|A(Z_{n+1})\vartheta_n + b(Z_{n+1})\|^2 - \gamma^2\vartheta_n^T A_*^2\vartheta_n,\\
s_{n+1} &= -4\lambda_{\min}\gamma^2\vartheta_n^T(\Pi\tilde{A})(Z_n)\vartheta_n - 4\lambda_{\min}\gamma^2\vartheta_n^T(\Pi\tilde{\xi})(Z_n),
\end{aligned}$$

while

$$u_n = 2\gamma\vartheta_n^T(\Pi\tilde{A})(Z_n)\vartheta_n,$$
$$v_n = 2\gamma\vartheta_n^T(\Pi\tilde{\xi})(Z_n)$$

and $a_n = \|\vartheta_n\|^2 + u_n + v_n$.

Outline of the Results of Section 4: The main result of Section 4 is the proof of Theorem 1. The proof of Theorem 1 is crucially based on the inequality (61). This inequality also provides an obvious explanation for why $\gamma < \gamma_{\min}^{-1}$ has to hold in order for (12) to be satisfied. The inequality (61) is essentially based on the decomposition (58) of $\|\vartheta'_{n+1}\|^2$ and Lemma 6. Lemma 6 directly follows from the results of Lemma 5 and basic properties of conditional expectations, while the decomposition (58) is crucially based on the Poisson eqs. (31), (32). Moreover, Lemma 6 (i.e., the right-hand sides of (56), (57)) provides an obvious explanation for the selection of $h_Q(\cdot)$ in (12). On the other hand, Lemma 5 provides upper bounds on the conditional expectations of $A(\cdot)$, $\tilde{A}(\cdot)$, $(\Pi\tilde{A})(\cdot)$, $\xi(\cdot\ )$, $\tilde{\xi}(\cdot)$, $(\Pi\tilde{\xi})(\cdot)$, while its proof uses only the Cauchy-Schwartz and Minkowski inequalities, and the results of Lemma 4.

Lemma 5. *Let A1 – A4 hold. Suppose that $\theta_* \in\ Q$. Then,*

$$\max\{(E_x\|A(Z_n)\|^2)^{1/2}, (E_x\|\tilde{A}(Z_n)\|^2)^{1/2}, (E_x\|(\Pi\tilde{A})(Z_n)\|^2)^{1/2}\} \leq \tilde{K}_Q(f^2(x) + g(x)) \quad (43)$$

$$\max\{(E_x\|\xi(Z_n)\|^2)^{1/2}, (E_x\|\tilde{\xi}(Z_n)\|^2)^{1/2}, (E_x\|(\Pi\tilde{\xi})(Z_n)\|^2)^{1/2}\} \leq \tilde{K}_Q(f^2(x) + g(x)) \quad (44)$$

for all $x \in\ R^{d'}$, $n \geq 0$.

Proof: Due to the Jensen inequality and A2,

$$(E_x(f^4(X_n)))^{1/4} = ((P^n f^4)(x))^{1/4} \leq K^{1/2} f(x), \tag{45}$$

$$\begin{aligned}(E_x|c(X_n, X_{n+1})|^4)^{1/4} &= \left(E_x\left(\int |c(X_n, x')|^4 P(X_n, dx')\right)\right)^{1/4} \\ &\leq ((P^n f^4)(x))^{1/4} \leq K^{1/2} f(x),\end{aligned} \tag{46}$$

$$(E_x(g^2(X_n)))^{1/2} = ((P^n g^2)(x))^{1/2} \leq Lg(x) \tag{47}$$

for all $x \in R^{d'}$, $n \geq 0$, while the Minkowski inequality, A2 and (45) yield

$$\begin{aligned}(E_x\|e_{n+1}\|^4)^{1/4} &\leq \left(E_x\left(\sum_{i=0}^{n}(\alpha\lambda)^{n-i} f(X_i)\right)^4\right)^{1/4} \\ &\leq \sum_{i=0}^{n}(\alpha\lambda)^{n-i}(E_x(f^4(X_i)))^{1/4} \leq K^{1/2}(1-\alpha\lambda)^{-1} f(x)\end{aligned} \tag{48}$$

for all $x \in R^{d'}$, $n \geq 0$. Using the Cauchy-Schwartz and Minkowski inequality, it can easily be deduced from A2 and (45) – (48) that

$$\begin{aligned}(E_x\|A(Z_{n+1})\|^2)^{1/2} &\leq (E_x((f(X_n) + f(X_{n+1}))\|e_{n+1}\|)^2)^{1/2} \\ &\leq (E_x(f^2(X_n)\|e_{n+1}\|^2))^{1/2} + (E_x(f^2(X_{n+1})\|e_{n+1}\|^2))^{1/2} \\ &\leq ((E_x(f^4(X_n))^{1/4}(E_x\|e_{n+1}\|^4)^{1/4} + (E_x(f^4(X_{n+1}))^{1/4})(E_x\|e_{n+1}\|^4)^{1/4} \\ &\leq 2K(1-\alpha\lambda)^{-1} f^2(x),\end{aligned} \tag{49}$$

$$\begin{aligned}(E_x\|b(Z_{n+1})\|^2)^{1/2} &\leq (E_x(|c(X_n, X_{n+1})|^2\|e_{n+1}\|^2))^{1/2} \\ &\leq (E_x|c(X_n, X_{n+1})|^4)^{1/4}(E_x\|e_{n+1}\|^4)^{1/4} \\ &\leq K(1-\alpha\lambda)^{-1} f^2(x),\end{aligned} \tag{50}$$

$$\begin{aligned}(E_x(h^2(Z_{n+1})))^{1/2} &\leq M((E_x(f^4(X_{n+1})))^{1/2} + (E_x(g^2(X_{n+1}))^{1/2} \\ &\quad + (E_x\|e_{n+1}\|^4)^{1/2}) \leq M^2(f^2(x) + g(x))\end{aligned} \tag{51}$$

for all $x \in R^{d'}$, $n \geq 0$. On the other hand the Cauchy-Schwartz inequality and Lemma 4 yield that

$$E_x\|\tilde{A}(Z_{n+1})\|^2 = E_x\left(\int \|\tilde{A}(z)\|^2 \Pi(Z_n, dz)\right) \leq E_x(h^2(Z_n)), \tag{52}$$

$$E_x\|\tilde{b}(Z_{n+1})\|^2 = E_x\left(\int \|\tilde{b}(z)\|^2\Pi(Z_n, dz)\right) \leq E_x(h^2(Z_n)), \tag{53}$$

$$\begin{aligned} E_x\|(\Pi\tilde{A})(Z_n)\|^2 &= E_x\left\|\int \tilde{A}(z)\Pi(Z_n, dz)\right\|^2 \\ &\leq E_x\left(\int \|\tilde{A}(z)\|^2\Pi(Z_n, dz)\right) \leq E_x(h^2(Z_{n+1})), \end{aligned} \tag{54}$$

$$\begin{aligned} E_x\|(\Pi\tilde{b})(Z_n)\|^2 &= E_x\left\|\int \tilde{b}(z)\Pi(Z_n, dz)\right\|^2 \leq E_x\left(\int \|\tilde{b}(z)\|^2\Pi(Z_n, dz)\right) \\ &\leq E_x(h^2(Z_{n+1})) \end{aligned} \tag{55}$$

for all $x \in R^{d'}$, $n \geq 0$. Since $\|\theta_*\| \leq \rho_Q$ and

$$\|\xi(Z_n)\| \leq \|A(Z_n)\|\|\theta_*\| + \|b(Z_n)\| \leq (1+\rho_Q)(\|A(Z_n)\| + \|b(Z_n)\|),$$

$$\|\tilde{\xi}(Z_n)\| \leq \|\tilde{A}(Z_n)\|\|\theta_*\| + \|\tilde{b}(Z_n)\| \leq (1+\rho_Q)(\|\tilde{A}(Z_n)\| + \|\tilde{b}(Z_n)\|),$$

$$\|(\Pi\tilde{\xi})(Z_n)\| \leq \|(\Pi\tilde{A})(Z_n)\|\|\theta_*\| + \|(\Pi\tilde{b})(Z_n)\| \leq (1+\rho_Q)(\|(\Pi\tilde{A})(Z_n)\| + \|(\Pi\tilde{b})(Z_n)\|)$$

for $n \geq 0$, (43) and (44) follow directly from (49) – (55). □

Lemma 6. *Let A1 – A4 hold. Suppose that $\theta_* \in Q$. Then, $E_x(p_n) = 0$ for all $x \in R^{d'}$, $n \geq 1$. Moreover,*

$$\overline{\lim_{n\to\infty}} E_x(|q_n| + |r_n| + |s_n|) \leq \tilde{L}_Q\gamma^2(f^4(x) + g^2(x)), \tag{56}$$

$$\overline{\lim_{n\to\infty}} E_x(|u_n| + |v_n|) \leq \tilde{L}_Q\gamma(f^4(x) + g^2(x)) \tag{57}$$

for all $x \in R^{d'}$.

Proof: It is straightforward to verify that

$$\begin{aligned} q_{n+1} &= 2\gamma(\vartheta_{n+1} - \vartheta_n)^T(\Pi\tilde{A})(Z_{n+1})\vartheta_n + 2\gamma\vartheta_n^T(\Pi\tilde{A})(Z_{n+1})(\vartheta_{n+1} - \vartheta_n) \\ &\quad + 2\gamma(\vartheta_{n+1} - \vartheta_n)^T(\Pi\tilde{A})(Z_{n+1})(\vartheta_{n+1} - \vartheta_n) + 2\gamma(\vartheta_{n+1} - \vartheta_n)^T(\Pi\tilde{\xi})(Z_{n+1}) \end{aligned}$$

for $n \geq 0$. Since

$$\|\vartheta_n\| = \|\theta_n - \theta_*\| \leq \rho_Q,$$

$$\|\vartheta_{n+1} - \vartheta_n\| = \|\theta_{n+1} - \theta_n\| \leq \rho_Q,$$

$$\|\vartheta_{n+1} - \vartheta_n\| = \|\theta_{n+1} - \theta_n\| \leq \|\theta'_{n+1} - \theta_n\| = \|\vartheta'_{n+1} - \vartheta_n\|$$

for $n \geq 1$ (notice that $\|P_Q(\theta') - \theta \| \leq \|\theta' - \theta\|$ for all $\theta \in Q$, $\theta' \in R^d$; see e.g., [(Pflug, 1996, Appendix E]), it can easily be deduced that for $n \geq 1$,

$$\begin{aligned}
|p_{n+1}| &\leq 2\rho_Q^2 \gamma (\|\tilde{A}(Z_{n+1})\| + \|(\Pi\tilde{A})(Z_n)\| + 2\rho_Q^2\gamma\|\tilde{\xi}(Z_{n+1})\| + \|(\Pi\tilde{\xi})(Z_n)\|),\\
|q_{n+1}| &\leq \gamma\|\vartheta_{n+1} - \vartheta_n\|(6\rho_Q\|(\Pi\tilde{A})(Z_{n+1})\| + 2\|(\Pi\tilde{\xi})(Z_{n+1})\|)\\
&\leq \gamma^2(\|A(Z_{n+1})\|\|\vartheta_n\| + \|\xi(Z_{n+1})\|)\cdot(6\rho_Q\|(\Pi\tilde{A})(Z_{n+1})\| + 2\|(\Pi\tilde{\xi})(Z_{n+1})\|)\\
&\leq 6(1+\rho_Q)^2\gamma^2(\|A(Z_{n+1})\|^2 + \|\xi(Z_{n+1})\|^2 + 6(1+\rho_Q)^2\gamma^2\|(\Pi\tilde{A})(Z_{n+1})\|^2\\
&\quad + \|(\Pi\tilde{\xi})(Z_{n+1})\|^2),\\
|r_{n+1}| &\leq \gamma^2(\|A(Z_{n+1})\|\|\vartheta_n\| + \|\xi(Z_{n+1})\|)^2 + \gamma^2\|A_*\|^2\|\vartheta_n\|^2\\
&\leq 2(1+\rho_Q)\gamma^2(\|A(Z_{n+1})\|^2 + \|\xi(Z_{n+1})\|^2) + M^2\rho_Q^2\gamma^2,\\
|s_{n+1}| &\leq 4\rho_Q^2\lambda_{\min}\gamma^2(\|(\Pi\tilde{A})(Z_n)\| + \|(\Pi\tilde{\xi})(Z_n)\|),
\end{aligned}$$

as well as

$$\begin{aligned}
|u_n| &\leq 2\rho_Q^2\gamma\|(\Pi\tilde{A})(Z_n)\|,\\
|v_n| &\leq 2\rho_Q\gamma\|(\Pi\tilde{\xi})(Z_n)\|.
\end{aligned}$$

Then, Lemma 5 implies that (56) and (57) hold, as well as that $E_x|p'_n| < \infty$ and $E_x|p'_n| < \infty, \forall\, x \in R^{d'}\}$, $n \geq 1$.

Let $\mathcal{F}_n = \sigma[\theta_0, X_0, \ldots, X_n]$, $n \geq 0$. Since $\sigma\{\vartheta_n\} \subseteq \mathcal{F}_n$ and

$$\begin{aligned}
E_x(\tilde{A}(Z_{n+1}) \mid \mathcal{F}_n) &= (\Pi\tilde{A})(Z_n),\\
E_x(\tilde{\xi}(Z_{n+1}) \mid \mathcal{F}_n) &= (\Pi\tilde{\xi})(Z_n) \qquad ()
\end{aligned}$$

for all $x \in R^{d'}$, $n \geq 0$, it can easily be deduced that $E_x(p_{n+1} \mid \mathcal{F}_n) = 0$ for all $x \in R^{d'}$, $n \geq 0$. Consequently, $E_x(p_n) = 0$ for $n \geq 0$. □

Proof of Theorem 1: *It is straightforward to verify that*

$$\|\vartheta'_{n+1}\|^2 = \vartheta_n^T(I + \gamma A_*)^2\vartheta_n + p_{n+1} + q_{n+1} + r_{n+1} + u_n + v_n - u_{n+1} - v_{n+1}, \quad (58)$$

$$a_n \geq \|\vartheta_n\|^2 - |u_n| - |v_n| \quad (59)$$

for $n \geq 1$. Since

$$\begin{aligned}
\vartheta_n^T(I + \gamma A_*)^2\vartheta_n &\leq (1 - \lambda_{\min}\gamma)^2\|\vartheta\|^2,\\
\|\vartheta_n\| = \|\theta_n - \theta_*\| &\leq \|\theta'_n - \theta_*\| = \|\vartheta'_n\|
\end{aligned}$$

for $n \geq 1$ (notice that $\|P_Q(\theta') - \theta\| \leq \|\theta' - \theta\|$ for all $\theta \in Q$, $\theta \in R^d$; for details see e.g.[(Pflug 1996, Appendix E)], it can easily be deduced from (58) that for $n \geq 1$

$$\|\vartheta_{n+1}\|^2 \leq (1 - \lambda_{\min}\gamma)^2\|\vartheta_n\|^2 + p_{n+1} + q_{n+1} + r_{n+1} + u_n + v_n - u_{n+1} - v_{n+1}. \quad (60)$$

Consequently,

$$\begin{aligned} a_{n+1} &\leq (1 - \lambda_{\min}\gamma)^2\|\vartheta_n\|^2 + p_{n+1} + q_{n+1} + r_{n+1} + u_n + v_n \\ &= (1 - \lambda_{\min}\gamma)^2 a_n + p_{n+1} + q_{n+1} + r_{n+1} + s_{n+1} \end{aligned}$$

for $n \geq 1$. Then, Lemma 6 and (59) imply that for all $x \in R^{d'}$, $n \geq 1$,

$$\begin{aligned} E_x(a_{n+1}) &\leq (1 - \lambda_{\min}\gamma)^2 E_x(a_n) + E_x(|q_{n+1}| + |r_{n+1}| + |s_{n+1}|) \\ &\leq (1 - \lambda_{\min}\gamma)^2 E_x(a_n) + \tilde{L}_Q\gamma^2(f^4(x) + g^2(x)), \end{aligned} \quad (61)$$

while Lemma 6 and (60) yield that

$$\overline{\lim_{n\to\infty}} E_x\|\vartheta_n\|^2 \leq \overline{\lim_{n\to\infty}} E_x|a_n| + \tilde{L}_Q\gamma(f^4(x) + g^2(x)) \quad (62)$$

for all $x \in R^{d'}$. Due to (61),

$$E(a_n) \leq (1 - \lambda_{\min}\gamma)^{2n}E(a_0) + \tilde{L}_Q\gamma^2(f^4(x) + g^2(x))\sum_{i=0}^{n-1}(1 - \lambda_{\min}\gamma)^{2i} \quad (63)$$

for $n \geq 1$, while (63) yields that for all $x \in R^{d'}$,

$$\begin{aligned} \overline{\lim_{n\to\infty}} E_x(a_n) &\leq \tilde{L}_Q\gamma^2(1 - (1 - \lambda_{\min}\gamma)^2)^{-1}(f^4(x) + g^2(x)) \\ &= \tilde{L}_Q\lambda_{\min}^{-1}\gamma(2 - \lambda_{\min}\gamma)^{-1}(f^4(x) + g^2(x)) \leq \tilde{L}_Q\lambda_{\min}^{-1}\gamma(f^4(x) + g^2(x)) (64) \end{aligned}$$

(notice that $2 - \lambda_{\min}\gamma > 1$ due to $\gamma < \lambda^{-1}{}_{\min}$). Owing to (62) and (64), for all $x \in R^{d'}$,

$$\overline{\lim_{n\to\infty}} E_x\|\vartheta_n\|^2 \leq \tilde{L}_Q(1 + \lambda_{\min}^{-1})\gamma(f^4(x) + g^2(x)),$$

wherefrom (12) directly follows for all $x \in R^{d'}$. □

5. Special Case

The results of this section correspond with a special case of A1 – A4 where $\{X_n\}_{n\geq 0}$ is geometrically ergodic. This case is analyzed because the geometric ergodicity is considered in practice as one of the most important types of stability of Markov chains (see e.g., Meyn & Tweedie, 1993). Furthermore, most of the existing asymptotic results on temporal-difference learning (as well as on reinforcement learning) either explicitly require the underlying chain $\{X_n\}_{n\geq 0}$ to be geometrically ergodic, or have been obtained under assumptions which are very close to geometric ergodicity (see Bertsekas & Tsitsiklis, 1996; Tsitsiklis & Van Roy, 1997 and references cited therein).

B1. $\{X_n\}_{n\geq 0}$ *has a unique invariant probability measure* $\pi(\cdot)$.
B2. *There exists a Borel-measurable function* $f: R^{d'} \rightarrow [1, \infty)$ *such that* $\int f^4(x)\,\pi(dx) < \infty$, $\|\phi(x)\| \leq f(x)$ *and*

$$\int |c(x, x')|^4 P(x, dx') \leq f^4(x)$$

for all $x \in R^{d'}$.
B3. *There exist constants* $K \in [1, \infty)$ and $\rho \in (0, 1)$ *such that*

$$\left| \int \varphi(x')(P^n - \pi)(x, dx') \right| \leq K\rho^n f^4(x) \tag{65}$$

for all $x \in R^{d'}$, $n \geq 0$, and for any Borel-measurable function $\varphi: R^{d'} \rightarrow R$ satisfying $0 \leq \varphi(x) \leq f^4(x)$ for all $x \in R^{d'}$.

Remark. *Assumption B3 is equivalent to the requirement that* $\{X_n\}_{n\geq 0}$ *is* f^4*-uniformly ergodic (for more details on this type of ergodicity, see* [(Meyns Tweedie, 1993, Section 16]).

Lemma 7. *Let B1 – B3 hold. Then, A1 – A3 are also satisfied.*

Proof: As $\{X_n\}_{n\geq 0}$ is geometrically ergodic (due to B1), A1 holds. Let

$$L = K + \int f^4(x)\pi(dx).$$

Due to the Jensen inequality and B2, B3,

$$(P^n f^4)(x) \leq \int f^4(x)\pi(dx) + \left| \int f^4(x')(P^n - \pi)(x, dx') \right| \leq Lf^4(x), \tag{66}$$

$$|\tilde{c}(x)| \leq \int |c(x, x')| P(x, dx') \leq \left(\int |c(x, x')|^4 P(x, dx') \right)^{1/4} \leq f(x) \tag{67}$$

for all $x \in R^{d'}$, $n \geq 0$. Owing the Jensen inequality and (66), for all $x \in R^{d'}$, $n \geq 0$,

$$(P^n f)(x) \leq ((P^n f^4)(x))^{1/4} \leq Lf(x) \tag{68}$$

while B2 and (67), (68) imply

$$\|\phi(x)(P^n\phi)(x)\| \leq f(x)(P^n f)(x) \leq 4Lf^2(x),$$
$$\|\phi(x)(P^n\tilde{c})(x)\| \leq f(x)(P^n f)(x) \leq 4Lf^2(x)$$

$x \in R^{d'}$, $n \geq 0$. Then, B3 yields

$$\left\| \int \phi(x')(P^m\phi^T)(x')(P^n - \pi)(x, dx') \right\| \leq KL\rho^n f^4(x),$$

$$\left\| \int \phi(x')(P^m \tilde{c})(x')(P^n - \pi)(x, dx') \right\| \leq KL\rho^n f^4(x)$$

$x \in R^{d'}$*, m, n* $\geq$ *0.* Consequently,

$$\sum_{n=0}^{\infty} \left\| \int \phi(x')(P^m \phi^T)(x')(P^n - \pi)(x, dx') \right\| \leq KL(1-\rho)^{-1} f^4(x),$$

$$\sum_{n=0}^{\infty} \left\| \int \phi(x')(P^m \tilde{c})(x')(P^n - \pi)(x, dx') \right\| \leq KL(1-\rho)^{-1} f^4(x)$$

$x \in R^{d'}$ *m* $\geq$ *0.* Then, it can easily be deduced that A2 and A3 hold, too. □

As an immediate consequence of Theorem 1, Lemma 7 and (13), the following corollary is obtained:

Corollary 1. *Let A4 and B1 – B3 hold. Suppose that* $\theta_* \in$ *Q. Then, there exists a Borel-measurable function* h_Q*:* $R^{d'} \rightarrow$ *[1,* ∞ *) such that*

$$\overline{\lim_{n\to\infty}} E(\|\theta_n - \theta_*\|^2 | X_0 = x) \leq h_Q(x)\gamma,$$

$$\overline{\lim_{n\to\infty}} E\|\theta_n - \theta_*\|^2 \leq \gamma \int h_Q(x')\pi(dx')$$

for all $\gamma \in (0, \lambda_{\min}^{-1})$*,* $x \in R^{d'}$ *(*$\lambda_{\min}$*is defined on page 7).*

6. Examples

In this section, the main results of the paper are illustrated with examples related to *M*/*G*/1 queues and nonlinear autoregressive (AR) models with Markov switching.

6.1. *M*/*G*/1 Queue

M/*G*/1 queues are models for service stations with the following properties:

(i) The times between arrivals of consecutive customers (interarrival times) are independent and identically distributed random variables.
(ii) Customers are severed on ‘first-come-first-served’ principle.
(iii) The times of serving customers (service times) are independent and identically distributed random variables.
(iv) Interarrival times are exponentially distributed.

Let Y_{n+1} is the number of customers in a *M/G/1* queue immediately after the completion of the service of the *n*-th customer. Moreover, let $X_n = \psi(Y_n)$, $n \geq 0$, where $\psi(\cdot)$ is a function mapping nonnegative integers into $R^{d'}$ and satisfying $\psi(m) = \psi(n)$ only if $m = n$. Then, it can easily be deduced that $\{X_n\}n \leq 0$ is a Markov process with values in $R^{d'}$ and the transition probability kernel $P(x, \cdot), x \in R^{d'}$, defined as

$$P(x, B) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\mathcal{P}(Y_1 = j \mid Y_0 = i)I_{\{\psi(i)\}}(x)I_B(\psi(j))$$
$$+\sum_{j=0}^{\infty}\mathcal{P}(Y_1 = j|Y_0 = 0)I_{\{\psi(i):i\geq 0\}^c}(x)I_B(\psi(j))$$

for $x \in R^{d'}, B \in \mathcal{B}^{d'}$.

Let μ be the mean of the interarrival times of the customers in the queue, while $\nu(\cdot)$ is the distribution of their service times (for details on $M/G/1$ queues see e.g., (Asmussen, 1987, Meyn, Tweedie, 1993) and references cited therein). The next lemma is a direct consequence of[(Meyn & Tweedie, 1993,Subsection 16.1.3]

Lemma 8. *Suppose that there exists a constant $s \in (0, \infty)$ such that $\int \exp(st)\, \nu(dt) < \infty$. Moreover, suppose that $\int t\nu(dt) < \mu$ and $\|\psi(n)\| \leq n$, $n \geq 0$. Then, $\{X_n\}_{n\geq 0}$ has a unique invariant probability measure $\pi(\cdot)$ (concentrated on $\{\psi(n) : n \geq 0\}$) and there exist constants $K \in [1, \infty)$, $L \in (0, \infty)$ and $\rho \in (0,1)$ such that $\int \exp(L\|x\|)\, \pi(dx) < \infty$ and*

$$\left|\int \phi(x')(P^n - \pi)(x, dx')\right| \leq K\rho^n \exp(L\|x\|)$$

for all $x \in R^{d'}$, $n \geq 0$, and any Borel-measurable function $\varphi: R^{d'} \to R$ satisfying $0 \leq \varphi(x) \leq \exp(L\|x\|)$ for all $x \in R^{d'}$.

As an immediate consequence of Theorem 1, Lemmas 7, 8, and (13), the following corollary is obtained.

Corollary 2. *Let A4 hold. Suppose that the conditions of Lemma 8 are satisfied and*

$$\|\phi(x)\| \leq M(1 + \|x\|^p),$$
$$|c(x, x')| \leq M(1 + \|x\|^p + \|x'\|^p)$$

for all $x, x' \in R^{d'}$, where $p, M \in [1, \infty)$ are constants. Moreover, suppose that $\theta_* \in Q$. Then, there exists a Borel-measurable function $h_Q: R^{d'} \to [1, \infty)$ such that

$$\overline{\lim_{n\to\infty}} E(\|\theta_n - \theta_*\|^2 | X_0 = x) \leq h_Q(x)\gamma,$$
$$\overline{\lim_{n\to\infty}} E\|\theta_n - \theta_*\|^2 \leq \gamma \int h_Q(x')\pi(dx')$$

for all $\gamma \in (0, \lambda_{\min}^{-1})$, $x \in R^{d'}$ ($\lambda_{\min}$ is defined on page 7).

6.2. Nonlinear Autoregressive Models with Markov Switching

Nonlinear autoregressive (AR) models with Markov switching are defined by the following difference equation:

$$X_{n+1} = F(X_n, s_n) + \xi_{n+1}, \quad n \geq 0. \tag{69}$$

$F : R^{d'} \times \{1, \ldots, d''\} \to R^{d'}$ is a Borel-measurable function, while $X_0 \in R^{d'}$ is a deterministic variable. $\{s_n\}_{n\geq 0}$ is a homogeneous Markov chain with values in $\{1, \ldots, d''\}$, while $\{\xi_n\}_{n\geq 0}$ is an $R^{d'}$-valued i.i.d. sequence independent of $\{s_n\}_{n\geq 0}$.

Nonlinear AR model with Markov switching (69) is a state-space model with the following properties:

(i) The model states have two components X_n and s_n.
(ii) The discrete-valued components $\{s_n\}_{n\geq 0}$ form a Markov chain which evolves independently of $\{X_n\}_{n\geq 0}$.
(iii) Conditionally on $\{s_n\}_{n\geq 0}$, continuously valued components $\{X_n\}_{n\geq 0}$ have a Markov property, i.e.,

$$\mathcal{P}(X_{n+1} \in B | X_0, \ldots, X_n, s_n) = \mathcal{P}(X_{n+1} \in B | X_n, s_n) \quad w.p.1$$

for all $B \in \mathcal{B}^{d'}$, $n \geq 0$.

Nonlinear AR models with Markov switching usually correspond to systems whose structure and dynamics (modeled by $\{X_n\}_{n\geq 0}$) is significantly influenced by certain exogenous events (modeled by $\{s_n\}_{n\geq 0}$). These models have found a great number of applications in areas such as automatic control, signal processing and econometrics (see e.g., Yao & Attoli, 2000).

Let $Y_n = (X_n, s_n)$, $n \geq 0$. Then, it can easily be deduced that $\{Y_n\}_{n\geq 0}$ is a Markov chain with values in $R^{d'} \times \{1, \ldots, d''\}$ and the transition probability kernel defined as

$$P(x, i, B \times \{j\}) = E(I_B(F(x, j) + \xi_0))p_{ij}$$

for $x \in R^{d'}$, $B \in \mathcal{B}^{d'}$, $1 \leq i, j \leq d''$, where

$$p_{ij} = \mathcal{P}(s_1 = j \mid s_0 = i).$$

The next lemma is a direct consequence of [Yao & Attali., 2000, Theorem 2]

Lemma 9. *Suppose that $\{s_n\}_{n\geq 0}$ is irreducible and aperiodic. Moreover, suppose that ξ_0 has everywhere positive density with respect to the Lebesgue measure. Furthermore, suppose that there exist a constant $p \in [1, \infty)$ and a sequence $\{f(i)\}_{1\leq i\leq d''}$ from* $[0, \infty)$ *such that*

$$\|F(x, i)\| \leq f(i)(1 + \|x\|^{4p}),$$

$$\sum_{j=0}^{d''} f^{4p}(j)p_{ij} < 1$$

for all $x \in R^{d'}$, $1 \le i \le d''$. Then, $\{Y_n\}_{n\ge 0}$ has an invariant probability measure $\pi(\cdot)$ and there exist constants $K \in [1, \infty)$, $\rho \in (0,1)$ such that

$$\left|\int \phi(x', s')(P^n - \pi)(x, i, dx', ds')\right| \le K\rho^n(1 + \|x\|^{4p})$$

for all $x \in R^{d'}$, $1 \le i \le d''$, $n \ge 0$, and any Borel-measurable function ϕ: $R^{d'} \times \{1, \ldots, d''\} \to R$ satisfying $\phi(x,i) \le 1 + \|x\|^{4p}$ for all $x \in R^{d'}$, $1 \le i \le d''$.

As an immediate consequence of Theorem 1, Lemmas 7, 9, and (13), the following corollary is obtained.

Corollary 4. *Let A4 hold. Suppose that the conditions of Lemma 9 are satisfied and*

$$\|\phi(x)\| \le M(1 + \|x\|^p),$$
$$|c(x, x')| \le M(1 + \|x\|^p + \|x'\|^p)$$

for all $x, x' \in R^{d'}$, where $M \in [1, \infty)$ is a constant. Moreover, suppose that $\theta_* \in Q$. Then, there exists a Borel-measurable function h_Q: $R^{d'} \to [1, \infty)$ such that

$$\overline{\lim_{n\to\infty}} E(\|\theta_n - \theta_*\|^2 | X_0 = x) \le h_Q(x)\gamma,$$

$$\lim_{n\to\infty} E\|\theta_n - \theta_*\|^2 \le \gamma \int h_Q(x')\pi(dx')$$

for all $\gamma \in (0, \lambda_{\min}{}^{-1})$, $x \in R^{d'}$ ($\lambda_{\min}$ is defined on page 7).

7. Conclusion

In this paper, the mean-square asymptotic behavior of temporal-difference learning algorithms with constant step-sizes and linear function approximation has been analyzed. The analysis has been carried out for the case of discounted cost function associated with a Markov chain with a finite dimensional state-space. Under mild conditions, it has been demonstrated that for sufficiently small step-size, the corresponding temporal-difference learning algorithm is stable in the mean-square sense and its asymptotic mean-square error is bounded by a linear function of the step-size. The main results of the paper are illustrated with examples related to $M/G/1$ queues and nonlinear AR models with Markov switching. The results of this paper are an extension of the results of (Tsitsiklis & Van Roy, 1997) and a continuation of the author's work presented in (Tadić, 2000).

References

Asmussen, S. (1987). *Applied probability and queues*, Wiley.

Benveniste, A., Métivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*, Springer Verlag.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Dayan, P. D. (1992). The convergence of *TD*(λ) for general λ, *Machine Learning, 8*, 341–362.

Dayan, P. D., & Sejnowski, T. J. (1994). *TD*(λ) converges with probability 1. *Machine Learning, 14*, 295–301.

Duflo, M. (1997). *Random iterative models*, Springer.

Goodwin, G. C., & Sin, K. S. (1984). *Adaptive filtering. Prediction and control*, Prentice Hall.

Jaakola, T, Jordan, M. I., & Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming. *Neural Computation, 6*, 1185–1201.

Konda, V. R. (2002). *Actor-critic algorithms*, PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic approximation algorithms and applications*. Springer Verlag.

Meyn, S. P., & Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer Verlag.

Nedić, A., & Bertsekas, D. P. (2003). Policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic System, 13*, 79–110.

Pflug, G. Ch. (1996). *Optimization of stochastic models: The interface between simulation and optimization*, Kluwer.

Solo, V., & Kong, X. (1995). *Adaptive signal processing algorithms: Stability and performance*, Prentice-Hall.

Sutton, R. S. (1988). Learning to predict by the method of temporal-difference learning. *Machine Learning 3*, 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Tadić, V. B. (2000). On the convergence of temporal-difference learning with linear function approximation. *Machine Learning, 42*, 241–267.

Tsitsiklis, J. N., & Van Roy. B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control, 42*, 674–690.

Yao J.-F. & Attali, J.-G. (2000). On stability of nonlinear AR processes with Markov switching. *Advances in Applied Probability, 32*, 394–407.