Journal of Machine Learning Research 17 (2016) 1-58 Submitted 5/16; Published 11/16

Weak Convergence Properties of Constrained Emphatic Temporal-difference Learning with Constant and Slowly Diminishing Stepsize

Huizhen Yu JANEY.HZYU@GMAIL.COM
Reinforcement Learning and Artificial Intelligence Group
Department of Computing Science, University of Alberta
Edmonton, AB, T6G 2E8, Canada

Editor: Shie Mannor

Abstract

We consider the emphatic temporal-difference (TD) algorithm, ETD(λ), for learning the value functions of stationary policies in a discounted, finite state and action Markov decision process. The ETD(λ) algorithm was recently proposed by Sutton, Mahmood, and White (2016) to solve a long-standing divergence problem of the standard TD algorithm when it is applied to off-policy training, where data from an exploratory policy are used to evaluate other policies of interest. The almost sure convergence of ETD(λ) has been proved in our recent work under general off-policy training conditions, but for a narrow range of diminishing stepsize. In this paper we present convergence results for constrained versions of ETD(λ) with constant stepsize and with diminishing stepsize from a broad range. Our results characterize the asymptotic behavior of the trajectory of iterates produced by those algorithms, and are derived by combining key properties of ETD(λ) with powerful convergence theorems from the weak convergence methods in stochastic approximation theory. For the case of constant stepsize, in addition to analyzing the behavior of the algorithms in the limit as the stepsize parameter approaches zero, we also analyze their behavior for a fixed stepsize and bound the deviations of their averaged iterates from the desired solution. These results are obtained by exploiting the weak Feller property of the Markov chains associated with the algorithms, and by using ergodic theorems for weak Feller Markov chains, in conjunction with the convergence results we get from the weak convergence methods. Besides ETD(λ), our analysis also applies to the off-policy TD(λ) algorithm, when the divergence issue is avoided by setting λ sufficiently large. It yields, for that case, new results on the asymptotic convergence properties of constrained off-policy TD(λ) with constant or slowly diminishing stepsize.

Keywords: Markov decision processes, approximate policy evaluation, reinforcement learning, temporal-difference methods, importance sampling, stochastic approximation, convergence

1. Introduction

We consider discounted finite state and action Markov decision processes (MDPs) and the problem of learning an approximate value function for a given policy from *off-policy* data, that is, from data due to a different policy. The first policy is called the *target policy* and the second the *behavior policy*. The case of *on-policy* learning, where the target and behavior policies are the same, has been well-studied and widely applied (see e.g.,

Sutton, 1988; Tsitsiklis and Van Roy, 1997; and the books Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). Off-policy learning provides additional flexibilities and is useful in many contexts. For example, one may want to avoid executing the target policy before estimating the potential risk for safety concerns, or one may want to learn value functions for many target policies in parallel from one exploratory behavior. These require off-policy learning. In addition, insofar as value functions (with respect to different reward/cost assignments) reflect statistical properties of future outcomes, off-policy learning can be used by an autonomous agent to build an experience-based internal model of the world in artificial intelligence applications (Sutton, 2009). Algorithms for off-policy learning are thus not only useful as model-free computational methods for solving MDPs, but can also potentially be a step toward the goal of making autonomous agents capable of learning over a long life-time, facing a sequence of diverse tasks.

In this paper we focus on a new off-policy learning algorithm proposed recently by Sutton, Mahmood, and White (2016): the emphatic temporal-difference (TD) learning algorithm, or ETD(λ). The algorithm is similar to the standard TD(λ) algorithm with linear function approximation (Sutton, 1988), but uses a novel scheme to resolve a long-standing divergence problem in TD(λ) when applied to off-policy data. Regarding the divergence problem, while TD(λ) was proved to converge for the on-policy case (Tsitsiklis and Van Roy, 1997), it was known quite early that the algorithm can diverge in other cases (Baird, 1995; Tsitsiklis and Van Roy, 1997).[1] The difficulty is intrinsic to sampling states according to an arbitrary distribution. Since then alternative algorithms without convergence issues have been sought for off-policy learning. In particular, in the off-policy LSTD(λ) algorithm (Bertsekas and Yu, 2009; Yu, 2012), which is an extension of the on-policy least-squares version of TD(λ) proposed by Bradtke and Barto (1996) and Boyan (1999), with higher computational complexity than TD(λ), the linear equation associated with TD(λ) is estimated from data and then solved.[2] In the gradient-TD algorithms (Sutton et al., 2008, 2009; Maei, 2011) and the proximal gradient-TD algorithms (Liu et al., 2009; Mahadevan and Liu, 2012; see also Mahadevan et al., 2014; Liu et al., 2015), the difficulty in TD(λ) is overcome by reformulating the approximate policy evaluation problem TD(λ) attempts to solve as optimization problems and then tackle them with optimization techniques. (See the surveys Geist and Scherrer, 2014 and Dann et al., 2014 for other algorithm examples.)

Compared to the algorithms just mentioned, ETD(λ) is closer to the standard TD(λ) algorithm and addresses the issue in TD(λ) more directly. It introduces a novel weighting scheme to re-weight the states when forming the eligibility traces in TD(λ), so that the weights reflect the occupation frequencies of the target policy rather than the behavior policy. An important result of this weighting scheme is that under natural conditions on the function approximation architecture, the average dynamics of ETD(λ) can be described by an affine function involving a negative definite matrix (Sutton et al., 2016; Yu, 2015a),[3]

1. For related discussions, see also Bertsekas and Tsitsiklis (1996); Sutton and Barto (1998); and Sutton et al. (2016).
2. An efficient algorithm for solving the estimated equations is the one given by Yao and Liu (2008) based on the line search method. It can also be applied to finding approximate solutions under additional penalty terms suggested by Pires and Szepesv́ari (2012).
3. Sutton et al. (2016) work with the negation of the matrix that we associate with ETD(λ) in this paper. The negative definiteness property we discuss here corresponds to the positive definiteness property discussed in their work.

which provides a desired stability property, similar to the case of convergent on-policy TD algorithms.

The almost sure convergence of ETD(λ), under general off-policy training conditions, has been shown in our recent work (Yu, 2015a) for diminishing stepsize. That result, however, requires the stepsize to diminish at the rate of $O(1/t)$, with t being the time index of the iterate sequence. This range of stepsize is too narrow for applications. In practice, algorithms tend to make progress too slowly if the stepsize becomes too small, and the environment may be non-stationary, so it is often preferred to use a much larger stepsize or constant stepsize.

The purpose of this paper is to provide an analysis of ETD(λ) for a broad range of stepsizes. Specifically, we consider constant stepsize and stepsize that can decrease at a rate much slower than $O(1/t)$. We will maintain general off-policy training conditions, without placing restrictions on the behavior policy. However, we will consider constrained versions of ETD(λ), which constrain the iterates to be in a bounded set, and a mode of convergence that is weaker than almost sure convergence. Constraining the ETD(λ) iterates is not only needed in analysis, but also a means to control the variances of the iterates, which is important in practice since off-policy learning algorithms generally have high variances. Almost sure convergence is no longer guaranteed for algorithms using large stepsizes; hence we analyze their behavior with respect to a weaker convergence mode.

We study a simple, basic version of constrained ETD(λ) and several variations of it, some of which are biased but can mitigate the variance issue better. To give an overview of our results, we shall refer to the first algorithm as the unbiased algorithm, and its biased variations as the biased variants. Two groups of results will be given to characterize the asymptotic behavior of the trajectory of iterates produced by these algorithms. The first group of results are derived by combining key properties of ETD(λ) with powerful convergence theorems from the weak convergence methods in stochastic approximation theory. The results show (roughly speaking) that:

(i) In the case of diminishing stepsize, under mild conditions, the trajectory of iterates produced by the unbiased algorithm eventually spends nearly all its time in an arbitrarily small neighborhood of the desired solution, with an arbitrarily high probability (Theorem 4); and the trajectory produced by the biased algorithms has a similar behavior, when the algorithmic parameters are set to make the biases sufficiently small (Theorem 6). These results entail the convergence in mean to the desired solution for the unbiased algorithm (Corollary 2), and the convergence in probability to some vicinity of the desired solution for the biased variants.

(ii) In the case of constant stepsize, imagine that we run the algorithms for all stepsizes; then conclusions similar to those in (i) hold in the limit as the stepsize parameter approaches zero (Theorems 5 and 7). In particular, a smaller stepsize parameter results in an increasingly longer segment of the trajectory to spend, with an increasing probability, nearly all its time in some neighborhood of the desired solution. The size of the neighborhood can be made arbitrarily small as the stepsize parameter approaches zero and, in the case of the biased variants, also as their biases are reduced.

The next group of results are for the constant-stepsize case and complement the results in (ii) by focusing on the asymptotic behavior of the algorithms for a fixed stepsize. Among others, they show (roughly speaking) that:

(iii) For any given stepsize parameter, asymptotically, the expected maximal deviation of multiple consecutive averaged iterates from the desired solution can be bounded in terms of the masses that the invariant probability measures of certain associated Markov chains assign to a small neighborhood of the desired solution. Those probability masses approach one when the stepsize parameter approaches zero and, in the case of the biased variants, also when their biases are sufficiently small (Theorems 8 and 9).

(iv) For a perturbed version of the unbiased algorithm and its biased variants, the maximal deviation of averaged iterates from the desired solution, under a given stepsize parameter, can be bounded almost surely in terms of those probability masses mentioned in (iii), for each initial condition (Theorems 10 and 11).

To derive the first group of results, we use powerful convergence theorems from the weak convergence methods in stochastic approximation theory (Kushner and Clark, 1978; Kushner and Shwartz, 1984; Kushner and Yin, 2003). This theory builds on the ordinary differential equation (ODE) based proof method, treats the trajectory of iterates as a whole, and studies its asymptotic behavior through the continuous-time processes corresponding to left-shifted and interpolated iterates. The probability distributions of these continuous-time interpolated processes are analyzed (as probability measures on a function space) by the weak convergence methods, leading to a characterization of their limiting distributions, from which asymptotic properties of the trajectory of iterates can be obtained.

Most of our efforts in the first part of our analysis are to prove that the constrained ETD(λ) algorithms satisfy the conditions required by the general convergence theorems just mentioned. We prove this by using key properties of ETD(λ) iterates, most importantly, the ergodicity and uniform integrability properties of the trace iterates, and the convergence of certain averaged processes which, intuitively speaking, describe the averaged dynamics of ETD(λ). Some of these properties were established earlier in our work (2015a) when analyzing the almost sure convergence of ETD(λ). Building upon that work, we prove the remaining properties needed in the analysis.

To derive the second group of results, we exploit the fact that in the case of constant stepsize, the iterates together with other random variables involved in the algorithms form weak Feller Markov chains, and such Markov chains have nice ergodicity properties. We use ergodic theorems for weak Feller Markov chains (Meyn, 1989; Meyn and Tweedie, 2009), together with the properties of ETD(λ) iterates and the convergence results we get from the weak convergence methods, in this second part of our analysis.

Besides ETD(λ), the analysis we give in the paper also applies to off-policy TD(λ), when the divergence issue mentioned earlier is avoided by setting λ sufficiently close to 1. The reason is that in that case the off-policy TD(λ) iterates have the same properties as the ones used in our analysis of ETD(λ) and therefore, the same conclusions hold for constrained versions of off-policy TD(λ), regarding their asymptotic convergence properties for constant or slowly diminishing stepsize (these results are new, to our knowledge). Similarly, our analysis also applies directly to the ETD(λ, β) algorithm, a variation of ETD(λ) recently proposed by Hallak et al. (2016).

Regarding practical performance of the algorithms, the biased ETD variant algorithms are much more robust than the unbiased algorithm despite the latter's superior asymptotic convergence properties. (This is not a surprise, for the biased algorithms are in fact defined

by using a well-known robustifying approach from stochastic approximation theory.) Their behavior is demonstrated by experiments in (Mahmood et al., 2015; Yu, 2016). In particular, the report (Yu, 2016) is our companion note for this paper and includes several simulation results to illustrate some of the theorems we give here regarding the behavior of multiple consecutive iterates of the biased algorithms.

The paper is organized as follows. In Section 2 we provide the background for the ETD(λ) algorithm. In Section 3 we present our convergence results on constrained ETD(λ) and several variants of it, and we give the proofs in Section 4. We conclude the paper in Section 5 with a brief discussion on direct applications of our convergence results to the off-policy TD(λ) algorithm and the ETD(λ, β) algorithm, as well as to ETD(λ) under relaxed conditions, followed by a discussion on several open issues. In Appendix A we include the key properties of the ETD(λ) trace iterates that are used in the analysis.

2. Preliminaries

In this section we describe the policy evaluation problem in the off-policy case, the ETD(λ) algorithm and its constrained version. We also review the results from our prior work (2015a) that are needed in this paper.

2.1 Off-policy Policy Evaluation

Let $\mathcal{S} = \{1, \ldots, N\}$ be a finite set of states, and let $\mathcal{A}$ be a finite set of actions. Without loss of generality we assume that for all states, every action in $\mathcal{A}$ can be applied. If $a \in \mathcal{A}$ is applied at state $s \in \mathcal{S}$, the system moves to state s' with probability $p(s' \mid s, a)$ and yields a random reward with mean $r(s, a, s')$ and bounded variance, according to a probability distribution $q(\cdot \mid s, a, s')$. These are the parameters of the MDP model we consider; they are unknown to the learning algorithms to be introduced.

A *stationary policy* is a time-invariant decision rule that specifies the probability of taking an action at each state. When actions are taken according to such a policy, the states and actions (S_t, A_t) at times $t \geq 0$ form a (time-homogeneous) Markov chain on the space $\mathcal{S} \times \mathcal{A}$, with the marginal state process $\{S_t\}$ being also a Markov chain.

Let π and π^o be two given stationary policies, with $\pi(a \mid s)$ and $\pi^o(a \mid s)$ denoting the probability of taking action a at state s under π and π^o, respectively. While the system evolves under the policy π^o, generating a stream of state transitions and rewards, we wish to use these observations to evaluate the performance of the policy π, with respect to a discounted reward criterion, the definition of which will be given shortly. Here π is the target policy and π^o the behavior policy. It is allowed that $\pi^o \neq \pi$ (the off-policy case), provided that at each state, all actions taken by π can also be taken by π^o (cf. Assumption 1(ii) below).

Let $\gamma(s) \in [0, 1]$, $s \in S$, be state-dependent discount factors, with $\gamma(s) < 1$ for at least one state. We measure the performance of π in terms of the expected discounted total rewards attained under π as follows: for each state $s \in \mathcal{S}$,

$$v_\pi(s) := \mathbb{E}^\pi \left[R_0 + \sum_{t=1}^{\infty} \gamma(S_1)\, \gamma(S_2) \cdots \gamma(S_t) \cdot R_t \,\middle|\, S_0 = s \right], \tag{1}$$

where R_t is the random reward received at time t, and $\mathbb{E}^\pi$ denotes expectation with respect to the probability distribution of the states, actions and rewards, (S_t, A_t, R_t), $t \geq 0$, generated under the policy π. The function v_π on $\mathcal{S}$ is called the *value function* of π. The special case of γ being a constant less than 1 corresponds to the γ-discounted reward criterion: $v_\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s\right]$. In the general case, by letting γ depend on the state, the formulation is able to also cover certain undiscounted total reward MDPs with termination;[4] however, for v_π to be well-defined (i.e., to have the right-hand side of Equation 1 well-defined for each state), a condition on the target policy is needed, which is stated below and will be assumed throughout the paper.

Let P_π denote the transition matrix of the Markov chain on $\mathcal{S}$ induced by π. Let Γ denote the $N \times N$ diagonal matrix with diagonal entries $\gamma(s), s \in \mathcal{S}$.

Assumption 1 (conditions on the target and behavior policies)

(i) *The target policy π is such that $(I - P_\pi \Gamma)^{-1}$ exists.*

(ii) *The behavior policy π^o induces an irreducible Markov chain on $\mathcal{S}$, and moreover, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\pi^o(a \mid s) > 0$ if $\pi(a \mid s) > 0$.*

Under Assumption 1(i), the value function v_π in (1) is well-defined, and furthermore, v_π satisfies uniquely the Bellman equation[5]

$$v_\pi = r_\pi + P_\pi \Gamma\, v_\pi, \qquad \text{i.e.,} \quad v_\pi = (I - P_\pi \Gamma)^{-1} r_\pi,$$

where r_π is the expected one-stage reward function under π (i.e., $r_\pi(s) = \mathbb{E}^\pi[R_0 \mid S_0 = s]$ for $s \in \mathcal{S}$).

2.2 The ETD(λ) Algorithm

Like the standard TD(λ) algorithm (Sutton, 1988; Tsitsiklis and Van Roy, 1997), the ETD(λ) algorithm (Sutton et al., 2016) approximates the value function v_π by a function of the form $v(s) = \phi(s)^\top \theta$, $s \in \mathcal{S}$, using a parameter vector $\theta \in \mathbb{R}^n$ and n-dimensional feature representations $\phi(s)$ for the states. (Here $\phi(s)$ is a column vector and $^\top$ stands for transpose.) In matrix notation, denote by Φ the $N \times n$ matrix with $\phi(s)^\top, s \in \mathcal{S}$, as its rows. Then the columns of Φ span the subspace of approximate value functions, and the approximation problem is to find in that subspace a function $v = \Phi\theta \approx v_\pi$.

We focus on a general form of the ETD(λ) algorithm, which uses state-dependent λ values specified by a function $\lambda : \mathcal{S} \to [0, 1]$. Inputs to the algorithm are the states, actions and rewards, $\{(S_t, A_t, R_t)\}$, generated under the behavior policy π^o, where R_t is the random reward received upon the transition from state S_t to S_{t+1} with action A_t. The algorithm can access the following functions, in addition to the features $\phi(s)$:

4. We may view $v_\pi(s)$ as the expected (undiscounted) total rewards attained under π starting from the state s and up to a random termination time $\tau \geq 1$ that depends on the states in a Markovian way. In particular, if at time $t \geq 1$, the state is s and termination has not occurred yet, the probability of $\tau = t$ (terminating at time t) is $1 - \gamma(s)$. Then $v_\pi(s)$ can be equivalently written as $v_\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\tau-1} R_t \mid S_0 = s\right]$.

5. One can verify this Bellman equation directly. It also follows from the standard MDP theory, as by definition v_π here can be related to a value function in a discounted MDP where the discount factors depend on state transitions, similar to discounted semi-Markov decision processes (see e.g., Puterman, 1994).

(i) the state-dependent discount factor $\gamma(s)$ that defines v_π, as described earlier;

(ii) $\lambda : \mathcal{S} \to [0,1]$, which determines the single or multi-step Bellman equation for the algorithm (cf. the subsequent Equations 6-7 and Footnote 7);

(iii) $\rho : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ given by $\rho(s,a) = \pi(a \mid s)/\pi^o(a \mid s)$ (with $0/0 = 0$), which gives the likelihood ratios for action probabilities that can be used to compensate for sampling states and actions according to the behavior policy π^o instead of the target policy π;

(iv) $i : \mathcal{S} \to \mathbb{R}_+$, which gives the algorithm additional flexibility to weigh states according to the degree of "interest" indicated by $i(s)$.

The algorithm also uses a sequence $\alpha_t > 0, t \geq 0$, as stepsize parameters. We shall consider only deterministic $\{\alpha_t\}$.

To simplify notation, let

$$\rho_t = \rho(S_t, A_t), \qquad \gamma_t = \gamma(S_t), \qquad \lambda_t = \lambda(S_t).$$

ETD(λ) calculates recursively $\theta_t \in \mathbb{R}^n$, $t \geq 0$, according to

$$\theta_{t+1} = \theta_t + \alpha_t \, e_t \cdot \rho_t \left(R_t + \gamma_{t+1} \phi(S_{t+1})^\top \theta_t - \phi(S_t)^\top \theta_t \right), \tag{2}$$

where $e_t \in \mathbb{R}^n$, called the "eligibility trace," is calculated together with two nonnegative scalar iterates (F_t, M_t) according to[6]

$$F_t = \gamma_t \, \rho_{t-1} \, F_{t-1} + i(S_t), \tag{3}$$

$$M_t = \lambda_t \, i(S_t) + (1 - \lambda_t) \, F_t, \tag{4}$$

$$e_t = \lambda_t \, \gamma_t \, \rho_{t-1} \, e_{t-1} + M_t \, \phi(S_t). \tag{5}$$

For $t = 0$, (e_0, F_0, θ_0) are given as an initial condition of the algorithm.

We recognize that the iteration (2) has the same form as TD(λ), but the trace e_t is calculated differently, involving an "emphasis" weight M_t on the state S_t, which itself evolves along with the iterate F_t, called the "follow-on" trace. If M_t is always set to 1 regardless of F_t and $i(\cdot)$, then the iteration (2) reduces to the off-policy TD(λ) algorithm in the case where γ and λ are constants.

2.3 Associated Bellman Equations and Approximation and Convergence Properties of ETD(λ)

Let Λ denote the diagonal matrix with diagonal entries $\lambda(s), s \in \mathcal{S}$. Associated with ETD($\lambda$) is a generalized multistep Bellman equation of which v_π is the unique solution (Sutton, 1995):[7]

$$v = r^\lambda_{\pi,\gamma} + P^\lambda_{\pi,\gamma} \, v. \tag{6}$$

6. The definition (5) we use here differs slightly from the original definition of e_t used by Sutton et al. (2016), but the two are equivalent and (5) appears to be more convenient for our analysis.

7. For the details of this Bellman equation, we refer the readers to the early work (Sutton, 1995; Sutton and Barto, 1998) and the recent work (Sutton et al., 2016). We remark that similar to the standard one-step Bellman equation, which is a recursive relation that expresses v_π in terms of the expected one-stage reward and the expected total future rewards given by v_π itself, one can use the strong Markov property to derive other recursive relations satisfied by v_π, in which the expected one-stage reward is replaced by the expected rewards attained by π up to some random stopping time. This gives rise to a general class of Bellman equations, of which (6) is one example. Earlier works on using such equations in TD

Here $P^{\lambda}_{\pi,\gamma}$ is an $N \times N$ substochastic matrix, $r^{\lambda}_{\pi,\gamma} \in \mathbb{R}^N$ is a vector of expected discounted total rewards attained by π up to some random time depending on the function λ, and they can be expressed in terms of P_π and r_π as

$$P^{\lambda}_{\pi,\gamma} = I - (I - P_\pi \Gamma \Lambda)^{-1} (I - P_\pi \Gamma), \qquad r^{\lambda}_{\pi,\gamma} = (I - P_\pi \Gamma \Lambda)^{-1} r_\pi. \tag{7}$$

ETD(λ) aims to solve a projected version of the Bellman equation (6) (Sutton et al., 2016), which takes the following forms in the space of approximate value functions and in the space of the θ-parameters, respectively:

$$v = \Pi\big(r^{\lambda}_{\pi,\gamma} + P^{\lambda}_{\pi,\gamma}\, v\big), \quad v \in \text{column-space}(\Phi), \qquad \Longleftrightarrow \qquad C\theta + b = 0, \quad \theta \in \mathbb{R}^n. \tag{8}$$

Here Π is a projection onto the approximation subspace with respect to a weighted Euclidean norm or seminorm, under a condition on the approximation architecture that will be explained shortly. The weights that define this norm also define the diagonal entries $\bar{M}_{ss}, s \in \mathcal{S}$, of a diagonal matrix $\bar{M}$, which are given by

$$diag(\bar{M}) = d^{\top}_{\pi^o,i}(I - P^{\lambda}_{\pi,\gamma})^{-1}, \qquad \text{with} \quad d_{\pi^o,i} \in \mathbb{R}^N,\, d_{\pi^o,i}(s) = d_{\pi^o}(s) \cdot i(s),\ s \in \mathcal{S}, \tag{9}$$

where $d_{\pi^o}(s) > 0$ denotes the steady state probability of state s for the behavior policy π^o, under Assumption 1(ii). For the corresponding linear equation in the θ-space in (8),

$$C = -\Phi^{\top} \bar{M} \,(I - P^{\lambda}_{\pi,\gamma})\, \Phi, \qquad b = \Phi^{\top} \bar{M}\, r^{\lambda}_{\pi,\gamma}. \tag{10}$$

From the expression (9) of the diagonal matrix $\bar{M}$, the most important difference between the earlier TD algorithms and ETD(λ) can be seen. For on-policy TD(λ), in stead of (9), the diagonal matrix $\bar{M}$ is determined by the steady state probabilities of the states under the target policy π under an ergodicity assumption (Tsitsiklis and Van Roy, 1997), and for off-policy TD(λ), it is determined by the steady state probabilities $d_{\pi^o}(s)$ under the behavior policy π^o. Here, due to the emphatic weighting scheme (3)-(5), the diagonals of $\bar{M}$ given by (9) reflect the occupation frequencies (with respect to $P^{\lambda}_{\pi,\gamma}$) of the target policy rather than the behavior policy.

Let $|\cdot|$ denote the (unweighted) Euclidean norm. The matrix C is said to be *negative definite* if there exists $c > 0$ such that $\theta^{\top} C \theta \le -c|\theta|^2$ for all $\theta \in \mathbb{R}^n$; and *negative semidefinite* if in the preceding inequality $c = 0$. A salient property of ETD(λ) is that the matrix C is always negative semidefinite (Sutton et al., 2016), and under natural and mild conditions, C is negative definite. This is proved in our work (2015a) and summarized below.

Call those states s with $\bar{M}_{ss} > 0$ *emphasized states* (define this set of states to be empty if $\bar{M}$ given by Equation 9 is ill-defined, a case we will not encounter).

Assumption 2 (condition on the approximation architecture)
The set of feature vectors of emphasized states, $\{\phi(s) \mid s \in \mathcal{S}, \bar{M}_{ss} > 0\}$, contains n linearly independent vectors.

learning include the paper (Sutton, 1995) and Chap. 5.3 of the book (Bertsekas and Tsitsiklis, 1996). Recently, Ueno et al. (2011) considered an even broader class of Bellman equations using the concept of estimating equations from statistics, and Yu and Bertsekas (2012) focused on a special class of generalized Bellman equations and discussed their potential advantages from an approximation viewpoint. But an in-depth study of the application of such equations is still lacking currently. Because generalized Bellman equations offer flexible ways to address the bias vs. variance problem in learning the value functions of a policy, they are especially important and deserve further study, in our opinion.

Theorem 1 (Yu, 2015a, Prop. C.2) *Under Assumption 1, the matrix C is negative definite if and only if Assumption 2 holds.*

Assumption 2, which implies the linear independence of the columns of Φ, is satisfied in particular if the set of feature vectors, $\{\phi(s) \mid s \in \mathcal{S}, i(s) > 0\}$, contains n linearly independent vectors, since states with positive interest $i(s)$ are among the emphasized states.[8] So this assumption can be easily satisfied in reinforcement learning without model knowledge.[9]

In view of Theorem 1, under Assumptions 1-2, the equation $C\theta + b = 0$ has a unique solution θ^*; equivalently, $\Phi\theta^*$ is the unique solution to the projected Bellman equation (7):

$$\Phi\theta^* = \Pi\big(r^\lambda_{\pi,\gamma} + P^\lambda_{\pi,\gamma}\,\Phi\theta^*\big),$$

where Π is a well-defined projection operator that projects a vector in $\mathbb{R}^N$ onto the approximation subspace with respect to the seminorm on $\mathbb{R}^N$ given by

$$\sqrt{\textstyle\sum_{s\in\mathcal{S}} \bar{M}_{ss}\cdot v(s)^2}, \quad \forall\, v \in \mathbb{R}^N$$

(which is a norm if $\bar{M}_{ss} > 0$ for all $s \in \mathcal{S}$). The relation between the approximate value function $v = \Phi\theta^*$ and the desired value function v_π, in particular, the approximation error, can be characterized by using the oblique projection viewpoint (Scherrer, 2010) for projected Bellman equations.[10]

The almost sure convergence of ETD(λ) to θ^* is proved in (Yu, 2015a, Theorem 2.2) under Assumptions 1 and 2, for diminishing stepsize satisfying $\alpha_t = O(1/t)$ and $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$. Despite this convergence guarantee, the stepsize range is too narrow for applications, as we discussed in the introduction. In this paper we will focus on constrained ETD(λ) algorithms that restrict the θ-iterates in a bounded set, but can operate with much larger stepsizes and also suffer less from the issue of high variance in off-policy learning. We will analyze their behavior under Assumptions 1 and 2, although our analysis extends to the case without Assumption 2 (see the discussion in Section 5.1).

8. This follows from the definition (9) of the diagonals $\bar{M}_{ss}$. Since $(I - P^\lambda_{\pi,\gamma})^{-1} = I + \sum_{k=1}^\infty (P^\lambda_{\pi,\gamma})^k \geq I$, we have $diag(\bar{M}) = d^\top_{\pi^o,i}(I - P^\lambda_{\pi,\gamma})^{-1} \geq d^\top_{\pi^o,i}$. Hence $i(s) > 0$ implies $\bar{M}_{ss} \geq d_{\pi^o}(s)\cdot i(s) > 0$.

9. There is another way to verify Assumption 2 without calculating $\bar{M}$. Suppose ETD(λ) starts from a state S_0 with $i(S_0) > 0$. Then it can be shown that if $S_t = s$ and $M_t > 0$, we must have $\bar{M}_{ss} > 0$. This means that as soon as we find among states S_t with emphasis weights $M_t > 0$ n states that have linearly independent feature vectors, we can be sure that Assumption 2 is satisfied.

10. Briefly speaking, Scherrer (2010) showed that the solutions of projected Bellman equations are oblique projections of v_π on the approximation subspace. An oblique projection is defined by two nonorthogonal subspaces of equal dimensions and is the projection onto the first subspace orthogonally to the second (Saad, 2003). In the special case of ETD(λ), the first of these two subspaces is the approximation subspace $\{v \in \mathbb{R}^N \mid v = \Phi\theta \text{ for some } \theta \in \mathbb{R}^n\}$, and the second is the image of the approximation subspace under the linear transformation $(I - P^\lambda_{\pi,\gamma})^\top \bar{M}$. Essentially it is the angle between the two subspaces that determines the approximation bias $\Phi\theta^* - \Pi v_\pi$ in the worst case, for a worst-case choice of $r^\lambda_{\pi,\gamma}$. (For details, see also Yu and Bertsekas 2012, Sec. 2.2.) Recently, for the case of constant λ, i and γ, Hallak et al. (2016) derived bounds on the approximation bias that are based on contraction arguments and are comparable to the bound for on-policy TD(λ) (Tsitsiklis and Van Roy, 1997). These bounds lie above the bounds given by the oblique projection view (cf. Yu and Bertsekas, 2010; Yu and Bertsekas, 2012, Sec. 2.2); however, they are expressed in terms of λ and γ, so they give us explicit numbers instead of analytical expressions to bound the approximation bias.

2.4 Constrained ETD(λ), Averaged Processes and Mean ODE

We consider first a constrained version of ETD(λ) that simply scales the θ-iterates, if necessary, to keep them bounded:

$$\theta_{t+1} = \Pi_B\Big(\theta_t + \alpha_t\, e_t \cdot \rho_t\big(R_t + \gamma_{t+1}\phi(S_{t+1})^\top\theta_t - \phi(S_t)^\top\theta_t\big)\Big), \tag{11}$$

where Π_B is the Euclidean projection onto a closed ball $B \subset \mathbb{R}^n$ at the origin with radius r_B: $B = \{\theta \in \mathbb{R}^n \mid |\theta| \le r_B\}$. Under Assumptions 1 and 2, when the radius r_B is sufficiently large (greater than the threshold given in Lemma 1 below), from any given initial (e_0, F_0, θ_0), the algorithm (11) converges almost surely to θ^*, for diminishing stepsize $\alpha_t = O(1/t)$ (Yu, 2015a, Theorem 4.1).

Our interest in this paper is to apply (11) with a much larger range of stepsize, in particular, constant stepsize or stepsize that diminishes much more slowly than $O(1/t)$. In Sections 3 and 4, we will analyze the algorithm (11) and its two variants for such stepsizes. To prepare for the analysis, in the rest of this section, we review several results from our prior work (2015a) that will be needed.

First, we discuss about the "mean ODE" that we wish to associate with (11). It is the projected ODE

$$\dot{x} = \bar{h}(x) + z, \qquad z \in -\mathcal{N}_B(x), \tag{12}$$

where the function $\bar{h}$ is the left-hand side of the equation $Cx + b = 0$ we want to solve:

$$\bar{h}(x) = Cx + b; \tag{13}$$

$\mathcal{N}_B(x)$ is the normal cone of B at x (i.e., $\mathcal{N}_B(x) = \{0\}$ for x in the interior of B and $\mathcal{N}_B(x) = \{ax \mid a \ge 0\}$ for x on the boundary of B); and z is the boundary reflection term that cancels out the component of $\bar{h}(x)$ in $\mathcal{N}_B(x)$ (i.e., $z = -y$ where y is the projection of $\bar{h}(x)$ on $\mathcal{N}_B(x)$), and it is the "minimal force" needed to keep the solution $x(\cdot)$ of (12) in B (Kushner and Yin, 2003, Chap. 4.3).

The negative definiteness of the matrix C ensures that when the radius of B is sufficiently large, the boundary reflection term is zero for all $x \in B$ and the projected ODE (12) has no stationary points other than θ^* (see Yu 2015a, Sec. 4.1 for a simple proof):

Lemma 1 *Let $c > 0$ be such that $x^\top Cx \le -c|x|^2$ for all $x \in \mathbb{R}^n$. Suppose B has a radius $r_B > |b|/c$. Then θ^* lies in the interior of B; a solution $x(\tau), \tau \in [0, \infty)$, to the projected ODE (12) for an initial condition $x(0) \in B$ coincides with the unique solution to $\dot{x} = \bar{h}(x)$, with the boundary reflection term being $z(\cdot) \equiv 0$; and the only solution $x(\tau), \tau \in (-\infty, +\infty)$, of (12) in B is $x(\cdot) \equiv \theta^*$.*

Informally speaking, suppose we have proved that (12) is the mean ODE for the algorithm (11) under stepsizes of our interest. Then applying powerful convergence theorems from stochastic approximation theory (Kushner and Yin, 2003), we can assert that the iterates θ_t will eventually "follow closely" a solution of the mean ODE. This together with the solution property of the mean ODE given in Lemma 1 will then give us a characterization of the asymptotic behavior of the algorithm (11) for a constraint set B with sufficiently large radius.

Several properties of the ETD(λ) iterates will be important in proving that (12) is indeed the mean ODE for (11) and reflects its average dynamics. We now discuss two such properties (other key properties will be given in Appendix A). They concern the ergodicity of the Markov chain $\{(S_t, A_t, e_t, F_t)\}$ on the joint space of states, actions and traces, and the convergence of certain averaged sequences associated with the algorithm (11). They will also be useful in analyzing variants of (11).

Let $Z_t = (S_t, A_t, e_t, F_t)$, $t \geq 0$. It was shown in (Yu, 2015a) that under Assumption 1, $\{Z_t\}$ is a weak Feller Markov chain[11] on the infinite state space $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$ and is ergodic. Specifically, on a metric space, a sequence of probability measures $\{\mu_t\}$ is said to *converge weakly* to a probability measure μ if for any bounded continuous function f, $\int f d\mu_t \to \int f d\mu$ as $t \to \infty$ (Dudley, 2002, Chap. 9.3). We are interested in the weak convergence of the occupation probability measures of the process $\{Z_t\}$, where for each initial condition $Z_0 = z$, the *occupation probability measures* $\mu_{z,t}$, $t \geq 0$, are defined by $\mu_{z,t}(D) = \frac{1}{t+1} \sum_{k=0}^{t} \mathbb{1}(Z_k \in D)$ for any Borel subset D of $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$, with $\mathbb{1}(\cdot)$ denoting the indicator function.

Theorem 2 (ergodicity of $\{Z_t\}$; Yu, 2015a, Theorem 3.2) *Under Assumption 1, the Markov chain $\{Z_t\}$ has a unique invariant probability measure ζ, and for each initial condition $Z_0 = z$, the sequence $\{\mu_{z,t}\}$ of occupation probability measures converges weakly to ζ, almost surely.*

Let $\mathbb{E}_\zeta$ denote expectation with respect to the stationary process $\{Z_t\}$ with ζ as its initial distribution. By the definition of weak convergence, the weak convergence of $\{\mu_{z,t}\}$ given in Theorem 2 implies that for each given initial condition of Z_0, the averages $\frac{1}{t} \sum_{k=0}^{t-1} f(Z_k)$ converge almost surely to $\mathbb{E}_\zeta\{f(Z_0)\}$ for any bounded continuous function f.[12] To study the average dynamics of the algorithm (11), however, we need to also consider unbounded functions. In particular, the function related to both (11) and the unconstrained ETD(λ) is $h : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$,

$$h(\theta, \xi) = e \cdot \rho(s, a) \big(r(s, a, s') + \gamma(s')\, \phi(s')^\top \theta - \phi(s)^\top \theta\big), \tag{14}$$

where

$$\xi = (e, F, s, a, s') \in \Xi := \mathbb{R}^{n+1} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

Writing ξ_t for the traces and transition at time t: $\xi_t = (e_t, F_t, S_t, A_t, S_{t+1})$, we can express the recursion (11) equivalently as

$$\theta_{t+1} = \Pi_B\big(\theta_t + \alpha_t\, h(\theta_t, \xi_t) + \alpha_t\, e_t \cdot \tilde{\omega}_{t+1}\big), \tag{15}$$

where $\tilde{\omega}_{t+1} = \rho_t\,(R_t - r(S_t, A_t, S_{t+1}))$ is the noise part of the observed reward.

The convergence to $\bar{h}(\theta)$ of the averaged sequence $\frac{1}{t} \sum_{k=0}^{t-1} h(\theta, \xi_k)$, with θ held fixed and t going to infinity, will be needed to prove that (12) is the mean ODE of (11). Since

11. See Section 4.3.1 or the book by Meyn and Tweedie (2009, Chap. 6) for the definition and properties of weak Feller Markov chains.

12. With the usual discrete topology for the finite space $\mathcal{S} \times \mathcal{A}$ and the usual topology for the Euclidean space $\mathbb{R}^{n+1}$, the space $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$ equipped with the product topology is metrizable. A continuous function $f(s, a, e, F)$ on this space is a function that is continuous in (e, F) for each $(s, a) \in \mathcal{S} \times \mathcal{A}$.

$\bar{h}(\theta) = C\theta + b$, this convergence for each fixed θ can be identified with the convergence of the matrix and vector iterates calculated by ELSTD(λ)—the least-squares version of ETD(λ)—to approximate the left-hand side of the equation $C\theta + b = 0$. It was proved in our work (2015a) as a special case of the convergence of averaged sequences for a larger set of functions including $h(\theta, \cdot)$. Since this general result will be needed in analyzing variants of (11), we give its formulation here.

Throughout the rest of the paper, we let $\|\cdot\|$ denote the infinity norm of a Euclidean space, and we use this notation for both vectors and matrices (viewed as vectors). For $\mathbb{R}^m$-valued random variables X_t, we say $\{X_t\}$ converges to a random variable X in mean if $\mathbb{E}[\|X_t - X\|] \to 0$ as $t \to \infty$.

Consider a vector-valued function $g : \Xi \to \mathbb{R}^m$ such that with $\xi = (e, F, s, a, s')$, $g(\xi)$ is Lipschitz continuous in (e, F) uniformly in (s, a, s'). That is, there exists a finite constant L_g such that for any $(e, F), (\hat{e}, \hat{F}) \in \mathbb{R}^{n+1}$,

$$\big\|g(e, F, s, a, s') - g(\hat{e}, \hat{F}, s, a, s')\big\| \leq L_g \big\|(e, F) - (\hat{e}, \hat{F})\big\|, \quad \forall\, (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}. \quad (16)$$

For each $\theta \in \mathbb{R}^n$, the function $h(\theta, \cdot)$ in (14) is a special case of g. The convergence of the averaged sequence $\frac{1}{t}\sum_{k=0}^{t-1} g(\xi_k)$ is given in the theorem below; the part on convergence in mean will be used frequently later in this paper. The convergence of $\frac{1}{t}\sum_{k=0}^{t-1} h(\theta, \xi_k)$ then follows as a special case.

Theorem 3 (convergence of averaged sequences; Yu, 2015a, Theorems 3.1-3.3) *Let g be a vector-valued function satisfying the Lipschitz condition (16). Then under Assumption 1, $\mathbb{E}_\zeta\big[\|g(\xi_0)\|\big] < \infty$ and for any given initial $(e_0, F_0) \in \mathbb{R}^{n+1}$, as $t \to \infty$, $\frac{1}{t}\sum_{k=0}^{t-1} g(\xi_k)$ converges to $\bar{g} = \mathbb{E}_\zeta\big[g(\xi_0)\big]$ in mean and almost surely.*

Corollary 1 (Yu, 2015a, Theorem 2.1) *Under Assumption 1, for the functions $\bar{h}, h$ given in (13), (14) respectively, the following hold: For each $\theta \in \mathbb{R}^n$, $\mathbb{E}_\zeta\big[\|h(\theta, \xi_0)\|\big] < \infty$ and $\bar{h}(\theta) = \mathbb{E}_\zeta\big[h(\theta, \xi_0)\big]$; and for any given initial $(e_0, F_0) \in \mathbb{R}^{n+1}$, as $t \to \infty$, $\frac{1}{t}\sum_{k=0}^{t-1} h(\theta, \xi_k)$ converges to $\bar{h}(\theta)$ in mean and almost surely.*

3. Convergence Results for Constrained ETD(λ)

In this section we present the convergence properties of the constrained ETD(λ) algorithm (11) and several variants of it, for constant stepsize and for stepsize that diminishes slowly. We will explain briefly how the results are obtained, leaving the detailed analyses to Section 4. The first set of results about the algorithm (11) will be given first in Section 3.1, followed by similar results in Section 3.2 for two variant algorithms that have biases but can mitigate the variance issue in off-policy learning better. These results are obtained through applying two general convergence theorems from (Kushner and Yin, 2003), which concern weak convergence of stochastic approximation algorithms for diminishing and constant stepsize. Finally, the constant-stepsize case will be analyzed further in Section 3.3, in order to refine some results of Sections 3.1-3.2 so that the asymptotic behavior of the algorithms for a fixed stepsize can be characterized explicitly. In that subsection, besides the three algorithms just mentioned, we will also discuss another variant algorithm with perturbation.

Regarding notation, recall that $\mathbb{1}(\cdot)$ is the indicator function, $|\cdot|$ stands for the usual (unweighted) Euclidean norm and $\|\cdot\|$ the infinity norm for $\mathbb{R}^m$. We denote by $N_\delta(D)$ the δ-neighborhood of a set $D \subset \mathbb{R}^m$: $N_\delta(D) = \{x \in \mathbb{R}^m \mid \inf_{y\in D} |x-y| \le \delta\}$, and we write $N_\delta(\theta^*)$ for the δ-neighborhood of θ^*. For the iteration index t, the notation $t \in [k_1, k_2]$ or $t \in [k_1, k_2)$ will be used to mean that the range of t is the set of integers in the interval $[k_1, k_2]$ or $[k_1, k_2)$. More definitions and notation will be introduced later where they are needed.

3.1 Main Results

We consider first the algorithm (11) for diminishing stepsize. Let the stepsize change slowly in the following sense.

Assumption 3 (condition on diminishing stepsize) *The (deterministic) nonnegative sequence $\{\alpha_t\}$ satisfies that $\sum_{t\ge 0}\alpha_t = \infty$, $\alpha_t \to 0$ as $t \to \infty$, and for some sequence of integers $m_t \to \infty$,*

$$\lim_{t\to\infty} \sup_{0\le j\le m_t} \left| \frac{\alpha_{t+j}}{\alpha_t} - 1 \right| = 0. \tag{17}$$

The condition (17) is the condition A.8.2.8 in (Kushner and Yin, 2003, Chap. 8) and allows stepsizes much larger than $O(1/t)$. We can have $\alpha_t = O(t^{-\beta})$, $\beta \in (0, 1]$, and even larger stepsizes are possible. For example, partition the time interval $[0, \infty)$ into increasingly longer intervals $I_k, k \ge 0$, and set α_t to be constant within each interval I_k. Then the condition (17) can be fulfilled by letting the constants for each I_k decrease as $O(k^{-\beta})$, $\beta \in (0, 1]$.

We now state the convergence result. For any $T > 0$, let $m(k, T) = \min\{t \ge k \mid \sum_{j=k}^{t+1} \alpha_j > T\}$. If we draw a continuous timeline and put each iteration of the algorithm at a specific moment, with the stepsize α_j being the length of time between iterations j and $j+1$, then $m(k, T)$ is the latest iteration before time T has elapsed since the k-th iteration. If $\alpha_t = O(t^{-\beta})$, $\beta \in (0, 1]$, for example, then for fixed T, there are $O(k^\beta)$ iterates between the k-th and $m(k, T)$-th iteration.

Recall that Assumption 1, Assumption 2, and Lemma 1 are given in Sections 2.1, 2.3, and 2.4, respectively.

Theorem 4 (convergence of constrained ETD with diminishing stepsize) *Suppose Assumptions 1-2 hold and the radius of B exceeds the threshold given in Lemma 1. Let $\{\theta_t\}$ be generated by the algorithm (11) with stepsize $\{\alpha_t\}$ satisfying Assumption 3, from any given initial condition (e_0, F_0). Then there exists a sequence $T_k \to \infty$ such that for any $\delta > 0$,*

$$\limsup_{k\to\infty} \mathbf{P}\Big(\theta_t \notin N_\delta(\theta^*), \text{ some } t \in \big[\, k,\, m(k, T_k) \,\big] \Big) = 0.$$

This theorem implies $\theta_t \to \theta^*$ in probability. Since $\{\theta_t\}$ is bounded, by (Dudley, 2002, Theorem 10.3.6), θ_t must also converge to θ^* in mean:

Corollary 2 (convergence in mean) *In the setting of Theorem 4, $\mathbb{E}\big[\|\theta_t - \theta^*\|\big] \to 0$ as $t \to \infty$.*

Another important note is that the conclusion of Theorem 4 is much stronger than that $\theta_t \to \theta^*$ in probability. Here as $k \to \infty$, we consider an increasingly longer segment $[k, m(k, T_k)]$ of iterates, and are able to conclude that the probability of that *entire segment* being inside an arbitrarily small neighborhood of θ^* approaches 1. This is the power of the weak convergence methods (Kushner and Clark, 1978; Kushner and Shwartz, 1984; Kushner and Yin, 2003), by which our conclusion is obtained.

In the case of constant stepsize, we consider all the trajectories that can be produced by the algorithm (11) using some constant stepsize, and we ask what the properties of these trajectories are in the limit as the stepsize parameter approaches 0. Here there is a common timeline used in relating trajectories generated with different stepsizes (and it comes from the ODE-based analysis): we imagine again a continuous timeline, along which we put the iterations at moments that are evenly separated in time by α, if the stepsize parameter is α. The scalars T, T_α in the theorem below represent amounts of time with respect to this continuous timeline.

Theorem 5 (convergence of constrained ETD with constant stepsize)
Suppose Assumptions 1-2 hold and the radius of B exceeds the threshold given in Lemma 1. For each $\alpha > 0$, let $\{\theta_t^\alpha\}$ be generated by the algorithm (11) with constant stepsize α, from any given initial condition (e_0, F_0). Let $\{k_\alpha \mid \alpha > 0\}$ be any sequence of nonnegative integers that are nondecreasing as $\alpha \to 0$. Then the following hold:

(i) *For any $\delta > 0$,*

$$\lim_{T\to\infty} \lim_{\alpha\to 0} \frac{1}{T/\alpha} \sum_{t=k_\alpha}^{k_\alpha + \lfloor T/\alpha \rfloor} \mathbb{1}\big(\theta_t^\alpha \in N_\delta(\theta^*)\big) = 1 \quad \textit{in probability.}$$

(ii) *Let $\alpha k_\alpha \to \infty$ as $\alpha \to 0$. Then there exists a sequence $\{T_\alpha \mid \alpha > 0\}$ with $T_\alpha \to \infty$ as $\alpha \to 0$, such that for any $\delta > 0$,*

$$\limsup_{\alpha\to 0} \mathbf{P}\Big(\theta_t^\alpha \notin N_\delta(\theta^*), \textit{ some } t \in \big[\, k_\alpha,\, k_\alpha + T_\alpha/\alpha \,\big] \Big) = 0.$$

Part (ii) above is similar to Theorem 4. Here as $\alpha \to 0$, an increasingly longer segment $[k_\alpha, k_\alpha + T_\alpha/\alpha]$ of the tail of the trajectory $\{\theta_t^\alpha\}$ is considered, and it is concluded that the probability of that *entire segment* being inside an arbitrarily small neighborhood of θ^* approaches 1. Part (i) above, roughly speaking, says that as α diminishes, within the segment $[k_\alpha, k_\alpha + T/\alpha]$, the fraction of iterates θ_t^α that lie in a small δ-neighborhood of θ^* approaches 1 for sufficiently large T.

We give the proofs of Theorems 4-5 in Section 4.1. As mentioned earlier, most of our efforts will be to use the properties of ETD iterates to show that the conditions of two general convergence theorems from stochastic approximation theory (Kushner and Yin, 2003, Theorems 8.2.2, 8.2.3) are satisfied by the algorithm (11). After that we can specialize the conclusions of those theorems to obtain Theorems 4-5. Specifically, after furnishing their conditions, applying (Kushner and Yin, 2003, Theorems 8.2.2, 8.2.3) will give us directly the desired conclusions in Theorems 4-5 with $N_\delta(L_B)$ in place of $N_\delta(\theta^*)$, where $N_\delta(L_B)$ is the δ-neighborhood of the *limit set* L_B for the projected ODE (12). This limit set is defined as follows:

$$L_B := \cap_{\bar\tau > 0} \overline{\cup_{x(0)\in B} \{x(\tau),\, \tau \geq \bar\tau\}}$$

where $x(\tau)$ is a solution of the projected ODE (12) with initial condition $x(0)$, the union is over all the solutions with initial $x(0) \in B$, and $\overline{D}$ for a set D denotes taking the closure of D. It can be shown that $L_B = \{\theta^*\}$ under our assumptions, so Theorems 4-5 will then follow as special cases of (Kushner and Yin, 2003, Theorems 8.2.2, 8.2.3).

Remark 1 (on weak convergence methods) The theorems from the book (Kushner and Yin, 2003) which we will apply are based on the weak convergence methods. While it is beyond the scope of this paper to explain these powerful methods, let us mention here a few basic facts about them to elucidate the origin of the convergence theorems we gave above. In the framework of (Kushner and Yin, 2003), one studies a trajectory of iterates produced by an algorithm by working with continuous-time processes that are piecewise constant or linear interpolations of the iterates. (Often one also left-shifts a trajectory of iterates to bring the "asymptotic part" of the trajectory closer to the origin of the continuous time axis.) In the case of our problem, for example, for diminishing stepsize, these continuous-time processes are $x^k(\tau), \tau \in [0, \infty)$, indexed by $k \geq 0$, where for each k, x^k is a piecewise constant interpolation of $\theta_{k+t}, t \geq 0$, given by $x^k(\tau) = \theta_k$ for $\tau \in [0, \alpha_k)$ and $x^k(\tau) = \theta_{k+t}$ for $\tau \in [\sum_{m=0}^{t-1} \alpha_{k+m}, \sum_{m=0}^{t} \alpha_{k+m})$, $t \geq 1$. Similarly, for constant stepsize, the continuous-time processes involved are $x^\alpha(\tau), \tau \in [0, \infty)$, indexed by $\alpha > 0$, and for each α, x^α is a piecewise constant interpolation of $\theta^\alpha_{k_\alpha+t}, t \geq 0$, given by $x^\alpha(\tau) = \theta_{k_\alpha+t}$ for $\tau \in [t\alpha, (t+1)\alpha)$. The behavior of the sequence $\{x^k\}$ or $\{x^\alpha\}$ as $k \to \infty$ or $\alpha \to 0$, tells us the asymptotic properties of the algorithm as the number of iterations grows to infinity or as the stepsize parameter approaches 0. With the weak convergence methods, one considers the probability distributions of the continuous-time processes in such sequences, and analyze the convergence of these probability distributions and their limiting distributions along any subsequences. Here each continuous-time process takes values in a space of vector-valued functions on $[0, \infty)$ or $(-\infty, \infty)$ that are right-continuous and have left-hand limits, and this function space equipped with an appropriate metric, known as the Skorohod metric, is a complete separable metric space (Kushner and Yin, 2003, p. 238-240). On this space, one analyzes the weak convergence of the probability distributions of the continuous-time processes. Under certain conditions on the algorithm, the general conclusions from (Kushner and Yin, 2003, Theorems 8.2.2, 8.2.3) are that any subsequence of these probability distributions contains a further subsequence which is convergent, and that all the limiting probability distributions must assign the full measure 1 to the set of solutions of the mean ODE associated with the algorithm. This general weak convergence property then yields various conclusions about the asymptotic behavior of the algorithm and its relation with the mean ODE solutions. When further combined with the solution properties of the mean ODE, it leads to specific results such as the theorems we give in this section. ■

3.2 Two Variants of Constrained ETD(λ) with Biases

We now consider two simple variants of (11). They constrain the ETD iterates even more, at a price of introducing biases in this process, so that unlike (11), they can no longer get to θ^* arbitrarily closely. Instead they aim at a small neighborhood of θ^*, the size of which depends on how they modify the ETD iterates. On the other hand, because the trace iterates $\{(e_t, F_t)\}$ can have unbounded variances and are also naturally unbounded in

common off-policy situations (see discussions in Yu, 2012, Prop. 3.1 and Footnote 3, p. 3320-3322 and Yu, 2015a, Remark A.1, p. 23), these variant algorithms have the advantage that they make the θ-iterates more robust against the drastic changes that can occur to the trace iterates. Indeed our definition of the variant algorithms below follows a well-known approach to "robustifying" algorithms in stochastic approximation theory (see discussions in Kushner and Yin, 2003, p. 23 and p. 141).

The two variant algorithms are defined as follows. For each $K > 0$, let $\psi_K : \mathbb{R}^n \to \mathbb{R}^n$ be a bounded Lipschitz continuous function such that

$$\|\psi_K(x)\| \le \|x\| \quad \forall\, x \in \mathbb{R}^n, \quad \text{and} \quad \psi_K(x) = x \ \text{ if } \|x\| \le K. \tag{18}$$

(For instance, let $\psi_K(x) = \bar{r}x/|x|$ if $|x| \ge \bar{r}$ and $\psi_K(x) = x$ otherwise, for $\bar{r} = \sqrt{n}K$; or let $\psi_K(x)$ be the result of truncating each component of x to be within $[-K, K]$.) For the first variant of the algorithm (11), we replace e_t in (11) by $\psi_K(e_t)$:

$$\theta_{t+1} = \Pi_B\Big(\theta_t + \alpha_t\, \psi_K(e_t) \cdot \rho_t\big(R_t + \gamma_{t+1}\phi(S_{t+1})^\top\theta_t - \phi(S_t)^\top\theta_t\big)\Big). \tag{19}$$

For the second variant, we apply ψ_K to bound the entire increment in (11) before it is multiplied by the stepsize α_t and added to θ_t:

$$\theta_{t+1} = \Pi_B\left(\theta_t + \alpha_t\, \psi_K(Y_t)\right), \quad \text{where } \ Y_t = e_t \cdot \rho_t\big(R_t + \gamma_{t+1}\phi(S_{t+1})^\top\theta_t - \phi(S_t)^\top\theta_t\big). \tag{20}$$

As will be proved later, these two algorithms are associated with mean ODEs of the form,

$$\dot{x} = \bar{h}_K(x) + z, \qquad z \in -\mathcal{N}_B(x), \tag{21}$$

where $\bar{h}_K : \mathbb{R}^n \to \mathbb{R}^n$ is determined by each algorithm and deviates from the function $\bar{h}(x) = Cx + b$ due to the alterations introduced by ψ_K. This ODE is similar to the projected ODE (12), except that since $\bar{h}_K$ is an approximation of $\bar{h}$, θ^* is no longer a stable or stationary point for the mean ODE (21). The two variant algorithms thus have a bias in their θ-iterates, and the bias can be made smaller by choosing a larger K. This is reflected in the two convergence theorems given below. They are similar to the previous two theorems for the algorithm (11), except that now given a desired small neighborhood of θ^*, a sufficiently large K needs to be used in order for the θ-iterates to reach that neighborhood of θ^* and exhibit properties similar to those shown in the previous case.

Theorem 6 (constrained ETD variants with diminishing stepsize)
In the setting of Theorem 4, let $\{\theta_t\}$ be generated instead by the algorithm (19) or (20), with a bounded Lipschitz continuous function ψ_K satisfying (18), and with stepsize $\{\alpha_t\}$ satisfying Assumption 3. Then for each $\delta > 0$, there exists $K_\delta > 0$ such that if $K \ge K_\delta$, then it holds for some sequence $T_k \to \infty$ that

$$\limsup_{k\to\infty} \mathbf{P}\Big(\, \theta_t \notin N_\delta(\theta^*), \text{ some } t \in \big[\, k,\, m(k, T_k)\,\big]\Big) = 0.$$

Theorem 7 (constrained ETD variants with constant stepsize)
In the setting of Theorem 5, let $\{\theta_t^\alpha\}$ be generated instead by the algorithm (19) or (20), with a bounded Lipschitz continuous function ψ_K satisfying (18) and with constant stepsize $\alpha > 0$. Let $\{k_\alpha \mid \alpha > 0\}$ be any sequence of nonnegative integers that are nondecreasing as $\alpha \to 0$. Then for each $\delta > 0$, there exists $K_\delta > 0$ such that the following hold if $K \ge K_\delta$:

(i)
$$\lim_{T\to\infty}\lim_{\alpha\to 0}\frac{1}{T/\alpha}\sum_{t=k_\alpha}^{k_\alpha+\lfloor T/\alpha\rfloor}\mathbb{1}\big(\theta_t^\alpha\in N_\delta(\theta^*)\big)=1\quad \textit{in probability.}$$

(ii) *Let $\alpha k_\alpha\to\infty$ as $\alpha\to 0$. Then there exists a sequence $\{T_\alpha\mid\alpha>0\}$ with $T_\alpha\to\infty$ as $\alpha\to 0$, such that*
$$\limsup_{\alpha\to 0}\mathbf{P}\Big(\theta_t^\alpha\notin N_\delta(\theta^*),\ \textit{some } t\in\big[\,k_\alpha,\,k_\alpha+T_\alpha/\alpha\,\big]\Big)=0.$$

We give the proofs of the above two theorems in Section 4.2. Because the proofs are similar for the two variant algorithms, we include in this paper only the proofs for the first variant—the proofs for the second variant can be found in the arXiv version of this paper (Yu, 2015b).

The proof arguments are largely the same as those that we will use first in Section 4.1 to prove Theorems 4-5 for the algorithm (11). Indeed, for all the three algorithms, the main proof step is the same, which is to apply the general conclusions of (Kushner and Yin, 2003, Theorems 8.2.2, 8.2.3) to establish the connection between the iterates of an algorithm and the solutions of an associated mean ODE, and this step does not concern what the solutions of the ODE are actually. (For the two variant algorithms, verifying that the conditions of Theorems 8.2.2, 8.2.3 in Kushner and Yin, 2003 are met is, in fact, easier because various functions involved in the analysis become bounded due to the use of the bounded function ψ_K.) For the two variant algorithms, the result of this step is that the same conclusions given in Theorems 4-5 hold with $N_\delta(L_B)$ in place of $N_\delta(\theta^*)$, where L_B is the limit set of the projected mean ODE (21) associated with each variant algorithm. To attain Theorems 6-7, we then combine this with the fact that by choosing K sufficiently large, one can make the limit set $L_B\subset N_\delta(\theta^*)$ for an arbitrarily small δ.

3.3 More about the Constant-stepsize Case

For the constant-stepsize case, the results given in Theorems 5 and 7 bear similarities to their counterparts for the diminishing stepsize case given in Theorems 4 and 6. However, they characterize the behavior of the iterates in the limit as the stepsize parameter approaches 0, and deal with only a finite segment of the iterates for each stepsize (although in their part (ii) both the segment's length $T_\alpha/\alpha\to\infty$ and its starting position $k_\alpha\to\infty$ as $\alpha\to 0$). So unlike in the diminishing stepsize case, these results do not tell us explicitly about the behavior of θ_t^α for a fixed stepsize α as we take t to infinity.

The purpose of the present subsection is to analyze further the case of a fixed stepsize just mentioned. We observe that for a fixed stepsize α, the iterates θ_t^α together with $Z_t=(S_t,A_t,e_t,F_t)$ form a weak Feller Markov chain $\{(Z_t,\theta_t^\alpha)\}$ (see Lemma 4, Section 4.3.1). Thus we can apply several ergodic theorems for weak Feller Markov chains (Meyn, 1989; Meyn and Tweedie, 2009) to analyze the constant-stepsize case and combine the implications from these theorems with the results we obtained previously using the weak convergence methods from stochastic approximation theory.

We now present our results using this approach. Let $\mathcal{M}_\alpha$ denote the set of invariant probability measures of the Markov chain $\{(Z_t,\theta_t^\alpha)\}$. This set depends on the particular

algorithm used to generate the θ-iterates, but we shall use the notation $\mathcal{M}_\alpha$ for all the algorithms we discuss here, for notational simplicity. We know that $\{Z_t\}$ has a unique invariant probability measure (Theorem 2, Section 2.4), but it need not be so for the Markov chain $\{(Z_t, \theta_t^\alpha)\}$ when $\{\theta_t^\alpha\}$ is generated by the algorithm (11) or its two variants. The set $\mathcal{M}_\alpha$ can therefore have multiple elements (it is nonempty; see Prop. 6, Section 4.3.2). We denote by $\bar{\mathcal{M}}_\alpha$ the set that consists of the marginal of μ on B (the space of the θ's), for all the invariant probability measures $\mu \in \mathcal{M}_\alpha$.

As in the previous analysis, we are interested in the behavior of multiple consecutive θ-iterates. In order to characterize that, we consider for each $m \geq 1$, the Markov chain

$$\left\{\left((Z_t, \theta_t^\alpha), (Z_{t+1}, \theta_{t+1}^\alpha), \ldots, (Z_{t+m-1}, \theta_{t+m-1}^\alpha)\right)\right\}_{t\geq 0}$$

(i.e., each state now consists of m consecutive states of the chain $\{(Z_t, \theta_t^\alpha)\}$). We shall refer to this chain as the *m-step version* of $\{(Z_t, \theta_t^\alpha)\}$. Similar to $\mathcal{M}_\alpha$, denote by $\mathcal{M}_\alpha^m$ the set of invariant probability measures of the m-step version of $\{(Z_t, \theta_t^\alpha)\}$, and correspondingly define $\bar{\mathcal{M}}_\alpha^m$ to be the set of marginals of μ on B^m for all $\mu \in \mathcal{M}_\alpha^m$. The set $\mathcal{M}_\alpha^m$ is, of course, determined by $\mathcal{M}_\alpha$, since each invariant probability measure in $\mathcal{M}_\alpha^m$ is just the m-dimensional distribution of a stationary Markov chain $\{(Z_t, \theta_t^\alpha)\}$.

Our first result, given in Theorem 8 below, says that for the algorithm (11), as the stepsize α approaches zero, the invariant probability measures in $\mathcal{M}_\alpha^m$ will concentrate their masses on an arbitrarily small neighborhood of $(\theta^*, \ldots, \theta^*)$ (m copies of θ^*). Moreover, for a fixed stepsize, as the number of iterations grows to infinity, the expected maximal deviation of the m consecutive averaged iterates from θ^* can be bounded in terms of the masses those invariant probability measures assign to the vicinities of $(\theta^*, \ldots, \theta^*)$. Here by averaged iterates, we mean

$$\bar{\theta}_t^\alpha = \frac{1}{t}\sum_{k=0}^{t-1} \theta_k^\alpha, \qquad \forall\, t \geq 1, \tag{22}$$

and we shall refer to $\{\bar{\theta}_t^\alpha\}$ as the *averaged sequence* corresponding to $\{\theta_t^\alpha\}$. This iterative averaging is also known as "Polyak-averaging" when it is applied to accelerate the convergence of the θ-iterates (see Polyak and Juditsky, 1992; Kushner and Yin, 2003, Chap. 10; and the references therein). This is not the role of the averaging operation here, however. The purpose here is to bring to bear the ergodic theorems for weak Feller Markov chains, in particular, the weak convergence of certain averaged probability measures or occupation probability measures to the invariant probability measures of the m-step version of $\{(Z_t, \theta_t^\alpha)\}$. (For the details see Section 4.3, where the proofs of the results of this subsection will be given.) It can also be seen that for a sequence $\{\beta_t\}$ with $\beta_t \in [0, 1), \beta_t \to 0$ as $t \to \infty$, if we drop a fraction β_t of the terms in (22) when averaging the θ's at each time t, the resulting differences in the averaged iterates $\bar{\theta}_t^\alpha$ are asymptotically negligible. Therefore, although our results below will be stated for (22), they apply to a variety of averaging schemes.

Recall that $N_\delta(\theta^*)$ denotes the closed δ-neighborhood of θ^*. In what follows, $N'_\delta(\theta^*)$ denotes the open δ-neighborhood of θ^*, i.e., the open ball around θ^* with radius δ. We write $[N_\delta(\theta^*)]^m$ or $[N'_\delta(\theta^*)]^m$ for the Cartesian product of m copies of $N_\delta(\theta^*)$ or $N'_\delta(\theta^*)$. Recall also that r_B is the radius of the constraint set B.

Theorem 8 *In the setting of Theorem 5, let $\{\theta_t^\alpha\}$ be generated by the algorithm (11) with constant stepsize $\alpha > 0$, and let $\{\bar{\theta}_t^\alpha\}$ be the corresponding averaged sequence. Then the following hold for any $\delta > 0$ and $m \geq 1$:*

(i) $\liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha^m} \mu\big([N_\delta(\theta^*)]^m\big) = 1$, *and more strongly, with $m_\alpha = \lceil \frac{m}{\alpha} \rceil$,*

$$\liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha^{m_\alpha}} \mu\big([N_\delta(\theta^*)]^{m_\alpha}\big) = 1.$$

(ii) *For each stepsize α and any initial condition of $(e_0, F_0, \theta_0^\alpha)$,*

$$\limsup_{k\to\infty} \mathbb{E}\Big[\sup_{k\leq t<k+m} \big|\bar{\theta}_t^\alpha - \theta^*\big|\Big] \leq \delta\,\kappa_{\alpha,m} + 2r_B\,(1-\kappa_{\alpha,m}),$$

where $\kappa_{\alpha,m} = \inf_{\mu\in\bar{\mathcal{M}}_\alpha^m} \mu([N'_\delta(\theta^*)]^m)$.

Note that in part (ii) above, $\kappa_{\alpha,m} \to 1$ as $\alpha \to 0$ by part (i). Note also that for $m = 1$, the conclusions from the preceding theorem take the simplest form:

$$\liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha} \mu\big(N_\delta(\theta^*)\big) = 1,$$

$$\limsup_{t\to\infty} \mathbb{E}\big[\big|\bar{\theta}_t^\alpha - \theta^*\big|\big] \leq \delta\,\kappa_\alpha + 2r_B\,(1-\kappa_\alpha), \qquad \text{for } \kappa_\alpha = \inf_{\mu\in\bar{\mathcal{M}}_\alpha} \mu\big(N'_\delta(\theta^*)\big).$$

The conclusions for $m > 1$ are, however, much stronger. They also suggest that in practice, instead of simply choosing the last iterate of the algorithm as its final output at the end of its run, one can base that choice on the behavior of multiple consecutive $\bar{\theta}_t^\alpha$ during the run.

For the two variant algorithms (19) and (20), we have a similar result given in Theorem 9 below. Here the neighborhood of $(\theta^*, \ldots, \theta^*)$ around which the masses of the invariant probability measures are concentrated, depends not only on the stepsize α but also on the biases of these algorithms. The proofs of Theorems 8-9 are given in Section 4.3.2.

Theorem 9 *In the setting of Theorem 5, let $\{\theta_t^\alpha\}$ be generated instead by the algorithm (19) or (20), with constant stepsize $\alpha > 0$ and with a bounded Lipschitz continuous function ψ_K satisfying (18). Let $\{\bar{\theta}_t^\alpha\}$ be the corresponding averaged sequence. Then the following hold:*

(i) *For any given $\delta > 0$, there exists $K_\delta > 0$ such that for all $K \geq K_\delta$,*

$$\liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha^m} \mu\big([N_\delta(\theta^*)]^m\big) = 1, \qquad \forall\, m \geq 1,$$

and more strongly, with $m_\alpha = \lceil \frac{m}{\alpha} \rceil$,

$$\liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha^{m_\alpha}} \mu\big([N_\delta(\theta^*)]^{m_\alpha}\big) = 1, \qquad \forall\, m \geq 1.$$

(ii) *Regardless of the choice of K, given any $\delta > 0, m \geq 1$ and stepsize α, for each initial condition of $(e_0, F_0, \theta_0^\alpha)$,*

$$\limsup_{k\to\infty} \mathbb{E}\Big[\sup_{k\leq t<k+m} \big|\bar{\theta}_t^\alpha - \theta^*\big|\Big] \leq \delta\,\kappa_{\alpha,m} + 2r_B\,(1-\kappa_{\alpha,m}),$$

where $\kappa_{\alpha,m} = \inf_{\mu\in\bar{\mathcal{M}}_\alpha^m} \mu([N'_\delta(\theta^*)]^m)$.

Finally, we consider a simple modification of the preceding algorithms, for which the conclusions of Theorems 8(ii) and 9(ii) can be strengthened. This is our motivation for introducing the modification, but we shall postpone the discussion till Remark 2 at the end of this subsection.

For any of the algorithms (11), (19) or (20), if the original recursion under a constant stepsize α can be written as

$$\theta^\alpha_{t+1} = \Pi_B\big(\theta^\alpha_t + \alpha Y^\alpha_t\big),$$

we now modify this recursion formula by adding a perturbation term $\alpha\Delta^\alpha_{\theta,t}$ as follows. Let

$$\theta^\alpha_{t+1} = \Pi_B\big(\theta^\alpha_t + \alpha Y^\alpha_t + \alpha\Delta^\alpha_{\theta,t}\big), \tag{23}$$

where for each $\alpha > 0$, $\Delta^\alpha_{\theta,t}, t \geq 0$, are $\mathbb{R}^n$-valued random variables such that[13]

(i) they are independent of each other and also independent of the process $\{Z_t\}$;

(ii) they are identically distributed with zero mean and finite variance, where the variance can be bounded uniformly for all α; and

(iii) they have a positive continuous density function with respect to the Lebesgue measure.

Below we refer to (23) as the perturbed version of the algorithm (11), (19) or (20).

Theorem 10 *In the setting of Theorem 5, let $\{\theta^\alpha_t\}$ be generated instead by the perturbed version (23) of the algorithm (11) for a constant stepsize $\alpha > 0$, and let $\{\bar\theta^\alpha_t\}$ be the corresponding averaged sequence. Then the conclusions of Theorems 5 and 8 hold. Furthermore, let the stepsize α be given. Then the Markov chain $\{(Z_t, \theta^\alpha_t)\}$ has a unique invariant probability measure μ_α, and for any $\delta > 0, m \geq 1$ and initial condition of $(e_0, F_0, \theta^\alpha_0)$, almost surely,*

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1} \mathbb{1}\Big(\sup_{k\leq j<k+m} \big|\theta^\alpha_j - \theta^*\big| < \delta\Big) \geq \bar\mu^{(m)}_\alpha\big([N'_\delta(\theta^*)]^m\big)$$

and

$$\limsup_{t\to\infty} \big|\bar\theta^\alpha_t - \theta^*\big| \leq \delta\,\kappa_\alpha + 2r_B\,(1-\kappa_\alpha), \qquad \text{with } \kappa_\alpha = \bar\mu_\alpha\big(N'_\delta(\theta^*)\big),$$

where $\bar\mu^{(m)}_\alpha$ is the unique element in $\bar{\mathcal{M}}^m_\alpha$, and $\bar\mu_\alpha$ is the marginal of μ_α on B.

Theorem 11 *In the setting of Theorem 5, let $\{\theta^\alpha_t\}$ be generated instead by the perturbed version (23) of the algorithm (19) or (20), with a constant stepsize $\alpha > 0$ and with a bounded Lipschitz continuous function ψ_K satisfying (18). Let $\{\bar\theta^\alpha_t\}$ be the corresponding averaged sequence. Then the conclusions of Theorems 7 and 9 hold. Furthermore, for any given stepsize α, the conclusions of the second part of Theorem 10 also hold.*

Note that in the second part of Theorem 10, both $\bar\mu^{(m)}_\alpha\big([N'_\delta(\theta^*)]^m\big)$ and κ_α approach 1 as $\alpha \to 0$, since by the first part of the theorem, the conclusions of Theorem 8 hold. For the second part of Theorem 11, the same is true provided that K is sufficiently large (so that $N_\delta(L_B) \subset N_\delta(\theta^*)$ where L_B is the limit set of the ODE associated with the algorithm),

13. We adopt these conditions for simplicity. They are not the weakest possible for our purpose, and our proof techniques can be applied to other types of perturbations as well. For related discussions, see Remark 2 at the end of this section, as well as Remark 3 and the discussion before Prop. 8 in Section 4.3.3.

and this can be seen from the conclusions of Theorem 9(i), which holds for the perturbed version (23) of the two variant algorithms, as the first part of Theorem 11 says. The proofs of Theorems 10-11 are given in Section 4.3.3.

Remark 2 (on the role of perturbation) At first sight it may seem counter-productive to add noise to the θ-iterates in the algorithm (23). Our motivation for such random perturbations of the θ-iterates is that this can ensure that the Markov chain $\{(Z_t, \theta_t^\alpha)\}$ has a unique invariant probability measure (see Prop. 9, Section 4.3.3). The uniqueness allows us to invoke a result of Meyn (1989) on the convergence of the occupation probability measures of a weak Feller Markov chain, so that we can bound the deviation of the averaged iterates from θ^* not only in an expected sense as before, but also for almost all sample paths under each initial condition, as in the second part of Theorems 10-11. For the unperturbed algorithms, we can only prove such pathwise bounds on $\limsup_{t\to\infty} |\bar{\theta}_t^\alpha - \theta^*|$ for a subset of the initial conditions of (Z_0, θ_0^α). A more detailed discussion of this is given in Remark 3 at the end of Section 4.3.3, after the proofs of the preceding theorems.

Regarding other effects of the perturbation, intuitively, larger noise terms may help the Markov chain "mix" faster, but they can also result in less probability mass $\bar{\mu}_\alpha\big(N'_\delta(\theta^*)\big)$ around θ^* than in the case without perturbation. What is a suitable amount of noise to add to achieve a desired balance? We do not yet have an answer. It seems reasonable to us to let the magnitude of the variance of the perturbation terms $\Delta_{\theta,t}^\alpha$ be approximately $\alpha^{2\epsilon}$ for some $\epsilon \in (0, 1]$, so that a typical perturbation $\alpha\Delta_{\theta,t}^\alpha$ is at a smaller scale relative to the "signal part" αY_t^α in an iteration. Further investigation is needed. ■

4. Proofs for Section 3

We now prove the theorems given in the preceding section. We shall use KY as an abbreviation for the book (Kushner and Yin, 2003), which we will refer to frequently below.

4.1 Proofs for Theorems 4 and 5

In this subsection we prove Theorems 4 and 5 (Section 3.1) on convergence properties of the constrained ETD(λ) algorithm (11). We will apply two theorems from (KY), Theorems 8.2.2 and 8.2.3, which concern weak convergence of stochastic approximation algorithms for constant and diminishing stepsize, respectively. This requires us to show that the conditions of those theorems are satisfied by our algorithm. The major conditions concern the uniform integrability, tightness, and convergence in mean of certain sequences of random variables involved in the algorithm. Our proofs will rely on many properties of the ETD iterates that we have established in (2015a) when analyzing the almost sure convergence of the algorithm.

4.1.1 Conditions to Verify

We need some definitions and notation, before describing the conditions required. For some index set $\mathcal{K}$, let $\{X_k\}_{k\in\mathcal{K}}$ be a set of random variables taking values in a metric space $\mathbf{X}$ (in our context $\mathbf{X}$ will be $\mathbb{R}^m$ or Ξ). The set $\{X_k\}_{k\in\mathcal{K}}$ is said to be *tight* or *bounded in probability*, if there exists for each $\delta > 0$ a compact set $D_\delta \subset \mathbf{X}$ such that

$$\inf_{k\in\mathcal{K}} \mathbf{P}(X_k \in D_\delta) \geq 1 - \delta.$$

For $\mathbb{R}^m$-valued X_k, the set $\{X_k\}_{k\in\mathcal{K}}$ is said to be *uniformly integrable* (u.i.) if

$$\lim_{a\to\infty} \sup_{k\in\mathcal{K}} \mathbb{E}\left[\|X_k\| \mathbb{1}\big(\|X_k\| \geq a\big)\right] = 0.$$

To analyze the constrained ETD(λ) algorithm (11), which is given by

$$\theta_{t+1} = \Pi_B(\theta_t + \alpha_t Y_t), \qquad \text{where } Y_t := e_t \cdot \rho_t\big(R_t + \gamma_{t+1}\phi(S_{t+1})^\top\theta_t - \phi(S_t)^\top\theta_t\big),$$

let $\mathbb{E}_t$ denote expectation conditioned on $\mathcal{F}_t$, the sigma-algebra generated by $\theta_m, \xi_m, m \leq t$, where we recall $\xi_m = (e_m, F_m, S_m, A_m, S_{m+1})$ and its space $\mathbb{R}^{n+1}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ is denoted by Ξ. By writing $Y_t = \mathbb{E}_t[Y_t] + (Y_t - \mathbb{E}_t[Y_t])$, we have the equivalent form of (11) given in (15):

$$\theta_{t+1} = \Pi_B(\theta_t + \alpha_t\, h(\theta_t, \xi_t) + \alpha_t\, e_t \cdot \tilde{\omega}_{t+1})\,.$$

In other words, $h(\theta_t, \xi_t) = \mathbb{E}_t[Y_t]$ and $e_t \cdot \tilde{\omega}_{t+1} = Y_t - \mathbb{E}_t[Y_t]$, a noise term that satisfies $\mathbb{E}_t[e_t \cdot \tilde{\omega}_{t+1}] = 0$.

This algorithm belongs to the class of stochastic approximation algorithms with "exogenous noises" studied in the book (KY) – the term "exogenous noises" reflects the fact that the evolution of $\{\xi_t\}$ is not driven by the θ-iterates. Theorems 4 and 5 will follow as special cases from Theorems 8.2.3 and 8.2.2 of (KY, Chap. 8), respectively, if we can show that the algorithm (11) satisfies the following conditions.

Conditions for the case of diminishing stepsize:

(i) The sequence $\{Y_t\} = \{h(\theta_t, \xi_t) + e_t \cdot \tilde{\omega}_{t+1}\}$ is u.i. (This corresponds to the condition A.8.2.1 in KY.)

(ii) The function $h(\theta, \xi)$ is continuous in θ uniformly in $\xi \in D$, for each compact set $D \subset \Xi$. (This corresponds to the condition A.8.2.3 in KY.)

(iii) The sequence $\{\xi_t\}$ is tight. (This corresponds to the condition A.8.2.4 in KY.)

(iv) The sequence $\{h(\theta_t, \xi_t)\}$ is u.i., and so is $\{h(\theta, \xi_t)\}$ for each fixed $\theta \in B$. (This corresponds to the condition A.8.2.5 in KY.)

(v) There is a continuous function $\bar{h}(\cdot)$ such that for each $\theta \in B$ and each compact set $D \subset \Xi$,

$$\lim_{k\to\infty, t\to\infty} \frac{1}{k} \sum_{m=t}^{t+k-1} \mathbb{E}_t\left[h(\theta, \xi_m) - \bar{h}(\theta)\right] \mathbb{1}\big(\xi_t \in D\big) = 0 \qquad \text{in mean},$$

where k and t are taken to ∞ in any way possible. In other words, if we denote the average on the left-hand side by $X_{k,t}$, then the requirement "$\lim_{k\to\infty,t\to\infty} X_{k,t} = 0$ in mean" means that along any subsequences $k_j \to \infty, t_j \to \infty$, we must have $\lim_{j\to\infty} \mathbb{E}[\|X_{k_j,t_j}\|] = 0$. (This condition corresponds to the condition A.8.2.7 in KY.)

For the case of constant stepsize, we consider the iterates that could be generated by the algorithm for all stepsizes. To distinguish between the iterates associated with different stepsizes, in the conditions below, the superscript α is attached to the variables involved in the algorithm with stepsize α, and similarly, the conditional expectation $\mathbb{E}_t$ is denoted by $\mathbb{E}_t^\alpha$ instead.

Conditions for the case of constant stepsize:

In addition to the condition (ii) above (which corresponds to the condition A.8.1.6 in KY for the case of constant stepsize), the following conditions are required.

(i′) The set $\{Y_t^\alpha \mid t \geq 0, \alpha > 0\} := \{h(\theta_t^\alpha, \xi_t^\alpha) + e_t^\alpha \cdot \tilde{\omega}_{t+1}^\alpha \mid t \geq 0, \alpha > 0\}$ is u.i. (This corresponds to the condition A.8.1.1 in KY.)

(iii′) The set $\{\xi_t^\alpha \mid t \geq 0, \alpha > 0\}$ is tight. (This corresponds to the condition A.8.1.7 in KY.)

(iv′) The set $\{h(\theta_t^\alpha, \xi_t^\alpha) \mid t \geq 0, \alpha > 0\}$ is u.i., in addition to the uniform integrability of $\{h(\theta, \xi_t^\alpha) \mid t \geq 0, \alpha > 0\}$ for each $\theta \in B$. (This corresponds to the condition A.8.1.8 in KY.)

(v′) There is a continuous function $\bar{h}(\cdot)$ such that for each $\theta \in B$ and each compact set $D \subset \Xi$,

$$\lim_{k\to\infty, t\to\infty, \alpha\to 0} \frac{1}{k} \sum_{m=t}^{t+k-1} \mathbb{E}_t^\alpha \left[h(\theta, \xi_m^\alpha) - \bar{h}(\theta)\right] \mathbb{1}\big(\xi_t^\alpha \in D\big) = 0 \qquad \text{in mean,}$$

where α is taken to 0 and k, t are taken to ∞ in any way possible. (This condition corresponds to the condition A.8.1.9 in KY, and it is in fact stronger than the latter condition but is satisfied by our algorithms as we will show.)

The preceding conditions allow ξ_t^α and θ_t^α to be generated under different initial conditions for different α. While we will need this generality later in Section 4.3, here we will focus on a common initial condition for all stepsizes, for simplicity. Then, the preceding conditions for the constant-stepsize case are essentially the same as those for the diminishing stepsize case, because except for the θ-iterates, all the other variables (such as ξ_t and $\tilde{\omega}_t$) involved in the algorithm have identical probability distributions for all stepsizes α and are not affected by the θ-iterates. For this reason, in the proofs below, except for the θ-iterates, we simply omit the superscript α for other variables in the case of constant stepsize, and to verify the two sets of conditions above, we shall treat the case of diminishing stepsize and the case of constant stepsize simultaneously.

As mentioned in Section 2.4, these conditions are to ensure that the projected ODE (12), $\dot{x} = \bar{h}(x) + z, z \in -\mathcal{N}_B(x)$, is the mean ODE for the algorithm (11) and reflects its average dynamics. Among the proofs for these conditions given next, the proof for the convergence in mean condition (v) and (v′) will be the most involved.

4.1.2 Proofs

The condition (ii) is clearly satisfied. In what follows, we prove that the rest of the conditions are satisfied as well. We start with the tightness conditions (iii) and (iii′), as they are immediately implied by a property of the trace iterates we already know. We then tackle the uniform integrability conditions (i), (i′), (iv) and (iv′), before we address the convergence in mean required in (v) and (v′). The proofs build upon several key properties of the ETD iterates we have established in (2015a) and recounted in Section 2.4 and Appendix A.

First, we show that the tightness conditions (iii) and (iii′) are satisfied. This is implied by the following property of traces: $\sup_{t\geq 0} \mathbb{E}\big[\big\|(e_t, F_t)\big\|\big] < \infty$ for any given initial condition (e_0, F_0) (see Prop. 11, Appendix A).

Proposition 1 *Under Assumption 1, for each given initial $(e_0, F_0) \in \mathbb{R}^{n+1}$, $\{(e_t, F_t)\}$ is tight and hence $\{\xi_t\}$ is tight.*

Proof By Prop. 11, $c := \sup_{t\geq 0} \mathbb{E}\big[\big\|(e_t, F_t)\big\|\big] < \infty$. Then, by the Markov inequality, for $a > 0$, $\sup_{t\geq 0} \mathbf{P}\big(\big\|(e_t, F_t)\big\| \geq a\big) \leq c/a \to 0$ as $a \to \infty$. This implies that $\{(e_t, F_t)\}$ is tight. Since the space $\mathcal{S} \times \mathcal{A} \times S$ is finite and $\xi_t = (e_t, F_t, S_t, A_t, S_{t+1})$, $\{\xi_t\}$ is also tight. ■

We now handle the uniform integrability conditions (i), (i′), (iv) and (iv′). The uniform integrability of the trace sequence $\{e_t\}$, as we will prove, is important here.

Proposition 2 *Under Assumption 1, for each given initial $(e_0, F_0) \in \mathbb{R}^{n+1}$, the following sets of random variables are u.i.:*

(i) $\{e_t\}$;

(ii) $\{h(\theta, \xi_t)\}$ *for each fixed* $\theta \in B$;

(iii) $\{h(\theta_t, \xi_t)\}$ *in the case of diminishing stepsize; and* $\{h(\theta_t^\alpha, \xi_t) \mid t \geq 0, \alpha > 0\}$ *in the case of constant stepsize;*

(iv) $\{h(\theta_t, \xi_t) + e_t\, \tilde{\omega}_{t+1}\}$ *in the case of diminishing stepsize; and* $\{h(\theta_t^\alpha, \xi_t) + e_t\, \tilde{\omega}_{t+1} \mid t \geq 0, \alpha > 0\}$ *in the case of constant stepsize.*

The proof of Prop. 2 will use facts about u.i. sequences of random variables given in the lemma below. This lemma basically follows from the definition of uniform integrability. We therefore omit its proof, which can be found in the arXiv version of this paper (Yu, 2015b).

Lemma 2 *Let $X_k, Y_k, k \in \mathcal{K}$ (some index set) be real-valued random variables with X_k and Y_k defined on a common probability space for each k.*

(i) *If $\{X_k\}_{k\in\mathcal{K}}, \{Y_k\}_{k\in\mathcal{K}}$ are u.i., then $\{X_k + Y_k\}_{k\in\mathcal{K}}$ is u.i.*

(ii) *If $\{X_k\}_{k\in\mathcal{K}}$ is u.i. and for all k, $|Y_k| \leq |X_k|$ a.s., then $\{Y_k\}_{k\in\mathcal{K}}$ is u.i.*

(iii) *If $\{X_k\}_{k\in\mathcal{K}}, \{Y_k\}_{k\in\mathcal{K}}$ are u.i. and for some $c \geq 0$, $\mathbb{E}[|Y_k|\,|\,X_k] \leq c$ a.s. for all k, then $\{X_k Y_k\}_{k\in\mathcal{K}}$ is u.i.*

We now proceed to prove Prop. 2. The proof will involve auxiliary variables, which we call truncated traces. They are defined similarly to the trace iterates (e_t, F_t), but instead of depending on all the past states and actions, they only depend on a certain number of the most recent states and actions. Specifically, for each integer $K \geq 1$, we define truncated traces $(\tilde{e}_{t,K}, \tilde{F}_{t,K})$ as follows:

$$(\tilde{e}_{t,K}, \tilde{F}_{t,K}) = (e_t, F_t) \quad \text{for} \quad t \leq K,$$

and for $t \geq K + 1$, with the shorthand $\beta_t := \rho_{t-1}\gamma_t\lambda_t$,

$$\tilde{F}_{t,K} = \sum_{k=t-K}^{t} i(S_k) \cdot \big(\rho_k \gamma_{k+1} \cdots \rho_{t-1}\gamma_t\big), \tag{24}$$

$$\tilde{M}_{t,K} = \lambda_t\, i(S_t) + (1 - \lambda_t)\tilde{F}_{t,K}, \tag{25}$$

$$\tilde{e}_{t,K} = \sum_{k=t-K}^{t} \tilde{M}_{k,K} \cdot \phi(S_k) \cdot \big(\beta_{k+1} \cdots \beta_t\big). \tag{26}$$

Note that when $t \geq 2K+1$, the traces $(\tilde{e}_{t,K}, \tilde{F}_{t,K})$ no longer depend on the initial (e_0, F_0); being functions of the states and actions between time $t-2K$ and t only, they lie in a bounded set determined by K, since the state and action spaces are finite. For $t = 0, \ldots, 2K$, $(\tilde{e}_{t,K}, \tilde{F}_{t,K})$ also lie in a bounded set, which is determined by K and the initial (e_0, F_0). We will use these bounded truncated traces to approximate the original traces $\{(e_t, F_t)\}$ in the analysis.

An important approximation property, given in Prop. 13 (Appendix A), is that for each K and any initial (e_0, F_0) from a given bounded set E,

$$\sup_{t\geq 0} \mathbb{E}\left[\left\|(e_t, F_t) - (\tilde{e}_{t,K}, \tilde{F}_{t,K})\right\|\right] \leq L_K,$$

where L_K is a finite constant that depends on K and E and decreases monotonically to 0 as K increases:

$$L_K \downarrow 0 \quad \text{as } K \to \infty.$$

We will use this property in the following analysis.

Proof of Prop. 2 First, we prove $\{e_t\}$ is u.i. We then use this to show the uniform integrability of the other sets required in parts (ii)-(iv).

(i) To prove $\{e_t\}$ is u.i., we shall exploit its relation with the truncated traces, $\tilde{e}_{t,K}, t \geq 0$ for integers $K \geq 1$. Note that since the state and action spaces are finite, the truncated traces $\{\tilde{e}_{t,K}\}$ lie in a bounded set (this set depends on K and the initial (e_0, F_0)), so there exists a constant a_K such that $\|\tilde{e}_{t,K}\| \leq a_K$ for all t. This fact will greatly simplify the analysis. Let us first fix K and consider $a \geq a_k$. Denote $\bar{a} = a - a_K \geq 0$. Then

$$\begin{aligned}
\|e_t\|\,\mathbb{1}\big(\|e_t\| \geq a\big) &\leq \|e_t\|\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big) \\
&\leq \|e_t - \tilde{e}_{t,K}\|\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big) + \|\tilde{e}_{t,K}\|\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big) \\
&\leq \|e_t - \tilde{e}_{t,K}\|\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big) + a_K\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big). \qquad (27)
\end{aligned}$$

For the second term on the right-hand side, we can bound its expectation by

$$\mathbb{E}\left[a_K\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big)\right] = a_K \mathbf{P}(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}) \leq a_K \cdot L_K/\bar{a}, \qquad \forall\, t, \qquad (28)$$

where in the last inequality L_K is a constant that depends on K (and the initial (e_0, F_0)) and has the property that $L_K \downarrow 0$ as $K \to \infty$, and this inequality is derived by combing the Markov inequality $\mathbf{P}(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}) \leq \mathbb{E}[\|e_t - \tilde{e}_{t,K}\|]/\bar{a}$ with Prop. 13, which bounds $\sup_{t\geq 0} \mathbb{E}[\|e_t - \tilde{e}_{t,K}\|]$ by L_K. Similarly, for the first term on the right-hand side of (27), using Prop. 13, we can bound its expectation by L_K:

$$\mathbb{E}\left[\|e_t - \tilde{e}_{t,K}\|\,\mathbb{1}\big(\|e_t - \tilde{e}_{t,K}\| \geq \bar{a}\big)\right] \leq \mathbb{E}[\|e_t - \tilde{e}_{t,K}\|] \leq L_K, \qquad \forall\, t. \qquad (29)$$

From (27)-(29) it follows that

$$\sup_{t\geq 0} \mathbb{E}\left[\|e_t\|\,\mathbb{1}\big(\|e_t\| \geq a\big)\right] \leq L_K + a_K \cdot L_K/(a - a_K),$$

so for fixed K, by taking $a \to \infty$, we obtain

$$\lim_{a\to\infty} \sup_{t\geq 0} \mathbb{E}\left[\|e_t\|\,\mathbb{1}\big(\|e_t\| \geq a\big)\right] \leq L_K.$$

Since $L_K \downarrow 0$ as $K \to \infty$ (Prop. 13), this implies $\lim_{a\to\infty} \sup_{t\geq 0} \mathbb{E}\left[\|e_t\| \mathbb{1}\big(\|e_t\| \geq a\big)\right] = 0$, which proves the uniform integrability of $\{e_t\}$.

(ii) We now prove for each θ, $\{h(\theta, \xi_t)\}$ is u.i. Since the state and action spaces are finite and θ is given, using the expression of $h(\theta, \xi_t)$, we can bound it as $\|h(\theta, \xi_t)\| \leq L\|e_t\|$ for some constant L. As just proved, $\{e_t\}$ is u.i. (equivalently $\{\|e_t\|\}$ is u.i.) and thus $\{L\|e_t\|\}$ is u.i., so by Lemma 2(ii), $\{h(\theta, \xi_t)\}$ is u.i. (since this is by definition equivalent to $\{\|h(\theta, \xi_t)\|\}$ being u.i., which is true by Lemma 2(ii)).

(iii) The uniform integrability of $\{h(\theta_t, \xi_t)\}$ in the case of diminishing stepsize or $\{h(\theta_t^\alpha, \xi_t) \mid t \geq 0, \alpha > 0\}$ in the case of constant stepsize follows from the same argument given for (ii) above, because θ_t or θ_t^α for all $t \geq 0$ and $\alpha > 0$ lie in the bounded set B by the definition of the constrained ETD(λ) algorithm.

(iv) Consider first the case of diminishing stepsize. We prove that $\{h(\theta_t, \xi_t) + e_t\, \tilde{\omega}_{t+1}\}$ is u.i. (recall $\tilde{\omega}_{t+1} = \rho_t \left(R_t - r(S_t, A_t, S_{t+1})\right)$ is the noise part of the observed reward). Since we already showed that $\{h(\theta_t, \xi_t)\}$ is u.i., by Lemma 2(i), it is sufficient to prove that $\{e_t\, \tilde{\omega}_{t+1}\}$ is u.i. Now $\{e_t\}$ is u.i. by part (i). Since the random rewards R_t in our model have bounded variances, the noise variables $\tilde{\omega}_{t+1}, t \geq 0$, also have bounded variances. This implies that $\{\tilde{\omega}_{t+1}\}$ is u.i. (Billingsley, 1968, p. 32) and that $\mathbb{E}[|\tilde{\omega}_{t+1}| \mid e_t] < c$ for some constant c (independent of t). It then follows from Lemma 2(iii) that $\{e_t\, \tilde{\omega}_{t+1}\}$ is u.i., and hence $\{h(\theta_t, \xi_t) + e_t\, \tilde{\omega}_{t+1}\}$ is u.i.

Similarly, in the case of constant stepsize, it follows from Lemma 2(i) that the set $\{h(\theta_t^\alpha, \xi_t) + e_t\, \tilde{\omega}_{t+1} \mid t \geq 0, \alpha > 0\}$ is u.i., because $\{h(\theta_t^\alpha, \xi_t) \mid t \geq 0, \alpha > 0\}$ is u.i. by part (iii) proved earlier and $\{e_t\, \tilde{\omega}_{t+1}\}$ is u.i. as we just proved. ■

Finally, we handle the conditions (v) and (v′) stated in Section 4.1.1. The two conditions are the same condition in the case here, because they concern each fixed θ, whereas $\{\xi_t\}$ is not affected by the stepsize and the θ-iterates. So we can focus just on the condition (v) in presenting the proof, for notational simplicity. For the algorithm (11), the continuous function $\bar{h}$ required in the condition is the function $\bar{h}(\theta) = C\theta + b$ associated with the desired mean ODE (12). We now prove the required convergence in mean by using the properties of trace iterates and the convergence results given in Theorem 3 and Corollary 1 (Section 2.4).

Proposition 3 *Let Assumption 1 hold. For each $\theta \in B$ and each compact set $D \subset \Xi$,*

$$\lim_{k\to\infty, t\to\infty} \frac{1}{k} \sum_{m=t}^{t+k-1} \mathbb{E}_t \left[h(\theta, \xi_m) - \bar{h}(\theta)\right] \mathbb{1}\big(\xi_t \in D\big) = 0 \qquad \textit{in mean.}$$

Proof Denote $X_{k,t} = \frac{1}{k} \sum_{m=t}^{t+k-1} \big(h(\theta, \xi_m) - \bar{h}(\theta)\big) \mathbb{1}\big(\xi_t \in D\big)$. Since $\mathbb{E}\big[\big\|\mathbb{E}_t\{X_{k,t}\}\big\|\big] \leq \mathbb{E}[\|X_{k,t}\|]$, to prove $\lim_{k,t} \mathbb{E}\big[\big\|\mathbb{E}_t\{X_{k,t}\}\big\|\big] = 0$ (here and in what follows we simply write "k, t" under a limit symbol for "$k \to \infty, t \to \infty$"), it is sufficient to prove $\lim_{k,t} \mathbb{E}[\|X_{k,t}\|] = 0$, that is, to prove

$$\lim_{k,t} \frac{1}{k} \sum_{m=t}^{t+k-1} \big(h(\theta, \xi_m) - \bar{h}(\theta)\big) \mathbb{1}\big(\xi_t \in D\big) = 0 \qquad \text{in mean.} \tag{30}$$

Furthermore, since $\limsup_{k,t} \mathbb{E}\big[\|X_{k,t}\|\,\mathbb{1}\big(\xi_t \in D\big)\big]$ is upper-bounded by

$$\sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \limsup_{k,t}\ \mathbb{E}\left[\|X_{k,t}\|\,\mathbb{1}\big(\xi_t \in D, (S_t, A_t, S_{t+1}) = (s,a,s')\big)\right],$$

it is sufficient in the proof to consider only those compact sets D of the form $D = E \times \{(s,a,s')\}$, for each compact set $E \subset \mathbb{R}^{n+1}$ and each $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Henceforth, let us fix a compact set E together with a triplet (s,a,s') as the set D under consideration, and for this set D, we proceed to prove (30).

To show (30), what we need to show is that for two arbitrary subsequences of integers $k_j \to \infty$, $t_j \to \infty$,

$$\lim_{j\to\infty} \frac{1}{k_j} \sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m) - \bar{h}(\theta)\big)\ \mathbb{1}\big(\xi_{t_j} \in D\big) = 0 \qquad \text{in mean.} \tag{31}$$

To this end, we first define auxiliary trace variables to decompose each difference term $h(\theta,\xi_m) - \bar{h}(\theta)$ into two difference terms as follows:

(a) Fix a point $(\bar{e}, \bar{F}) \in E$.

(b) For each $j \geq 1$, define a sequence of trace pairs, (e_m^j, F_m^j), $m \geq t_j$, by using the same recursion (3)-(5) that defines the traces $\{(e_t, F_t)\}$, based on the same trajectory $\{(S_t, A_t)\}$, but starting at time $m = t_j$ with the initial $(e_{t_j}^j, F_{t_j}^j) = (\bar{e}, \bar{F})$.

Denote $\xi_m^j = (e_m^j, F_m^j, S_m, A_m, S_{m+1})$ for $m \geq t_j$; it differs from ξ_m only in the two trace components. Next, for each m, we write $h(\theta,\xi_m) - \bar{h}(\theta) = (h(\theta,\xi_m^j) - \bar{h}(\theta)) + (h(\theta,\xi_m) - h(\theta,\xi_m^j))$ and correspondingly, we write

$$\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m) - \bar{h}(\theta)\big) = \frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m^j) - \bar{h}(\theta)\big) + \frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m) - h(\theta,\xi_m^j)\big).$$

We see that for (31) to hold, it is sufficient that

$$\lim_{j\to\infty} \frac{1}{k_j} \sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m^j) - \bar{h}(\theta)\big)\ \mathbb{1}\big(\xi_{t_j} \in D\big) = 0 \qquad \text{in mean,} \tag{32}$$

and

$$\lim_{j\to\infty} \frac{1}{k_j} \sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m) - h(\theta,\xi_m^j)\big)\ \mathbb{1}\big(\xi_{t_j} \in D\big) = 0 \qquad \text{in mean.} \tag{33}$$

Let us now prove these two statements.

Proof of (32): Since the set $D = E \times \{(s,a,s')\}$ and $\mathbb{1}\big(\xi_{t_j} \in D\big) \leq \mathbb{1}\big((S_{t_j}, A_{t_j}, S_{t_j+1}) = (s,a,s')\big)$, we can remove ξ_{t_j} from consideration and show instead

$$\lim_{j\to\infty} \frac{1}{k_j} \sum_{m=t_j}^{t_j+k_j-1} \big(h(\theta,\xi_m^j) - \bar{h}(\theta)\big)\ \mathbb{1}\big((S_{t_j}, A_{t_j}, S_{t_j+1}) = (s,a,s')\big) = 0 \qquad \text{in mean,} \tag{34}$$

which will imply (32). By definition $\xi_m^j, m \geq t_j$, are generated from the initial trace pairs $(\bar{e}, \bar{F})$ and initial transition $(S_{t_j}, A_{t_j}, S_{t_j+1})$ at time $m = t_j$. So if $(S_{t_j}, A_{t_j}, S_{t_j+1}) = (s, a, s')$, then conditioned on this transition at t_j, the sequence $\{\xi_m^j, m \geq t_j\}$ has the same probability distribution as a sequence $\hat{\xi}_m, m \geq 0$, where $\hat{\xi}_m = (\hat{e}_m, \hat{F}_m, \hat{S}_m, \hat{A}_m, \hat{S}_{m+1})$ is generated from the initial condition $\hat{\xi}_0 = (\bar{e}, \bar{F}, s, a, s')$ by the same recursion (3)-(5) and a trajectory $\{(\hat{S}_m, \hat{A}_m)\}$ of states and actions under the behavior policy. This shows that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\Big(h(\theta,\xi_m^j)-\bar{h}(\theta)\Big)\mathbb{1}\Big((S_{t_j},A_{t_j},S_{t_j+1})=(s,a,s')\Big)\right\|\right]\\
&\leq \mathbb{E}\left[\left\|\frac{1}{k_j}\sum_{m=0}^{k_j-1}\Big(h(\theta,\hat{\xi}_m)-\bar{h}(\theta)\Big)\right\|\right],
\end{aligned}$$

from which we see that the convergence in mean stated by (34) holds if we have

$$\lim_{k\to\infty}\frac{1}{k}\sum_{m=0}^{k-1}\Big(h(\theta,\hat{\xi}_m)-\bar{h}(\theta)\Big)=0 \qquad \text{in mean.} \tag{35}$$

Now since for each θ, the function $h(\theta,\cdot)$ is Lipschitz continuous in e uniformly in the other arguments, (35) holds by Theorem 3 and its implication Corollary 1 (Section 2.4). Consequently, (34) holds, and this implies (32).

Proof of (33): Using the expression of h and the finiteness of the state and action spaces, we can bound the difference $h(\theta,\xi_m) - h(\theta,\xi_m^j)$ by

$$\big\|h(\theta,\xi_m)-h(\theta,\xi_m^j)\big\| \leq c\cdot\big\|e_m-e_m^j\big\|$$

for some constant c (independent of m, j). Let us show

$$\lim_{j\to\infty}\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|e_m-e_m^j\big\|\,\mathbb{1}\big(\xi_{t_j}\in D\big)=0 \qquad \text{in mean,} \tag{36}$$

which will imply (33).

To prove (36), similarly to the preceding proof, we first decompose each difference term $e_m - e_m^j$ in (36) into several difference terms, by using truncated traces $\{(\tilde{e}_{m,K}, \tilde{F}_{m,K})\}$ and $\{(\tilde{e}_{m,K}^j, \tilde{F}_{m,K}^j) \mid m \geq t_j\}$, $j \geq 1, K \geq 1$, which we now introduce. Specifically, for each $K \geq 1$, $\{(\tilde{e}_{m,K}, \tilde{F}_{m,K})\}$ are defined by (24)-(26). For each $j \geq 1$ and $K \geq 1$, the truncated traces $\{(\tilde{e}_{m,K}^j, \tilde{F}_{m,K}^j) \mid m \geq t_j\}$ are also defined by (24)-(26), except that the initial time is set to be t_j (instead of 0) and for $m \leq t_j + K$, $(\tilde{e}_{m,K}^j, \tilde{F}_{m,K}^j)$ is set to be (e_m^j, F_m^j) (instead of (e_m, F_m)).

Let us fix K for now. We bound the difference $e_m - e_m^j$ by the sum of three difference terms as

$$\big\|e_m-e_m^j\big\| \leq \big\|e_m-\tilde{e}_{m,K}\big\| + \big\|e_m^j-\tilde{e}_{m,K}^j\big\| + \big\|\tilde{e}_{m,K}-\tilde{e}_{m,K}^j\big\|, \tag{37}$$

and correspondingly, we consider the following three sequences of variables, as j tends to ∞:

$$\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|e_m-\tilde{e}_{m,K}\big\|\,\mathbb{1}\big(\xi_{t_j}\in D\big),\qquad \frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|e^j_m-\tilde{e}^j_{m,K}\big\|, \tag{38}$$

and

$$\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|\tilde{e}_{m,K}-\tilde{e}^j_{m,K}\big\|\,\mathbb{1}\big(\xi_{t_j}\in D\big). \tag{39}$$

In what follows, we will bound their expected values as $j\to\infty$ and then take $K\to\infty$; this will lead to (36).

The analyses for the two sequences in (38) are similar. Recall $D=E\times\{(s,a,s')\}$, so $\xi_{t_j}\in D$ implies $(e_{t_j},F_{t_j})\in E$. Since the set E is bounded, if $(e_{t_j},F_{t_j})\in E$, then we can use Prop. 13 (Appendix A) to bound the expectation of $\|e_m-\tilde{e}_{m,K}\|$ for $m\geq t_j$ conditioned on $\mathcal{F}_{t_j}$, and this gives us the bound

$$\sup_{m\geq t_j}\mathbb{E}_{t_j}\left[\big\|e_m-\tilde{e}_{m,K}\big\|\right]\,\mathbb{1}\big(\xi_{t_j}\in D\big)\leq L_K$$

where L_K is a constant that depends on K and the set E, and has the property that $L_K\downarrow 0$ as $K\to\infty$. From this bound, we obtain

$$\mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|e_m-\tilde{e}_{m,K}\big\|\,\mathbb{1}\big(\xi_{t_j}\in D\big)\right]\leq L_K,\qquad \forall\, j\geq 1. \tag{40}$$

Similarly, for the second sequence in (38), by Prop. 13 we have

$$\mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|e^j_m-\tilde{e}^j_{m,K}\big\|\right]\leq L_K,\qquad \forall\, j\geq 1, \tag{41}$$

where L_K is some constant that can be chosen to be the same constant in (40) (because the point $(\bar{e},\bar{F})$, which is the initial trace pair for (e^j_m,F^j_m) at time $m=t_j$, lies in E).

Consider now the sequence in (39). As discussed after the definition (24)-(26) of truncated traces, because of truncation, these traces lie in a bounded set determined by K and the set in which the initial trace pair lies. Therefore, there exists a finite constant c_K which depends on K and E, such that for all $m\geq t_j$,

$$\|\tilde{e}^j_{m,K}\|\leq c_K,\qquad \text{and}\quad \|\tilde{e}_{m,K}\|\leq c_K\ \text{ if } (e_{t_j},F_{t_j})\in E.$$

Also by their definition, once m is sufficiently large, the truncated traces do not depend on the initial trace pairs; in particular,

$$\tilde{e}^j_{m,K}=\tilde{e}_{m,K},\qquad \forall\, m\geq t_j+2K+1.$$

From these two arguments it follows that

$$\mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1}\big\|\tilde{e}_{m,K}-\tilde{e}^j_{m,K}\big\|\,\mathbb{1}\big(\xi_{t_j}\in D\big)\right]\leq\frac{(2K+1)\cdot 2c_K}{k_j}\to 0\quad\text{as } j\to\infty. \tag{42}$$

Finally, combining (40)-(42) with (37), we obtain

$$\begin{aligned}
&\limsup_{j\to\infty} \mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \left\|e_m - e_m^j\right\| \mathbb{1}\big(\xi_{t_j}\in D\big)\right] \\
&\le \limsup_{j\to\infty} \mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \left\|e_m - \tilde{e}_{m,K}\right\| \mathbb{1}\big(\xi_{t_j}\in D\big)\right] + \limsup_{j\to\infty} \mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \left\|e_m^j - \tilde{e}_{m,K}^j\right\|\right] \\
&\quad + \lim_{j\to\infty} \mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \left\|\tilde{e}_{m,K} - \tilde{e}_{m,K}^j\right\| \mathbb{1}\big(\xi_{t_j}\in D\big)\right] \\
&\le 2L_K.
\end{aligned}$$

Since $L_K \downarrow 0$ as $K\to\infty$ (Prop. 13, Appendix A), by taking $K\to\infty$, we obtain

$$\lim_{j\to\infty} \mathbb{E}\left[\frac{1}{k_j}\sum_{m=t_j}^{t_j+k_j-1} \left\|e_m - e_m^j\right\| \mathbb{1}\big(\xi_{t_j}\in D\big)\right] = 0.$$

This proves (36), which implies (33). ■

With Props. 1-3, we have furnished all the conditions required in order to apply (KY, Theorems 8.2.2, 8.2.3) to the constrained ETD algorithm (11), so we can now specialize the conclusions of these two theorems to our problem. In particular, they tell us that the projected ODE (12) is the mean ODE for (11), and furthermore, by (KY, Theorem 8.2.3) (respectively, KY, Theorem 8.2.2), the conclusions of Theorem 4 (respectively, Theorem 5) hold with $N_\delta(L_B)$ in place of $N_\delta(\theta^*)$, where $N_\delta(L_B)$ is the δ-neighborhood of the limit set L_B for the projected ODE (12). Recall that this limit set is given by

$$L_B = \cap_{\bar{\tau}>0}\, \overline{\cup_{x(0)\in B}\{x(\tau),\, \tau\ge\bar{\tau}\}}$$

where $x(\tau)$ is a solution of the projected ODE (12) with initial condition $x(0)$, the union is over all the solutions with initial $x(0)\in B$, and $\overline{D}$ for a set D denotes the closure of D.

Now when the matrix C is negative definite (as implied by Assumptions 1-2) and when the radius of B exceeds the threshold given in Lemma 1, by the latter lemma, the solutions $x(\tau), \tau\in[0,\infty)$, of the ODE (12) coincide with the solutions of $\dot{x} = \bar{h}(x) = Cx + b$ for all initial $x(0)\in B$. Then from the negative definiteness of C (Theorem 1, Section 2.3), it follows that as $\tau\to\infty$, $x(\tau)\to\theta^*$ uniformly in the initial condition, and consequently, $L_B = \{\theta^*\}$.[14] Thus $N_\delta(L_B) = N_\delta(\theta^*)$ and we obtain Theorems 4 and 5.

14. The details for this statement are as follows. Since $\bar{h}$ is bounded on B and the boundary reflection term $z(\cdot)\equiv 0$ under our assumptions (Lemma 1, Section 2.4), a solution $x(\cdot)$ of (12) is Lipschitz continuous on $[0,\infty)$. We calculate $\dot{V}(\tau)$ for the Lyapunov function $V(\tau) = |x(\tau)-\theta^*|^2$. By the negative definiteness of the matrix C, for some $c>0$, $x^\top Cx \le -c|x|^2$ for all $x\in\mathbb{R}^n$. Then, since $\bar{h}(x) = Cx+b = C(x-\theta^*)$, we have $\dot{V}(\tau) = 2\left\langle x(\tau)-\theta^*,\, \bar{h}(x(\tau))\right\rangle \le -2c\left|x(\tau)-\theta^*\right|^2$, and hence for any $\delta>0$, there exists $\epsilon>0$ such that $\dot{V}(\tau)\le-\epsilon$ if $V(\tau) = |x(\tau)-\theta^*|^2\ge\delta^2$. This together with the continuity of the solution $x(\cdot)$ implies that for any $x(0)\in B$, within time $\bar{\tau} = r_B^2/\epsilon$, the trajectory $x(\tau)$ must reach $N_\delta(\theta^*)$ and stay in that set thereafter. By the definition of the limit set and the arbitrariness of δ, this implies $L_B = \{\theta^*\}$.

4.2 Proofs for Theorems 6 and 7

In this subsection we prove the part of Theorems 6-7 for the first variant of the constrained ETD(λ) algorithm given in (19), Section 3.2. The proof for the second variant algorithm (20) is similar and can be found in the arXiv version of this paper (Yu, 2015b). Like in the previous subsection, we will apply (KY, Theorems 8.2.2, 8.2.3) and show that the required conditions are met. Using the properties of the mean ODE of the variant algorithm, we will then specialize the conclusions of those theorems to obtain the desired results.

Consider the first variant algorithm (19):

$$\theta_{t+1} = \Pi_B\Big(\theta_t + \alpha_t\, \psi_K(e_t) \cdot \rho_t\big(R_t + \gamma_{t+1}\phi(S_{t+1})^\top\theta_t - \phi(S_t)^\top\theta_t\big)\Big).$$

We define a function $h_K : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$ by

$$h_K(\theta,\xi) = \psi_K(e)\cdot\rho(s,a)\left(r(s,a,s') + \gamma(s')\,\phi(s')^\top\theta - \phi(s)^\top\theta\right), \quad \text{for } \xi = (e,F,s,a,s'), \tag{43}$$

and write (19) equivalently as

$$\theta_{t+1} = \Pi_B\Big(\theta_t + \alpha_t\, h_K(\theta_t,\xi_t) + \alpha_t\,\psi_K(e_t)\cdot\tilde{\omega}_{t+1}\Big)$$

with $\tilde{\omega}_{t+1} = \rho_t(R_t - r(S_t,A_t,S_{t+1}))$ as before. Note that $\mathbb{E}_t\left[\psi_K(e)\,\tilde{\omega}_{t+1}\right] = 0$, and the algorithm is similar to the algorithm (11)—equivalently (15)—except that we have h_K and $\psi_K(e_t)$ in place of h and e_t, respectively.

We note two properties of the function h_K. They follow from direct calculations and will be useful in our analysis shortly:

(a) Using the Lipschitz continuity of the function ψ_K (cf. Equation 18, Section 3.2), we have that for each $\theta \in \mathbb{R}^n$, there exists a finite $c > 0$ such that with $\xi = (e,F,s,a,s')$ and $\xi' = (e',F',s,a,s')$,

$$\|h_K(\theta,\xi) - h_K(\theta,\xi')\| \le c\,\|e - e'\|, \qquad \forall\,(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}. \tag{44}$$

Thus $h_K(\theta,\cdot)$ is Lipschitz continuous in (e,F) uniformly in (s,a,s').

(b) Since the set B is bounded, we can bound the difference $h_K(\theta,\xi) - h(\theta,\xi)$ for all θ in B as follows. For some finite constant $c > 0$,

$$\|h_K(\theta,\xi) - h(\theta,\xi)\| \le c\,\|\psi_K(e) - e\| \le 2c\,\|e\|\cdot\mathbb{1}(\|e\| \ge K), \qquad \forall\,\theta \in B, \tag{45}$$

where the last inequality follows from the property (18) of ψ_K:

$$\|\psi_K(x)\| \le \|x\| \;\; \forall\, x \in \mathbb{R}^n, \quad \text{and} \quad \psi_K(x) = x \;\; \text{if } \|x\| \le K.$$

We now apply (KY, Theorems 8.2.2, 8.2.3) to obtain the desired conclusions in Theorems 6-7 for the algorithm (19). This requires us to show that the conditions (i)-(v) and (i′)-(v′) given in Section 4.1.1 are still satisfied when we replace e_t by $\psi_K(e_t)$ and h by h_K. The uniform integrability conditions (i), (i′), (iv) and (iv′) require the following sets to be u.i.: $\{h_K(\theta_t,\xi_t) + \psi_K(e_t)\cdot\tilde{\omega}_{t+1}\}$ and $\{h_K(\theta_t^\alpha,\xi_t) + \psi_K(e_t)\cdot\tilde{\omega}_{t+1} \mid t \ge 0, \alpha > 0\}$, $\{h_K(\theta_t,\xi_t)\}$ and $\{h_K(\theta_t^\alpha,\xi_t) \mid t \ge 0, \alpha > 0\}$, and $\{h_K(\theta,\xi_t)\}$ for each θ. These conditions are evidently

satisfied, in view of the boundedness of the functions ψ_K and $h_K(\theta,\cdot)$ for each θ, the boundedness of the θ-iterates due to constraints, and the finite variances of $\{\tilde{\omega}_t\}$. The condition (ii) on the continuity of $h_K(\cdot,\xi)$ uniformly in $\xi \in D$, for each compact set $D \subset \Xi$, is also clearly satisfied, whereas the condition (iii) (equivalently (iii′)) on the tightness of $\{\xi_t\}$ was already verified earlier in Prop. 1 (Section 4.1.2).

What remains is the condition (v) (which is equivalent to (v′), for the same reason as discussed immediately before Prop. 3, Section 4.1.2). It requires the existence of a continuous function $\bar{h}_K : \mathbb{R}^n \to \mathbb{R}^n$ such that for each $\theta \in B$ and each compact set $D \subset \Xi$,

$$\lim_{k\to\infty,t\to\infty} \frac{1}{k} \sum_{m=t}^{t+k-1} \mathbb{E}_t\big[h_K(\theta,\xi_m) - \bar{h}_K(\theta)\big]\, \mathbb{1}\big(\xi_t \in D\big) = 0 \qquad \text{in mean.} \tag{46}$$

If this condition is satisfied as well, then the mean ODE for the algorithm (19) is given by

$$\dot{x} = \bar{h}_K(x) + z, \qquad z \in -\mathcal{N}_B(x). \tag{47}$$

To furnish the condition (v), we first identify the function $\bar{h}_K(\theta)$ to be $\mathbb{E}_\zeta[h_K(\theta,\xi_0)]$, the expectation of $h_K(\theta,\xi_0)$ under the stationary distribution of the process $\{Z_t\}$ with the invariant probability measure ζ as its initial distribution. We relate the functions $\bar{h}_K, K > 0$, to $\bar{h}$ in the proposition below, and we will use it to characterize the bias of the algorithm (19) later.

Proposition 4 *Let Assumption 1 hold. Consider the setting of the algorithm (19), and for each $\theta \in \mathbb{R}^n$, let $\bar{h}_K(\theta) = \mathbb{E}_\zeta[h_K(\theta,\xi_0)]$. Then the function $\bar{h}_K$ is Lipschitz continuous on $\mathbb{R}^n$, and*

$$\sup_{\theta\in B} \|\bar{h}_K(\theta) - \bar{h}(\theta)\| \to 0 \quad \textit{as } K \to \infty. \tag{48}$$

Proof For each θ, the function $h_K(\theta,\cdot)$ is by definition bounded. Under Assumption 1, the Markov chain $\{(S_t, A_t, e_t, F_t)\}$ has a unique invariant probability measure ζ (Theorem 2, Section 2.4). Therefore, $\bar{h}_K(\theta)$ is well-defined and finite. Let $c_1 = \sup_{e\in\mathbb{R}^n} \|\psi_K(e)\| < \infty$ (since ψ_K is bounded). For any θ, θ', using the definition of h_K, a direct calculation shows that for some $c_2 > 0$, $\|h_K(\theta,\xi) - h_K(\theta',\xi)\| \le c_1 c_2 \|\theta - \theta'\|$ for all $\xi \in \Xi$, from which it follows that

$$\|\bar{h}_K(\theta) - \bar{h}_K(\theta')\| \le \mathbb{E}_\zeta\left[\|h_K(\theta,\xi_0) - h_K(\theta',\xi_0)\|\right] \le c_1 c_2 \|\theta - \theta'\|.$$

This shows that $\bar{h}_K$ is Lipschitz continuous. We now prove (48). Since $\bar{h}_K(\theta) = \mathbb{E}_\zeta[h_K(\theta,\xi_0)]$ by definition and $\bar{h}(\theta) = \mathbb{E}_\zeta[h(\theta,\xi_0)]$ by Corollary 1 (Section 2.4), it is sufficient to prove the following statement, which entails (48):

$$\sup_{\theta\in B} \mathbb{E}_\zeta\left[\big\|h_K(\theta,\xi_0) - h(\theta,\xi_0)\big\|\right] \to 0 \ \text{ as } K \to \infty. \tag{49}$$

By (45), for some constant $c > 0$,

$$\|h_K(\theta,\xi_0) - h(\theta,\xi_0)\| \le 2c\,\|e_0\| \cdot \mathbb{1}(\|e_0\| \ge K), \qquad \forall\, \theta \in B,$$

and therefore,

$$\sup_{\theta\in B} \mathbb{E}_\zeta \left[\left\|h_K(\theta,\xi_0) - h(\theta,\xi_0)\right\|\right] \le 2c\, \mathbb{E}_\zeta\big[\|e_0\| \cdot \mathbb{1}(\|e_0\| \ge K)\big].$$

By Theorem 3 (Section 2.4), $\mathbb{E}_\zeta[\|e_0\|] < \infty$ and hence $\mathbb{E}_\zeta[\|e_0\|\cdot\mathbb{1}(\|e_0\| \ge K)] \to 0$ as $K \to \infty$. Together with the preceding inequality, this implies (49), which in turn implies (48). ■

We now show that the convergence in mean required in (46) is satisfied.

Proposition 5 *Under Assumption 1, the conclusion of Prop. 3 (Section 4.1.2) holds in the setting of the algorithm (19), with the functions h_K and $\bar{h}_K$ in place of h and $\bar{h}$, respectively.*

Proof The same arguments given in the proof of Prop. 3 apply here, with the functions $h_K, \bar{h}_K$ in place of $h, \bar{h}$, respectively. Only two details are worth noting here. The proof relies on the Lipschitz continuity property of h_K given in (44). As mentioned earlier, this property implies that for each θ, with $\xi = (e, F, s, a, s')$, $h_K(\theta, \xi)$ is Lipschitz continuous in (e, F) uniformly in (s, a, s'), so we can apply Theorem 3 to conclude that (35) and hence (32) hold in this case (for $h_K, \bar{h}_K$ instead of $h, \bar{h}$). The property (44) also allows us to obtain (33) in this case, by exactly the same proof given earlier. ■

Thus we have furnished all the conditions required by (KY, Theorems 8.2.2, 8.2.3). As in the case of the algorithm (11), by these two theorems, the assertions of Theorems 4-5 hold for the variant algorithm (19) with $N_\delta(L_B)$ in place of $N_\delta(\theta^*)$, where L_B is the limit set of the projected mean ODE associated with (19):

$$\dot{x} = \bar{h}_K(x) + z, \qquad z \in -\mathcal{N}_B(x).$$

To finish the proof for Theorems 6-7, it is now sufficient to show that for any given $\delta > 0$, we can choose a number K_δ large enough so that $L_B \subset N_\delta(\theta^*)$ for all $K \ge K_\delta$. We prove this below, using Prop. 4. Note that the set L_B reflects the bias of the constrained algorithm (19), so what we are showing now is that this bias decreases as K increases.

Lemma 3 *Let Assumptions 1-2 hold, and let the radius of the set B exceed the threshold given in Lemma 1. Then for all K sufficiently large, given any initial condition $x(0) \in B$, a solution to the projected ODE (47) coincides with the unique solution to $\dot{x} = \bar{h}_K(x)$, with the boundary reflection term being $z(\cdot) \equiv 0$. Given $\delta > 0$, there exists K_δ such that for $K \ge K_\delta$, the limit set L_B of (47) satisfies $L_B \subset N_\delta(\theta^*)$.*

Proof Under Assumptions 1-2, the matrix C is negative definite (Theorem 1, Section 2.3), and when the radius of the set B exceeds the threshold given in Lemma 1, there exists a constant $\epsilon > 0$ such that for all boundary points x of B, $\langle x, \bar{h}(x)\rangle < -\epsilon$. At such points x, the normal cone $\mathcal{N}_B(x) = \{ax \mid a \ge 0\}$, and

$$\langle x, \bar{h}_K(x)\rangle = \langle x, \bar{h}(x)\rangle + \langle x, \bar{h}_K(x) - \bar{h}(x)\rangle < -\epsilon + \langle x, \bar{h}_K(x) - \bar{h}(x)\rangle.$$

By (48) in Prop. 4, $\langle x, \bar{h}_K(x) - \bar{h}(x)\rangle \to 0$ uniformly on B as $K \to \infty$. Thus when K is sufficiently large, at all boundary points x of B, $\langle x, \bar{h}_K(x)\rangle < 0$; i.e., $\bar{h}_K(x)$ points inside B

and the boundary reflection term $z = 0$. It then follows that for such K, given an initial condition $x(0) \in B$, a solution to (47) coincides with the unique solution to $\dot{x} = \bar{h}_K(x)$, where the uniqueness is ensured by the Lipschitz continuity of $\bar{h}_K$ proved in Prop. 4 (cf. Borkar, 2008, Chap. 11.2).

To prove the second statement concerning the limit set of the projected ODE, let K be large enough so that the conclusion of the first part holds. Let $x(\tau), \tau \in [0, \infty)$, be the solution of (47) for a given initial $x(0) \in B$. Since $\bar{h}_K$ is bounded on B, $x(\cdot)$ is Lipschitz continuous on $[0, \infty)$. Let $V(\tau) = |x(\tau) - \theta^*|^2$, and we calculate $\dot{V}(\tau)$. Since for all x, $\bar{h}(x) = Cx + b = C(x - \theta^*)$ and $x^\top Cx \leq -c|x|^2$ for some $c > 0$ by the negative definiteness of C, a direct calculation shows that

$$\begin{aligned}
\dot{V}(\tau) &= 2\left\langle x(\tau) - \theta^* \,,\, \bar{h}_K(x(\tau))\right\rangle \\
&= 2\left\langle x(\tau) - \theta^* \,,\, \bar{h}(x(\tau))\right\rangle + 2\left\langle x(\tau) - \theta^*, \bar{h}_K(x(\tau)) - \bar{h}(x(\tau))\right\rangle \\
&\leq -2c\left|x(\tau) - \theta^*\right|^2 + 2\left|x(\tau) - \theta^*\right| \cdot \left|\bar{h}_K(x(\tau)) - \bar{h}(x(\tau))\right|.
\end{aligned}$$

By (48) in Prop. 4, $\sup_{x \in B} |h_K(x) - \bar{h}(x)| \to 0$ as $K \to \infty$. It then follows that for any $\delta > 0$, there exist $\epsilon > 0$ and $K_\delta > 0$ such that for all $K \geq K_\delta$, $\dot{V}(\tau) \leq -\epsilon$ if $V(\tau) = |x(\tau) - \theta^*|^2 \geq \delta^2$. This together with the continuity of the solution $x(\cdot)$ shows that for any $x(0) \in B$, within time $\bar{\tau} = r_B^2/\epsilon$ (where r_B is the radius of B), the trajectory $x(\tau)$ must reach $N_\delta(\theta^*)$ and stay in that set thereafter. Consequently, for all $K \geq K_\delta$, the limit set $L_B = \cap_{\bar{\tau} \geq 0} \overline{\cup_{x(0) \in B}\{x(\tau), \tau \geq \bar{\tau}\}} \subset N_\delta(\theta^*)$. ■

This completes the proofs of Theorems 6 and 7 for the first variant algorithm (19).

4.3 Further Analysis of the Constant-stepsize Case

We now consider again the case of constant stepsize, and prove Theorems 8-11 given in Section 3.3. The proofs will be based on combining the results we obtained earlier by using stochastic approximation theory, with the ergodic theorems of weak Feller Markov chains. As before the proofs will also rely on the key properties of the ETD iterates.

4.3.1 Weak Feller Markov Chains

We shall focus on Markov chains on complete separable metric spaces. For such a Markov chain $\{X_t\}$ with state space $\mathbf{X}$, let $P(\cdot, \cdot)$ denote its transition kernel, that is, $P : \mathbf{X} \times \mathcal{B}(\mathbf{X}) \to [0, 1]$,

$$P(x, D) = \mathbf{P}_x(X_1 \in D), \qquad \forall\, x \in \mathbf{X},\ D \in \mathcal{B}(\mathbf{X}),$$

where $\mathcal{B}(\mathbf{X})$ denotes the Borel sigma-algebra on $\mathbf{X}$, and $\mathbf{P}_x$ denotes the probability distribution of $\{X_t\}$ conditioned on $X_0 = x$. Multiple-step transition kernels will also be needed. For $t \geq 1$, the t-step transition kernel $P^t(\cdot, \cdot) : \mathbf{X} \times \mathcal{B}(\mathbf{X}) \to [0, 1]$ is given by

$$P^t(x, D) = \mathbf{P}_x(X_t \in D), \qquad \forall\, x \in \mathbf{X},\ D \in \mathcal{B}(\mathbf{X}),$$

and for $t = 0$, P^0 is defined as $P^0(x, \cdot) = \delta_x$, the Dirac measure that assigns probability 1 to the point x, for each $x \in \mathbf{X}$. Define averaged probability measures $\bar{P}_k(x, \cdot)$ for $k \geq 1$ and

$x \in \mathbf{X}$, as

$$\bar{P}_k(x,\cdot) = \frac{1}{k}\sum_{t=0}^{k-1} P^t(x,\cdot).$$

The Markov chain $\{X_t\}$ has the *weak Feller* property if for every bounded continuous function f on $\mathbf{X}$,

$$Pf(x) := \int f(y)P(x,dy) = \mathbb{E}\big[f(X_1) \mid X_0 = x\big]$$

is a continuous function of x (Meyn and Tweedie, 2009, Prop. 6.1.1). Weak Feller Markov chains have nice properties. In our analysis, we will use in particular several properties relating to the invariant probability measures of these chains and convergence of certain probability measures to the invariant probability measures.

Recall that if μ and $\mu_t, t \geq 0$, are probability measures on $\mathbf{X}$, $\{\mu_t\}$ is said to converge weakly to μ if $\int f d\mu_t \to \int f d\mu$ for every bounded continuous function f on $\mathbf{X}$. For $\{\mu_t\}$ that is not necessarily convergent, we shall call the limiting probability measure of any of its convergent subsequence, in the sense of weak convergence, a *weak limit* of $\{\mu_t\}$. For an (arbitrary) index set $\mathcal{K}$, a set of probability measures $\{\mu_k\}_{k\in\mathcal{K}}$ on $\mathbf{X}$ is said to be *tight* if for every $\delta > 0$, there exists a compact set $D_\delta \subset \mathbf{X}$ such that $\mu_k(D_\delta) \geq 1-\delta$ for all $k \in \mathcal{K}$. An important fact is that on a complete separable metric space, any tight sequence of probability measures has a further subsequence that converges weakly to some probability measure (Dudley, 2002, Theorem 11.5.4).

For weak Feller Markov chains, their averaged probability measures $\{\bar{P}_k(x,\cdot)\}_{k\geq 1}$ are known to have the following property; see e.g., the proof of Lemma 4.1 in (Meyn, 1989). It will be needed in our proofs of Theorems 8-9.

Lemma 4 *Let $\{X_t\}$ be a weak Feller Markov chain with transition kernel $P(\cdot,\cdot)$ on a metric space $\mathbf{X}$. For each $x \in \mathbf{X}$, any weak limit of $\{\bar{P}_k(x,\cdot)\}_{k\geq 1}$ is an invariant probability measure of $\{X_t\}$.*

Recall that the occupation probability measures of $\{X_t\}$, denoted $\{\mu_{x,t}\}$ for each initial condition $x \in \mathbf{X}$, are defined as follows:

$$\mu_{x,t}(D) := \frac{1}{t}\sum_{k=0}^{t-1} \mathbb{1}(X_k \in D), \qquad \forall\, D \in \mathcal{B}(\mathbf{X}),$$

where the chain $\{X_t\}$ starts from $X_0 = x$, and each $\mu_{x,t}$ is a random variable taking values in the space of probability measures on $\mathbf{X}$. Let "$\mathbf{P}_x$-a.s." stand for "almost surely with respect to $\mathbf{P}_x$." The next lemma concerns the convergence of occupation probability measures of a weak Feller Markov chain. It is a result of Meyn (1989) and will be needed in our proofs of Theorems 10-11.

Lemma 5 (Meyn, 1989, Prop. 4.2) *Let $\{X_t\}$ be a weak Feller Markov chain with transition kernel $P(\cdot,\cdot)$ on a complete separable metric space $\mathbf{X}$. Suppose that*

(i) *$\{X_t\}$ has a unique invariant probability measure μ;*

(ii) *for each compact set $E \subset \mathbf{X}$, the set $\{\bar{P}_k(x,\cdot) \mid x \in E,\, k \geq 1\}$ is tight; and*

(iii) *for all initial conditions $x \in \mathbf{X}$, there exists a sequence of compact sets $E_k \uparrow \mathbf{X}$ (that is $E_k \subset E_{k+1}$ for all k and $\cup_k E_k = \mathbf{X}$) such that*

$$\lim_{k\to\infty} \liminf_{t\to\infty} \mu_{x,t}(E_k) = 1, \qquad \mathbf{P}_x\text{-a.s.}$$

Then, for each initial condition $x \in \mathbf{X}$, the sequence $\{\mu_{x,t}\}$ of occupation probability measures converges weakly to μ, $\mathbf{P}_x$-almost surely.

The condition (iii) above is equivalent to that the sequence $\{\mu_{x,t}\}$ of occupation probability measures is almost surely tight for each initial condition.

4.3.2 Proofs of Theorems 8 and 9

In this subsection we prove Theorem 8 for the algorithm (11) and Theorem 9 for its two variants (19) and (20). We also show that the conclusions of Theorems 8-9 hold for the perturbed version (23) of these algorithms as well. The proof arguments are largely the same for all the algorithms we consider here. So except where noted otherwise, it will be taken for granted through out this subsection that $\{\theta_t^\alpha\}$ *is generated by either of the six algorithms just mentioned*, for a constant stepsize $\alpha > 0$.

We start with some preliminary analysis given in the next two lemmas. Recall $Z_t = (S_t, A_t, e_t, F_t)$ and $\{Z_t\}$ is a weak Feller Markov chain on $\mathcal{Z} := \mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$ (Yu, 2015a, Sec. 3.1), and its evolution is not affected by the θ-iterates. We consider the Markov chain $\{(Z_t, \theta_t^\alpha)\}$ on the state space $\mathcal{Z} \times B$ (note that this is a complete separable metric space). This chain also has the weak Feller property:

Lemma 6 *Let Assumption 1(ii) hold. The process $\{(Z_t, \theta_t^\alpha)\}$ is a weak Feller Markov chain.*

The proof of the preceding lemma is a straightforward verification using the definition of the weak Feller property. It is included in the arXiv version of this paper (Yu, 2015b) but omitted here due to space limit.

In order to study the behavior of multiple consecutive θ-iterates, we consider for $m \geq 1$, the m-step version of $\{(Z_t, \theta_t^\alpha)\}$, that is, the Markov chain $\{X_t\}$ on $(\mathcal{Z} \times B)^m$ where each state X_t consists of m consecutive states of the original chain $\{(Z_t, \theta_t^\alpha)\}$:

$$X_t = \big((Z_t, \theta_t^\alpha), \ldots, (Z_{t+m-1}, \theta_{t+m-1}^\alpha)\big).$$

Similarly to Lemma 6, it is straightforward to show that the m-step version of a weak Feller Markov chain is a weak Feller chain as well. Thus the m-step version of $\{(Z_t, \theta_t^\alpha)\}$ is also a weak Feller Markov chain, and we can apply the ergodic theorems for weak Feller Markov chains to analyze it. In particular, in this subsection we will use Lemma 4 to prove Theorems 8-9; in the next subsection we will also use Lemma 5.

In analyzing the m-step version of $\{(Z_t, \theta_t^\alpha)\}$, sometimes it will be more convenient for us to take as its initial condition the condition of just (Z_0, θ_0^α)—instead of $(Z_0, \theta_0^\alpha), \ldots, (Z_{m-1}^\alpha, \theta_{m-1}^\alpha)$—and to work with the following objects that are essentially equivalent to the averaged probability measures $\{\bar{P}_k(x, \cdot)\}$ and the occupation probability measures $\{\mu_{x,t}\}$ defined earlier for a general Markov chain $\{X_t\}$. Specifically, with $\{X_t\}$ denoting the m-step

version of $\{(Z_t, \theta_t^\alpha)\}$, for each $(z,\theta) \in \mathcal{Z} \times B$, we define probability measures $\bar{P}_{(z,\theta)}^{(m,k)}, k \geq 1$, on the space $\mathbf{X} = (\mathcal{Z} \times B)^m$, by

$$\bar{P}_{(z,\theta)}^{(m,k)}(D) := \frac{1}{k} \sum_{t=0}^{k-1} \mathbf{P}_{(z,\theta)}\big(X_t \in D\big), \qquad \forall\, D \in \mathcal{B}(\mathbf{X}). \tag{50}$$

Similarly, we define occupation probability measures $\{\mu_{(z,\theta),t}^{(m)}\}$ for each $(z,\theta) \in \mathcal{Z} \times B$ by

$$\mu_{(z,\theta),t}^{(m)}(D) := \frac{1}{t} \sum_{k=0}^{t-1} \mathbb{1}\big(X_k \in D\big), \qquad \forall\, D \in \mathcal{B}(\mathbf{X}), \tag{51}$$

where the initial $(Z_0, \theta_0^\alpha) = (z,\theta)$. Compared with the definitions of $\{\bar{P}_k(x,\cdot)\}$ and $\{\mu_{x,t}\}$ for $\{X_t\}$, apparently, all the previous conclusions given in Section 4.3.1 for $\{\bar{P}_k(x,\cdot)\}$ and $\{\mu_{x,t}\}$ hold for $\big\{\bar{P}_{(z,\theta)}^{(m,k)}\big\}$ and $\big\{\mu_{(z,\theta),t}^{(m)}\big\}$ as well; therefore we can use the objects $\big\{\bar{P}_{(z,\theta)}^{(m,k)}\big\}$ and $\{\bar{P}_k(x,\cdot)\}$, and $\big\{\mu_{(z,\theta),t}^{(m)}\big\}$ and $\{\mu_{x,t}\}$, interchangeably in our analysis.

Lemma 7 *Let Assumption 1 hold. For $m \geq 1$, let $\{X_t\}$ be the m-step version of $\{(Z_t, \theta_t^\alpha)\}$ on $\mathbf{X} = (\mathcal{Z} \times B)^m$, with transition kernel $P(\cdot,\cdot)$. Then $\{X_t\}$ satisfies the conditions (ii)-(iii) of Lemma 5.*

Proof To show that the condition (ii) of Lemma 5 is satisfied, fix a compact set $E \subset \mathbf{X}$ and let us first show that the set $\{P^t(x,\cdot) \mid x \in E, t \geq 0\}$ is tight. Since the set B is compact and the state and action spaces are finite, of concern here is just the tightness of the marginals of these probability measures on the space of the trace components $(e_t, F_t, \ldots, e_{t+m-1}, F_{t+m-1})$ of the state X_t. By Prop. 11 (Appendix A), for all initial conditions of (e_0, F_0) in a given bounded subset of $\mathbb{R}^{n+1}$, $\sup_{t\geq 0} \mathbb{E}[\|(e_t, F_t)\|] \leq L$ for a constant L (that depends on the subset). So for the set E, applying the Markov inequality together with the union bound, we have that there exists a constant $L > 0$ such that for all $x \in E$ and $a > 0$, $\mathbf{P}_x\big(\sup_{k\leq t<k+m} \|(e_t, F_t)\| \geq a\big) \leq mL/a$ for all $k \geq 0$. Now for any given $\delta > 0$, let a be large enough so that $mL/a < \delta$ and let D_a be the closed ball in $\mathbb{R}^{n+1}$ centered at the origin with radius a. Then for the compact set $D = (\mathcal{S} \times \mathcal{A} \times D_a \times B)^m$, we have $P^k(x, D) = \mathbf{P}_x\big(\sup_{k\leq t<k+m} \|(e_t, F_t)\| \leq a\big) \geq 1 - \delta$ for all $x \in E$ and all $k \geq 0$. This shows that the set $\{P^t(x,\cdot) \mid x \in E, t \geq 0\}$ is tight. Consequently, the averages of the probability measures in this set must also form a tight set; in particular, the set $\{\bar{P}_k(x,\cdot) \mid x \in E, k \geq 1\}$ must be tight. Hence $\{X_t\}$ satisfies the condition (ii) of Lemma 5.

Consider now the condition (iii) of Lemma 5. For positive integers k, let E_k in that condition be the compact set $(\mathcal{S} \times \mathcal{A} \times D_k \times B)^m$, where D_k is the closed ball of radius k in $\mathbb{R}^{n+1}$ centered at the origin. We wish to show that for each initial condition $x \in \mathbf{X}$,

$$\lim_{k\to\infty} \liminf_{t\to\infty} \mu_{x,t}(E_k) = 1, \qquad \mathbf{P}_x\text{-a.s.}$$

Since the θ-iterates do not affect the evolution of Z_t, they can be neglected in the proof. It is sufficient to consider instead the m-step version of $\{Z_t\}$ and show that for the compact sets $\hat{E}_k = (\mathcal{S} \times \mathcal{A} \times D_k)^m$, it holds for any initial condition $z \in \mathcal{Z}$ of Z_0 that

$$\lim_{k\to\infty} \liminf_{t\to\infty} \hat{\mu}_{z,t}^{(m)}\big(\hat{E}_k\big) = 1, \qquad \mathbf{P}_z\text{-a.s.}, \tag{52}$$

where $\{\hat{\mu}^{(m)}_{z,t}\}$ are the occupation probability measures of the m-step version of $\{Z_t\}$, defined analogously to (51) with $(Z_t, \ldots, Z_{t+m-1})$ in place of X_t.

To prove (52), consider $\{Z_t\}$ first and its occupation probability measures $\{\hat{\mu}_{z,t}\}$ for each initial condition $Z_0 = z \in \mathcal{Z}$. By Theorem 2 (Section 2.4), $\mathbf{P}_z$-almost surely, $\{\hat{\mu}_{z,t}\}$ converges weakly to ζ (the unique invariant probability measure of $\{Z_t\}$). So by (Dudley, 2002, Theorem 11.1.1), for the open set $\tilde{D}_k = \mathcal{S} \times \mathcal{A} \times D^o_k$, where D^o_k denotes the interior of D_k (i.e., D^o_k is the open ball with radius k), almost surely,

$$\liminf_{t\to\infty} \hat{\mu}_{z,t}\big(\tilde{D}_k\big) \geq \zeta(\tilde{D}_k), \qquad \text{and hence} \quad \lim_{k\to\infty} \liminf_{t\to\infty} \hat{\mu}_{z,t}\big(\tilde{D}_k\big) = 1. \tag{53}$$

Now for the m-step version of $\{Z_t\}$, with $[\tilde{D}_k]^m$ denoting the Cartesian product of m copies of $\tilde{D}_k$, we have

$$\hat{\mu}^{(m)}_{z,t}\big([\tilde{D}_k]^m\big) := \frac{1}{t}\sum_{j=0}^{t-1} \mathbb{1}\big(Z_{j+j'} \in \tilde{D}_k,\, 0 \leq j' < m\big) \geq 1 - \sum_{j'=0}^{m-1} \frac{1}{t}\sum_{j=0}^{t-1} \mathbb{1}\big(Z_{j+j'} \notin \tilde{D}_k\big). \tag{54}$$

For each $j' < m$, by the definition of $\hat{\mu}_{z,t}$, we have $\limsup_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} \mathbb{1}\big(Z_{j+j'} \notin \tilde{D}_k\big) = \limsup_{t\to\infty} \hat{\mu}_{z,t}\big(\tilde{D}^c_k\big)$, where $\tilde{D}^c_k$ denotes the complement of $\tilde{D}_k$ in $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$. By (53), $\lim_{k\to\infty} \limsup_{t\to\infty} \hat{\mu}_{z,t}\big(\tilde{D}^c_k\big) = 0$ almost surely. Hence for each $j' < m$, we have $\lim_{k\to\infty} \limsup_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} \mathbb{1}\big(Z_{j+j'} \notin \tilde{D}_k\big) = 0$ almost surely. We then obtain from (54), by taking the limits as $t \to \infty$ and $k \to \infty$, that

$$\liminf_{k\to\infty} \liminf_{t\to\infty} \hat{\mu}^{(m)}_{z,t}\big([\tilde{D}_k]^m\big) \geq 1 - \sum_{j'=0}^{m-1} \limsup_{k\to\infty} \limsup_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} \mathbb{1}\big(Z_{j+j'} \notin \tilde{D}_k\big) = 1$$

almost surely. The desired equality (52) then follows, since $[\tilde{D}_k]^m \subset \hat{E}_k$. ■

Recall that $\mathcal{M}^m_\alpha$ is the set of invariant probability measures of the m-step version of $\{(Z_t, \theta^\alpha_t)\}$. By Lemma 7 the latter Markov chain satisfies the condition (ii) of Lemma 5, and this implies that the set $\big\{\bar{P}^{(m,k)}_{(z,\theta)}\big\}_{k\geq 1}$ is tight for each initial condition $(Z_0, \theta^\alpha_0) = (z, \theta)$. Recall that any subsequence of a tight sequence has a further convergent subsequence (Dudley, 2002, Theorem 11.5.4). For $\big\{\bar{P}^{(m,k)}_{(z,\theta)}\big\}_{k\geq 1}$, all the weak limits (i.e., the limits of its convergent subsequences) must be invariant probability measures in $\mathcal{M}^m_\alpha$, by the property of weak Feller Markov chains given in Lemma 4:

Proposition 6 *Under Assumption 1, consider the m-step version of $\{(Z_t, \theta^\alpha_t)\}$ for $m \geq 1$. For each $(z, \theta) \in \mathcal{Z} \times B$, the sequence $\big\{\bar{P}^{(m,k)}_{(z,\theta)}\big\}_{k\geq 1}$ of probability measures is tight, and any weak limit of this sequence is an invariant probability measure of the m-step version of $\{(Z_t, \theta^\alpha_t)\}$. (Thus $\mathcal{M}^m_\alpha \neq \emptyset$.)*

We are now ready to prove Theorems 8-9. The idea is to use the conclusions on the θ-iterates that we can obtain by applying (KY, Theorem 8.2.2), to infer the concentration of the mass around a small neighborhood of $(\theta^*, \ldots, \theta^*)$ (m copies of θ^*) for the marginals

of all the invariant probability measures in the set $\mathcal{M}_\alpha^m$, when α is sufficiently small. This can then be combined with Prop. 6 above to prove the desired conclusions on the θ-iterates for a given stepsize.

Recall that $\mathcal{M}_\alpha$ is the set of invariant probability measures of $\{(Z_t, \theta_t^\alpha)\}$. Recall also that $\bar{\mathcal{M}}_\alpha^m$ denotes the set of marginals of the invariant probability measures in $\mathcal{M}_\alpha^m$, on the space of the θ's.

Proposition 7 *In the setting of Theorem 5, for each $\alpha > 0$, let $\{\theta_t^\alpha\}$ be generated instead by the algorithm (11) or its perturbed version (23), with constant stepsize α and under the condition that the initial (Z_0, θ_0^α) is distributed according to some invariant probability measure in $\mathcal{M}_\alpha$. Then the conclusions of Theorem 5 continue to hold.*

Proof The proof arguments are the same as those for Theorem 5 given in Section 4.1. We only need to show that the conditions (ii) and (i′)-(v′) given in Section 4.1.1 for applying (KY, Theorem 8.2.2) are still satisfied under our present assumptions.

For the algorithm (11), the only difference from the previous assumptions in Theorem 5 is that here for each stepsize α, the initial (Z_0, θ_0^α) has a distribution $\mu_\alpha \in \mathcal{M}_\alpha$. The condition (ii) does not depend on such initial conditions, so it continues to hold. For the other conditions, note that since $\{Z_t\}$ has a unique invariant probability measure ζ (Theorem 2), regardless of the choice of μ_α, for all α, $\{Z_t\}$ is stationary and has the same distribution. Then the tightness condition (iii′) trivially holds because as $\{\xi_t\}$ is also stationary and unaffected by the stepsize, each ξ_t^α in (iii′) has the same distribution. Similarly, since $\{e_t\}$ is stationary and unaffected by the stepsize, and each e_t has the same distribution with the mean of $\|e_t\|$ given by $\mathbb{E}_\zeta[\|e_t\|] < \infty$ (Theorem 3), we obtain that $\{e_t\}$ is u.i. From this the uniform integrability required in the conditions (i′) and (iv′) follows as a consequence, as shown in the proof of Prop. 2(ii)-(iv). Lastly, the convergence in mean condition (v′) continues to hold (by the same proof given for Prop. 3). This is because $\{\xi_t\}$ has the same distribution regardless of the stepsize, and because the condition (v′) is for each compact set D and concerns tails of a trajectory starting at instants t with $\xi_t \in D$, which renders any initial condition on Z_0 ineffective. Thus all the required conditions are met, and we obtain the same conclusions on the θ-iterates as given in Theorem 5.

For the perturbed version (23) of the algorithm (11), the only difference to (11) under the present assumptions is the perturbation variables $\Delta_{\theta,t}^\alpha$ involved in each iteration. But by definition these variables have conditional zero mean: $\mathbb{E}_t^\alpha[\Delta_{\theta,t}^\alpha] = 0$, so the only condition in which they appear is the uniform integrability condition (i′): $\{Y_t^\alpha \mid t \geq 0, \alpha > 0\}$ is u.i., where Y_t^α is now given by $Y_t^\alpha = h(\theta_t^\alpha, \xi_t) + e_t \cdot \tilde{\omega}_{t+1} + \Delta_{\theta,t}^\alpha$. By definition $\Delta_{\theta,t}^\alpha$ for all α and t have bounded variance, and hence $\{\Delta_{\theta,t}^\alpha\}$ is u.i. (Billingsley, 1968, p. 32). The set $\{h(\theta_t^\alpha, \xi_t) + e_t \cdot \tilde{\omega}_{t+1} \mid t \geq 0, \alpha > 0\}$ is u.i., which follows from the u.i. of $\{e_t\}$, as we just verified in the case of the algorithm (11). Therefore, by Lemma 2(i), $\{Y_t^\alpha \mid t \geq 0, \alpha > 0\}$ is u.i. and the condition (i′) is satisfied. Since the perturbed version (23) meets all the required conditions, and shares with (11) the same mean ODE, the same conclusions given in Theorem 5 hold for this algorithm as well. ■

We now prove Theorem 8 for the algorithm (11). We prove its part (i) and part (ii) separately, as the arguments are different. Our proofs below also apply to the perturbed

version (23) of the algorithm (11), and together with the preceding proposition, they establish the first part of Theorem 10 (which says that the conclusions of both Theorem 5 and Theorem 8 hold for the perturbed algorithm).

Proof of Theorem 8(i) Proof by contradiction. Consider the statement of Theorem 8(i):

$$\forall\, \delta > 0, \quad \liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha^{m_\alpha}} \mu\big([N_\delta(\theta^*)]^{m_\alpha}\big) = 1, \qquad \text{where } m_\alpha = \lceil \tfrac{m}{\alpha} \rceil.$$

Suppose it is not true. Then there exist $\delta, \epsilon > 0$, $m \geq 1$, a sequence $\alpha_k \to 0$, and a sequence $\mu_{\alpha_k} \in \bar{\mathcal{M}}_{\alpha_k}^{m_k}$, where $m_k = m_{\alpha_k}$, such that

$$\mu_{\alpha_k}([N_\delta(\theta^*)]^{m_k}) \leq 1 - \epsilon, \qquad \forall\, k \geq 0. \tag{55}$$

Each μ_{α_k} corresponds to an invariant probability measure of $\{(Z_t, \theta_t^{\alpha_k})\}$ in $\mathcal{M}_{\alpha_k}$, which we denote by $\hat{\mu}_{\alpha_k}$. For each $k \geq 0$, generate the iterates $\{\theta_t^{\alpha_k}\}$ using $\hat{\mu}_{\alpha_k}$ as the initial distribution of $(Z_0, \theta_0^{\alpha_k})$. For other values of α, generate the iterates $\{\theta_t^{\alpha}\}$ using some $\hat{\mu}_\alpha \in \mathcal{M}_\alpha$ as the initial distribution of (Z_0, θ_0^{α}). By Prop. 7, the conclusions of Theorem 5 hold:

$$\limsup_{\alpha\to 0} \mathbf{P}\Big(\theta_t^\alpha \notin N_\delta(\theta^*), \text{ some } t \in \big[\, k_\alpha,\, k_\alpha + T_\alpha/\alpha \,\big] \Big) = 0,$$

where $T_\alpha \to \infty$ as $\alpha \to 0$, and this implies for the given m,

$$\limsup_{\alpha\to 0} \mathbf{P}\Big(\theta_t^\alpha \notin N_\delta(\theta^*), \text{ some } t \in \big[\, k_\alpha,\, k_\alpha + \lceil \tfrac{m}{\alpha} \rceil \,\big) \Big) = 0. \tag{56}$$

But for each $\alpha > 0$, the process $\{(Z_t, \theta_t^\alpha)\}$ with the initial distribution $\hat{\mu}_\alpha$ is stationary, so the probability in the left-hand side of (56) is just $1 - \mu_\alpha([N_\delta(\theta^*)]^{m_\alpha})$, for the marginal probability measure $\mu_\alpha \in \bar{\mathcal{M}}_\alpha^{m_\alpha}$ that corresponds to the invariant probability measure $\hat{\mu}_\alpha$. Therefore, by (56), $\liminf_{\alpha\to 0} \mu_\alpha([N_\delta(\theta^*)]^{m_\alpha}) = 1$. On the other hand, by (55), $\liminf_{\alpha\to 0} \mu_\alpha([N_\delta(\theta^*)]^{m_\alpha}) \leq \liminf_{k\to\infty} \mu_{\alpha_k}([N_\delta(\theta^*)]^{m_k}) < 1$, a contradiction. Thus the statement of Theorem 8(i) recounted at the beginning of this proof must hold.

This also proves the other statement of Theorem 8(i), $\liminf_{\alpha\to 0} \inf_{\mu\in\bar{\mathcal{M}}_\alpha^m} \mu\big([N_\delta(\theta^*)]^m\big) = 1$, because for $\alpha < 1$, by the correspondences between those invariant probability measures in $\mathcal{M}_\alpha^m$ and those in $\mathcal{M}_\alpha^{m_\alpha}$, $\inf_{\mu\in\bar{\mathcal{M}}_\alpha^m} \mu\big([N_\delta(\theta^*)]^m\big) \geq \inf_{\mu\in\bar{\mathcal{M}}_\alpha^{m_\alpha}} \mu\big([N_\delta(\theta^*)]^{m_\alpha}\big)$. This completes the proof. ■

Proof of Theorem 8(ii) We suppress the superscript α of θ_t^α in the proof. The statement is trivially true if $\delta \geq 2r_B$, so consider the case $\delta < 2r_B$. Let $(z, \theta) \in Z \times B$ be the initial condition of (Z_0, θ_0). By convexity of the Euclidean norm, $\big|\bar{\theta}_t - \theta^*\big| \leq \frac{1}{t}\sum_{j=0}^{t-1} |\theta_j - \theta^*|$, and therefore, for all $k \geq 1$,

$$\sup_{k\leq t<k+m} \big|\bar{\theta}_t - \theta^*\big| \leq \frac{1}{k} \sum_{j=0}^{k-1} \sup_{j\leq t<j+m} |\theta_t - \theta^*|, \tag{57}$$

and

$$\mathbb{E}\left[\sup_{k\leq t<k+m} \big|\bar{\theta}_t - \theta^*\big| \right] \leq \frac{1}{k} \sum_{j=0}^{k-1} \mathbb{E}\left[\sup_{j\leq t<j+m} |\theta_t - \theta^*| \right]. \tag{58}$$

With $N'_\delta(\theta^*)$ denoting the open δ-neighborhood of θ^*, we have

$$\begin{aligned}
&\frac{1}{k}\sum_{j=0}^{k-1}\mathbb{E}\left[\sup_{j\le t<j+m}|\theta_t-\theta^*|\right]\\
&\le \frac{1}{k}\sum_{j=0}^{k-1}\mathbb{E}\left[\Big(\sup_{j\le t<j+m}|\theta_t-\theta^*|\Big)\cdot\mathbb{1}\big(\theta_t\in N'_\delta(\theta^*),\, j\le t<j+m\big)\right]\\
&\quad+\frac{1}{k}\sum_{j=0}^{k-1}\mathbb{E}\left[\Big(\sup_{j\le t<j+m}|\theta_t-\theta^*|\Big)\cdot\mathbb{1}\big(\theta_t\notin N'_\delta(\theta^*),\,\text{some } t\in[j,j+m)\big)\right]\\
&\le \delta\cdot\bar{P}^{(m,k)}_{(z,\theta)}(D_\delta)+2r_B\cdot\big(1-\bar{P}^{(m,k)}_{(z,\theta)}(D_\delta)\big),
\end{aligned}\tag{59}$$

where $D_\delta=\big\{\big(z^1,\theta^1,\ldots,z^m,\theta^m\big)\in(\mathcal{Z}\times B)^m\;\big|\;\sup_{1\le j\le m}\big|\theta^j-\theta^*\big|<\delta\big\}$, and the second inequality follows from the definition (50) of the averaged probability measure $\bar{P}^{(m,k)}_{(z,\theta)}$.

By Prop. 6, $\big\{\bar{P}^{(m,k)}_{(z,\theta)}\big\}_{k\ge1}$ is tight and all its weak limits are in $\mathcal{M}^m_\alpha$, the set of invariant probability measure of the m-step version of $\{(Z_t,\theta_t)\}$. There is also the fact that on a metric space, if a sequence of probability measures μ_k converges to some probability measure μ weakly, then $\liminf_{k\to\infty}\mu_k(D)\ge\mu(D)$ for any open set D (Dudley, 2002, Theorem 11.1.1). From these two arguments we have that for the set D_δ, which is open with respect to the topology on $(\mathcal{Z}\times B)^m$,

$$\liminf_{k\to\infty}\bar{P}^{(m,k)}_{(z,\theta)}(D_\delta)\ge\inf_{\mu\in\mathcal{M}^m_\alpha}\mu(D_\delta)=\inf_{\mu\in\bar{\mathcal{M}}^m_\alpha}\mu\big([N'_\delta(\theta^*)]^m\big)=:\kappa_{\alpha,m}.\tag{60}$$

Combining the three inequalities (58)-(60), and using also the relation $\delta<2r_B$, we obtain

$$\limsup_{k\to\infty}\mathbb{E}\left[\sup_{k\le t<k+m}\big|\bar{\theta}_t-\theta^*\big|\right]\le\delta\,\kappa_{\alpha,m}+2r_B\big(1-\kappa_{\alpha,m}\big).$$

This complete the proof. ■

We prove Theorem 9 in exactly the same way as we proved Theorem 8, so we omit the details and only outline the proof here. First, for the variant algorithms (19) and (20) as well as their perturbed version (23), we consider fixed K and ψ_K. Similar to Prop. 7, we show that if for each stepsize α, the initial (Z_0,θ_0^α) is distributed according to some invariant probability measure in $\mathcal{M}_\alpha$, then the algorithms continue to satisfy the conditions given in Section 4.1.1, so we can apply (KY, Theorem 8.2.2) to assert that the conclusions of Theorem 5 continue to hold with $N_\delta(\theta^*)$ replaced by the limit set $N_\delta(L_B)$ of the mean ODE associated with each algorithm. (Recall Theorem 7 is also obtained in this way.) Subsequently, with $N_\delta(L_B)$ in place of $N_\delta(\theta^*)$ again, and with K and ψ_K still held fixed, we use the same proof for Theorem 8(i) to obtain that for any $\delta>0$ and $m\ge1$,

$$\liminf_{\alpha\to0}\inf_{\mu\in\bar{\mathcal{M}}^{m_\alpha}_\alpha}\mu\big([N_\delta(L_B)]^{m_\alpha}\big)=1,\qquad\text{where } m_\alpha=\lceil\tfrac{m}{\alpha}\rceil.$$

Finally, we combine this with the fact that given any $\delta > 0$, the limit set $N_\delta(L_B) \subset N_\delta(\theta^*)$ for all K sufficiently large (see Lemma 3 in Section 4.2, which holds for (19) and (20), as well as their perturbed version (23) since the latter has the same mean ODE as the original algorithm). Theorem 9(i) then follows: given $\delta > 0$, for all K sufficiently large,

$$\liminf_{\alpha \to 0} \inf_{\mu \in \bar{\mathcal{M}}_\alpha^{m_\alpha}} \mu\big([N_\delta(\theta^*)]^{m_\alpha}\big) = 1.$$

The proof for Theorem 9(ii) is exactly the same as that for Theorem 8(ii) given earlier. In particular, this proof relies solely on the weak Feller property of the Markov chain $\{(Z_t, \theta_t^\alpha)\}$ and the convergence property of the averaged probability measures of the m-step version of $\{(Z_t, \theta_t^\alpha)\}$, all of which have shown to hold for the algorithms (19) and (20) and their perturbed version (23) in this subsection.

The preceding arguments also show that the first part of Theorem 11 holds; that is, the conclusions of Theorem 7 and Theorem 9 hold for the perturbed version (23) of the algorithm (19) or (20) as well.

4.3.3 Proofs of Theorems 10 and 11

In this subsection we establish completely Theorems 10 and 11 regarding the perturbed version (23) of the algorithms (11), (19) and (20). We have already proved the first part of both of these theorems in the previous subsection. Below we tackle their second part, which, as we recall, is stronger than the corresponding part of Theorems 8 and 9 in that for a fixed stepsize α, the deviation of the averaged iterates $\{\bar{\theta}_t^\alpha\}$ from θ^* in the limit as $t \to \infty$ is now characterized not in an expected sense but for almost all sample paths.

To simplify the presentation, except where noted otherwise, it will be taken for granted throughout this subsection that $\{\theta_t^\alpha\}$ *is generated by the perturbed version (23) of any of the three algorithms (11), (19) and (20).* Recall that when updating θ_t^α to θ_{t+1}^α, the perturbed algorithm (23) adds the perturbation term $\alpha\Delta_{\theta,t}$ to the iterate before the projection Π_B, where $\Delta_{\theta,t}, t \geq 0$, are assumed to be i.i.d. $\mathbb{R}^n$-valued random variables that have zero mean and bounded variances and have a positive continuous density function with respect to the Lebesgue measure. (Here and in what follows, we omit the superscript α of the noise terms $\Delta_{\theta,t}$ since we deal with a fixed stepsize α in this part of the analysis.) As mentioned in Section 3.3, these conditions are not as weak as possible. Indeed, the purpose of the perturbation is to make the invariant probability measure of $\{(Z_t, \theta_t^\alpha)\}$ unique so that we can invoke the ergodic theorem for weak Feller Markov chains given in Lemma 5, Section 4.3.1. Therefore, any conditions that can guarantee the uniqueness of the invariant probability measure can be used. In the present paper, for simplicity, we focus on the conditions we assumed earlier on $\Delta_{\theta,t}$, and prove the uniqueness just mentioned under these conditions, although our proof arguments can be useful for weaker conditions as well.

Proposition 8 *Under Assumption 1,* $\{(Z_t, \theta_t^\alpha)\}$ *has a unique invariant probability measure.*

The next two lemmas are the intermediate steps to prove Prop. 8. We need the notion of a stochastic kernel, of which the transition kernel of a Markov chain is one example. For two topological spaces $\mathbf{X}$ and $\mathbf{Y}$, a function $Q : \mathcal{B}(\mathbf{X}) \times \mathbf{Y} \to [0, 1]$ is a (Borel measurable) *stochastic kernel on* $\mathbf{X}$ *given* $\mathbf{Y}$, if for each $y \in \mathbf{Y}$, $Q(\cdot \mid y)$ is a probability measure

on $\mathcal{B}(\mathbf{X})$ and for each $D \in \mathcal{B}(\mathbf{X})$, $Q(D \mid y)$ is a Borel measurable function on $\mathbf{Y}$. For the algorithms we consider, the iteration that generates $(Z_{t+1}, \theta^\alpha_{t+1})$ from (Z_t, θ^α_t) can be equivalently described in terms of stochastic kernels. In particular, the transition from Z_t to Z_{t+1} is described by the transition kernel of the Markov chain $\{Z_t\}$, and the probability distribution of θ^α_{t+1} given θ^α_t and $\xi_t = (e_t, F_t, S_t, A_t, S_{t+1})$ is described by another stochastic kernel, which will be our focus in the analysis below.

Lemma 8 *Let Assumption 1(ii) hold. Let $Q(d\theta' \mid \xi, \theta)$ be the stochastic kernel (on B given $\Xi \times B$) that describes the probability distribution of θ^α_{t+1} given $\xi_t = \xi, \theta^\alpha_t = \theta$. Then for each bounded set $E \subset \Xi$, there exist $\beta \in (0, 1]$ and a probability measure Q_1 on B such that*

$$Q(d\theta' \mid \xi, \theta) \geq \beta\, Q_1(d\theta'), \qquad \forall\, \xi \in E,\ \theta \in B. \tag{61}$$

Proof We consider only the case where $\{\theta^\alpha_t\}$ is generated by the perturbed version of the algorithm (11); the proof for the perturbed version of the two other algorithms (19) and (20) follows exactly the same arguments. In the proof below we use the notation that for a scalar c and a set $D \subset \mathbb{R}^n$, the set $cD = \{cx \mid x \in D\}$.

By the definitions of the algorithms (11) and (23), for $\xi = (e, F, s, a, s') \in \Xi$ and $\theta \in B$, we can express $Q(\cdot \mid \xi, \theta)$ as

$$Q(D \mid \xi, \theta) = \int\int \mathbb{1}\Big(\Pi_B\big(\theta + \alpha f(\xi, \theta, r) + \alpha\Delta\big) \in D\Big)\, p(d\Delta)\, q(dr \mid s, a, s'), \quad \forall\, D \in \mathcal{B}(B), \tag{62}$$

where $f(\xi, \theta, r) = e \cdot \rho(s, a)\big(r + \gamma(s')\phi(s')^\top\theta - \phi(s)^\top\theta\big)$, and $p(\cdot)$ is the common distribution of the perturbation variables $\Delta_{\theta,t}$. Let $\bar{r} > 0$ be large enough so that for some $c > 0$, $q([-\bar{r}, \bar{r}] \mid \bar{s}, \bar{a}, \bar{s}') \geq c$ for all $(\bar{s}, \bar{a}, \bar{s}') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Let E be an arbitrary bounded subset of Ξ. For all $\xi \in E, \theta \in B$ and $r \in [-\bar{r}, \bar{r}]$, since E and B are bounded, $g(\xi, \theta, r) := (\theta + \alpha f(\xi, \theta, r))/\alpha$ lies in a compact subset of $\mathbb{R}^n$, which we denote by $\bar{D}$. Let $\epsilon \in (0, r_B/\alpha]$ and let $\bar{D}_\epsilon$ be the ϵ-neighborhood of $\bar{D}$. By our assumption on the perturbation variables involved in the algorithm (23), $p(\cdot)$ has a positive continuous density function with respect to the Lebesgue measure $\ell(\cdot)$. Therefore, there exists some $c' > 0$ such that for any Borel subset D of the compact set $-\bar{D}_\epsilon := \{-x \mid x \in \bar{D}_\epsilon\}$, $p(D) \geq c'\ell(D)$.

Now consider an arbitrary $\xi \in E$, $\theta \in B$, and $r \in [-\bar{r}, \bar{r}]$. We have $y := g(\xi, \theta, r) \in \bar{D}$. Let $B_\epsilon(-y)$ be the ϵ-neighborhood of $-y$, and let B_ϵ denote the closed ball in $\mathbb{R}^n$ centered at the origin with radius ϵ. If $\Delta \in B_\epsilon(-y)$, then $\theta + \alpha f(\xi, \theta, r) + \alpha\Delta = \alpha y + \alpha\Delta \in \alpha B_\epsilon \subset B$ (since $\alpha\epsilon \leq r_B$). Therefore, for any $D \in \mathcal{B}(B)$,

$$\begin{aligned}
\int \mathbb{1}\Big(\Pi_B\big(\theta + \alpha f(\xi, \theta, r) + \alpha\Delta\big) \in D\Big)\, p(d\Delta) &\geq \int_{B_\epsilon(-y)} \mathbb{1}\big(\alpha y + \alpha\Delta \in D\big)\, p(d\Delta) \\
&\geq c' \int_{B_\epsilon(-y)} \mathbb{1}\big(\alpha y + \alpha\Delta \in D\big)\, \ell(d\Delta) \\
&= c'\ell\big(\tfrac{1}{\alpha}D \cap B_\epsilon\big),
\end{aligned} \tag{63}$$

where in the second inequality we used the fact that $B_\epsilon(-y) \subset -\bar{D}_\epsilon$ and restricted to $\mathcal{B}(-\bar{D}_\epsilon)$, $p(d\Delta) \geq c'\ell(d\Delta)$, as discussed earlier.

To finish the proof, define the probability measure Q_1 on B by $Q_1(D) = \ell(\frac{1}{\alpha}D \cap B_\epsilon)/\ell(B_\epsilon)$ for all $D \in \mathcal{B}(B)$. Then for all $\xi \in E$ and $\theta \in B$, using (62) and (63) and our choice of $\bar{r}$, we have

$$Q(D \mid \xi, \theta) \geq \int_{[-\bar{r},\bar{r}]} c'\ell(B_\epsilon) \cdot Q_1(D)\, q(dr \mid s, a, s') \geq c \cdot c'\ell(B_\epsilon) \cdot Q_1(D), \qquad D \in \mathcal{B}(B),$$

and the desired inequality (61) then follows by letting $\beta = cc'\ell(B_\epsilon) > 0$ (we must have $\beta \leq 1$ since we can choose $D = B$ in the inequality above). ■

We will use the preceding result in the proof of the next lemma.

Lemma 9 *Let Assumption 1 hold. Let $\{\mu_{x,t}\}$ be the sequence of occupation probability measures of $\{(Z_t, \theta_t^\alpha)\}$ for each initial condition $x \in \mathcal{Z} \times B$. Suppose that for some $x = (z, \theta) \in \mathcal{Z} \times B$ and $\mu \in \mathcal{M}_\alpha$, $\{\mu_{x,t}\}$ converges weakly to μ, $\mathbf{P}_x$-almost surely. Then for each $\theta' \in B$ and $x' = (z, \theta')$, $\{\mu_{x',t}\}$ also converges weakly to μ, $\mathbf{P}_{x'}$-almost surely.*

Proof We use a coupling argument to prove the statement. In the proof, we suppress the superscript α of θ_t^α. Let $\{X_t\}$ denote the process $\{(Z_t, \theta_t)\}$ with initial condition $x = (z, \theta)$, and let $\{X'_t\}$ denote the process $\{(Z_t, \theta_t)\}$ with initial condition $x' = (z, \theta')$, for an arbitrary $\theta' \in B$. In what follows, we first define a sequence

$$\{(Z_t, \tilde{\theta}_t, \tilde{\theta}'_t)\} \quad \text{with} \quad (Z_0, \tilde{\theta}_0, \tilde{\theta}'_0) = (z, \theta, \theta'),$$

in such a way that the two marginal processes $\{(Z_t, \tilde{\theta}_t)\}$ and $\{(Z_t, \tilde{\theta}'_t)\}$ have the same probability distributions as $\{X_t\}$ and $\{X'_t\}$, respectively. We then relate the occupation probability measures $\{\mu_{x,t}\}$, $\{\mu_{x',t}\}$ to those of the marginal processes, $\{\tilde{\mu}_{x,t}\}$, $\{\tilde{\mu}_{x',t}\}$, which are defined as

$$\tilde{\mu}_{x,t}(D) = \frac{1}{t}\sum_{k=0}^{t-1} \mathbb{1}\big((Z_k, \tilde{\theta}_k) \in D\big), \quad \tilde{\mu}_{x',t}(D) = \frac{1}{t}\sum_{k=0}^{t-1} \mathbb{1}\big((Z_k, \tilde{\theta}'_k) \in D\big), \quad \forall\, D \in \mathcal{B}(\mathcal{Z} \times B).$$

We now define $\{(Z_t, \tilde{\theta}_t, \tilde{\theta}'_t)\}$. First, let $\{Z_t\}$ be generated as before with $Z_0 = z$. Denote $\xi_t = (e_t, F_t, S_t, A_t, S_{t+1})$ as before, and let Q be the stochastic kernel that describes the evolution of θ_{t+1} given (ξ_t, θ_t). By Lemma 7, the occupation probability measures of $\{Z_t\}$ is almost surely tight for each initial condition. This implies the existence of a compact set $\bar{E} \subset \mathbb{R}^{n+1}$ such that for the compact set $E = \bar{E} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \subset \Xi$, the sequence $\{\xi_t\}$ visits E infinitely often with probability one. For this set E, by Lemma 8, there exist some $\beta \in (0, 1]$ and probability measure Q_1 on B such that $Q(\cdot \mid \bar{\xi}, \bar{\theta}) \geq \beta Q_1(\cdot)$ for all $\bar{\xi} \in E$ and $\bar{\theta} \in B$. Therefore, on $E \times B$, we can write $Q(\cdot \mid \bar{\xi}, \bar{\theta})$ as the convex combination of Q_1 and another stochastic kernel Q_0 as follows:

$$Q(\cdot \mid \bar{\xi}, \bar{\theta}) = \beta\, Q_1(\cdot) + (1 - \beta)\, Q_0(\cdot \mid \bar{\xi}, \bar{\theta}), \qquad \forall\, \bar{\xi} \in E,\ \bar{\theta} \in B, \tag{64}$$

where $Q_0(\cdot \mid \bar{\xi}, \bar{\theta}) = \big[Q(\cdot \mid \bar{\xi}, \bar{\theta}) - \beta\, Q_1(\cdot)\big]/(1 - \beta)$ and Q_0 is a stochastic kernel on B given $E \times B$.

Next, independently of $\{Z_t\}$, generate a sequence $\{Y_t\}_{t\geq 1}$ of i.i.d., $\{0,1\}$-valued random variables such that $Y_t = 1$ with probability β and $Y_t = 0$ with probability $1-\beta$. Set $Y_0 = 0$. Let

$$t_Y = \min\{t \geq 1 \mid Y_t = 1, \xi_{t-1} \in E\}.$$

Then $t_Y < \infty$ with probability one. (Since $\{\xi_t\}$ visits E infinitely often and the process $\{Y_t\}$ is independent of $\{\xi_t\}$, this follows easily from applying the Borel-Cantelli lemma to $\{(\xi_{t_k}, Y_{t_k+1})\}_{k\geq 1}$, where t_k is when the k-th visit to E by $\{\xi_t\}$ occurs.)

Now for each $t \geq 0$, let us define the pair $(\tilde{\theta}_{t+1}, \tilde{\theta}'_{t+1})$ according to the following rule, based on the values of $(\xi_0, \tilde{\theta}_0, \tilde{\theta}'_0), \ldots, (\xi_t, \tilde{\theta}_t, \tilde{\theta}'_t)$ and $(Y_0, \ldots, Y_t, Y_{t+1})$:

(i) In the case $t < t_Y$ and $\xi_t \notin E$, generate $\tilde{\theta}_{t+1}$ and $\tilde{\theta}'_{t+1}$ according to $Q(\cdot \mid \xi_t, \tilde{\theta}_t)$ and $Q(\cdot \mid \xi_t, \tilde{\theta}'_t)$ respectively.

(ii) In the case $t < t_Y$ and $\xi_t \in E$, if $Y_{t+1} = 0$, generate $\tilde{\theta}_{t+1}$ and $\tilde{\theta}'_{t+1}$ according to $Q_0(\cdot \mid \xi_t, \tilde{\theta}_t)$ and $Q_0(\cdot \mid \xi_t, \tilde{\theta}'_t)$ respectively; if $Y_{t+1} = 1$, generate $\tilde{\theta}_{t+1}$ according to $Q_1(\cdot)$ and let $\tilde{\theta}'_{t+1} = \tilde{\theta}_{t+1}$.

(iii) In the case $t \geq t_Y$, generate $\tilde{\theta}_{t+1}$ according to $Q(\cdot \mid \xi_t, \tilde{\theta}_t)$ and let $\tilde{\theta}'_{t+1} = \tilde{\theta}_{t+1}$.

In view of (64), it can be verified directly by induction on t that the marginal process $\{(Z_t, \tilde{\theta}_t)\}$ (resp. $\{(Z_t, \tilde{\theta}'_t)\}$) in the preceding construction has the same probability distribution as $\{X_t\}$ (resp. $\{X'_t\}$). This implies that $\{\mu_{x,t}\}$ (resp. $\{\mu_{x',t}\}$) converges weakly to μ with probability one if and only if $\{\tilde{\mu}_{x,t}\}$ (resp. $\{\tilde{\mu}_{x',t}\}$) converges weakly to μ with probability one. On the other hand, by construction $\tilde{\theta}_t = \tilde{\theta}'_t$ for $t \geq t_Y$, where $t_Y < \infty$ with probability one, so except on a null set, $\{\tilde{\mu}_{x,t}\}$ and $\{\tilde{\mu}_{x',t}\}$ have the same weak limits. Combining these two arguments with the assumption that $\{\mu_{x,t}\}$ converges weakly to μ with probability one, it follows that the three sequences $\{\tilde{\mu}_{x,t}\}$, $\{\tilde{\mu}_{x',t}\}$, and $\{\mu_{x',t}\}$ must all converge weakly to μ with probability one. ■

Proof of Prop. 8 We suppress the superscript α of θ^α_t in the proof. Let $\{X_t\} = \{(Z_t, \theta_t)\}$. By Prop. 6, the set $\mathcal{M}_\alpha$ of invariant probability measures of $\{X_t\}$ is nonempty. Recall also that since the evolution of $\{Z_t\}$ is not affected by the θ-iterates, the marginal of any $\mu \in \mathcal{M}_\alpha$ on the space $\mathcal{Z}$ must equal ζ, the unique invariant probability measure of $\{Z_t\}$ (Theorem 2).

Suppose $\{X_t\}$ has multiple invariant probability measures; i.e., there exist $\mu, \mu' \in \mathcal{M}_\alpha$ with $\mu \neq \mu'$. Then by (Dudley, 2002, Theorem 11.3.2) there exists a bounded continuous function f on $\mathcal{Z} \times B$ such that

$$\int f \, d\mu \neq \int f \, d\mu'. \tag{65}$$

On the other hand, since μ is an invariant probability measure of $\{X_t\}$, applying a strong law of large numbers for stationary processes (Doob, 1953, Chap. X, Theorem 2.1; see also Meyn and Tweedie, 2009, Lemma 17.1.1 and Theorem 17.1.2) to the stationary Markov chain $\{X_t\}$ with initial distribution μ, we have that there exist a set $D_1 \subset \mathcal{Z} \times B$ with $\mu(D_1) = 1$ and a measurable function g_f on $\mathcal{Z} \times B$ such that

(i) for each $x \in D_1$, with the initial condition $X_0 = x$, $\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} f(X_k) = g_f(x)$, $\mathbf{P}_x$-a.s.;

(ii) $\mathbb{E}_\mu[g_f(X_0)] = \mathbb{E}_\mu[f(X_0)]$ (i.e., $\int g_f d\mu = \int f d\mu$).

The same is true for the invariant probability measure μ': there exist a set $D_2 \subset \mathcal{Z} \times B$ with $\mu'(D_2) = 1$ and a measurable function $g'_f(x)$ such that

(i) for each $x \in D_2$, with the initial condition $X_0 = x$, $\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} f(X_k) = g'_f(x)$, $\mathbf{P}_x$-a.s.;

(ii) $\mathbb{E}_{\mu'}[g'_f(X_0)] = \mathbb{E}_{\mu'}[f(\mathbf{X}_0)]$ (i.e., $\int g'_f d\mu' = \int f d\mu'$).

Also, since $\{X_t\}$ is a weak Feller Markov chain (Lemma 6), by (Meyn, 1989, Prop. 4.1), for a set of initial conditions x with μ-measure 1, the occupation probability measures $\{\mu_{x,t}\}$ of $\{X_t\}$ converge weakly, $\mathbf{P}_x$-almost surely, to some (nonrandom) $\tilde{\mu}_x \in \mathcal{M}_\alpha$ that depends on the initial x. The same is true for μ'. So by excluding from D_1 a μ-null set and from D_2 a μ'-null set if necessary, we can assume that the sets D_1, D_2 above also satisfy that for each $x \in D_1 \cup D_2$, the occupation probability measures $\{\mu_{x,t}\}$ converge weakly to an invariant probability measure $\tilde{\mu}_x$ almost surely. Then since $\frac{1}{t} \sum_{k=0}^{t-1} f(X_k)$ is the same as $\int f d\mu_{x,t}$ for $X_0 = x$, we have, by the weak convergence of $\{\mu_{x,t}\}$ just discussed, that

$$g_f(x) = \int f d\tilde{\mu}_x \quad \text{for each } x \in D_1, \qquad g'_f(x) = \int f d\tilde{\mu}_x \quad \text{for each } x \in D_2. \tag{66}$$

Certainly we must have $g_f(x) = g'_f(x)$ on $D_1 \cap D_2$. We now relate the values of these two functions at points that share the same z-component. In particular, let $\text{proj}(D_1)$ denote the projection of D_1 on $\mathcal{Z}$: $\text{proj}(D_1) = \{z \in \mathcal{Z} \mid \exists\, \theta \text{ with } (z, \theta) \in D_1\}$, and let $D_{1,z}$ be the vertical section of D_1 at z: $D_{1,z} = \{\theta \mid (z, \theta) \in D_1\}$. Define $\text{proj}(D_2)$ and $D_{2,z}$ similarly. If $x = (z, \theta) \in D_1 \cup D_2$ and $x' = (z, \theta') \in D_1 \cup D_2$, then in view of Lemma 9 and the weak convergence of $\{\mu_{x,t}\}$ and $\{\mu_{x',t}\}$, we must have $\tilde{\mu}_x = \tilde{\mu}_{x'}$. Consequently, by (66), for each $z \in \text{proj}(D_1)$, $g_f(z, \cdot)$ is constant on $D_{1,z}$; for each $z \in \text{proj}(D_2)$, $g'_f(z, \cdot)$ is constant on $D_{2,z}$; and for each $z \in \text{proj}(D_1) \cap \text{proj}(D_2)$, the constants that $g_f(z, \cdot)$, $g'_f(z, \cdot)$ take on $D_{1,z}$, $D_{2,z}$, respectively, are the same.

We now show $\int f d\mu = \int f d\mu'$ to contradict (65) and finish the proof. Since $\mu(D_1) = \mu'(D_2) = 1$ and by Theorem 2 (Section 2.4) μ, μ' have the same marginal distribution on $\mathcal{Z}$, which is ζ, there exists a Borel set $E \subset \text{proj}(D_1) \cap \text{proj}(D_2)$ with $\zeta(E) = 1$. Consider the sets $(E \times B) \cap D_1$ and $(E \times B) \cap D_2$, which have μ-measure 1 and μ'-measure 1, respectively. By (Dudley, 2002, Prop. 10.2.8), we can decompose μ, μ' into the marginal ζ on $\mathcal{Z}$ and the conditional distributions $\mu(d\theta \mid z), \mu'(d\theta \mid z)$ for $z \in \mathcal{Z}$. Then

$$1 = \mu\big((E \times B) \cap D_1\big) = \int_E \int_{D_{1,z}} \mu(d\theta \mid z)\, \zeta(dz), \;\; 1 = \mu'\big((E \times B) \cap D_2\big) = \int_E \int_{D_{2,z}} \mu'(d\theta \mid z)\, \zeta(dz),$$

where the equality for the iterated integral in each relation follows from (Dudley, 2002, Theorem 10.2.1(ii)). These relations imply that for some set $E_0 \subset E$ with $\zeta(E_0) = 0$,

$$\int_{D_{1,z}} \mu(d\theta \mid z) = \int_{D_{2,z}} \mu'(d\theta \mid z) = 1, \qquad \forall\, z \in E \setminus E_0. \tag{67}$$

We now calculate $\int g_f d\mu$ and $\int g'_f d\mu'$. We have

$$\int g_f \, d\mu = \int_{(E\times B)\cap D_1} g_f \, d\mu = \int_E \int_{D_{1,z}} g_f(z, \theta)\, \mu(d\theta \mid z)\, \zeta(dz), \tag{68}$$

$$\int g'_f \, d\mu' = \int_{(E\times B)\cap D_2} g'_f \, d\mu' = \int_E \int_{D_{2,z}} g'_f(z, \theta)\, \mu'(d\theta \mid z)\, \zeta(dz), \tag{69}$$

where the equality for the iterated integral in each relation also follows from (Dudley, 2002, Theorem 10.2.1(ii)). As discussed earlier, for each $z \in E \subset \text{proj}(D_1) \cap \text{proj}(D_2)$, the two constant functions, $g_f(z,\cdot)$ on $D_{1,z}$ and $g'_f(z,\cdot)$ on $D_{2,z}$, have the same value. Using this together with (67), we conclude that

$$\int_{D_{1,z}} g_f(z,\theta)\,\mu(d\theta \mid z) = \int_{D_{2,z}} g'_f(z,\theta)\,\mu'(d\theta \mid z), \qquad \forall\, z \in E \setminus E_0. \tag{70}$$

Since $\zeta(E_0) = 0$, we obtain from (68)-(70) that $\int g_f d\mu = \int g'_f d\mu'$. But $\int g_f d\mu = \int f d\mu$ and $\int g'_f d\mu' = \int f d\mu'$ (as we obtained at the beginning of the proof), so $\int f d\mu = \int f d\mu'$, a contradiction to (65). This proves that $\{X_t\}$ must have a unique invariant probability measure. ■

Proposition 8 implies that for every $m \geq 1$, the m-step version of $\{(Z_t, \theta_t^\alpha)\}$ has a unique invariant probability measure. This together with Lemma 7 (Section 4.3.2) furnishes the conditions (A1)-(A3) of (Meyn, 1989, Prop. 4.2) for weak Feller Markov chains (these conditions are the conditions (i)-(iii) of our Lemma 5). We can therefore apply the conclusions of (Meyn, 1989, Prop. 4.2) (see Lemma 5 in our Section 4.3.1) to the m-step version of $\{(Z_t, \theta_t^\alpha)\}$ here, and the result is the following proposition:

Proposition 9 *Under Assumption 1, for each $m \geq 1$, the m-step version of $\{(Z_t, \theta_t^\alpha)\}$ has a unique invariant probability measure $\mu^{(m)}$, and the occupation probability measures $\mu^{(m)}_{(z,\theta),t}, t \geq 1$, as defined by (51), converge weakly to $\mu^{(m)}$ almost surely, for each initial condition $(z,\theta) \in \mathcal{Z} \times B$ of (Z_0, θ_0^α).*

With Prop. 9 we can proceed to prove the second part of Theorems 10 and 11. Given that we have already established their first part in the previous subsection, the arguments for their second part are the same for both theorems and are given below. The proof is similar to that for Theorem 8(ii) in Section 4.3.2, except that here, instead of working with the averaged probability measures $\{\bar{P}^{(m,k)}_{(z,\theta)}\}$, Prop. 9 allows us to work with the occupation probability measures.

Proof of the second part of both Theorem 10 and Theorem 11 We suppress the superscript α of θ_t^α in the proof. By Prop. 9, $\{(Z_t, \theta_t)\}$ has a unique invariant probability measure μ_α, and its m-step version has a corresponding unique invariant probability measure $\mu_\alpha^{(m)}$. We prove first the statement that for each initial condition $(z,\theta) \in \mathcal{Z} \times B$, almost surely,

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \mathbb{1}\Big(\sup_{k \leq j < k+m} \big|\theta_j - \theta^*\big| < \delta \Big) \geq \bar{\mu}_\alpha^{(m)}\big([N'_\delta(\theta^*)]^m\big), \tag{71}$$

where $\bar{\mu}_\alpha^{(m)}$ is the unique element in $\bar{\mathcal{M}}_\alpha^m$, and $N'_\delta(\theta^*)$ is the open δ-neighborhood of θ^*. For each t, by the definition (51) of the occupation probability measure $\mu^{(m)}_{(z,\theta),t}$, the average in the left-hand side above is the same as $\mu^{(m)}_{(z,\theta),t}(D_\delta)$, where $D_\delta = \big\{\big(z^1, \theta^1, \ldots, z^m, \theta^m\big) \in (\mathcal{Z} \times B)^m \;\big|\; \sup_{1 \leq j \leq m} \big|\theta^j - \theta^*\big| < \delta\big\}$. By Prop. 9, $\mathbf{P}_{(z,\theta)}$-almost surely, $\{\mu^{(m)}_{(z,\theta),t}\}$ converges

weakly to $\mu_\alpha^{(m)}$, and therefore, except on a null set of sample paths, we have by (Dudley, 2002, Theorem 11.1.1) that for the open set D_δ,

$$\liminf_{t\to\infty} \mu_{(z,\theta),t}^{(m)}(D_\delta) \geq \mu_\alpha^{(m)}(D_\delta) = \bar{\mu}_\alpha^{(m)}\big([N'_\delta(\theta^*)]^m\big).$$

This proves (71).

We now prove the statement that for each initial condition $(z,\theta) \in \mathcal{Z}\times B$, almost surely,

$$\limsup_{t\to\infty} \big|\bar{\theta}_t - \theta^*\big| \leq \delta\,\kappa_\alpha + 2r_B\,(1-\kappa_\alpha), \qquad \text{where } \kappa_\alpha = \bar{\mu}_\alpha\big(N'_\delta(\theta^*)\big), \tag{72}$$

and $\bar{\mu}_\alpha$ is the marginal of μ_α on B. The statement is trivially true if $\delta \geq 2r_B$, so consider the case $\delta < 2r_B$. Fix an initial condition $(z,\theta) \in \mathcal{Z}\times B$ for (Z_0,θ_0), and let $\{\mu_{(z,\theta),t}\}$ be the corresponding occupation probability measures of $\{(Z_t,\theta_t)\}$. For the averaged sequence $\{\bar{\theta}_t\}$, by convexity of the norm $|\cdot|$,

$$\big|\bar{\theta}_t - \theta^*\big| \leq \frac{1}{t}\sum_{k=0}^{t-1} |\theta_k - \theta^*|. \tag{73}$$

We have

$$\begin{aligned}\frac{1}{t}\sum_{k=0}^{t-1} |\theta_k-\theta^*| &\leq \frac{1}{t}\sum_{k=0}^{t-1} |\theta_k-\theta^*|\cdot \mathbb{1}\big(\theta_k \in N'_\delta(\theta^*)\big) + \frac{1}{t}\sum_{k=0}^{t-1} |\theta_t-\theta^*|\cdot \mathbb{1}\big(\theta_k \notin N'_\delta(\theta^*)\big)\\ &\leq \delta\cdot \mu_{(z,\theta),t}(D_\delta) + 2r_B\cdot\big(1-\mu_{(z,\theta),t}(D_\delta)\big),\end{aligned} \tag{74}$$

where $D_\delta = \big\{(z^1,\theta^1)\in\mathcal{Z}\times B \;\big|\; |\theta^1-\theta^*|<\delta\big\}$. By Prop. 9, $\mathbf{P}_{(z,\theta)}$-almost surely, $\{\mu_{(z,\theta),t}\}$ converges weakly to μ_α. Therefore, except on a null set of sample paths, we have by (Dudley, 2002, Theorem 11.1.1) that for the open set D_δ,

$$\liminf_{t\to\infty} \mu_{(z,\theta),t}(D_\delta) \geq \mu_\alpha(D_\delta) = \bar{\mu}_\alpha\big(N'_\delta(\theta^*)\big). \tag{75}$$

Combining the three inequalities (73)-(75), and using also the relation $\delta < 2r_B$, we obtain that (72) holds almost surely for each initial condition $(z,\theta)\in\mathcal{Z}\times B$. ■

Remark 3 (on the role of perturbation again) As mentioned before Prop. 8, our purpose of perturbing the constrained ETD algorithms is to guarantee that the Markov chain $\{(Z_t,\theta_t^\alpha)\}$ has a unique invariant probability measure. Without the perturbation, this cannot be ensured, so we cannot apply the ergodic theorem given in Lemma 5 to exploit the convergence of occupation probability measures, as we did in the preceding proof, even though $\{(Z_t,\theta_t^\alpha)\}$ satisfies the remaining two conditions required by that ergodic theorem (cf. Lemma 7, Section 4.3.2).

In connection with this discussion, let us clarify a point. We know that the occupation probability measures of $\{Z_t\}$ converge weakly to its unique invariant probability measure ζ almost surely for each initial condition of Z_0 (Theorem 2). But this fact alone cannot rule out the possibility that $\{(Z_t,\theta_t^\alpha)\}$ has multiple invariant probability measures and that its occupation probability measures do not converge for some initial condition (z,θ).

Finally, another property of weak Feller Markov chains and its implication for our problem are worth noting here. By (Meyn, 1989, Prop. 4.1), for a weak Feller Markov chain $\{X_t\}$, provided that an invariant probability measure μ exists, we have that for a set of initial conditions x with μ-measure 1, the occupation probability measures $\{\mu_{x,t}\}$ converge weakly, $\mathbf{P}_x$-almost surely, to an invariant probability measure μ_x that depends on the initial condition. Thus, for the unperturbed algorithms (11), (19) and (20), despite the possibility of $\{(Z_t, \theta_t^\alpha)\}$ having multiple invariant probability measures, the preceding proof can be applied to those initial conditions from which the occupation probability measures converge almost surely. In particular, this argument leads to the following conclusion. In the case of the algorithm (11), (19) or (20), under the same conditions as in Theorem 8 or 9, it holds for any invariant probability measure μ of $\{(Z_t, \theta_t^\alpha)\}$ that for each initial condition (z, θ) from some set of initial conditions with μ-measure 1,

$$\limsup_{t\to\infty} \left|\bar{\theta}_t^\alpha - \theta^*\right| \leq \delta\,\kappa_\alpha + 2r_B\,(1-\kappa_\alpha) \quad \mathbf{P}_{(z,\theta)}\text{-a.s.},$$

where $\kappa_\alpha = \inf_{\mu\in\bar{\mathcal{M}}_\alpha} \mu(N'_\delta(\theta^*))$. The limitation of this result, however, is that the set of initial conditions involved is unknown and can be small. ■

5. Discussion

In this section we discuss direct applications of our convergence results to ETD(λ) under relaxed conditions and to two other algorithms, the off-policy TD(λ) algorithm and the ETD(λ, β) algorithm (Hallak et al., 2016). We then discuss several open issues to conclude the paper.

5.1 The Case without Assumption 2

Let Assumption 1 hold. Recall from Section 2.3 that ETD(λ) aims to solve the equation $C\theta + b = 0$, where

$$b = \Phi^\top \bar{M}\, r_{\pi,\gamma}^\lambda, \qquad C = -\Phi^\top G\,\Phi \quad \text{with} \quad G = \bar{M}(I - P_{\pi,\gamma}^\lambda).$$

In this paper we have focused on the case where Assumption 2 holds and C is negative definite (Theorem 1, Section 2.3). If Assumption 2 does not hold, then either there are less than n emphasized states (i.e., states s with $\bar{M}_{ss} > 0$), or the feature vectors of emphasized states are not rich enough to contain n linearly independent vectors. In either case the function approximation capacity is not fully utilized. It is hence desirable to fulfill Assumption 2 by adding more states with positive interest weights $i(s)$ or by enriching the feature representation.

Nevertheless, suppose Assumption 2 does not hold (in which case C is negative semidefinite as shown by Sutton et al., 2016). This essentially has no effects on the convergence properties of the constrained or unconstrained ETD(λ) algorithms, because of the emphatic weighting scheme (3)-(5), as we explain now.

Let there be at least one state s with interest weight $i(s) > 0$ (the case is vacuous otherwise). Partition the state space into the set of emphasized states and the set of non-emphasized states:

$$\mathcal{J}_1 = \{s \in \mathcal{S} \mid \bar{M}_{ss} > 0\}, \qquad \mathcal{J}_0 = \{s \in \mathcal{S} \mid \bar{M}_{ss} = 0\}.$$

Corresponding to the partition, by rearranging the indices of states if necessary, we can write

$$\Phi = \begin{bmatrix} \Phi_1 \\ \Phi_0 \end{bmatrix}, \qquad r^{\lambda}_{\pi,\gamma} = \begin{bmatrix} r_1 \\ r_0 \end{bmatrix}, \qquad \bar{M} = \begin{bmatrix} \hat{M} & 0_{|\mathcal{J}_1|\times|\mathcal{J}_0|} \\ 0_{|\mathcal{J}_0|\times|\mathcal{J}_1|} & 0_{|\mathcal{J}_0|\times|\mathcal{J}_0|} \end{bmatrix},$$

where $0_{m\times m'}$ denotes an $m \times m'$ zero matrix, $\hat{M}$ is a diagonal matrix with $\bar{M}_{ss}, s \in \mathcal{J}_1$, as its diagonals. Let $\hat{Q}$ be the sub-matrix of $P^{\lambda}_{\pi,\gamma}$ that consists of the entries whose row/column indices are in $\mathcal{J}_1$. For the equation $C\theta + b = 0$, clearly $b = \Phi_1^\top \hat{M} r_1$. Consider now the matrix C. It is shown in the proof of Prop. C.2 in (Yu, 2015a) that G has a block-diagonal structure with respect to the partition $\{\mathcal{J}_1, \mathcal{J}_0\}$,

$$G = \begin{bmatrix} \hat{G} & 0_{|\mathcal{J}_1|\times|\mathcal{J}_0|} \\ 0_{|\mathcal{J}_0|\times|\mathcal{J}_1|} & 0_{|\mathcal{J}_0|\times|\mathcal{J}_0|} \end{bmatrix},$$

where the block corresponding to $\mathcal{J}_0$ is a zero matrix as shown above, and the block $\hat{G}$ corresponding to $\mathcal{J}_1$ is a positive definite matrix given by

$$\hat{G} = \hat{M}(I - \hat{Q}), \tag{76}$$

and $\hat{M}$ can be expressed explicitly as

$$diag(\hat{M}) = {d^1_{\pi^o,i}}^\top (I - \hat{Q})^{-1}, \qquad d^1_{\pi^o,i} \in \mathbb{R}^{|\mathcal{J}_1|},\ d^1_{\pi^o,i}(s) = d_{\pi^o}(s) \cdot i(s),\ s \in \mathcal{J}_1. \tag{77}$$

Thus the matrix C has a special structure:

Theorem 12 (structure of the matrix C; Yu, 2015a, Appendix C.2, p. 41-44) *Let Assumption 1 hold, and let $i(s) > 0$ for at least one state $s \in \mathcal{S}$. Then*

$$C = -\Phi_1^\top \hat{G} \Phi_1, \qquad \textit{where } \hat{G} = \hat{M}(I - \hat{Q}) \textit{ is positive definite.}$$

Let range(A) denote the range space of a matrix A. By the positive definiteness of the matrix $\hat{G}$ given in the preceding theorem, the negative semidefinite matrix C possesses the following properties (we omit the straightforward proof):

Proposition 10 *Let Assumption 1 hold, and let $i(s) > 0$ for at least one state $s \in \mathcal{S}$. Then the matrix C satisfies that*

(i) *range(C) = range$(C^\top)$ = span$\{\phi(s) \mid s \in \mathcal{J}_1\}$; and*

(ii) *there exists $c > 0$ such that for all $x \in$ span$\{\phi(s) \mid s \in \mathcal{J}_1\}$, $x^\top C x \le -c\,|x|^2$.*

Two observations then follow immediately:

(i) Since $b = \Phi_1^\top \hat{M} r_1 \in \text{span}\{\phi(s) | s \in \mathcal{J}_1\}$, Prop. 10(i) shows that the equation $C\theta + b = 0$ admits a solution, and a unique one in $\text{span}\{\phi(s) \mid s \in \mathcal{J}_1\}$, which we denote by θ^*.[15]

15. From the structures of G, $P^{\lambda}_{\pi,\gamma}$, $\hat{Q}$ and $\hat{M}$ shown in (Yu, 2015a, Appendix C.2, p. 41-44), which give rise to (76)-(77), we also have the following facts. The approximate value function $v = \Phi_1\theta^*$ for the emphasized states $\mathcal{J}_1$ is the unique solution of the projected Bellman equation $v = \Pi(r_1 + \hat{Q}v)$, where Π is the projection onto the column space of Φ_1 with respect to the weighted Euclidean norm on $\mathbb{R}^{|\mathcal{J}_1|}$ defined by the weights $\bar{M}_{ss}, s \in \mathcal{J}_1$ (the diagonals of $\hat{M}$). The equation $v = r_1 + \hat{Q}v$ is indeed a generalized Bellman equation for the emphasized states only, and has $v_\pi(s), s \in \mathcal{J}_1$, as its unique solution. Then for the emphasized states, the relation between the approximate value function $\Phi_1\theta^*$ and v_π on $\mathcal{J}_1$, in particular the approximation error, can again be characterized using the oblique projection viewpoint (Scherrer, 2010), similar to the case with Assumption 2 discussed in Section 2.3.

(ii) Prop. 10(ii) shows that C acts like a negative definite matrix on the space of feature vectors, $\text{span}\{\phi(s)|s \in \mathcal{J}_1\}$, that the ETD($\lambda$) algorithms naturally operate on.[16]

We remark that for an arbitrary negative semidefinite matrix C, neither of these conclusions holds. They hold here as direct consequences of the positive definiteness of the matrix $\hat{G}$ that underlies C, and this positive definiteness property is due to the emphatic weighting scheme (3)-(5) employed by ETD(λ).

Now let us discuss the behavior of the constrained ETD(λ) algorithms starting from some state S_0 of interest (i.e., $i(S_0) > 0$), in the absence of Assumption 2. Recall that earlier we did not need Assumption 2 when applying the two general convergence theorems from (Kushner and Yin, 2003), and we used the negative definiteness of C implied by this assumption only near the end of our proofs to get the solution properties of the mean ODE associated with each algorithm. In the absence of Assumption 2, for the unperturbed algorithms (11), (19) and (20), we can simply restrict attention to the subspace $\text{span}\{\phi(s)|s \in \mathcal{J}_1\}$ and use the property in Prop. 10(ii) in lieu of negative definiteness. After all, the θ-iterates of these algorithms always lie in the span of the feature vectors if the initial $\theta_0, e_0 \in \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$ and in the case of the two biased algorithms (19) and (20), if the function $\psi_K(x)$ does not change the direction of x. On the subspace $\text{span}\{\phi(s)|s \in \mathcal{J}_1\}$, in view of Prop. 10(ii), the function $|\theta - \theta^*|^2$ serves again as a Lyapunov function for analyzing the ODE solutions in exactly the same way as before. Thus, in the absence of Assumption 2, for the algorithms (11), (19) and (20) that set θ_0, e_0 and ψ_K as just described, and for $r_B > |b|/c$ where c is as in Prop. 10(ii), the conclusions of Theorems 4-9 in Section 3 continue to hold with $N_\delta(\theta^*)$ or $N'_\delta(\theta^*)$ replaced by $N_\delta(\theta^*) \cap \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$ or $N'_\delta(\theta^*) \cap \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$.

The same is true for the almost sure convergence of the unconstrained ETD(λ) algorithm (2) under diminishing stepsize: with $i(S_0) > 0$ and $\theta_0, e_0 \in \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$, the conclusion of (Yu, 2015a, Theorem 2.2) continues to hold in the absence of Assumption 2; that is, for $\alpha_t = O(1/t)$ with $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$, $\theta_t \overset{a.s.}{\to} \theta^*$.

It can be seen now that without Assumption 2, complications can only arise through initializing the algorithms outside the desired subspace. We discussed such situations in the arXiv version of this paper (Yu, 2015b, Sec. 5.1), but we shall omit them here in part because it does not seem natural to initialize θ_0, e_0 with a component perpendicular to $\text{span}\{\phi(s)|s \in \mathcal{J}_1\}$ in the first place.

As a final note, in the absence of Assumption 2, any solution $\bar{\theta}$ of $C\theta + b = 0$ gives *the same approximate value function for emphasized states*, but the approximate values $\Phi_0\bar{\theta}$ for non-emphasized states in $\mathcal{J}_0$ are *different* for different solutions $\bar{\theta}$. Thus one needs to be cautious in using the approximate values $\Phi_0\bar{\theta}$. They correspond to different extrapolations from the approximate values $\Phi_1\theta^*$ for the emphasized states, whereas $\Phi_1\theta^*$ is not defined to take into account approximation errors for those states in $\mathcal{J}_0$, although its approximation error for emphasized states can be well characterized (cf. Footnote 15).

16. Start ETD(λ) from a state S_0 with $i(S_0) > 0$. It can be verified that the emphatic weighting scheme dictates that if $S_t \in \mathcal{J}_0$, then the emphasis weight M_t for that state must be zero. Consequently, e_t is a linear combination of the features of the emphasized states and the initial e_0. So when $e_0 \in \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$, $e_t \in \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$ always, and if in addition $\theta_0 \in \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$, then $\theta_t \in \text{span}\{\phi(s)|s \in \mathcal{J}_1\}$ always. This is very similar to the case of TD(λ) with possibly linearly dependent features discussed in (Tsitsiklis and Van Roy, 1997).

5.2 Off-policy TD(λ) and ETD(λ, β)

Applying TD(λ) to off-policy learning by using importance sampling techniques was first proposed in (Precup et al., 2000, 2001), and the focus there was on episodic data. The analysis we gave in this paper applies directly to the (non-episodic) off-policy TD(λ) algorithm studied in (Bertsekas and Yu, 2009; Yu, 2012; Dann et al., 2014), when its divergence issue is avoided by setting λ sufficiently large. Specifically, we consider constant $\gamma \in [0, 1)$ and constant $\lambda \in [0, 1]$, and an infinitely long trajectory generated by the behavior policy as before. The algorithm is the same as TD(λ) except for incorporating the importance sampling weight ρ_t:[17]

$$\theta_{t+1} = \theta_t + \alpha_t \, e_t \cdot \rho_t \left(R_t + \gamma \phi(S_{t+1})^\top \theta_t - \phi(S_t)^\top \theta_t \right),$$

where

$$e_t = \lambda \gamma \rho_{t-1} \, e_{t-1} + \phi(S_t).$$

The constrained versions of the algorithm are defined similarly to those for ETD(λ).

Under Assumption 1(ii), the associated projected Bellman equation is the same as that for on-policy TD(λ) (Tsitsiklis and Van Roy, 1997) except that the projection norm is the weighted Euclidean norm with weights given by the steady state probabilities $d_{\pi^o}(s), s \in \mathcal{S}$. Assuming Φ has full column rank, the corresponding equation in the θ-space, $C\theta + b = 0$, has the desired property that the matrix C is negative definite, if λ is sufficiently large (in particular if $\lambda = 1$) (Bertsekas and Yu, 2009). For that case, the conclusions given in this paper for constrained ETD(λ) all hold for the corresponding versions of off-policy TD(λ). (Similarly, for the case of C being negative semidefinite due to Φ having rank less than n, the discussion given in the previous subsection for ETD(λ) also applies.) The reason is that besides the property of C, the other properties of the iterates that we used in our analysis, which are given in Section 2 and Appendix A, all hold for off-policy TD(λ). In fact, some of these properties were first derived for off-policy LSTD(λ) and TD(λ) in (Yu, 2012) and extended later in (Yu, 2015a) to ETD(λ).

For the same reason, the convergence analyses we gave in (2015a) and this paper for ETD also apply to a variation of the ETD algorithm, ETD(λ, β), proposed recently by Hallak et al. (2016), when the parameter β is set in an appropriate range.

5.3 Open Issues

A major difficulty in applying off-policy TD learning, especially with $\lambda > 0$, is the high variances of the iterates. For ETD(λ), off-policy TD(λ) and their least-squares versions, because of the growing variances of products of the importance sampling weights $\rho_t \rho_{t+1} \cdots$ along a trajectory, and because of the amplifying effects these weights can have on the traces, the variances of the traces iterates can grow unboundedly with time, severely affecting the behavior of the algorithms in practice. (The problem of growing variances when applying

17. It is not necessary to multiply the term $\phi(S_t)^\top \theta_t$ by ρ_t, and that version of the algorithm was the one given in (Bertsekas and Yu, 2009; Yu, 2012). The experimental results in (Dann et al., 2014) suggest to us that each version can have less variance than the other in some occasions, however. As far as convergence analysis is concerned, the two versions are essentially the same and the analyses given in (Yu, 2012, 2015a) and this paper indeed apply simultaneously to both versions of the algorithm.

importance sampling to simulate Markov systems was also known earlier and discussed in prior works; see e.g., Glynn and Iglehart, 1989; Randhawa and Juneja, 2004.) The two biased constrained algorithms discussed in this paper were motivated by the need to mitigate the variance problem, and their robust behavior has been observed in our experiments (Mahmood et al., 2015; Yu, 2016). However, beyond simply constraining the iterates, more variance reduction techniques are needed, such as control variates (Randhawa and Juneja, 2004; Ahamed et al., 2006) and weighted importance sampling (Precup et al., 2000, 2001; Mahmood et al., 2014; Mahmood and Sutton, 2015). To overcome the variance problem in off-policy learning, further research is required.

Regarding convergence analysis of ETD(λ), the results we gave in (2015a) and this paper concern only the convergence properties and not the rates of convergence. For on-policy TD(λ) and LSTD(λ), convergence rate analyses are available (Konda, 2002, Chap. 6). Such analyses in the off-policy case will give us better understanding of the asymptotic behavior of the off-policy algorithms. Finally, besides asymptotic behavior of the algorithms, their finite-time or finite-sample properties (such as those considered by Munos and Szepesvári, 2008; Antos et al., 2008; Lazaric et al., 2012; Liu et al., 2015), and their large deviations properties are also worth studying.

Acknowledgments

I thank Professors Richard Sutton and Csaba Szepesvári for helpful discussions, and I thank the anonymous reviewers for their helpful feedback. This research was supported by a grant from Alberta Innovates—Technology Futures.

Appendix A. Key Properties of Trace Iterates

In this appendix we list four key properties of trace iterates $\{(e_t, F_t)\}$ generated by the ETD(λ) algorithm. Three of them were derived in (Yu, 2015a, Appendix A), and used in the convergence analysis of ETD(λ) in both (Yu, 2015a) and the present paper.

As discussed in Section 3.2, $\{(e_t, F_t)\}$ can have unbounded variances and is naturally unbounded in common off-policy situations. However, as the proposition below shows, $\{(e_t, F_t)\}$ is bounded in a stochastic sense.

Proposition 11 *Under Assumption 1, given a bounded set $E \subset \mathbb{R}^{n+1}$, there exists a constant $L < \infty$ such that if the initial $(e_0, F_0) \in E$, then $\sup_{t\geq 0} \mathbb{E}\big[\big\|(e_t, F_t)\big\|\big] < L$.*

The preceding proposition is the same as (Yu, 2015a, Prop. A.1) except that the conclusion is for all the initial (e_0, F_0) from the set E, instead of a fixed initial (e_0, F_0). By making explicit the dependence of the constant L on the initial (e_0, F_0), the same proof of (Yu, 2015a, Prop. A.1) (which is a relatively straightforward calculation) applies to the preceding proposition.

We note that Prop. 11 does not imply the *uniform integrability of* $\{(e_t, F_t)\}$—this stronger property does hold for the trace iterates, as we proved in Prop. 2(i), Section 4.1.2. (The latter and its proof focus on $\{e_t\}$ only, but the same argument applies to $\{(e_t, F_t)\}$.)

The next proposition concerns the change in the trace iterates due to the change in its initial condition. It is the same as (Yu, 2015a, Prop. A.2); its proof is more involved than the proofs of the two other properties of the trace iterates and uses, among others, a theorem for nonnegative random processes (Neveu, 1975). We did not use this proposition directly in the analysis of the present paper, but it is important in establishing that the Markov chain $\{Z_t\}$ has a unique invariant probability measure (Theorem 2, Section 2.4), which the results of the present paper rely on. In addition, it is helpful for understanding the behavior of the trace iterates.

Let $(\hat{e}_t, \hat{F}_t)$, $t \geq 1$, be defined by the same recursion (3)-(5) that defines (e_t, F_t), using the same state and action random variables $\{(S_t, A_t)\}$, but with a different initial condition $(\hat{e}_0, \hat{F}_0)$. We write a zero vector in any Euclidean space as $\mathbf{0}$.

Proposition 12 *Under Assumption 1, for any two given initial conditions* (e_0, F_0) *and* $(\hat{e}_0, \hat{F}_0)$*,*

$$F_t - \hat{F}_t \overset{a.s.}{\to} 0, \qquad e_t - \hat{e}_t \overset{a.s.}{\to} \mathbf{0}.$$

The third proposition below concerns approximating the trace iterates (e_t, F_t) by truncated traces that depend on a fixed number of the most recent states and actions only. First, let us express the traces (e_t, F_t), by using their definitions (cf. Equations 3-5), as

$$F_t = F_0 \cdot \big(\rho_0\gamma_1 \cdots \rho_{t-1}\gamma_t\big) + \sum_{k=1}^{t} i(S_k) \cdot \big(\rho_k\gamma_{k+1} \cdots \rho_{t-1}\gamma_t\big), \tag{78}$$

$$e_t = e_0 \cdot \big(\beta_1 \cdots \beta_t\big) + \sum_{k=1}^{t} M_k \cdot \phi(S_k) \cdot \big(\beta_{k+1} \cdots \beta_t\big), \tag{79}$$

where $\beta_k = \rho_{k-1}\gamma_k\lambda_k$ and

$$M_k = \lambda_k\, i(S_k) + (1 - \lambda_k)\, F_k.$$

For each integer $K \geq 1$, the truncated traces $(\tilde{e}_{t,K}, \tilde{F}_{t,K})$ are defined by limiting the summations in (78)-(79) to be over $K + 1$ terms only as follows:

$$(\tilde{e}_{t,K}, \tilde{F}_{t,K}) = (e_t, F_t) \quad \text{for} \;\; t \leq K,$$

and for $t \geq K + 1$,

$$\tilde{F}_{t,K} = \sum_{k=t-K}^{t} i(S_k) \cdot \big(\rho_k\gamma_{k+1} \cdots \rho_{t-1}\gamma_t\big), \tag{80}$$

$$\tilde{M}_{t,K} = \lambda_t\, i(S_t) + (1 - \lambda_t)\tilde{F}_{t,K}, \tag{81}$$

$$\tilde{e}_{t,K} = \sum_{k=t-K}^{t} \tilde{M}_{k,K} \cdot \phi(S_k) \cdot \big(\beta_{k+1} \cdots \beta_t\big). \tag{82}$$

We have the following approximation property for truncated traces, in which the notation "$L_K \downarrow 0$" means that L_K decreases monotonically to 0 as $K \to \infty$.

Proposition 13 *Let Assumption 1 hold. Given a bounded set $E \subset \mathbb{R}^{n+1}$, there exist constants $L_K, K \geq 1$, with $L_K \downarrow 0$ as $K \to \infty$, such that if the initial $(e_0, F_0) \in E$, then*

$$\sup_{t \geq 0} \mathbb{E}\left[\left\|(e_t, F_t) - (\tilde{e}_{t,K}, \tilde{F}_{t,K})\right\|\right] \leq L_K.$$

The preceding proposition is the same as (Yu, 2015a, Prop. A.3(i)), except that the initial (e_0, F_0) can be from a bounded set E instead of being fixed. The proof given in (Yu, 2015a) applies here as well, similar to the case of Prop. 11. This proposition about truncated traces was used in (Yu, 2015a) to obtain the convergence in mean given in Theorem 3 (Section 2.4) and allowed us to work with simple finite-space Markov chains, instead of working with the infinite-space Markov chain $\{Z_t\}$ directly, in that proof. In the present paper, it has expedited our proofs of Props. 2-3 (Section 4.1.2) regarding the uniform integrability and convergence in mean conditions for constrained ETD(λ).

Finally, the uniform integrability of $\{(e_t, F_t)\}$ (proved in Prop. 2(i) in this paper, as already mentioned) is important both for convergence analysis and for understanding the behavior of the trace iterates.

References

T. P. Ahamed, V. S. Borkar, and S. Juneja. Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*, 54:489–504, 2006.

A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with Bellman residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

L. C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *The 12th International Conference on Machine Learning (ICML)*, 1995.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

D. P. Bertsekas and H. Yu. Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*, 227(1):27–50, 2009.

P. Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, New York, 1968.

V. S. Borkar. *Stochastic Approximation: A Dynamic Viewpoint.* Cambridge University Press, Cambridge, 2008.

J. A. Boyan. Least-squares temporal difference learning. In *The 16th International Conference on Machine Learning (ICML)*, 1999.

S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(2):33–57, 1996.

C. Dann, G. Neumann, and J. Peters. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

J. L. Doob. *Stochastic Processes*. John Wiley & Sons, New York, 1953.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2002.

M. Geist and B. Scherrer. Off-policy learning with eligibility traces: A survey. *Journal of Machine Learning Research*, 15:289–333, 2014.

P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35:1367–1392, 1989.

A. Hallak, A. Tamar, R. Munos, and S. Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. In *The 30th AAAI Conference on Artificial Intelligence*, 2016.

V. R. Konda. *Actor-Critic Algorithms*. PhD thesis, MIT, 2002.

H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, 1978.

H. J. Kushner and A. Shwartz. Weak convergence and asymptotic properties of adaptive filters with constant gains. *IEEE Transactions on Information Theory*, 30:177–182, 1984.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York, 2nd edition, 2003.

A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

B. Liu, S. Mahadevan, and J. Liu. Regularized off-policy TD-learning. In *Advances in Neural Information Processing Systems (NIPS) 22*, 2009.

B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *The 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

H. R. Maei. *Gradient Temporal-Difference Learning Algorithms*. PhD thesis, University of Alberta, 2011.

S. Mahadevan and B. Liu. Sparse Q-learning with mirror descent. In *The 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

S. Mahadevan, B. Liu, P. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, and J. Liu. Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces, 2014. arXiv:1405.6757.

A. R. Mahmood and R. S. Sutton. Off-policy learning based on weighted importance sampling with linear computational complexity. In *The 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

A. R. Mahmood, H. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems (NIPS) 27*, 2014.

A. R. Mahmood, H. Yu, M. White, and R. S. Sutton. Emphatic temporal-difference learning. In *European Workshops on Reinforcement Learning (EWRL)*, 2015.

S. Meyn. Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function. *SIAM Journal on Control and Optimization*, 27:1409–1439, 1989.

S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009.

R. Munos and C. Szepesvári. Finite time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

J. Neveu. *Discrete-Parameter Martingales*. North-Holland, Amsterdam, 1975.

B. A. Pires and C. Szepesvári. Statistical linear estimation with penalized estimators: An application to reinforcement learning. In *The 29th International Conference on Machine Learning (ICML)*, 2012.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.

D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *The 17th International Conference on Machine Learning (ICML)*, 2000.

D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *The 18th International Conference on Machine Learning (ICML)*, 2001.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

R. S. Randhawa and S. Juneja. Combining importance sampling and temporal difference control variates to simulate Markov chains. *ACM Transactions on Modeling and Computer Simulation*, 14(1):1–30, 2004.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2nd edition, 2003.

B. Scherrer. Should one compute the temporal difference fix point or minimize the Bellman residual? The unified oblique projection view. In *The 27th International Conference on Machine Learning (ICML)*, 2010.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

R. S. Sutton. TD models: Modeling the world at a mixture of time scales. In *The 12th International Conference on Machine Learning (ICML)*, 1995.

R. S. Sutton. The grand challenge of predictive empirical abstract knowledge. In *IJCAI Workshop on Grand Challenges for Reasoning from Experiences*, 2009.

R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998.

R. S. Sutton, C. Szepesvári, and H. Maei. A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. In *Advances in Neural Information Processing Systems (NIPS) 21*, 2008.

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *The 26th International Conference on Machine Learning (ICML)*, 2009.

R. S. Sutton, A. R. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73): 1–29, 2016.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

T. Ueno, S. Maeda, M. Kawanabe, and S. Ishii. Generalized TD learning. *Journal of Machine Learning Research*, 12:1977–2020, 2011.

H. S. Yao and Z. Q. Liu. Preconditioned temporal difference learning. In *The 25th International Conference on Machine Learning (ICML)*, 2008.

H. Yu. Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 50:3310–3343, 2012.

H. Yu. On convergence of emphatic temporal-difference learning, 2015a. `http://arxiv.org/abs/1506.02582`; a shorter version appeared in *The 28th Annual Conference on Learning Theory (COLT)*, 2015.

H. Yu. Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize, 2015b. `http://arxiv.org/abs/1511.07471`.

H. Yu. Some simulation results for emphatic temporal-difference learning algorithms, 2016. `http://arxiv.org/abs/1605.02099`.

H. Yu and D. P. Bertsekas. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329, 2010.

H. Yu and D. P. Bertsekas. Weighted Bellman equations and their applications in approximate dynamic programming. LIDS Technical Report 2876, MIT, 2012.