Non-asymptotic Analysis of Biased Stochastic Approximation Scheme

Belhal Karimi  
CMAP, École Polytechnique.

Blazej Miasojedow  
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland.

Eric Moulines  
CMAP, École Polytechnique.

Hoi-To Wai  
Department of SEEM, The Chinese University of Hong Kong, Hong Kong.*

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Stochastic approximation (SA) is a key method used in statistical learning. Recently, its non-asymptotic convergence analysis has been considered in many papers. However, most of the prior analyses are made under restrictive assumptions such as unbiased gradient estimates and convex objective function, which significantly limit their applications to sophisticated tasks such as online and reinforcement learning. These restrictions are all essentially relaxed in this work. In particular, we analyze a general SA scheme to minimize a non-convex, smooth objective function. We consider update procedure whose drift term depends on a state-dependent Markov chain and the mean field is not necessarily of gradient type, covering approximate second-order method and allowing asymptotic bias for the one-step updates. We illustrate these settings with the online EM algorithm and the policy-gradient method for average reward maximization in reinforcement learning.

Keywords: biased stochastic approximation, state-dependent Markov chain, non-convex optimization, policy gradient, online expectation-maximization

1. Introduction

Stochastic Approximation (SA) schemes are sequential (online) methods for finding a zero of a function when only noisy observations of the function values are available. Consider the recursion:

\[ \eta_{n+1} = \eta_n - \gamma_{n+1} H_{\eta_n}(X_{n+1}), \quad n \in \mathbb{N} \]  

where \( \eta_n \in \mathcal{H} \subset \mathbb{R}^d \) denotes the \( n \)th iterate, \( \gamma_n > 0 \) is the step size and \( H_{\eta_n}(X_{n+1}) \) is the \( n \)th stochastic update (a.k.a. drift term) depending on a random element \( X_{n+1} \) taking its values in a measurable space \( X \). In the simplest setting, \( \{X_n, \ n \in \mathbb{N}\} \) is an i.i.d. sequence of random vectors and \( H_{\eta_n}(X_{n+1}) \) is a conditionally unbiased estimate of the so-called mean-field \( h(\eta_n) \), i.e.,

\[ \mathbb{E} [H_{\eta_n}(X_{n+1}) | \mathcal{F}_n] = h(\eta_n) \]  

where \( \mathcal{F}_n \) denotes the filtration generated by the random variables \( \{\eta_n, \{X_m\}_{m \leq n}\} \). In such case, \( e_{n+1} = H_{\eta_n}(X_{n+1}) - h(\eta_n) \) is a martingale difference. In more sophisticated settings, \( \{X_n, \ n \in \mathbb{N}\} \) is a state-dependent (or controlled) Markov chain, i.e., for any bounded measurable function \( f : X \rightarrow \mathbb{R} \),

\[ \mathbb{E} [f(X_{n+1}) | \mathcal{F}_n] = P_{\eta_n} f(X_n) = \int f(x) P_{\eta_n}(X_n, dx), \]  

\(^*\) Authors listed in alphabetical order.

where \( P_\eta : X \times \mathcal{X} \rightarrow \mathbb{R}_+ \) is a Markov kernel such that, for each \( \eta \in \mathcal{H} \), \( P_\eta \) has a unique stationary distribution \( \pi_\eta \). In such case, the mean field for the SA is defined as:

\[
h(\eta) = \int H_\eta(x) \pi_\eta(dx),
\]

where we have assumed that \( \int \|H_\eta(x)\| \pi_\eta(dx) < \infty \).

Throughout this paper, we assume that the mean field \( h \) is ‘related’ (to be defined precisely later) to a smooth Lyapunov function \( V : \mathbb{R}^d \rightarrow \mathbb{R} \), where \( V(\eta) > -\infty \). The aim of the SA scheme (1) is to find a minimizer or stationary point of the possibly non-convex Lyapunov function \( V \).

Though more than 60 years old (Robbins and Monro, 1951), SA is now of renewed interest as it covers a wide range of applications at the heart of many successes with statistical learning. This includes in particular the stochastic gradient (SG) method and its variants as surveyed in (Bottou, 1998; Bottou et al., 2018), but also in reinforcement learning (Williams, 1992; Peters and Schaal, 2008; Sutton and Barto, 2018). Most convergence analyses assume that \{\eta_n, n \in \mathbb{N}\} is bounded with probability one or visits a prescribed compact set infinitely often. Under such global stability or recurrence conditions [and appropriate regularity conditions on the mean field \( h \)], the SA sequences might be seen as approximation of the ordinary differential equation \( \dot{\eta} = h(\eta) \). Most results available as of today [see for example (Benveniste et al., 1990), (Kushner and Yin, 2003, Chapter 5, Theorem 2.1) or (Borkar, 2009)] have an asymptotic flavor. The focus is to establish that the stationary point of the sequence \{\eta_n, n \in \mathbb{N}\} belongs to a stable attractor of its limiting ODE.

To gain insights on the difference among statistical learning algorithms, non-asymptotic analysis of SA scheme has been considered only recently. In particular, SG methods whose mean field is the gradient of the objective function, i.e., \( h(\eta) = \nabla V(\eta) \), are considered by Moulines and Bach (2011) for strongly convex function \( V \) and martingale difference noise; see (Bottou et al., 2018) for a recent survey on the topic. Extensions to stationary dependent noise have been considered in (Duchi et al., 2012; Agarwal and Duchi, 2013). Meanwhile, many machine learning models can lead to non-convex optimization problems. To this end, SG methods for non-convex, smooth objective function \( V \) have been first studied in (Ghadimi and Lan, 2013) with martingale noise (see (Bottou et al., 2018), Section 4), and it was extended in (Sun et al., 2018) to the case where \{X_n, n \in \mathbb{N}\} is a state-independent Markov chain, i.e., the Markov kernel in (2) does not depend on \( \eta \).

Of course, SA schemes go far beyond SG methods. In fact, in many important applications, the drift term of the SA is not a noisy version of the gradient, i.e., the mean field \( h \) is not the gradient of \( V \). Obvious examples include second-order methods, which aim at combatting the adverse effects of high non-linearity and ill-conditioning of the objective function through stochastic quasi-Newton algorithms. Another closely related example is the online Expectation Maximization (EM) algorithm introduced by Cappé and Moulines (2009) and is further developed in (Balakrishnan et al., 2017; Chen et al., 2018). In many cases, the mean field of the drift term may even be asymptotically biased with the random element \{X_n, n \in \mathbb{N}\} drawn from a Markov chain with state-dependent transition probability. Examples for this situation are common in reinforcement learning such as Q-learning (Jaakkola et al., 1994), policy gradient (Baxter and Bartlett, 2001) and temporal difference learning (Bhandari et al., 2018; Lakshminarayanan and Szepesvari, 2018; Dalal et al., 2018b,a).

Surprisingly enough, we are not aware of non-asymptotic convergence results of the general SA (1) comparable to (Ghadimi and Lan, 2013) and (Bottou et al., 2018, Section 4,5) when (a) the drift term \( H_\eta(x) \) in (1) is not the noisy gradient of the objective function \( V \) and is potentially biased, and/or (b) the sequence \{X_n, n \in \mathbb{N}\} is a state-dependent Markov chain. To this end, the main objective of this work is to fill this gap in the literature by establishing non-asymptotic convergence...
of SA under the above settings. Our main assumption is the existence of a smooth function $V$ satisfying for all $\eta \in \mathcal{H}$, $c_0 + c_1 \langle \nabla V(\eta) \mid h(\eta) \rangle \geq \|h(\eta)\|^2$ there exists $c_1 > 0$, $c_0 \geq 0$; see Section 2 and A1. If $c_0 = 0$, then $\langle \nabla V(\eta) \mid h(\eta) \rangle > 0$ as soon as $h(\eta) \neq 0$ in which case $V$ is a Lyapunov function for the ODE $\dot{\eta} = h(\eta)$. Assuming $c_0 > 0$ allows us to consider situations in which the estimate of the mean field is biased, a situation which has been first studied in Tadić and Doucet (2017). To summarize, our contributions are two-fold:

1. We provide non-asymptotic convergence analysis for (1) with a potentially biased mean field $h$ under two cases — (Case 1) $\{X_n, n \in \mathbb{N}\}$ is an i.i.d. sequence; (Case 2) $\{X_n, n \in \mathbb{N}\}$ is a state-dependent Markov chain. For these two cases, we provide non asymptotic bounds such that for all $n \in \mathbb{N}$, $\mathbb{E}[\|h(X_n)\|^2] = O(c_0 + \log(n)/\sqrt{n})$, for some random index $N \in \{1, \ldots, n\}$ and $c_0 \geq 0$ characterizes the (potential) bias of the mean field $h$.

2. We illustrate our findings by analyzing popular statistical learning algorithms such as the online expectation maximization (EM) algorithm (Cappé and Moulines, 2009) and the average-cost policy-gradient method (Sutton and Barto, 2018). Our findings provide new insights into the non-asymptotic convergence behavior of these algorithms.

Our theory significantly extends the results reported in (Bottou et al., 2018, Sections 4.5) and (Ghadimi and Lan, 2013, Theorem 2.1). When focused on the Markov noise setting, our result is a nontrivial relaxation of (Sun et al., 2018), which considers Markov noise that is not state dependent and the mean field satisfies $h(\eta) = \nabla V(\eta)$; and of (Tadić and Doucet, 2017) which shows asymptotic convergence of (1) under the uniform boundedness assumption on iterates.

Notation Let $(X, \mathcal{X})$ be a measurable space. A Markov kernel $R$ on $X \times \mathcal{X}$ is a mapping $R : X \times \mathcal{X} \to [0,1]$ satisfying the following conditions: (a) for every $x \in X$, $R(x, \cdot) : A \mapsto R(x,A)$ is a probability measure on $\mathcal{X}$ (b) for every $A \in \mathcal{X}$, $R(\cdot,A) : x \mapsto R(x,A)$ is a measurable function. For any probability measure $\lambda$ on $(X, \mathcal{X})$, we define $\lambda R$ by $\lambda R(A) = \int_X \lambda(dx) R(x,A)$. For all $k \in \mathbb{N}^*$, we define the Markov kernel $R^k$ recursively by $R^1 = R$ and for all $x \in X$ and $A \in \mathcal{X}$, $R^{k+1}(x,A) = \int_X R^k(x,dx') R(x',A)$. A probability measure $\bar{\pi}$ is invariant for $R$ if $\pi R = \pi$. $\| \cdot \|$ denotes the standard Euclidean norm (for vectors) or the operator norm (for matrices).

2. Stochastic Approximation Schemes and Their Convergence

Consider the following assumptions:

A1 For all $\eta \in \mathcal{H}$, there exists $c_0 \geq 0, c_1 > 0$ such that $c_0 + c_1 \langle \nabla V(\eta) \mid h(\eta) \rangle \geq \|h(\eta)\|^2$.

A2 For all $\eta \in \mathcal{H}$, there exists $d_0 \geq 0, d_1 > 0$ such that $d_0 + d_1\|h(\eta)\| \geq \|\nabla V(\eta)\|$.

A3 Lyapunov function $V$ is $L$-smooth. For all $(\eta, \eta') \in \mathcal{H}^2$, $\|\nabla V(\eta) - \nabla V(\eta')\| \leq L\|\eta - \eta'\|$.

A1,A2 assume that the mean field $h(\eta)$ [cf. (2)] is indirectly related to the Lyapunov function $V(\eta)$ where it needs not be the same as $\nabla V(\eta)$. In particular, the constants $c_0, d_0$ characterize the ‘bias’ between the mean field and the gradient of the Lyapunov function. From an optimization perspective, we note that the Lyapunov function $V$ can be non-convex under A3. In light of A1, A2, we study the convergence of the non-negative quantity $\|h(\eta_n)\|^2$, where $\eta_n$ is produced by (1). If $c_0 = d_0 = 0$ in A1,A2, then $h(\eta_n) = 0$ implies that $\|\nabla V(\eta_n)\| = 0$, i.e., the point $\eta_n$ is a stationary point of the deterministic recursion $\bar{\eta}_n = \eta_n - \gamma_n h(\eta_n)$. As a convention, for any $\epsilon \geq 0$, we say that $\eta_n$ is an $\epsilon$-quasi-stationary point if $\|h(\eta_n)\|^2 \leq \epsilon$. 

3
As a common step in analyzing SA scheme for smooth but non-convex Lyapunov function (e.g., (Ghadimi and Lan, 2013)), we shall adopt a randomized stopping rule. For any \( n \geq 1 \), let \( N \in \{0, \ldots, n\} \) be a discrete random variable (independent of \( \mathcal{F}_n, n \in \mathbb{N} \)) with
\[
\mathbb{P}(N = \ell) := \left( \sum_{k=0}^{n} \gamma_{k+1} \right)^{-1} \gamma_{\ell+1} ,
\]
where \( N \) serves as the terminating iteration for \( (1) \). Throughout this paper, we focus on analyzing \( \mathbb{E} \left( \| \nabla h(\eta_N) \|^2 \right) \) where the expectation is taken over \( N \) and the stochastic updates in SA. We consider two settings for the noise in SA scheme. Define the following noise vector:
\[
e_{n+1} := H\eta_n(X_{n+1}) - h(\eta_n) ,
\]
where \( h(\eta_n) \) was defined in (3). Our settings and convergence results are in order.

Case 1. \( \{ e_n \}_{n \geq 1} \) is a Martingale Difference Sequence. We first consider a case similar to the classical SG method analyzed by Ghadimi and Lan (2013). In particular,

A4 The sequence of noise vectors is a Martingale difference sequence with, for any \( n \in \mathbb{N} \),
\[
\mathbb{E} \left[ e_{n+1} | \mathcal{F}_n \right] = 0, \quad \mathbb{E} \left[ \| e_{n+1} \|^2 | \mathcal{F}_n \right] \leq \sigma_0^2 + \sigma_1^2 \| h(\eta_n) \|^2 \text{ with } \sigma_0^2, \sigma_1^2 \in [0, \infty) .
\]

As a concrete example, A4 can be satisfied when \( H\eta_n(X_{n+1}) = h(\eta_n) + X_{n+1} \) where \( X_{n+1} \) is an i.i.d., zero-mean random vector with bounded variance. We show:

**Theorem 1** Let A1, A3, A4 hold and \( \gamma_{n+1} \leq (2c_1L(1 + \sigma_1^2))^{-1} \) for all \( n \geq 0 \). We have
\[
\mathbb{E} \left[ \| h(\eta_N) \|^2 \right] \leq 2c_1 \left( V_{0,n} + \sigma_0^2 L \sum_{k=0}^{n} \gamma_{k+1}^2 \right) + 2c_0 ,
\]
where \( N \) is distributed according to (4) and we have defined \( V_{0,n} := \mathbb{E} \left[ V(\eta_0) - V(\eta_{n+1}) \right] \).

If we set \( \gamma_k = (2c_1L(1 + \sigma_1^2)\sqrt{k})^{-1} \) for all \( k \geq 1 \), then the right hand side in (6) evaluates to \( O(c_0 + \log n/\sqrt{n}) \) for any \( n \geq 1 \). Therefore, the SA scheme (1) finds an \( \mathcal{O}(c_0 + \log n/\sqrt{n}) \) quasi-stationary point within \( n \) iterations.

Case 2. \( \{ e_n \}_{n \geq 1} \) is State-dependent Markov Noise. Next, we consider a general scenario when \( X_{n+1} \) is drawn from a state-dependent Markov process. For any bounded measurable function \( \varphi \) and \( n \in \mathbb{N} \), we have \( \mathbb{E} \left[ \varphi(X_{n+1}) | \mathcal{F}_n \right] = \mathbb{E} \left[ \varphi(X_{n+1}) \right] \), where \( \mathbb{E} \left[ \varphi(X_{n+1}) \right] \) is a Markov kernel on \( \mathcal{X} \times \mathcal{X} \). We assume that for each \( \eta \in \mathcal{H} \), \( P_{\eta} \) has a unique stationary distribution \( \pi_{\eta} \), i.e., \( \pi_{\eta} P_{\eta} = \pi_{\eta} \). In addition, for each \( \eta \in \mathcal{H} \), we have \( \int \| H\eta(x) \| \pi_{\eta}(dx) < \infty \) and \( h(\eta) = \int H\eta(x) \pi_{\eta}(dx) \).

A5 There exists a Borel measurable function \( \hat{H} : \mathcal{H} \times \mathcal{X} \to \mathcal{H} \) where for each \( \eta \in \mathcal{H}, x \in \mathcal{X}, \)
\[
\hat{H}\eta(x) - P_{\eta} \hat{H}\eta(x) = H\eta(x) - h(\eta) .
\]

A6 There exists \( L_{PH}^{(0)} < \infty \) and \( L_{PH}^{(1)} < \infty \) such that, for all \( \eta \in \mathcal{H} \) and \( x \in \mathcal{X} \), one has \( \| \hat{H}\eta(x) \| \leq L_{PH}^{(0)} \| P_{\eta} H\eta(x) \| \leq L_{PH}^{(0)} . \) Moreover, for \( (\eta, \eta') \in \mathcal{H}^2 \),
\[
\sup_{x \in \mathcal{X}} \| P_{\eta} \hat{H}\eta(x) - P_{\eta'} \hat{H}\eta(x) \| \leq L_{PH}^{(1)} \| \eta - \eta' \| .
\]

A7 The stochastic update is bounded, i.e., \( \sup_{\eta \in \mathcal{H}, x \in \mathcal{X}} \| H\eta(x) - h(\eta) \| \leq \sigma \).

Assumption A5 requires that for each \( \eta \in \mathcal{H} \), the Poisson equation associated with the Markov kernel \( P_{\eta} \) and the function \( H\eta(\cdot) \) has a solution. Assumption A6 implies that for each \( x \in \mathcal{X} \), the function \( \eta \mapsto H\eta(x) \) is Lipshitz and that the Lipshitz constant is uniformly bounded in \( x \in \mathcal{X} \). We provide in Appendix D conditions upon which these assumptions hold. Lastly, Assumption A7 assumes that the drift terms are bounded uniformly. Our main result reads as follows:
### Theorem 2
Let $A1–A3$, $A5–A7$ hold. Suppose that the step sizes satisfy

$$
\gamma_{n+1} \leq \gamma_n, \quad \gamma_n \leq a \gamma_{n+1}, \quad \gamma_n - \gamma_{n+1} \leq a' \gamma_n^2, \quad \gamma_1 \leq 0.5 \left( c_1 (L + C_h) \right)^{-1},
$$

for some $a, a' > 0$ and all $n \geq 0$. We have

$$
E[h(\eta_N)]^2 \leq \frac{2c_1 (V_0, n + C_0, n + (\sigma^2 L + C_\gamma) \sum_{k=0}^{\infty} \gamma_{k+1}^2)}{\sum_{k=0}^{\infty} \gamma_{k+1}^2} + 2c_0,
$$

where $N$ is distributed according to (4), $V_0, n := E[V(\eta_0) - V(\eta_{n+1})]$, and the constants are:

$$
C_h := (L_{PH}^{(1)}(d_0 + d_1 a + a d_1 \gamma + a \delta_{PH}(L + d_1 \{1 + a'\})),
$$

$$
C_\gamma := L_{PH}^{(1)}(d_0 + d_0 \gamma + d_1 \sigma) + LL_{PH}^{(0)}(1 + \sigma),
$$

$$
C_0, n := L_{PH}^{(0)}((1 + d_0)(\gamma_1 - \gamma_{n+1}) + d_0(\gamma_1 + \gamma_{n+1}) + 2d_1).
$$

Similar to the case with Martingale difference noise, if we set $\gamma_k = (2c_1 L(1 + C_h)\sqrt{k})^{-1}$ for all $k \geq 1$, then the step size satisfies (9) with $a = \sqrt{2}$ and $a' = \sqrt{\frac{2}{\sqrt{2}}}(2c_1 L(1 + C_h))$, and the right hand side in (10) evaluates to $O(c_0 + \log n/\sqrt{n})$ for any $n \geq 1$. We obtain a similar convergence rate as in Theorem 1. In fact, if we consider a special case when for all $\eta \in \mathcal{H}$ and $x \in X$, $P_{\eta}(x, \cdot) = \pi(\cdot)$, we have $L_{PH}^{(0)} = L_{PH}^{(1)} = 0$. The constants evaluate to $C_h = C_\gamma = C_0, n = 0$ and our Theorem 2 can be reduced into Theorem 1. We remark that Theorem 2 cannot be treated as a strict generalization of Theorem 1 as $A4$ does not imply the uniform boundedness $A7$. Our analysis [cf. Lemma 2] relies on a new decomposition of the error terms, which controls the growth of $E[\|h(\eta_N)\|^2]$ without explicitly assuming that $\{\eta_n\}_{n \geq 0}$ is bounded.

In Appendix A.3, we provide a lower bound on the rate of SA scheme (1), (4) such that $E[\|h(\eta_N)\|^2] = \Omega(\log n/\sqrt{n})$. This shows that our analysis in Theorem 1, 2 is tight.

### Related Studies
Non-asymptotic analysis of biased SA schemes can be found in the literature on temporal difference (TD) learning (Bhandari et al., 2018; Lakshminarayanan and Szepesvari, 2018; Dalal et al., 2018a,b), which analyzed a special case of linear SA. Their assumptions can essentially be satisfied by our $A1–A3$ with $V(\eta) = \|\eta - \eta^*\|_q^2$, e.g., (Bhandari et al., 2018, Lemma 3) shows that the TD learning has a mean field which satisfies $A1$. Furthermore, the above mentioned analysis requires a strongly convex Lyapunov function, which is not needed in our results.

For Case 1, our results generalizes (Ghadimi and Lan, 2013, Theorem 2.1) by accounting for biased SA updates. In fact we recover the latter result with $h(\eta) = \nabla V(\eta)$, $A1 [c_0 = 0, c_1 = 1]$.

For Case 2, our assumptions $A1–A3$, $A5–A7$ are similar to (Tadić and Doucet, 2017, Section 3). The exception is $A7$ which is used in place of the assumption $\sup_{n \in \mathbb{N}} \|\eta_n\| < \infty$ in (Tadić and Doucet, 2017). We note that the two conditions are neither stronger nor weaker than the other.

### 2.1. Convergence Analysis
The detailed proofs in this section are in Appendix A. To simplify notations, we denote $h_n := \|h(\eta_n)\|^2$ from now on. We first describe an intermediate result that holds under just $A1$, $A3$:

### Lemma 1
Let $A1$, $A3$ hold. It holds for all $n \geq 1$ that:

$$
\sum_{k=0}^{n} \gamma_{k+1} \left( 1 - c_1 L_{\gamma_{k+1}} \right) h_k \leq V(\eta_0) - V(\eta_{n+1}) + L \sum_{k=0}^{n} \gamma_{k+1}^2 \|e_{k+1}\|^2 + \sum_{k=0}^{n} \gamma_{k+1} \left( c_1 \gamma_{k+1} \|e_{k+1}\|^2 - \langle \nabla V(\eta_k) | e_{k+1} \rangle \right).
$$

5
Having established Lemma 1, our main convergence results can be obtained as follows.

**Proof of Theorem 1** With Martingale difference noise, the expected value of \( \langle \nabla V(\eta_k) \mid e_{k+1} \rangle \) is zero when conditioned on \( \mathcal{F}_k \). Therefore, taking total expectation on both sides of (14) yields:

\[
\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1} (1 - c_1 L \gamma_{k+1}) \mathbb{E}[h_k] \leq V_{0,n} + L \sum_{k=0}^{n} (\gamma_{k+1}^2 \mathbb{E}[\|e_{k+1}\|^2] + \gamma_{k+1}^2 \frac{\sigma_0^2}{c_1})  \\
\leq V_{0,n} + L \sigma_0^2 \sum_{k=0}^{n} \gamma_{k+1}^2 + L \sigma_1^2 \sum_{k=0}^{n} \gamma_{k+1} \mathbb{E}[h_k] + \gamma_{k+1} \frac{\sigma_0}{c_1} ,
\]

where the last inequality is due to A4. Rearranging terms yields:

\[
\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1} (1 - c_1 L (1 + \sigma_1^2) \gamma_{n+1}) \mathbb{E}[h_k] \leq V_{0,n} + \sigma_0^2 L \sum_{k=0}^{n} \gamma_{k+1}^2 + \frac{\sigma_0}{c_1} \sum_{k=0}^{n} \gamma_{k+1} .
\]

Consequently, using (4) and noting that \( 1 - c_1 L (1 + \sigma_1^2) \gamma_{n+1} \geq \frac{1}{2} \), we obtain

\[
\mathbb{E}[h_N] = \sum_{n'=0}^{n} \gamma_{n'+1} \mathbb{E}[h_{n'}] \leq \frac{2c_1 (V_{0,n} + \sigma_0^2 L \sum_{k=0}^{n} \gamma_{k+1})}{\sum_{k=0}^{n} \gamma_{k+1}} + 2c_0 ,
\]

**Proof of Theorem 2** In the case with state-dependent Markovian noise. Under A7, one has

\[
\sum_{k=0}^{n} \gamma_{k+1}^2 \mathbb{E}[\|e_{k+1}\|^2] \leq \sum_{k=0}^{n} \gamma_{k+1}^2 \sigma_1^2 .
\]

Unlike in Theorem 1, the expected value of the inner product \( \langle \nabla V(\eta_k) \mid e_{k+1} \rangle \) is non-zero in general. Fortunately, as we show next in Lemma 2, this issue can be mitigated.

**Lemma 2** Let A1–A3, A5–A7 hold and the step sizes satisfy (9). It holds:

\[
\mathbb{E}[-\sum_{k=0}^{n} \gamma_{k+1} \langle \nabla V(\eta_k) \mid e_{k+1} \rangle] \leq C_h \sum_{k=0}^{n} \gamma_{k+1}^2 \mathbb{E}[\|h(\eta_k)\|^2] + C_\gamma \sum_{k=0}^{n} \gamma_{k+1}^2 + C_{0,n} ,
\]

where \( C_h, C_\gamma \) and \( C_{0,n} \) are defined in (11), (12), (13).

Finally, to prove the theorem, we combine Lemma 1, (17) and Lemma 2 to obtain:

\[
\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1} (1 - c_1 L \gamma_{k+1}) \mathbb{E}[h_k]  \\
\leq V_{0,n} + C_{0,n} + \left( \sigma_2^2 L + C_\gamma \right) \sum_{k=0}^{n} \gamma_{k+1}^2 + C_h \sum_{k=0}^{n} \gamma_{k+1} \mathbb{E}[h_k] + \frac{\sigma_0}{c_1} \sum_{k=0}^{n} \gamma_{k+1} .
\]

Repeating a similar argument as in (16) using the distribution (4) shows the desired bound (10).

3. Applications

We present several applications and provide new non-asymptotic convergence rate for them.

3.1. Regularized Online Expectation Maximization

Expectation-Maximization (EM) (Dempster et al., 1977) is a powerful tool for learning latent variable models, which can be inefficient due to the high storage cost. This has motivated the development of online version of the EM which makes it possible to estimate the parameters of latent variables model without storing the data; the online EM algorithm analyzed below was introduced in (Cappé and Moulines, 2009) and later developed by many authors: see for example (Chen et al., 2018) and the references therein. The online EM algorithm sticks closely to the principles of the batch-mode EM algorithm. Each iteration of the online EM algorithm is decomposed into two steps, where the first one is a stochastic approximation version of the E-step aimed at incorporating the information brought by the newly available observation, and, the second step consists in the maximization program that appears in the M-step of the traditional EM algorithm.
The latent variable statistical model postulates the existence of a latent variable $X$ distributed under $f(x; \theta)$ where $\{f(x; \theta); \theta \in \Theta\}$ is a parametric family of probability density functions and $\Theta$ is an open convex subset of $\mathbb{R}^d$. The observation $Y \in \mathcal{Y}$ is a deterministic function of $X$. We denote by $g(y; \theta)$ the (observed) likelihood function. The notations $\mathbb{E}_{\theta} [\cdot]$ and $\mathbb{E}_{\theta} [\cdot | Y]$ are used to denote the expectation and conditional expectation under the statistical model $\{f(x; \theta); \theta \in \Theta\}$. We denote by $\pi$ the probability density function of the observation $Y$: the model might be misspecified, that is, the "true" distribution of the observations may not belong to the family $\{g(y; \theta), \theta \in \Theta\}$. The notations $\mathbb{E}_\pi$ is used below to denote the expectation under the actual distribution of the observations.

Let $S$ be a convex open subset of $\mathbb{R}^m$ and $S : X \rightarrow S$ be a measurable function. We assume that the complete data-likelihood function belongs to the curved exponential family

$$f(x; \theta) = h(x) \exp \left( \langle S(x) | \phi(\theta) \rangle - \psi(\theta) \right) ,$$

where $\psi : \Theta \rightarrow \mathbb{R}$ is twice differentiable and convex and $\phi : \Theta \rightarrow S \subset \mathbb{R}^m$ is concave and differentiable. In this setting, $S$ is the complete data sufficient statistics. For any $\theta \in \Theta$ and $y \in \mathcal{Y}$, we assume that the conditional expectation

$$\mathbb{E}(y; \theta) = \mathbb{E}_{\theta} [S(X) | Y = y]$$

is well-defined and belongs to $S$. For any $s \in S$, we consider the penalized negated complete data log-likelihood:

$$\ell(s; \theta) := \psi(\theta) + R(\theta) - \langle s | \phi(\theta) \rangle ,$$

where $R : \Theta \rightarrow \mathbb{R}$ is a penalization term assumed to be twice differentiable. This penalty term is used to enforce constraints on the estimated parameter. If $\kappa : \Theta \rightarrow \mathbb{R}^m$ is a differentiable function, we denote by $\mathcal{J}_\kappa(\theta') \in \mathbb{R}^m \times d$ the Jacobian of the map $\kappa$ with respect to $\theta$ at $\theta'$. Consider:

A8 For all $s \in S$, the function $\theta \mapsto \ell(s; \theta)$ admits a unique global minimum $\overline{\theta}(s)$ in the interior of $\Theta$, characterized by

$$\nabla \psi(\overline{\theta}(s)) + \nabla R(\overline{\theta}(s)) - J_\phi(\overline{\theta}(s))^\top s = 0 .$$

In addition, for any $s \in S$, $J_\phi(\overline{\theta}(s))$ is invertible and the map $s \mapsto \overline{\theta}(s)$ is differentiable on $S$.

The regularized version of the online EM (ro-EM) method is an iterative procedure which alternately updates an estimate of the sufficient statistics and the estimated parameters as:

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1} (\overline{s}(Y_{n+1}; \hat{\theta}_n) - \hat{s}_n), \quad \hat{\theta}_{n+1} = \overline{\theta}(\hat{s}_{n+1}) .$$

In the following, we show that our non-asymptotic convergence result holds for the ro-EM. We establish convergence of the online method to a stationary point of the Lyapunov function defined as a regularized Kullback-Leibler (KL) divergence between $\pi$ and $g_\theta$. Precisely, we set

$$V(s) := \text{KL} (\pi, g(\cdot; \overline{\theta}(s))) + R(\overline{\theta}(s)), \quad \text{KL} (\pi, g(\cdot; \theta)) := \mathbb{E}_\pi [\log(\pi(Y))/g(Y; \theta)] .$$

We establish a few key results that relate the ro-EM method to an SA scheme seeking for a stationary point of $V(s)$. Denote by $\mathcal{F}_n$ the filtration generated by the random variables $\{\hat{s}_0, Y_k\}_{k \leq n}$. From (24) we can identify the drift term and its mean field respectively as

$$H_{\hat{s}_n}(Y_{n+1}) = \hat{s}_n - \overline{s}(Y_{n+1}; \overline{\theta}(\hat{s}_n)) ,$$

$$h(\hat{s}_n) = \mathbb{E}_\pi [H_{\hat{s}_n}(Y_{n+1}) | \mathcal{F}_n] = \hat{s}_n - \mathbb{E}_\pi [\overline{s}(Y_{n+1}; \overline{\theta}(\hat{s}_n))] .$$

and $e_{n+1} := H_{\hat{s}_n}(Y_{n+1}) - h(\hat{s}_n)$. Define by $H^\theta$ the Hessian of the function $\ell$ with respect to $\theta$. Our results are summarized by the following propositions, which proofs can be found in Appendix B:
Proposition 1 Assume A8. The following holds:

- If \( h(s^*) = 0 \) for some \( s^* \in S \), then \( \nabla_\theta \KL(\pi, g_{\theta^*}) + \nabla_\theta R(\theta^*) = 0 \) with \( \theta^* := \overline{\theta}(s^*) \).
- If \( \nabla_\theta \KL(\pi, g_{\theta^*}) + \nabla_\theta R(\theta^*) = 0 \) for some \( \theta^* \in \Theta \) then \( s^* = \mathbb{E}_\pi[S(Y, \theta^*)] \).

Proposition 2 Assume A8. We have \( \nabla_s V(s) = J^\theta_\phi(\overline{\theta}(s))(H^\theta_\phi(s; \theta)^{-1} J^\theta_\phi(\overline{\theta}(s))^\top h(s) \) for \( s \in S \).

Proposition 1 relates the root(s) of the mean field \( h(s) \) to the stationary condition of the regularized KL divergence. Moreover, if \( \lambda_{\min}(J^\theta_\phi(\overline{\theta}(s))(H^\theta_\phi(s; \theta)^{-1} J^\theta_\phi(\overline{\theta}(s))^\top) \geq v > 0 \) for all \( s \in S \), then Proposition 2 shows that the mean field of the stochastic update in (26) satisfies A1 with \( c_0 = 0 \) and \( c_1 = 1/v \). If we assume that the Lyapunov function in (25), and the stochastic update in (26) satisfy the assumptions in Case 1 [i.e., A4], then these results show that Theorem 1 applies. To further illustrate the above principles, we look at an example with Gaussian mixture model (GMM).

Example: GMM Inference Consider the inference problem of a mixture of \( M \) Gaussian distributions, each with a unit variance from an observation stream \( Y_1, Y_2, \ldots \). The likelihood is:

\[
g(y; \theta) \propto \left(1 - \sum_{m=1}^{M-1} \omega_m \right) \exp\left(-\frac{(y - \mu_m)^2}{2}\right) + \sum_{m=1}^{M-1} \omega_m \exp\left(-\frac{(y - \mu_m)^2}{2}\right). \tag{27}
\]

The parameters are denoted by \( \theta := (\omega_1, \ldots, \omega_{M-1}, \mu_1, \ldots, \mu_{M-1}, \mu_M) \in C \) where the parameter set is defined as \( C = \Delta_{M-1} \times \mathbb{R}^M \) with \( \Delta_{M-1} := \{ (\omega_1, \ldots, \omega_{M-1}) \in \mathbb{R}^{M-1}, \omega_m \geq 0, \sum_{m=1}^{M-1} \omega_m = 1 \} \). To apply the ro-EM method, we augment the \( n \)th data \( Y_n \) with the latent variable \( Z_n \in \{1, \ldots, M\} \). The log likelihood of the complete data tuple is

\[
\mathcal{L}(x; \theta) = \mathbb{I}_{\{z=M\}} \left[ \log\left(1 - \sum_{m=1}^{M-1} \omega_m \right) - \frac{(y - \mu_m)^2}{2}\right] + \sum_{m=1}^{M-1} \mathbb{I}_{\{z=m\}} \left[ \log(\omega_m) - \frac{(y - \mu_m)^2}{2}\right]. \tag{28}
\]

The above can be written in the standard curved exponential family form (20). In particular, we partition the sufficient statistics as \( S(x) = (S^{(1)}(x)^\top, S^{(2)}(x)^\top, S^{(3)}(x)) \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R} \), and partition \( \phi(\theta) = (\phi^{(1)}(\theta)^\top, \phi^{(2)}(\theta)^\top, \phi^{(3)}(\theta)) \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R} \). Using the fact that \( \mathbb{I}_{\{z=M\}} = 1 - \sum_{m=1}^{M-1} \mathbb{I}_{\{z=m\}} \), (28) can be expressed in the standard form as (20) with

\[
\begin{align*}
\mathbb{I}_{\{z=m\}}, & \quad \phi^{(1)}(\theta) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\}, \\
\mathbb{I}_{\{z=m\}} y, & \quad \phi^{(2)}(\theta) = \mu_m, \quad m = 1, \ldots, M - 1, \\
\mathbb{I}_{\{z=M\}}, & \quad \phi^{(3)}(\theta) = \mu_M,
\end{align*}
\]

and

\[
\psi(\theta) = -\left\{ \log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_j^2}{2\sigma^2} \right\}.
\]

We apply the ro-EM method to the above model. Following the partition of sufficient statistics and parameters in the above, we define \( \hat{s}_n = ((\hat{s}_n^{(1)})^\top, (\hat{s}_n^{(2)})^\top, (\hat{s}_n^{(3)})^\top) \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R} \), and \( \hat{\theta}_n = (\hat{\omega}_n, \hat{\mu}_n, \hat{\mu}_M) \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R} \). Also, define the conditional expected value:

\[
\tilde{\omega}_m(Y_{n+1}; \theta_n) := \mathbb{E}_{\theta_n}[1_{\{z=m\}}|Y = Y_{n+1}] = \frac{\tilde{\omega}_m \exp\left(-\frac{1}{2}(Y_{n+1} - \hat{\mu}_n)^2\right)}{\sum_{j=1}^{M} \tilde{\omega}_j \exp\left(-\frac{1}{2}(Y_{n+1} - \hat{\mu}_j)^2\right)}. \tag{30}
\]

With the above notations, the E-step’s update in (21) can be described with

\[
\pi(Y_{n+1}; \theta_n) = \left( \begin{array}{c}
(\tilde{\omega}_1(Y_{n+1}; \hat{\theta}_n), \ldots, \tilde{\omega}_{M-1}(Y_{n+1}; \hat{\theta}_n))^\top \\
(Y_{n+1}\tilde{\omega}_1(Y_{n+1}; \hat{\theta}_n), \ldots, Y_{n+1}\tilde{\omega}_{M-1}(Y_{n+1}; \hat{\theta}_n))^\top
\end{array} \right) = \left( \begin{array}{c}
\hat{s}_n^{(1)} \\
\hat{s}_n^{(2)} \\
\hat{s}_n^{(3)}
\end{array} \right). \tag{31}
\]
For the M-step, let \( \epsilon > 0 \) be a user designed parameter, we consider the following regularizer:

\[
R(\theta) = \epsilon \sum_{m=1}^{M} \left\{ \mu_m^2 / 2 - \log(\omega_m) \right\} - \epsilon \log \left( 1 - \sum_{m=1}^{M-1} \omega_m \right),
\]

(32)

For any \( s \) with \( s^{(1)} \geq 0 \), it can be shown that the regularized M-step in (24) evaluates to

\[
\hat{\theta}(s) = \left( (1 + \epsilon M)^{-1} (s^{(1)} + \epsilon, \ldots, s^{(1)}_{M-1} + \epsilon)^\top, \right.

\[ \left. \left( (s^{(1)} + \epsilon)^{-1} s^{(2)}_1, \ldots, (s^{(1)}_{M-1} + \epsilon)^{-1} s^{(2)}_{M-1} \right)^\top, \right)

\[ (1 - \sum_{m=1}^{M-1} s^{(1)}_m + \epsilon)^{-1} (s^{(3)} - \sum_{m=1}^{M-1} s^{(2)}_m) \right) \right).
\]

(33)

Note that, as opposed to an unregularized solution (i.e., with \( \epsilon = 0 \)), the regularized solution is numerically stable as it avoids issues such as division by zero.

To analyze the convergence of ro-EM, we verify that (24), (31), (33) yield a special case of an SA scheme on \( \hat{s}_n \) which satisfies A1, A3, A4. Assume the following on the observations \( \{Y_n\}_{n \geq 0} \)

**A9** Each observed sample \( Y_n \) is drawn i.i.d. and they are bounded as \( |Y_n| \leq \bar{Y} \) for any \( n \geq 0 \).

The ro-EM method can be initialized by setting \( \hat{s}_1 = (0, 0, 0)^\top \) and begun with the M-step. Note that under A9, the sufficient statistics \( \hat{s}_n \) lie in the compact set \( S = \Delta_{M-1} \times (-\bar{Y}, \bar{Y})^M \) for all \( n \geq 1 \), where \( \Delta_{M-1} := \{ s_1, \ldots, s_{M-1} : s_m \geq 0, \sum_{m=1}^{M-1} s_m \leq 1 \} \). We observe the following propositions that are proven in Appendix B:

**Proposition 3** Under A9, it holds that \( \mathbb{E}[\|\hat{s}(Y_{n+1}; \hat{\theta}_n) - \hat{s}_n\|^2 | F_n] \leq 2M \bar{Y}^2 \) for all \( n \geq 0 \).

**Proposition 4** Under A9 and the regularizer (32) set with \( \epsilon > 0 \), then for all \( (s, s') \in S^2 \), there exists positive constants \( v, \bar{Y}, \Psi \) such that:

\[
\langle \nabla V(s) | h(s) \rangle \geq v \|h(s)\|^2, \quad \|\nabla V(s) - \nabla V(s')\| \leq \Psi \|s - s'\|.
\]

(34)

The above propositions show that the ro-EM method applied to GMM is a special case of the SA scheme with Martingale difference noise, for which A1 [with \( c_0 = 0, c_1 = v^{-1} \)], and A3 [with \( L = \Psi \)], A4 [with \( \sigma_0^2 = 2M \bar{Y}^2 , \sigma_1^2 = 0 \)] are satisfied. As such, applying Theorem 1 shows that

**Corollary 1** Under A9 and set \( \gamma_k = \left( 2c_1 L (1 + \sigma_1^2) \sqrt{k} \right)^{-1} \). For any \( n \in \mathbb{N} \), let \( N \in \{0, \ldots, n\} \) be an independent discrete r.v. distributed according to (4). The ro-EM method for GMM (24), (31), (33) finds \( \hat{s}_N \) such that

\[
\mathbb{E}[\|\nabla V(\hat{s}_N)\|^2] = O(\log n / \sqrt{n})
\]

(35)

where \( V(\cdot) \) is defined in (25). The expectation is taken w.r.t. \( N \) and the observation law \( \pi \).

**Related Studies** Convergence analysis for the EM method in batch mode has been the focus of the classical work by Dempster et al. (1977); Wu (1983), in which asymptotic convergence has been established; also see the recent work by Wang et al. (2015); Xu et al. (2016). Several work has studied the convergence of stochastic EM with fixed data, e.g., Mairal (2015) studied the asymptotic convergence to a stationary point, Chen et al. (2018) studied the local linear convergence of a variance reduced method by assuming that the iterates are bounded. On the other hand, the online EM
method considered here, where a fresh sample is drawn at each iteration, has only been considered by a few work. Particularly, Cappé and Moulines (2009) showed the asymptotic convergence of the online EM method to a stationary point; Balakrishnan et al. (2017) analyzed non-asymptotic convergence for a variant of online EM method which requires a-priori the initial radius $\|\theta_0 - \theta^*\|$, where $\theta^*$ is the optimal parameter. To our best knowledge, the rate results in Corollary 1 is new.

3.2. Policy Gradient for Average Reward over Infinite Horizon

There has been a growing interest in policy-gradient methods for model-free planning in Markov decision process; see (Sutton and Barto, 2018) and the references therein. Consider a finite Markov Decision Process (MDP) $(S, A, R, P)$, where $S$ is a finite set of spaces (state-space), $A$ is a finite set of action (action-space), $R : S \times A \to [0, R_{\text{max}}]$ is a reward function and $P$ is the transition model, i.e., given an action $a \in A$, $P^a = \{P^a_{s,s'}\}$ is a matrix, $P^a_{s,s'}$ is the probability of transiting from the $s$th state to the $s'$th state upon taking action $a$. The agent’s decision is characterized by a parametric family of policies $\{ \Pi_\eta \}_{\eta \in \mathcal{H}}$: $\Pi_\eta(a; s)$ which is the probability of taking action $a$ when the current state is $s$ (a semi-column is used to distinguish the random variables from parameters of the distribution). The state-action sequence $\{(S_t, A_t)\}_{t \geq 1}$ forms an MC with the transition matrix:

\[
Q_\eta((s, a); (s', a')) := \Pi_\eta(a'; s') P^a_{s,s'}, \tag{36}
\]

where the above corresponds to the $(s, a)$th row, $(s', a')$th column of the matrix $Q_\eta$, and it denotes the transition probability from $(s, a) \in S \times A$ to $(s', a') \in S \times A$.

We assume that for each $\eta \in \mathcal{H}$, the policy $\Pi_\eta$ is ergodic, i.e., $Q_\eta$ has a unique stationary distribution $\nu$. Under this assumption, the average reward (or undiscounted reward) is given by

\[
J(\eta) := \sum_{s,a} \nu(s,a) R(s,a). \tag{37}
\]

The goal of the agent is to find a policy that maximizes the average reward over the class $\{ \Pi_\eta \}_{\eta \in \mathcal{H}}$. It can be verified (Sutton and Barto, 2018) that the gradient is evaluated by the limit:

\[
\nabla J(\eta) = \lim_{T \to \infty} E_\eta \left[ R(S_T, A_T) \sum_{i=0}^{T-1} \nabla \log \Pi_\eta(A_{T-i}; S_{T-i}) \right]. \tag{38}
\]

To approximate (38) with a numerically stable estimator, (Baxter and Bartlett, 2001) proposed the following gradient estimator. Let $\lambda \in [0, 1)$ be a discount factor and $T$ be sufficiently large, one has

\[
\hat{\nabla}_T J(\eta) := R(S_T, A_T) \sum_{i=0}^{T-1} \lambda^i \nabla \log \Pi_\eta(A_{T-i}; S_{T-i}) \approx \nabla J(\eta), \tag{39}
\]

where $(S_1, A_1, \ldots, S_T, A_T)$ is a realization of state-action sequence generated by the policy $\Pi_\eta$. This gradient estimator is biased and its bias is of order $O(1 - \lambda)$ as the discount factor $\lambda \uparrow 1$. The approximation above leads to the following policy gradient method (Baxter and Bartlett, 2001):

\[
\begin{align*}
G_{n+1} &= \lambda G_n + \nabla \log \Pi_{\eta_n}(A_{n+1}; S_{n+1}), \tag{40a} \\
\eta_{n+1} &= \eta_n + \gamma_{n+1} G_{n+1} R(S_{n+1}, A_{n+1}). \tag{40b}
\end{align*}
\]

We focus on a linear parameterization of the policy in the exponential family (or soft-max):

\[
\Pi_\eta(a; s) = \left\{ \sum_{a' \in A} \exp \left( \langle \eta | x(s, a') - x(s, a) \rangle \right) \right\}^{-1}, \tag{41}
\]

where $x(s, a) \in \mathbb{R}^d$ is a known feature vector. We make the following assumptions:
A10 For all \( s \in S, a \in A \), the feature vector \( x(s,a) \) and reward \( R(s,a) \) are bounded with \( \|x(s,a)\| \leq \bar{b}_x \), \( R(s,a) \leq R_{\text{max}} \).

A11 For all \( \eta \in \mathcal{H} \), the MC \( \{(S_t, A_t)\}_{t \geq 1} \), as governed by the transition matrix \( Q_\eta \) [cf. (36)], is uniformly geometrically ergodic: there exists \( \rho \in [0, 1) \), \( K_R < \infty \) such that, for all \( n \geq 0 \),

\[
\|Q^n_\eta - 1_{\mathcal{A}}\| \leq \rho^n K_R ,
\]

where \( v_\eta \in \mathbb{R}^{|\mathcal{A}|} \) is the stationary distribution of \( \{(S_t, A_t)\}_{t \geq 1} \). Moreover, there exists \( L_Q, L_v < \infty \) such that for any \( (\eta, \eta') \in \mathcal{H}^2 \),

\[
\|v_\eta - v_{\eta'}\| \leq L_Q\|\eta - \eta'\| , \quad \|J^n_{v_\eta}(\eta) - J^n_{v_{\eta'}}(\eta')\| \leq L_v\|\eta - \eta'\| ,
\]

where \( J^n_{v_\eta}(\eta) \) denotes the Jacobian of \( v_\eta \) w.r.t. \( \eta \).

Both A10 and A11 are regularity conditions on the MDP model that essentially hold as we focus on the finite state/action spaces setting. Under the uniform ergodicity assumption (42), the Lipschitz continuity conditions (43) can be implied using (Fort et al., 2011; Tadić and Doucet, 2017).

Our task is to verify that the policy gradient method (40) is an SA scheme with state-dependent Markovian noise [cf. Case 2 in Section 2]. To this end, we denote the joint state of this SA scheme as \( X_n = (S_n, A_n, G_n) \in \mathcal{X} := S \times A \times \mathbb{R}^d \), and notice that \( \{X_n\}_{n \geq 1} \) is a Markov chain. Adopting the same notation as in Section 2, the drift term and its mean field can be written as

\[
H_{\eta_n}(X_{n+1}) = G_{n+1} R(S_{n+1}, A_{n+1}) \quad \text{with} \quad h(\eta) = \lim_{T \to \infty} \mathbb{E}_{T \sim T \sim \mathbb{P}_{\eta}, S_1 \sim \mathbb{P}_{\eta}} [\nabla_T J(\eta)] ,
\]

where \( \nabla_T J(\eta) \) is defined in (39). Moreover, we let \( P_{\eta} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \) to be the Markov kernel associated with the MC \( \{X_n\}_{n \geq 1} \). Observe that

**Proposition 5** Under A10, it holds for any \( (\eta, \eta') \in \mathcal{H}^2 \), \( (s,a) \in S \times A \),

\[
\|\nabla \log \Pi_{\eta}(a; s)\| \leq 2\bar{b}_x, \quad \|\nabla \log \Pi_{\eta}(a; s) - \nabla \log \Pi_{\eta'}(a; s)\| \leq 8\bar{b}_x\|\eta - \eta'\| .
\]

Using the recursive update of (40a), we show that

\[
\|G_n\| = \|\lambda G_{n-1} + \nabla \log \Pi_{\eta}(A_n; S_n)\| \leq \lambda\|G_{n-1}\| + 2\bar{b} = \mathcal{O}(2\bar{b}\|G_0\|/(1 - \lambda)) ,
\]

for any \( n \geq 1 \), which then implies that the stochastic update \( H_{\eta_n}(X_{n+1}) \) in (40) is bounded since the reward is bounded using A10. The above proposition also implies that \( h(\eta) \) is bounded for all \( \eta \in \mathcal{H} \). Therefore, the assumption A7 is satisfied.

Next, with a slight abuse of notation, we shall consider the compact state space \( \mathcal{X} = S \times A \times G \), with \( G = \{g \in \mathbb{R}^d : \|g\| \leq C_0\bar{b}/(1 - \lambda)\} \) and \( C_0 \in [1, \infty) \), and analyze the policy gradient algorithm accordingly where \( \{X_{n+1}\}_{n \geq 0} \) is in \( \mathcal{X} \). Consider the following propositions whose proofs are adapted from (Fort et al., 2011; Tadić and Doucet, 2017) and can be found in Appendix C:

**Proposition 6** Under A10, A11, the following function is well-defined:

\[
\hat{H}_{\eta}(x) = \sum_{t=0}^{\infty} \left\{ P^n_\eta H_{\eta}(x) - h(\eta) \right\} ,
\]

and satisfies Eq. (7). For all \( x \in \mathcal{X} \), \( (\eta, \eta') \in \mathcal{H}^2 \), there exists constants \( L_{PH}^{(0)}, L_{PH}^{(1)} \) where

\[
\max\{\|P_\eta \hat{H}_{\eta}(x)\|, \|\hat{H}_{\eta}(x)\|\} \leq L_{PH}^{(0)}, \quad \|P_\eta \hat{H}_{\eta}(x) - P_{\eta'} \hat{H}_{\eta'}(x)\| \leq L_{PH}^{(1)}\|\eta - \eta'\| .
\]
Proposition 7 Under A10, A11, the gradient $\nabla J(\eta)$ is $\Upsilon$-Lipschitz continuous, where we defined $\Upsilon := \max |S||A|$. Moreover, for any $\eta \in \mathcal{H}$ and let $\Gamma := 2\tilde{b} R_{\max} K R (1 - \rho) \Upsilon$, it holds that

$$(1 - \lambda)^2 \Gamma^2 + 2 \langle \nabla J(\eta) | \text{h}(\eta) \rangle \geq \| \text{h}(\eta) \|^2, \quad \| \nabla J(\eta) \| \leq \| \text{h}(\eta) \| + (1 - \lambda) \Gamma.$$  \hspace{1cm} (49)

Proposition 6 verifies A5 and A6 for the policy gradient algorithm, while Proposition 7 implies A1 [with $c_0 = (1 - \lambda)^2 \Gamma^2$, $c_1 = 2$], A2 [with $d_0 = (1 - \lambda) \Gamma$, $d_1 = 1$], A3 [with $L = \Upsilon$]. As such, applying Theorem 2 gives

Corollary 2 Under A10, A11 and set $\gamma_k = (2c_1 L (1 + C_h) \sqrt{k})^{-1}$. For any $n \in \mathbb{N}$, let $N \in \{0, ..., n\}$ be an independent discrete r.v. distributed according to (4), the policy gradient algorithm (40) finds a policy, $\eta_N$, with

$$E[\| \nabla J(\eta_N) \|^2] = O \left( (1 - \lambda)^2 \Gamma^2 + \log n / \sqrt{n} \right),$$

where $J(\cdot)$ is defined in (37). The expectation is taken w.r.t. $N$ and action-state pairs $(A_n, S_n)$.

Related Studies The convergence of policy gradient method is typically studied for the episodic setting where the goal is to maximize the total reward over a finite horizon. The REINFORCE algorithm (Williams, 1992) has been analyzed as an SG method with unbiased gradient estimate in (Sutton et al., 2000), which proved an asymptotic convergence condition. A recent work (Papini et al., 2018) combined the variance reduction technique with the REINFORCE algorithm.

The infinite horizon setting is more challenging. To our best knowledge, the first asymptotically convergent policy gradient method is the actor-critic algorithm by Konda and Tsitsiklis (2003) which is extended to off-policy learning in (Degris et al., 2012). The analysis are based on the theory of two time-scales SA, which relies on controlling the ratio between the two set of step sizes used (Borkar, 1997). On the other hand, the algorithm which we have studied was a direct policy gradient method proposed by Baxter and Bartlett (2001), whose asymptotic convergence was proven only recently by Tadić and Doucet (2017). In comparison, our Corollary 2 provides the first non-asymptotic convergence for the policy gradient method. Of related interest, it is worthwhile to mention that (Fazel et al., 2018; Abbasi-Yadkori et al., 2018) have studied the global convergence for average reward maximization under the linear quadratic regulator setting where the state transition can be characterized by a linear dynamics and the reward is a quadratic function.

4. Conclusion

In this paper, we analyze under mild assumptions a general SA scheme with either zero-mean [cf. Case 1] or state-dependent/controlled Markovian [cf. Case 2] noise. We establish a novel non-asymptotic convergence analysis of this procedure without assuming convexity of the Lyapunov function. In both cases, our results highlight a convergence rate of order $O(\log(n) / \sqrt{n})$ under conservative assumptions. We verify our findings on two applications of growing interest: the online EM for learning an exponential family distribution (e.g., Gaussian Mixture Model) and the policy gradient method for maximizing an average reward.
Acknowledgement

HTW’s work is supported by the CUHK Direct Grant #4055113. The authors would like to thank the anonymous reviewers for valuable feedback.

References


Appendix A. Analysis of the SA Schemes

A.1. Proof of Lemma 1

**Lemma** Assume A1, A3. Then, for all \( n \geq 1 \), it holds that:

\[
\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1} (1 - c_1 L \gamma_{k+1}) h_k \\
\leq V(\eta_0) - V(\eta_{n+1}) + L \sum_{k=0}^{n} \gamma_{k+1}^2 \| e_{k+1} \|^2 + \sum_{k=0}^{n} \gamma_{k+1} (c_1^{-1} c_0 - \langle \nabla V(\eta_k) \mid e_{k+1} \rangle) .
\] (51)

**Proof** As the Lyapunov function \( V(\eta) \) is \( L \) smooth [cf. A3], we obtain:

\[
V(\eta_{k+1}) \leq V(\eta_k) - \gamma_{k+1} \langle \nabla V(\eta_k) \mid H_{\eta_k}(X_{k+1}) \rangle + \frac{L \gamma_{k+1}^2}{2} \| H_{\eta_k}(X_{k+1}) \|^2 \\
\leq V(\eta_k) - \gamma_{k+1} \langle \nabla V(\eta_k) \mid h(\eta_k) + e_{k+1} \rangle + L \gamma_{k+1}^2 (\| h(\eta_k) \|^2 + \| e_{k+1} \|^2) .
\] (52)

The above implies that

\[
\gamma_{k+1} \langle \nabla V(\eta_k) \mid h(\eta_k) \rangle \leq V(\eta_k) - V(\eta_{k+1}) - \gamma_{k+1} \langle \nabla V(\eta_k) \mid e_{k+1} \rangle + L \gamma_{k+1}^2 (\| h(\eta_k) \|^2 + \| e_{k+1} \|^2) .
\] (53)

Using A1, \( \langle \nabla V(\eta_k) \mid h(\eta_k) \rangle \geq \frac{1}{c_1} (h_k - c_0) \) and rearranging terms, we obtain

\[
\frac{\gamma_{k+1}}{c_1} (1 - c_1 L \gamma_{k+1}) h_k \leq V(\eta_k) - V(\eta_{k+1}) - \gamma_{k+1} \langle \nabla V(\eta_k) \mid e_{k+1} \rangle + L \gamma_{k+1}^2 \| e_{k+1} \|^2 + \frac{c_0}{c_1} \gamma_{k+1} .
\] (54)

Summing up both sides from \( k = 0 \) to \( k = n \) gives the conclusion (14).

\[\blacksquare\]

A.2. Proof of Lemma 2

**Lemma** Assume A1–A3, A5–A7 and the step sizes satisfy (9). Then:

\[
\mathbb{E} \left[ - \sum_{k=0}^{n} \gamma_{k+1} \langle \nabla V(\eta_k) \mid e_{k+1} \rangle \right] \leq C_h \sum_{k=0}^{n} \gamma_{k+1}^2 \mathbb{E} [\| h(\eta_k) \|^2] + C_\gamma \sum_{k=0}^{n} \gamma_{k+1}^2 + C_{0,n} ,
\] (55)

where \( C_h, C_\gamma \) and \( C_{0,n} \) are defined in (11), (12), (13).

**Proof** Under A5, A7, for any \( \eta \in \mathcal{H} \) there exists a bounded, measurable function \( x \to \hat{H}_\eta(x) \) such that the Poisson equation holds:

\[
e_{n+1} = H_{\eta_n}(X_{n+1}) - h(\eta_n) = \hat{H}_{\eta_n}(X_{n+1}) - P_{\eta_n} \hat{H}_{\eta_n}(X_{n+1}) .
\] (56)

The inner product on the left hand side of (18) can thus be decomposed as

\[
\mathbb{E} \left[ - \sum_{k=0}^{n} \gamma_{k+1} \langle \nabla V(\eta_k) \mid e_{k+1} \rangle \right] = \mathbb{E} [A_1 + A_2 + A_3 + A_4 + A_5] ,
\] (57)
with

\[ A_1 := -\sum_{k=1}^{n} \gamma_{k+1} \left< \nabla V(\eta_k) \mid \hat{H}_{\eta_k}(X_{k+1}) - P_{\eta_k} \hat{H}_{\eta_k}(X_k) \right>, \]

\[ A_2 := -\sum_{k=1}^{n} \gamma_{k+1} \left< \nabla V(\eta_k) \mid P_{\eta_k} \hat{H}_{\eta_k}(X_k) - P_{\eta_{k-1}} \hat{H}_{\eta_{k-1}}(X_k) \right>, \]

\[ A_3 := -\sum_{k=1}^{n} \gamma_{k+1} \left< \nabla V(\eta_k) - \nabla V(\eta_{k-1}) \mid P_{\eta_{k-1}} \hat{H}_{\eta_{k-1}}(X_k) \right>, \]

\[ A_4 := -\sum_{k=1}^{n} (\gamma_{k+1} - \gamma_k) \left< \nabla V(\eta_{k-1}) \mid P_{\eta_{k-1}} \hat{H}_{\eta_{k-1}}(X_k) \right>, \]

\[ A_5 := -\gamma_1 \left< \nabla V(\eta_0) \mid \hat{H}_{\eta_0}(X_1) \right> + \gamma_{n+1} \left< \nabla V(\eta_n) \mid P_{\eta_n} \hat{H}_{\eta_n}(X_{n+1}) \right>. \]

For \( A_1 \), we note that \( \hat{H}_{\eta_k}(X_{k+1}) - P_{\eta_k} \hat{H}_{\eta_k}(X_k) \) is a martingale difference sequence [cf. (2)] and therefore we have \( \mathbb{E}[A_1] = 0 \) by taking the total expectation.

For \( A_2 \), applying the Cauchy-Schwarz inequality and (8), we have

\[ A_2 \leq L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1} \| \nabla V(\eta_k) \| \| \eta_k - \eta_{k-1} \| \]

\[ = L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1} \gamma_k \| \nabla V(\eta_k) \| \| H_{\eta_{k-1}}(X_k) \| \]

\[ \overset{(a)}{\leq} L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1} \gamma_k \left( d_0 + d_1 \| h(\eta_k) \| \right) \left( \| h(\eta_{k-1}) \| + \sigma \right) \]

\[ \overset{(b)}{\leq} L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1} \gamma_k \left( d_0 \sigma + d_0 \| h(\eta_{k-1}) \| + d_1 \sigma \| h(\eta_k) \| + d_1 \| h(\eta_k) \| \| h(\eta_{k-1}) \| \right), \]

where (a) is due to \( A_2 \) on the norm of \( \nabla V(\eta_k) \) and \( A_7 \) on the norm of \( e_k \), (b) is obtained by expanding the scalar product. Using the inequality \( \| h(\eta_n) \| \leq 1 + \| h(\eta_n) \|^2 \) and \( 2 \| h(\eta_n) \| \| h(\eta_{n-1}) \| \leq \| h(\eta_n) \|^2 + \| h(\eta_{n-1}) \|^2 \), we obtain:

\[ A_2 \leq L_{PH}^{(1)} \left( (d_0 + d_0 \sigma + d_1 \sigma) \sum_{k=1}^{n} \gamma_k^2 \left( d_0 + \frac{d_1}{2} + ad_1 \sigma + \frac{ad_1}{2} \right) \sum_{k=0}^{n} \gamma_{k+1} \| h(\eta_k) \|^2 \right). \]

For \( A_3 \), we obtain

\[ A_3 \overset{(a)}{\leq} L \sum_{k=1}^{n} \gamma_{k+1} \gamma_k \| H_{\eta_{k-1}}(X_k) \| \| P_{\eta_{k-1}} \hat{H}_{\eta_{k-1}}(X_k) \| \]

\[ \overset{(b)}{\leq} LL_{PH}^{(0)} \sum_{k=1}^{n} \gamma_{k+1} \gamma_k \left( \| h(\eta_{k-1}) \| + \sigma \right) \]

\[ \leq LL_{PH}^{(0)} \left( (1 + \sigma) \sum_{k=1}^{n} \gamma_k^2 + \sum_{k=1}^{n} \gamma_k^2 \| h(\eta_{k-1}) \|^2 \right), \]
where (a) uses $A_3$, (b) uses $H_{\eta_{k-1}}(X_k) = h(\eta_{k-1}) + e_k$ and $A_6$.

For $A_4$, we have

$$A_4 \leq \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| (d_0 + d_1 \|h(\eta_{k-1})\|) \|P_{\eta_{k-1}} \hat{H}_{\eta_{k-1}}(X_k)\|^{(a)} \leq \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| + d_1 \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| \|h(\eta_{k-1})\|^2$$

$$= \sum_{k=1}^{n} (d_0 + 1) (\gamma_1 - \gamma_{n+1}) + d_1 \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| \|h(\eta_{k-1})\|^2$$

where (a) is again an application of $A_6$, and (b) uses the assumptions on step size $\gamma_{k+1} \leq \gamma_k$, $\gamma_k - \gamma_{k+1} \leq a' \gamma_k^2$. Finally, for $A_5$, we obtain

$$A_5 \leq \sum_{k=1}^{n} (d_0 + 1) (\gamma_1 - \gamma_{n+1}) + 2d_1 \|h(\eta_0)\|^2 + \gamma_{n+1} \|h(\eta_{n})\|^2 + d_1 \sum_{k=0}^{n} \gamma_{k+1}^2 \|h(\eta_{k})\|^2$$

where (a) is an application of $A_2$ and $A_6$, and (b) uses $a \leq 1 + a^2$. Gathering the relevant terms and taking expectations conclude the proof of this lemma.

### A.3. Lower bound on the rate of SA scheme

We provide a lower bound on $\mathbb{E}[\|h(\eta_N)\|]^{2}$ with the SA scheme (1) and (4):

**Lemma 3** Consider the SA scheme (1) with $h(\eta) = \nabla V(\eta)$. There exists a Lyapunov function $V(\eta)$ satisfying $A_3$ and a noise sequence $\{e_n\}_{n \geq 1}$ satisfying $A_4$-$A_7$ such that for any $n \geq 1$,

$$\mathbb{E}[\|h(\eta_N)\|^2] \geq \frac{\mathbb{E}[V(\eta_0) - V(\eta_{n+1})]}{\sum_{k=0}^{n} \gamma_{k+1}^2} + C_{lb} \sum_{k=0}^{n} \gamma_{k+1}^2$$

where $N$ is distributed according to (4), and $C_{lb} > 0$ is some constant independent of $n$.

For large $n$, setting $\gamma_k = c/\sqrt{k}$ minimizes the right hand side of (63), yielding $\mathbb{E}[\|h(\eta_N)\|^2] = \Omega(\log(n)/\sqrt{n})$. The considered SA scheme satisfies assumptions $A_1$-$A_7$, and the lower bound (63) matches the upper bounds in Theorem 1 & 2 (when $c_0 = 0$). The upper bounds are therefore tight.

We remark that our proof in Appendix A.3 uses the construction with a strongly convex Lyapunov function. It does not violate the known $\mathbb{E}[\|h(\eta_N)\|^2] = \mathcal{O}(1/n)$ rate in (Moulines and Bach, 2011) as the latter uses SA with a Polyak-Ruppert average estimator. To our best knowledge, it remains an open problem to lower bound the convergence rate of SA for smooth but non-convex Lyapunov function. We mention here a recent work (Fang et al., 2018, Remark 1) which shows $\mathbb{E}[\|h(\eta_N)\|^2] = \Omega(1/\sqrt{n})$ under different conditions than those satisfied in this paper.
Proof Our proof is achieved through constructing the Lyapunov and mean field function below. Consider a scalar parameter $\eta \in \mathbb{R}$ and set $V(\eta)$ to be a $\mu$-strongly convex and $L$-smooth function, where $0 < \mu \leq L < \infty$. Also, the mean field is set as
\[
h(\eta) = V'(\eta).
\]
Consider the following SA scheme (1) defined on the mean field $h$ as:
\[
\eta_{k+1} = \eta_k - \gamma_{k+1}(h(\eta_k) + e_{k+1}),
\]
where $e_k$ is i.i.d. and uniformly distributed on $[-\varepsilon, \varepsilon]$.

Clearly, the SA scheme (65) satisfies A1-A3 as we have set $V'(\eta) = h(\eta)$. The noise sequence is i.i.d. satisfying A4-A7. As $V$ is $\mu$-strongly convex, it can be shown
\[
V(\eta_{k+1}) \geq V(\eta_k) - \gamma_{k+1}V'(\eta_k)(h(\eta_k) + e_{k+1}) + \gamma_{k+1}^2 \frac{\mu}{2} (h(\eta_k) + e_{k+1})^2.
\]

Now by construction, we have $\mathbb{E}[\epsilon_{k+1} V'(\eta_k)|\mathcal{F}_k] = 0$, $\mathbb{E}[(h(\eta_k) + e_{k+1})^2|\mathcal{F}_k] \geq \frac{1}{3}\varepsilon^2$. Taking the total expectation on both sides gives
\[
\mathbb{E}[V(\eta_{k+1})] \geq \mathbb{E}[V(\eta_k)] - \gamma_{k+1} h^2(\eta_k) + \gamma_{k+1}^2 \frac{\mu\varepsilon^2}{6}.
\]

Denote $C_{lb} := \frac{\mu\varepsilon^2}{6}$. Using (4), we observe
\[
\mathbb{E}[(h(\eta_N))^2] = \frac{1}{\sum_{k=0}^{n} \gamma_{k+1}} \sum_{k=0}^{n} \gamma_{k+1} \mathbb{E}[(h(\eta_k))^2] \geq \frac{\mathbb{E}[V(\eta_0) - V(\eta_{n+1})] + C_{lb} \sum_{k=0}^{n} \gamma_{k+1}^2}{\sum_{k=0}^{n} \gamma_{k+1}}.
\]

This completes the proof of the lower bound.

Appendix B. Analysis of the ro-EM method

B.1. Proof of Proposition 1

Proposition Assume A8. Then

• If $h(s^*) = 0$ for some $s^* \in S$, then $\nabla_\theta \text{KL}(\pi, g_{\theta^*}) + \nabla_\theta R(\theta^*) = 0$ with $\theta^* = \overline{\theta}(s^*)$.

• If $\nabla_\theta \text{KL}(\pi, g_{\theta^*}) + \nabla_\theta R(\theta^*) = 0$ for some $\theta^* \in \Theta$ then $s^* = \mathbb{E}_\pi[S(Y, \theta^*)]$.

Proof We have
\[
\nabla_\theta \text{KL}(\pi, g; \theta) = -\nabla_\theta \mathbb{E}_\pi [\log g(Y; \theta)] = -\mathbb{E}_\pi [\nabla_\theta \log g(Y; \theta)],
\]
where the last equality assumes that we can exchange integration with differentiation. Furthermore, using the Fisher’s identity (Douc et al., 2014), it holds for any $y \in Y$ that
\[
\nabla_\theta \log g(y; \theta) = -\langle \nabla_\theta \psi(\theta) + J_\theta^0(\theta) \pi(y; \theta) \rangle = -\nabla_\theta \psi(\theta) + J_\theta^0(\theta) \mathbb{E}_\theta [S(X)|Y = y].
\]
Therefore, for any $s$, it holds that
\[
\nabla_{\theta} \text{KL}(\pi, g(\cdot; \mathbf{\bar{\theta}}(s))) + \nabla_{\theta} R(\mathbf{\bar{\theta}}(s)) = \nabla_{\theta} \psi(\mathbf{\bar{\theta}}(s)) + \nabla_{\theta} R(\mathbf{\bar{\theta}}(s)) - J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s)) \mathbb{E}_{\pi}[\mathbf{\bar{r}}(Y; \mathbf{\bar{\theta}}(s))] \\
\overset{(a)}{=} J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s)) \left( s - \mathbb{E}_{\pi}[\mathbf{\bar{r}}(Y; \mathbf{\bar{\theta}}(s))] \right) \overset{(b)}{=} J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s)) h(s). \tag{71}
\]

where we have used the assumption A8 in (a) and the definition of $h(s)$ in (b). The conclusion follows directly from the identity (71) since $J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s))$ is full rank. 

\[\]  

B.2. Proof of Proposition 2

**Proposition** Assume A8. Then, for $s \in \mathcal{S}$,
\[
\nabla_{s} V(s) = J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s)) \left( H_{\ell}^{\theta}(s; \theta) \right)^{-1} J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s))^{T} h(s). \tag{72}
\]

**Proof** Using chain rule and A8, we obtain
\[
\nabla_{s} V(s) = J_{\phi}^{s}(s) \left( \nabla_{\theta} \text{KL}(\pi, g(\cdot; \mathbf{\bar{\theta}}(s))) + \nabla_{\theta} R(\mathbf{\bar{\theta}}(s)) \right) \\
= J_{\phi}^{s}(s) J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s))^{T} h(s), \tag{73}
\]

where the last equality uses the identity in (71). Consider the following vector map:
\[
s \rightarrow \nabla_{\theta} \psi(\mathbf{\bar{\theta}}(s)) + \nabla_{\theta} R(\mathbf{\bar{\theta}}(s)) - J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s))^{T} s. \tag{74}
\]
Taking the gradient of the above map w.r.t. $s$ and note that the map is constant for all $s \in \mathcal{S}$, we show that:
\[
0 = -J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s)) + \left[ \nabla_{\theta}^{2} \psi(\theta) + R(\theta) - \langle \phi(\theta) | s \rangle \right] \bigg|_{\theta = \mathbf{\bar{\theta}}(s)} J_{\phi}^{s}(s) \tag{75}
\]
This implies $J_{\phi}^{s}(s) = \left( \Pi_{\ell}^{\theta}(s; \mathbf{\bar{\theta}}(s)) \right)^{-1} J_{\phi}^{\theta}(\mathbf{\bar{\theta}}(s))$. Substituting into (73) yields the conclusion. 

B.3. Proof of Proposition 3

**Proposition** Under A9, it holds that $\mathbb{E}[\|\tilde{\pi}(Y_{n+1}; \mathbf{\hat{\theta}}_{n}) - \tilde{s}_{n}\|^{2} | \mathcal{F}_{n}] \leq 2M^{2}$ for all $n \geq 0$.

**Proof** From (26), we note that the error term is given by
\[
e_{n+1} = H_{\delta_{n}}(Y_{n+1}) - h(\tilde{s}_{n}) = \left( \begin{array}{c} \mathbb{E}_{Y_{n+1} \sim \pi}[\mathbf{s}_{n}^{(1)} | \mathcal{F}_{n}] - \mathbf{s}_{n}^{(1)} \\
\mathbb{E}_{Y_{n+1} \sim \pi}[\mathbf{s}_{n}^{(2)} | \mathcal{F}_{n}] - \mathbf{s}_{n}^{(2)} \\
\mathbb{E}_{Y_{n+1} \sim \pi}[\mathbf{s}_{n}^{(3)} | \mathcal{F}_{n}] - \mathbf{s}_{n}^{(3)} \end{array} \right). \tag{76}
\]

Obviously, it holds that $\mathbb{E}[e_{n+1} | \mathcal{F}_{n}] = 0$. Furthermore, for all $m \in \{1, \ldots, M - 1\}$, the $m$th element of the first block in $e_{n+1}$ has a bounded conditional variance
\[
\mathbb{E}\left[ \|\mathbb{E}_{Y_{n+1} \sim \pi}[\omega_{m}(Y_{n+1}; \mathbf{\hat{\theta}}_{n})] - \omega_{m}(Y_{n+1}; \mathbf{\hat{\theta}}_{n})\|^{2} \right] \leq 1. \tag{77}
\]
For the second block in $e_{n+1}$, the conditional variance of its $m$th element is

$$
\mathbb{E}\left[|E_{Y_{n+1} \sim \pi}[Y_{n+1} \omega_m(Y_{n+1}; \hat{\theta}_n)] - Y_{n+1} \omega_m(Y_{n+1}; \hat{\theta}_n)|^2\right]
$$

$$
= \mathbb{E}\left[|Y_{n+1} \omega_m(Y_{n+1}; \hat{\theta}_n)|^2 - |E_{Y_{n+1} \sim \pi}[Y_{n+1} \omega_m(Y_{n+1}; \hat{\theta}_n)]|^2\right]
$$

$$
\leq \mathbb{E}\left[|Y_{n+1} \omega_m(Y_{n+1}; \hat{\theta}_n)|^2\right] \leq \mathbb{E}[|Y_{n+1}|^2] \leq \Upsilon^2.
$$

(78)

Lastly, we also have $\mathbb{E}[|E_{Y_{n+1} \sim \pi}[s_n^{(3)} | F_n] - s_n^{(3)}|^2] \leq \Upsilon^2$. Therefore, we conclude that $\mathbb{E}[|e_{n+1}|^2 | F_n] \leq M - 1 + M\Upsilon^2 < \infty$. 

\section*{B.4. Proof of Proposition 4}

\textbf{Proposition} Under A9 and the regularizer (32) set with $\epsilon > 0$, then for all $(s, s') \in S^2$, there exists positive constants $v, \Upsilon, \Psi$ such that:

$$
(\nabla V(s) | h(s)) \geq v\|h(s)\|^2, \quad \|\nabla V(s)\| \leq \Upsilon\|h(s)\|, \quad \|\nabla V(s) - \nabla V(s')\| \leq \Psi\|s - s'\|.
$$

(79)

\textbf{Proof} We first check that A8 is satisfied under A9. In particular, one observes that when $s \in S = \Delta_{M-1} \times [-\Upsilon, \Upsilon]^M$, the M-step update (33) is the unique solution satisfying the stationary condition of the minimization problem (24) and $\bar{\theta}(s) \in C$.

As A8 is satisfied, applying Proposition 2 shows that the gradient of the Lyapunov function is

$$
\nabla V(s) = \frac{\partial^0(\bar{\theta}(s))}{\partial \ell^0(\bar{\theta}(s))^\top} h(s).
$$

(80)

Using (29), we observe that for any given $\theta \in C$, the Jacobian of $\phi$ and the Hessian of $\ell(s, \theta)$ are given by

$$
J_0^\theta(\theta) = \left(\begin{array}{ccc}
\frac{1}{1 - \sum_{m=1}^{M-1} \omega_m} & \frac{1}{\omega_m} & -\text{Diag}(\frac{1}{\omega}) & -\text{Diag}(\mu) & \mu M^1
\end{array}\right),
$$

(81)

$$
H_\ell^\theta(s, \theta) = \left(\begin{array}{ccc}
\frac{1 + \epsilon - \sum_{m=1}^{M-1} s^{(1)}_m}{(1 - \sum_{m=1}^{M-1} \omega_m)2} & 11^\top & \text{Diag}\left(\frac{s^{(1)}_m + \epsilon}{\omega^2_m}\right) & 0 & 0 \\
0 & 11^\top & \text{Diag}(s^{(1)} + \epsilon 1) & 0 & 0 \\
0 & 0 & 1 + \epsilon - \sum_{m=1}^{M-1} s^{(1)}_m & 0 & 0
\end{array}\right),
$$

where we have denoted $\frac{s^{(1)}_m + \epsilon}{\omega^2_m}$ as the $(M - 1)$-vector $\left(\frac{s^{(1)}_1 + \epsilon}{\omega^2_1}, \ldots, \frac{s^{(1)}_{M-1} + \epsilon}{\omega^2_{M-1}}\right)$. Let us define $J_{11}, H_{11}$ as the top-left matrices in the above, evaluated at $\bar{\theta}(s)$, as follows

$$
J_{11} := \frac{1}{1 - \frac{1}{1 + \epsilon M}} 11^\top + \text{Diag}\left(\frac{1 + \epsilon M}{s^{(1)} + \epsilon 1}\right)
$$

(82)

$$
H_{11} := \frac{1 + \epsilon - \sum_{m=1}^{M-1} s^{(1)}_m}{(1 - \frac{1}{1 + \epsilon M})2} 11^\top + \text{Diag}\left(\frac{(1 + \epsilon M)^2}{s^{(1)} + \epsilon 1}\right).
$$

(83)
When $\epsilon > 0$, the above matrices, $J_{11}$ and $H_{11}$, are full rank and bounded if $s \in S$.

The matrix product $J_{\phi}(\overline{\theta}(s))(H_{\phi}(s, \overline{\theta}(s)))^{-1}J_{\phi}(\overline{\theta}(s))^T$ can hence be expressed as an outer product

$$J_{\phi}(\overline{\theta}(s))(H_{\phi}(s, \overline{\theta}(s)))^{-1}J_{\phi}(\overline{\theta}(s))^T = J(s)J(s)^T,$$

with

$$J(s) := J_{\phi}(\overline{\theta}(s)) \begin{pmatrix} H_{11}^{-\frac{1}{2}} & 0 & \frac{1}{\sqrt{s^{(1)}+1}} & 0 \\ 0 & \text{Diag} & 0 & \frac{1}{\sqrt{1+\epsilon - \sum_{m=1}^{M-1} s^{(1)}_m}} \\ 0 & 0 & \frac{s^{(2)}_{11}}{s^{(1)}+1} & \frac{1}{\sqrt{1+\epsilon - \sum_{m=1}^{M-1} s^{(1)}_m}} \\ 0 & \frac{1}{\sqrt{s^{(1)}+1}} & 0 & \frac{1}{\sqrt{1+\epsilon - \sum_{m=1}^{M-1} s^{(1)}_m}} \end{pmatrix}.$$ (85)

Under A9 and using the above structured form, it can be verified that $J(s)$ is a bounded and full rank matrix. As such, for all $s \in S$, there exists $v > 0$ such that

$$\langle \nabla V(s) \mid h(s) \rangle = \langle J(s)J(s)^T h(s) \mid h(s) \rangle \geq v \| h(s) \|^2.$$

The second part in (34) can be verified by observing that $J_{\phi}(\overline{\theta}(s))(H_{\phi}(s; \theta))^{-1}J_{\phi}(\overline{\theta}(s))^T$ is bounded due to A9.

For the third part in (34), again from (80) we obtain:

$$\nabla V(s) = J(s)J(s)^T h(s).$$ (87)

From (85), it can be seen that $J(s)J(s)^T$ is Lipschitz continuous in $s$ and bounded, i.e., there exists constants $L_J, C_J < \infty$ such that

$$\|J(s)J(s)^T - J(s')J(s')^T\| \leq L_J \|s - s'\|, \quad \|J(s)J(s)^T\| \leq C_J, \quad \forall s, s' \in S.$$ (88)

For example, the above can be checked by observing that the Hessian (w.r.t. $s$) of each entry in $J(s)J(s)^T$ is bounded for $s \in S$. On the other hand, the mean field $h(s)$ satisfies,

$$\|h(s) - h(s')\| = \|s - s' + \mathbb{E}_{Y \sim \pi}[\overline{\sigma}(Y; \overline{\theta}(s')) - \overline{\sigma}(Y; \overline{\theta}(s))]\| \leq \|s - s'\| + \mathbb{E}_{Y \sim \pi}\|\overline{\sigma}(Y; \overline{\theta}(s')) - \overline{\sigma}(Y; \overline{\theta}(s))\|,$$ (89)

where (a) uses the triangular inequality and the Jensen’s inequality. Moreover, we observe

$$\overline{\sigma}(Y; \overline{\theta}(s')) - \overline{\sigma}(Y; \overline{\theta}(s)) = \begin{pmatrix} \overline{\omega}(Y; \overline{\theta}(s')) - \overline{\omega}(Y; \overline{\theta}(s)) \\ Y(\overline{\omega}(Y; \overline{\theta}(s')) - \overline{\omega}(Y; \overline{\theta}(s))) \end{pmatrix},$$ (90)
where \( \tilde{\omega}(Y; \tilde{\theta}(s)) \) is a collection of the \( M - 1 \) terms \( \tilde{\omega}_m(Y; \tilde{\theta}(s)) \), \( m = 1, \ldots, M - 1 \) [cf. (30)]. Observe that

\[
\tilde{\omega}_m(Y; \tilde{\theta}(s)) = \frac{\frac{s_j^{(1)}}{s_j^{(1)} + \epsilon}}{\sum_{j=1}^M \frac{s_j^{(1)}}{s_j^{(1)} + \epsilon}} \exp(-\frac{1}{2}(Y - \frac{s_j^{(2)}}{s_j^{(1)} + \epsilon})^2).
\]  

(91)

Under A9 and the condition that \( s \in \mathcal{S} \), i.e., a compact set, there exists \( L_\omega < \infty \) such that

\[
|\tilde{\omega}_m(Y; \tilde{\theta}(s)) - \tilde{\omega}_m(Y; \tilde{\theta}(s'))|^2 \leq L_\omega^2 \| s - s' \|^2,
\]  

(92)

for all \( m = 1, \ldots, M - 1 \). Consequently, again using A9, we have

\[
\| \tilde{\pi}(Y; \tilde{\theta}(s')) - \tilde{\pi}(Y; \tilde{\theta}(s)) \| \leq (M - 1)(1 + \bar{Y})L_\omega \| s - s' \|,
\]  

(93)

and we have \( \| h(s) - h(s') \| \leq L_h \| s - s' \| \) for some \( L_h < \infty \). It can also be shown easily that \( \| h(s) \| \leq C_h \) for all \( s \in \mathcal{S} \). Finally, we observe the following chain:

\[
\| \nabla V(s) - \nabla V(s') \| = \| J(s)J(s)^\top h(s) - J(s')J(s')^\top h(s') \|
= \| J(s)J(s)^\top (h(s) - h(s')) + (J(s)J(s)^\top - J(s')J(s')^\top)h(s') \|
\leq (L_hC_J + L_JC_h) \| s - s' \|,
\]  

(94)

which concludes our proof.

**Appendix C. Analysis on the Policy Gradient Algorithm**

This section proves a few key lemmas that are modified from (Tadić and Doucet, 2017) which lead to the convergence of the policy gradient algorithm analyzed in Section 3.2.

Let \( \tilde{Q}_\eta := Q_\eta - 1v^\top_\eta \) and denote \( \tilde{Q}_\eta^t((s, a); (s', a')) \) to be the \((s, a); (s', a')\)th element of the \( t \)th power of \( \tilde{Q}_\eta^t \). Under A11, we observe that \( \| \tilde{Q}_\eta^t \| \leq \rho^t K_R \) for any \( t \geq 0 \). For \( i = 1, \ldots, d \), we also define the \((s, a)\)th element of the \(|S| \cdot |A|\)-dimensional gradient vector \( \nabla_i \Pi_\eta \) and reward vector \( r \), respectively as:

\[
\nabla_i \Pi_\eta(s, a) := \frac{\partial \log \Pi(a; s, \eta)}{\partial \eta_i}, \quad r(s, a) := R(s, a).
\]  

(95)

Using the above notations, the mean field in (44) can be evaluated as

\[
h(\eta) = \sum_{t=0}^\infty \sum_{(s,a),(s',a') \in \mathcal{S} \times \mathcal{A}} \lambda^t R(s', a')\tilde{Q}_\eta^t((s, a); (s', a')) \nabla \log \Pi(a; s, \eta) v_\eta(s, a).
\]  

(96)

In particular, its \( i \)th element can be expressed as

\[
h_i(\eta) = \sum_{t=0}^\infty \lambda^t v^\top_\eta \text{Diag}(\nabla_i \Pi_\eta) \tilde{Q}_\eta^t r.
\]  

(97)

We also define the difference between \( h(\eta) \) and \( \nabla J(\eta) \) as

\[
\Delta(\eta) := h(\eta) - \nabla J(\eta).
\]  

(98)
C.1. Useful Lemmas

Lemma 4  Let $A10, A11$ hold. For any $(\eta, \eta') \in \mathcal{H}^2$ and $t \geq 0$, one has

$$\|Q^t_\eta - Q^t_{\eta'}\| \leq C_1 \|\eta - \eta'\|, \quad \|\tilde{Q}^t_\eta - \tilde{Q}^t_{\eta'}\| \leq C_1 (t \rho^t) \|\eta - \eta'\|,$$

(99)

where we have set $C_1 := \rho K^2_R (2\tilde{b} + L_Q) + L_Q$ in the above.

Proof  For part 1), we observe that each entry of $Q_\eta$ is given by [cf. (36)]:

$$Q_\eta((s, a); (s', a')) := \Pi(a'; s', \eta) P^a_{s,s'},$$

which is Lipschitz continuous w.r.t. $\eta$ since

$$\nabla \Pi(a|s, \eta) = -\left( \sum_{a' \in A} \exp (\langle \eta | x(s, a') - x(s, a) \rangle) \right)^{-2} \sum_{a' \in A} \exp (\langle \eta | x(s, a') - x(s, a) \rangle) (x(s, a') - x(s, a))$$

is bounded by $\max_{s,a,a'} \|x(s, a') - x(s, a)\| \leq 2\tilde{b}$ [cf. A10]. This implies

$$|Q_\eta((s, a); (s', a')) - Q_{\eta'}((s, a); (s', a'))| \leq 2\tilde{b} |P^a_{s,s'}| \|\eta - \eta'\|.$$

(100)

Since $|P^a_{s,s'}| \leq 1$ for any $s, s', a$, we have $\|Q_\eta - Q_{\eta'}\| \leq 2\tilde{b} \|\eta - \eta'\|$.

For any $\eta \in \mathcal{H}$ and any $t \geq 0$, we have:

$$\tilde{Q}^{t+1}_\eta - \tilde{Q}^{t+1}_{\eta'} = \sum_{\tau=0}^{t} \tilde{Q}^\tau_\eta (\tilde{Q}_\eta - \tilde{Q}_{\eta'}) \hat{Q}^{t-\tau}_{\eta'}$$

$$= \sum_{\tau=0}^{t} \tilde{Q}^\tau_\eta (Q_\eta - Q_{\eta'} - 1 (v_\eta - v_{\eta'})^\top) \hat{Q}^{t-\tau}_{\eta'}.$$

(101)

As such,

$$\|\tilde{Q}^{t+1}_\eta - \tilde{Q}^{t+1}_{\eta'}\| \leq \sum_{\tau=0}^{t} \|\tilde{Q}^\tau_\eta\| \|Q_\eta - Q_{\eta'} - 1 (v_\eta - v_{\eta'})^\top\| \|\hat{Q}^{t-\tau}_{\eta'}\|$$

$$\leq K^2_R \sum_{\tau=0}^{t} \rho^\tau \rho^t \|Q_\eta - Q_{\eta'}\| + \|v_\eta - v_{\eta'}\|$$

$$\leq K^2_R (2\tilde{b} + L_Q) (t \rho^t) \|\eta - \eta'\|.$$

(102)

Consequently,

$$\|Q^{t+1}_\eta - Q^{t+1}_{\eta'}\| \leq \|\tilde{Q}^{t+1}_\eta - \tilde{Q}^{t+1}_{\eta'}\| + \|v_\eta - v_{\eta'}\|$$

$$\leq (K^2_R (t \rho^t) (2\tilde{b} + L_Q) + L_Q) \|\eta - \eta'\|.$$

(103)

Setting $C_1 := \rho K^2_R (2\tilde{b} + L_Q) + L_Q$ completes the proof.
Lemma 5  Let $A_{10}$, $A_{11}$ hold. The following statements are true:

1. The average reward $J(\eta)$ is differentiable and for any $(\eta, \eta') \in H^2$, one has
   \[
   \|\nabla J(\eta) - \nabla J(\eta')\| \leq R_{\max} |S||A|L_\upsilon \|\eta - \eta'\|. \tag{104}
   \]

2. For any $\eta \in H$, one has
   \[
   \|\Delta(\eta)\| \leq 2\tilde{b} R_{\max} K_R \frac{1 - \lambda}{(1 - \rho)^2}. \tag{105}
   \]

Proof  For part 1), we observe that
   \[
   J(\eta) = \mathbb{E}_{(S,A) \sim \upsilon_\eta} [\mathcal{R}(S,A)] = \sum_{(s,a) \in S \times A} \upsilon_\eta(s,a) R(s,a). \tag{106}
   \]

It follows from the Lipschitz continuity of $J^T_{\upsilon_\eta}(\eta)$ [cf. A11] that
   \[
   \|\nabla J(\eta) - \nabla J(\eta')\| \leq \sum_{(s,a) \in S \times A} |\mathcal{R}(s,a)| \|\nabla \upsilon_\eta(s,a) - \nabla \upsilon_\eta(s,a)\|
   \leq R_{\max} |S||A|L_\upsilon \|\eta - \eta'\|. \tag{107}
   \]

The above verifies (104).

For part 2), we define
   \[
   J_T(\eta, (s,a)) := \sum_{(s',a') \in S \times A} \mathcal{R}(s', a') Q^T_\eta((s,a); (s',a')), \tag{108}
   \]

\[
   g(\eta) := \sum_{t=0}^{\infty} \sum_{(s,a),(s',a') \in S \times A} \mathcal{R}(s,a) \tilde{Q}^t_\eta((s,a); (s',a')) \nabla \log \Pi(a; s, \eta) v_\eta(s,a). \tag{109}
   \]

As shown in (Tadić and Doucet, 2017, Lemma 8.2), we have $\lim_{T \to \infty} \nabla \eta J_T(\eta, (s,a)) = g(\eta)$ for all $\eta \in H$ and $(s,a) \in S \times A$. As such
   \[
   \Delta(\eta) = h(\eta) - g(\eta)
   \]

\[
= \sum_{t=0}^{\infty} \sum_{(s,a),(s',a') \in S \times A} (\lambda^t - 1) \mathcal{R}(s,a) \tilde{Q}^t_\eta((s,a); (s',a')) \nabla \log \Pi(a; s, \eta) v_\eta(s,a). \tag{110}
   \]

and in particular, the $i$th element is given by
   \[
   \Delta_i(\eta) = \sum_{t=0}^{\infty} \sum_{(s,a),(s',a') \in S \times A} (\lambda^t - 1) v^T_\eta \text{Diag}(\nabla_i \Pi_\eta) \tilde{Q}^t_\eta r, \tag{111}
   \]

which can be bounded as
   \[
   |\Delta_i(\eta)| \leq \sum_{t=0}^{\infty} (1 - \lambda^t) \|v_\eta\| \|\nabla_i \Pi_\eta\|_\infty \|\tilde{Q}^t_\eta\| \|r\|
   \leq 2\tilde{b} R_{\max} K_R \sum_{t=0}^{\infty} (1 - \lambda^t) \rho^t \leq 2\tilde{b} R_{\max} K_R \frac{1 - \lambda}{(1 - \rho)^2}. \tag{112}
   \]
where (a) uses A11, A10, and Proposition 5. The above implies that \( \| \Delta(\eta) \| \leq 2\overline{b} R_{\max} K_{R \frac{1-A}{1-\rho}} \). ■

**Lemma 6** Let A10, A11 hold. Denote the joint state \( x \) as \( x = (s, a, g) \in S \times A \times \mathbb{R}^d \). There exists \( \delta \in [0, 1) \), \( C_2 \in [1, \infty) \) such that for any \( t \geq 0 \),

\[
\begin{align*}
\| P^t_\eta H_\eta(x) - h(\eta) \| & \leq C_2 t \delta^t (1 + \| g \|) , \\
\big\| (P^t_\eta H_\eta(x) - h(\eta)) - (P^t_\eta H_\eta'(x) - h(\eta')) \big\| & \leq C_2 t \delta^t \| \eta - \eta' \| (1 + \| g \|) .
\end{align*}
\] (113)

**Proof** Denote the joint state as \( x = (s, a, g) \), we observe that

\[
P^t_\eta H_\eta(x) = \mathbb{E}_{\Pi_\eta} \left[ \mathcal{R}(S_t, A_t) G_t \big| (S_0, A_0) = (s, a), G_0 = g \right]
= \mathbb{E}_{\Pi_\eta} \left[ \mathcal{R}(S_t, A_t) \left( \lambda^t g + \sum_{i=1}^{t-1} \lambda^i \nabla \log \Pi(A_i; S_i, \eta) \right) \big| (S_0, A_0) = (s, a) \right]
= \sum_{i=0}^{t-1} \sum_{(s', a'), (s'', a'') \in S \times A} \lambda^i \mathcal{R}(s', a'') Q^t_\eta((s', a'); (s'', a'')) \nabla \log \Pi(a'; s', \eta) Q^{t-i}(s, a); (s', a')) + \lambda^t g \sum_{(s', a') \in S \times A} \mathcal{R}(s', a') Q^t_\eta((s, a); (s', a')) .
\]

The \( j \)th element of the above is thus given by

\[
[P^t_\eta H_\eta(x)]_j = \sum_{i=0}^{t-1} \lambda^i e_{(s,a)}^\top \mathcal{R}^{t-i}(\nabla_j \Pi_\eta) \mathcal{Q}^{t-i}_\eta r + \lambda^t g_j 1^\top \mathcal{Q}^{t}_\eta r ,
\] (114)

where \( g_j \) is the \( j \)th element of \( g \) and \( e_{(s,a)} \) is the \( (s, a) \)th coordinate vector. Moreover, we recall that

\[
h_j(\eta) = \sum_{t=0}^{\infty} \lambda^t v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) \mathcal{Q}^{t}_\eta r .
\] (115)

Note that

\[
v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) 1 = \sum_{(s,a) \in S \times A} v_\eta(s, a) \nabla j \log \Pi(a; s, \eta)
= \sum_{s \in S} \left( \sum_{a \in A} \Pi(a; s, \eta) \nabla_j \log \Pi(a; s, \eta) \right) \Pi_\eta(s) = 0 .
\] (116)

where we recalled that \( \Pi_\eta(s) \) is the stationary distribution for the MDP on the state. Using the decomposition \( \mathcal{Q}^t_\eta = \mathcal{Q}^t_\eta - 1 v^\top_\eta \), we observe

\[
h_j(\eta) = \sum_{i=0}^{t-1} \lambda^i \left\{ v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) \mathcal{Q}^i_\eta r - v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) 1 \right\} + \sum_{i=t}^{\infty} \lambda^i v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) \mathcal{Q}^i_\eta r
= \sum_{i=0}^{t-1} \lambda^i v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) \mathcal{Q}^i_\eta r + \sum_{i=t}^{\infty} \lambda^i v_\eta^\top \mathcal{D} \left( \nabla_j \Pi_\eta \right) \mathcal{Q}^i_\eta r .
\]

26
Using $A_{10}$, $A_{11}$ and notice that $\parallel \nabla \psi(x) \parallel \leq 2 \delta_0$, $\parallel \nabla \psi(x) \parallel \leq R$, $\parallel \nabla \psi(x) \parallel \leq R K R \sqrt{|S||A|} \delta_i$, we obtain
\[
\left| \quad [P_{\eta}^t H_\eta(x)]_j - h_j(\eta) \quad \right| \leq 2 \delta_0 R K R \sum_{i=0}^{t-1} \lambda^i R^{t-i} + \lambda^t |g_j| R + 2 \delta_0 R K R \sqrt{|S||A|} \sum_{i=t}^{\infty} \lambda^i \rho^i .
\]
Consequently, we obtain the upper bound as
\[
\left| [P_{\eta}^t H_\eta(x)]_j - h_j(\eta) \right| \leq \sum_{i=0}^{t-1} \lambda^i \parallel \nabla \psi(x) \parallel \left| \nabla \psi(x) \right| \leq 2 \delta_0 R K R \sum_{i=0}^{t-1} \lambda^i R^{t-i} + \lambda^t |g_j| R + 2 \delta_0 R K R \sqrt{|S||A|} \sum_{i=t}^{\infty} \lambda^i \rho^i .
\]

Observe that each of the above term decays geometrically with $t$, as such there exists $C_2' \in [1, \infty)$, $\delta \in [0, 1)$ such that
\[
\left| [P_{\eta}^t H_\eta(x)]_j - h_j(\eta) \right| \leq C_2' (t \delta^t) \left( 1 + \parallel g \parallel \right) ,
\]
which naturally implies the first equation in (113).

For the second equation in (113),
\[
\left[ P_{\eta}^t H_\eta(x) \right]_j - h_j(\eta) = \left[ P_{\eta}^t H_\eta(x) \right]_j - h_j(\eta)'
\]
\[
= \sum_{i=0}^{t-1} \lambda^i e_{(s,a)} \left\{ \nabla \psi(x) \right\}_j \left( \nabla \psi(x) \right)_j - \nabla \psi(x) \right\}_j \left( \nabla \psi(x) \right)_j - h_j(\eta)'
\]
\[
+ \lambda^i |g_j| R^i \left( \nabla \psi(x) \right)_j \left( \nabla \psi(x) \right)_j - \nabla \psi(x) \right\}_j \left( \nabla \psi(x) \right)_j - h_j(\eta)'
\]
\[
+ \lambda^i |g_j| R^i \left( \nabla \psi(x) \right)_j \left( \nabla \psi(x) \right)_j - \nabla \psi(x) \right\}_j \left( \nabla \psi(x) \right)_j - h_j(\eta)'
\]
1. Note that an exact characterization for $C_2'$ is also possible.
This leads to the upper bound:

\[
\left| P^t_{\eta} H_{\eta}(x) \right|_j - h_j(\eta) - \left\{ P^t_{\eta'} H_{\eta'}(x) \right|_j - h_j(\eta') \right\| \\
\leq \sqrt{|S||A|R} \sum_{i=0}^{\infty} \lambda^i \left\| \tilde{Q}^t_{\eta} - \tilde{Q}^t_{\eta'} \right\| + \lambda^i S||A||Q^t_{\eta} - Q^t_{\eta'}|| + \sqrt{|S||A|R} \sum_{i=0}^{\infty} \lambda^i \left\| v_{\eta}^T \text{Diag}(\nabla_j \Pi_{\eta}) \tilde{Q}^t_{\eta} - v_{\eta'}^T \text{Diag}(\nabla_j \Pi_{\eta'}) \tilde{Q}^t_{\eta'} \right\|
\]

(122)

Using the boundedness and Lipschitz continuity of \( \nabla_j \Pi_{\eta}, v_{\eta}, Q^t_{\eta}, \tilde{Q}^t_{\eta} \) [cf. Lemma 4], let \( C_{2,1}, C_{2,2} \in [1, \infty) \), the norms in the above can be bounded as

\[
\left\| \tilde{Q}^t_{\eta} - \tilde{Q}^t_{\eta'} \right\| \leq C_{2,1} \left( (t-i) \rho^i \right) \|\eta - \eta'\| \\
\left\| v_{\eta}^T \text{Diag}(\nabla_j \Pi_{\eta}) \tilde{Q}^t_{\eta} - v_{\eta'}^T \text{Diag}(\nabla_j \Pi_{\eta'}) \tilde{Q}^t_{\eta'} \right\| \leq C_{2,2} \left( i \rho^i \right) \|\eta - \eta'\| \\
\left\| Q^t_{\eta} - Q^t_{\eta'} \right\| \leq C \|\eta - \eta'\| .
\]

(123)

The above shows that the three terms in the right hand side of (122) are proportional to \( (1 + \|g\|) \|\eta - \eta'\| \) and decay geometrically with \( t \). This implies there exists \( C'' \in [1, \infty), \delta \in (0, 1) \) such that

\[
\left\| P^t_{\eta} H_{\eta}(x) - h(\eta) - \left\{ P^t_{\eta'} H_{\eta'}(x) - h(\eta') \right\} \right\| \leq C'' \left( 1 + \|\eta - \eta'\| \right).
\]

(124)

Setting \( C_2 = \max\{ C'_2, C'' \} \) concludes the proof of the current lemma.

C.2. Proof of Proposition 5

**Proposition** Under A10, it holds for any \( (\eta, \eta') \in \mathcal{H}^2, (s, a) \in S \times A \),

\[
\| \nabla \log \Pi_{\eta}(a; s) \| \leq 2\delta, \quad \| \nabla \log \Pi_{\eta}(a; s) - \nabla \log \Pi_{\eta'}(a; s) \| \leq 8\delta^2 \|\eta - \eta'\|. 
\]

(125)

**Proof** To simplify notations, let us define \( \Delta a := a - b \) as the difference between two features. The proof is straightforward as we observe that

\[
\nabla \log \Pi_{\eta}(a; s) = \frac{1}{\sum_{a' \in A} \exp(\langle \eta \mid \Delta a(a', a) \rangle)} \sum_{b \in A} \exp(\langle \eta \mid \Delta a(b, a) \rangle) \Delta a(a, b).
\]

(126)

Observe that

\[
\| \nabla \log \Pi_{\eta}(a; s) \| \leq \max_{a, b \in A} \| a - b \| \leq 2\delta.
\]

(127)

Moreover, the Hessian of the log policy can be evaluated as:

\[
\nabla^2 \log \Pi_{\eta}(a; s) = \frac{1}{\sum_{a' \in A} \exp(\langle \eta \mid \Delta a(a', a) \rangle)} \sum_{b \in A} \exp(\langle \eta \mid \Delta a(b, a) \rangle) \Delta a(a, b) \Delta a(b, a)^\top - \\
\left( \sum_{b \in A} \frac{\exp(\langle \eta \mid \Delta a(a, b) \rangle)}{\sum_{a' \in A} \exp(\langle \eta \mid \Delta a(a', a) \rangle)} \Delta a(a, b) \right) \left( \frac{\exp(\langle \eta \mid \Delta a(b, a) \rangle)}{\sum_{a' \in A} \exp(\langle \eta \mid \Delta a(a', a) \rangle)} \Delta a(a, b) \right)^\top.
\]

(128)
It can be checked that
\[ \| \nabla^2 \log \Pi_\eta(a; s) \| \leq \max_{a, b \in A} \| \Delta x(a, b) \Delta x(b, a)^\top \| + \left( \max_{a, b \in A} \| \Delta x(a, b) \| \right)^2 \leq 8 b^2. \] (129)

This implies smoothness condition in (45).

C.3. Proof of Proposition 6

Proposition
Under A10, A11, the function
\[ \hat{H}_\eta(x) = \sum_{t=0}^\infty \left\{ P^t_\eta H_\eta(x) - h(\eta) \right\}, \] (130)
is well defined and satisfies the Poisson equation (7). For all \( x \in \mathcal{X}, (\eta, \eta') \in \mathcal{H}^2 \), there exists constants \( L^{(0)}_{PH}, L^{(1)}_{PH} \) such that
\[ \max\{\| P_\eta \hat{H}_\eta(x) \|, \| \hat{H}_\eta(x) \| \} \leq L^{(0)}_{PH}, \quad \| P_\eta \hat{H}_\eta(x) - P_\eta' \hat{H}_\eta'(x) \| \leq L^{(1)}_{PH} \| \eta - \eta' \|. \] (131)

Proof
From Lemma 6, there exists \( C_2 \in [1, \infty), \delta \in [0, 1) \) such that
\[ \| P^t_\eta H_\eta(x) - h(\eta) \| \leq C_2 t \delta (1 + \| g \|), \quad \forall \ t \geq 1, \ \forall \ x \in \mathcal{X}, \] (132)

It follows that the solution to the Poisson equation \( \hat{H}_\eta(x) \) in (47) is well defined. Moreover, it satisfies (7) and
\[ \max\{\| \hat{H}_\eta(x) \|, \| P_\eta \hat{H}_\eta(x) \| \} \leq L^{(0)}_{PH}, \] (133)
for some \( L^{(0)}_{PH} < \infty \) (note that \( g \) is bounded as specified by the state space \( \mathcal{X} \)). As such, the first equation in (48) of the proposition is proven. Finally, applying the definition of \( \hat{H}_\eta(x) \) shows that
\[ P_\eta \hat{H}_\eta(x) - P_\eta \hat{H}_\eta'(x) = \sum_{t=1}^\infty \left\{ (P^t_\eta H_\eta(x) - h(\eta)) - (P^t_\eta' H_\eta'(x) - h(\eta')) \right\}. \] (134)

Using Lemma 6, this implies
\[ \| P_\eta \hat{H}_\eta(x) - P_\eta \hat{H}_\eta'(x) \| \leq \sum_{t=1}^\infty \left\| (P^t_\eta H_\eta(x) - h(\eta)) - (P^t_\eta' H_\eta'(x) - h(\eta')) \right\| \]
\[ \leq \sum_{t=1}^\infty \left\{ C_2 (t \delta^t) (1 + \| g \|) \| \eta - \eta' \| \right\}. \] (135)

As such, there exists \( L^{(1)}_{PH} \in [1, \infty) \) such that
\[ \| P_\eta \hat{H}_\eta(x) - P_\eta \hat{H}_\eta'(x) \| \leq L^{(1)}_{PH} \| \eta - \eta' \|, \] (136)
for all \( x \in \mathcal{X} \). This proves the second equation in (48) of the proposition.
C.4. Proof of Proposition 7

Proposition Under A10, A11, the gradient $\nabla J(\eta)$ is $R_{\text{max}} |S||A|$-Lipschitz continuous. Moreover, for any $\eta \in \mathcal{H}$, it holds that

$$(1 - \lambda)^2 \Gamma^2 + 2 \langle \nabla J(\eta) | h(\eta) \rangle \geq \|h(\eta)\|^2, \quad \|\nabla J(\eta)\| \leq \|h(\eta)\| + (1 - \lambda) \Gamma,$$

where $\Gamma := 2b R_{\text{max}} K_{P} \frac{1}{(1 - \rho)^2}$.

Proof The first statement is a direct application of part 1) in Lemma 5 which holds under A10, A11. To prove the second statement, let us define the error vector as

$$\Delta(\eta) := h(\eta) - \nabla J(\eta)$$

Applying Lemma 5 shows that $\sup_{\eta \in \mathcal{H}} \|\Delta(\eta)\|^2 \leq \Gamma^2 (1 - \lambda)^2$. We observe that

$$\langle \nabla J(\eta) | h(\eta) \rangle = \langle h(\eta) - \Delta(\eta) | h(\eta) \rangle = \|h(\eta)\|^2 - \langle \Delta(\eta) | h(\eta) \rangle$$

$$\geq \|h(\eta)\|^2 - \frac{1}{2} \left( \|h(\eta)\|^2 + \|\Delta(\eta)\|^2 \right).$$

This implies

$$\frac{\Gamma^2}{2} (1 - \lambda)^2 + \langle \nabla J(\eta) | h(\eta) \rangle \geq \frac{1}{2} \|h(\eta)\|^2.$$
Moreover, for \( \eta, \eta' \in \mathcal{H} \) and \( x \in \mathcal{X} \),

\[
\| P_\eta \tilde{H}_\eta(x) - P_{\eta'} \tilde{H}_{\eta'}(x) \| \leq L_{PH}^{(1)} \| \eta - \eta' \| ,
\]

where

\[
L_{PH}^{(1)} = \frac{K_P^2 \sigma L_P}{(1 - \rho)^2} (2 + K_P) + \frac{K_P}{1 - \rho} L_H .
\]

**Proof** Note that, under A14,

\[
\sum_{i=0}^{\infty} \left\| P_i^\eta (H_\eta(x) - h(\eta)) - \pi_\eta (H_\eta(\cdot) - h(\eta)) \right\|
\leq \| H_\eta(\cdot) - h(\eta) \| \sup_{\eta \in \mathcal{H}, x \in \mathcal{X}} \left\| P_\eta (x, \cdot) - P_{\eta'} (x, \cdot) \right\|_{TV}
\]

(147)

Therefore, for all \( \eta \in \mathcal{H} \) and \( x \in \mathcal{X} \), the series

\[
\sum_{i=0}^{\infty} P_i^\eta (H_\eta(x) - h(\eta)) - \pi_\eta (H_\eta(\cdot) - h(\eta))
\]

(148)

is uniformly converging and is a solution of the Poisson equation (7). In addition, (143) and (144) follow directly from (147). Under A14, applying a simple modification of (Fort et al., 2011, Lemma 4.2, 1st statement) shows that for any \( \eta, \eta' \in \mathcal{H} \), we have

\[
\| \pi_\eta - \pi_{\eta'} \|_{TV} \leq \frac{K_P (1 + K_P)}{1 - \rho} \sup_{x \in \mathcal{X}} \left\| P_\eta(x, \cdot) - P_{\eta'}(x, \cdot) \right\|_{TV} .
\]

(149)

Again using a simple modification of (Fort et al., 2011, Lemma 4.2, 2nd statement) shows that for any \( X \in \mathcal{X}, \eta, \eta' \in \mathbb{R}^d \), it holds

\[
\left\| P_\eta \tilde{H}_\eta(x) - P_{\eta'} \tilde{H}_{\eta'}(x) \right\|
\leq \frac{K_P^2}{(1 - \rho)^2} \left( \sup_{\eta \in \mathcal{H}, x \in \mathcal{X}} \| H_\eta(x) - h(\eta) \| \right) \left( \sup_{x \in \mathcal{X}} \left\| P_\eta(x, \cdot) - P_{\eta'}(x, \cdot) \right\|_{TV} \right)
+ \frac{K_P}{1 - \rho} \left( \sup_{\eta \in \mathcal{H}, x \in \mathcal{X}} \| H_\eta(x) - h(\eta) \| \right) \| \pi_\eta - \pi_{\eta'} \|_{TV}
+ \frac{K_P}{1 - \rho} \sup_{x \in \mathcal{X}} \| H_\eta(x) - H_{\eta'}(x) \|
\leq \left( \frac{K_P^2 \sigma L_P}{(1 - \rho)^2} (2 + K_P) + \frac{K_P}{1 - \rho} L_H \right) \| \eta - \eta' \| = L_{PH}^{(1)} \| \eta - \eta' \| ,
\]

where the last inequality is due to A12, A13, A7 and (149).

---

2. We note that under A14, the constants \( \rho_\theta, \rho_{\theta'} \) are the same in (Fort et al., 2011, Lemma 4.2) which simplifies the derivation and yields a tighter bound.

3. Note that we take the measurable function as \( V = 1 \) therein.