Proceedings of Machine Learning Research vol XX:1–30, 2019

Non-asymptotic Analysis of Biased Stochastic Approximation Schemes

Belhal Karimi, Błażej Miasojedow, Éric Moulines, Hoi-To Wai[*][†]

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, reinforcement learning, 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:

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n - \gamma_{n+1} H_{\boldsymbol{\eta}_n}(X_{n+1}), \quad n \in \mathbb{N} \tag{1}$$

where $\boldsymbol{\eta}_n \in \mathcal{H} \subset \mathbb{R}^d$ denotes the nth iterate, $\gamma_n > 0$ is the step size and $H_{\boldsymbol{\eta}_n}(X_{n+1})$ is the nth *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 $H_{\boldsymbol{\eta}_n}(X_{n+1})$ is a conditionally *unbiased* estimate of the so-called mean-field $h(\boldsymbol{\eta}_n)$, *i.e.,* $\mathbb{E}\left[H_{\boldsymbol{\eta}_n}(X_{n+1}) \,|\, \mathcal{F}_n\right] = h(\boldsymbol{\eta}_n)$ where $\mathcal{F}_n$ denotes the filtration generated by the random variables $(\boldsymbol{\eta}_0, \{X_m\}_{m \leq n})$. In such case, $\boldsymbol{e}_{n+1} = H_{\boldsymbol{\eta}_n}(X_{n+1}) - h(\boldsymbol{\eta}_n)$ 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 $f : \mathsf{X} \to \mathbb{R}$,

$$\mathbb{E}\left[f(X_{n+1}) \,|\, \mathcal{F}_n\right] = P_{\boldsymbol{\eta}_n} f(X_n) = \int f(x) P_{\boldsymbol{\eta}_n}(X_n, \mathrm{d}x)\ , \tag{2}$$

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

$$h(\boldsymbol{\eta}) = \int H_{\boldsymbol{\eta}}(x) \pi_{\boldsymbol{\eta}}(\mathrm{d}x)\ , \tag{3}$$

[*] Equal contributions with authors listed in alphabetical order.

[†] B. Karimi and E. Moulines are with CMAP, École Polytechnique, Palaiseau, France. B. Miasojedow is with Informatics and Mechanics, Faculty of Mathematics, University of Warsaw, Poland. H.-T. Wai is with Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong. E-mails: belhal.karimi@polytechnique.edu, bmiasojedow@gmail.com, eric.moulines@polytechnique.edu, htwai@se.cuhk.edu.hk

where we have assumed that $\int \|H_{\boldsymbol{\eta}}(x)\| \pi_{\boldsymbol{\eta}}(\mathrm{d}x) < \infty$. Throughout this paper, we assume that the mean field h is 'related' (to be defined precisely later) to a smooth Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}$, and the aim of the SA scheme (1) is to find a minimizer or stationary point of 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 $\{\boldsymbol{\eta}_n,\ n \in \mathbb{N}\}$ 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 $\dot{\boldsymbol{\eta}} = h(\boldsymbol{\eta})$. 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 $\{\boldsymbol{\eta}_n,\ n \in \mathbb{N}\}$ 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.,* $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$, 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 $\boldsymbol{\eta}$.

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, which include Q-learning (Jaakkola et al., 1994) and policy gradient method (Baxter and Bartlett, 2001).

Surprisingly enough, we are not aware of non-asymptotic convergence results of SA comparable to (Ghadimi and Lan, 2013) and (Bottou et al., 2018, Section 4,5) when (a) the drift term $H_{\boldsymbol{\eta}}(x)$ 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 $\boldsymbol{\eta} \in \mathcal{H}$, $c_0 + c_1 \left\langle \nabla V(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle \geq \|h(\boldsymbol{\eta})\|^2$ there exists $c_1 > 0, c_0 \geq 0$; see Section 2 and A1. If $c_0 = 0$, then $\left\langle \nabla V(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle > 0$ as soon as $h(\boldsymbol{\eta}) \neq \mathbf{0}$ in which case V is a Lyapunov function for the ODE $\dot{\boldsymbol{\eta}} = h(\boldsymbol{\eta})$. 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 $n \in \mathbb{N}$, $\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] = \mathcal{O}(c_0 + \log(n)/\sqrt{n})$, 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 $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$; and of (Tadić and Doucet, 2017) which shows asymptotic convergence of (1) under the uniform boundedness assumption on iterates.

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

2. Stochastic Approximation Schemes and Their Convergence

We consider the following assumptions:

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

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

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

A1,A2 assume that the mean field $h(\boldsymbol{\eta})$ [cf. (2)] is indirectly related to the Lyapunov function $V(\boldsymbol{\eta})$ where it needs not be the same as $\nabla V(\boldsymbol{\eta})$. 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 $\|h(\boldsymbol{\eta}_n)\|^2$, where $\boldsymbol{\eta}_n$ is produced by (1). If $c_0 = d_0 = 0$ in A1,A2, then $h(\boldsymbol{\eta}_*) = 0$ implies that $\|\nabla V(\boldsymbol{\eta}_*)\| = 0$, *i.e.,* the point $\boldsymbol{\eta}_*$ is a stationary point of the deterministic recursion $\bar{\boldsymbol{\eta}}_n = \bar{\boldsymbol{\eta}}_n - \gamma_{n+1} h(\bar{\boldsymbol{\eta}}_n)$. As a convention, for any $\epsilon \geq 0$, we say that $\boldsymbol{\eta}_*$ is an ϵ*-quasi-stationary point* if $\|h(\boldsymbol{\eta}_*)\|^2 \leq \epsilon$.

We consider two settings for the SA scheme. Define the noise vector, $\boldsymbol{e}_{n+1}$, as the difference between the stochastic update $H_{\boldsymbol{\eta}_n}(X_{n+1})$ and the mean field $h(\boldsymbol{\eta}_n)$ defined in (3):

$$\boldsymbol{e}_{n+1} := H_{\boldsymbol{\eta}_n}(X_{n+1}) - h(\boldsymbol{\eta}_n) \ . \tag{4}$$

The settings and the corresponding 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 $n \in \mathbb{N}$, $\mathbb{E}\left[\boldsymbol{e}_{n+1} \,|\, \mathcal{F}_n\right] = \boldsymbol{0}$, $\mathbb{E}\left[\|\boldsymbol{e}_{n+1}\|^2 \,\middle|\, \mathcal{F}_n\right] \leq \sigma_0^2 + \sigma_1^2 \|h(\boldsymbol{\eta}_n)\|^2$ with $\sigma_0^2, \sigma_1^2 \in [0, \infty)$.*

As a concrete example, A4 can be satisfied when $H_{\boldsymbol{\eta}_n}(X_{n+1}) = h(\boldsymbol{\eta}_n) + X_{n+1}$ 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 $\gamma_{n+1} \leq (2c_1 L(1+\sigma_1^2))^{-1}$ for all $n \geq 0$. 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 the distribution*

$$\mathbb{P}(N = \ell) = \Big\{ \sum_{k=0}^{n} \gamma_{k+1} \big(1 - c_1 L (1+\sigma_1^2) \gamma_{k+1}\big) \Big\}^{-1} \gamma_{\ell+1} \big(1 - c_1 L(1+\sigma_1^2)\gamma_{\ell+1}\big) \,. \tag{5}$$

We have

$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] \leq 2c_1 \big(\textstyle\sum_{k=0}^{n} \gamma_{k+1} \big)^{-1} \big(V_{0,n} + \sigma_0^2 L \textstyle\sum_{k=0}^{n} \gamma_{k+1}^2 \big) + 2c_0 \,, \tag{6}$$

where we have defined $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$.

If we set $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$ for all $k \geq 1$, then the right hand side in (6) evaluates to $\mathcal{O}(c_0 + \log n/\sqrt{n})$ 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. Note for the special case with $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$ [where A1 is satisfied with $c_0 = 0, c_1 = 1$], our result recovers (Ghadimi and Lan, 2013, Theorem 2.1).

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, *i.e.,* for any bounded measurable function φ and $n \in \mathbb{N}$, $\mathbb{E}\left[\varphi(X_{n+1}) \,|\, \mathcal{F}_n\right] = P_{\boldsymbol{\eta}_n}\varphi(X_n)$, where for any $\boldsymbol{\eta} \in \mathcal{H}$, $P_{\boldsymbol{\eta}}$ is a Markov kernel on $\mathsf{X} \times \mathcal{X}$. We assume that for each $\boldsymbol{\eta} \in \mathcal{H}$, $P_{\boldsymbol{\eta}}$ has a unique stationary distribution $\pi_{\boldsymbol{\eta}}$, *i.e.,* $\pi_{\boldsymbol{\eta}} P_{\boldsymbol{\eta}} = \pi_{\boldsymbol{\eta}}$. In addition, for each $\boldsymbol{\eta} \in \mathcal{H}$, $\int \|H_{\boldsymbol{\eta}}(x)\| \pi_{\boldsymbol{\eta}}(\mathrm{d}x) < \infty$ and $h(\boldsymbol{\eta}) = \int H_{\boldsymbol{\eta}}(x) \pi_{\boldsymbol{\eta}}(\mathrm{d}x)$. Consider the following assumptions:

A5 *There exists a Borel measurable function $\hat{H} : \mathcal{H} \times \mathsf{X} \to \mathcal{H}$ where for each $\boldsymbol{\eta} \in \mathcal{H}$, $x \in \mathsf{X}$,*

$$\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) = H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}) \,. \tag{7}$$

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

$$\sup_{x \in \mathsf{X}} \|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x)\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \,. \tag{8}$$

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

Basically, assumption A5 requires that for each $\boldsymbol{\eta} \in \mathcal{H}$, the Poisson equation associated with the Markov kernel $P_{\boldsymbol{\eta}}$ and the function $H_{\boldsymbol{\eta}}(\cdot)$ has a solution. Assumption A6 implies that for each $x \in \mathsf{X}$, the function $\boldsymbol{\eta} \mapsto H_{\boldsymbol{\eta}}(x)$ is Lipshitz and that the Lipshitz constant is uniformly bounded in $x \in \mathsf{X}$. 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*

$$\gamma_{n+1} \leq \gamma_n,\ \gamma_n \leq a\gamma_{n+1},\ \gamma_n - \gamma_{n+1} \leq a'\gamma_n^2,\ \gamma_1 \leq 0.5\big(c_1(L+C_h)\big)^{-1}\ , \tag{9}$$

for some $a, a' > 0$ and all $n \geq 0$. For any $n \geq 1$, let $N \in \{0, \ldots, n\}$ be a discrete r.v. (independent of $\{\mathcal{F}_n,\ n \in \mathbb{N}\}$) such that

$$\mathbb{P}(N = \ell) = \Big\{ \sum_{k=0}^{n} \gamma_{k+1}\big(1 - c_1\gamma_{k+1}(L + C_h)\big)\Big\}^{-1} \gamma_{\ell+1}\big(1 - c_1\gamma_{\ell+1}(L+C_h)\big)\ , \tag{10}$$

then

$$\mathbb{E}[h(\boldsymbol{\eta}_N)\|^2] \leq \frac{V_{0,n} + C_{0,n} + \big(\sigma^2 L + C_\gamma\big)\sum_{k=0}^{n}\gamma_{k+1}^2}{\sum_{k=0}^{n}\gamma_{k+1}/2} + 2c_0\ , \tag{11}$$

where we have $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$, and the constants are defined as:

$$C_h := \big(L_{PH}^{(1)}(d_0 + \frac{d_1}{2}(a+1) + ad_1\sigma) + L_{PH}^{(0)}\big(L + d_1\{1 + a'\}\big)\big)\ , \tag{12}$$

$$C_\gamma := L_{PH}^{(1)}(d_0 + d_0\sigma + d_1\sigma) + LL_{PH}^{(0)}(1+\sigma)\ , \tag{13}$$

$$C_{0,n} := L_{PH}^{(0)}\big((1+d_0)(\gamma_1 - \gamma_{n+1}) + d_0(\gamma_1 + \gamma_{n+1}) + 2d_1\big)\ . \tag{14}$$

Similar to the case with Martingale difference noise, if we set $\gamma_k = (2c_1L(1+C_h)\sqrt{k})^{-1}$ for all $k \geq 1$, then the step size satisfies (9) with $a = \sqrt{2}$ and $a' = \frac{\sqrt{2}-1}{\sqrt{2}}(2c_1L(1+C_h))$, and the right hand side in (11) evaluates to $\mathcal{O}(c_0 + \log n/\sqrt{n})$ 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 $\boldsymbol{\eta} \in \mathcal{H}$ and $x \in \mathsf{X}$, $P_{\boldsymbol{\eta}}(x, \cdot) = \pi_{\boldsymbol{\eta}}(\cdot)$, we have $L_{PH}^{(0)} = L_{PH}^{(1)} = 0$. The constants evaluates 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.

The novelty of our result lies on a new decomposition method of the error terms used in our analysis [cf. proof of Lemma 2], which allows us to control the growth of $\mathbb{E}[\|h(\boldsymbol{\eta}_n)\|^2]$ with $\boldsymbol{\eta}_n$ produced by the SA scheme, without explicitly assuming that $\{\boldsymbol{\eta}_n\}_{n\geq 0}$ is bounded.

2.1. Convergence Analysis

The detailed proofs in this section are in Appendix A. To simplify notations, we shall denote $h_n := \|h(\boldsymbol{\eta}_n)\|^2$ 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:*

$$\begin{aligned}&\textstyle\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}\big(1 - c_1L\gamma_{k+1}\big)h_k \\ &\leq V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1}) + L\textstyle\sum_{k=0}^{n}\gamma_{k+1}^2\|\boldsymbol{e}_{k+1}\|^2 + \sum_{k=0}^{n}\gamma_{k+1}\big(c_1^{-1}c_0 - \langle\nabla V(\boldsymbol{\eta}_k)\,|\,\boldsymbol{e}_{k+1}\rangle\big).\end{aligned} \tag{15}$$

Proof of Theorem 1 Having established Lemma 1, the convergence of SA with Martingale difference noise can be obtained. Particularly, the expected value of $\langle\nabla V(\boldsymbol{\eta}_k)\,|\,\boldsymbol{e}_{k+1}\rangle$ is zero when conditioned on $\mathcal{F}_k$. Therefore, taking total expectation on both sides of (15) yields:

$$\begin{aligned}\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}\big(1 - c_1 L\gamma_{k+1}\big)\mathbb{E}[h_k] &\leq V_{0,n} + L\sum_{k=0}^{n}\big(\gamma_{k+1}^2\mathbb{E}[\|\boldsymbol{e}_{k+1}\|^2] + \gamma_{k+1}\frac{c_0}{c_1}\big) \\ &\leq V_{0,n} + L\sigma_0^2\sum_{k=0}^{n}\gamma_{k+1}^2 + L\sigma_1^2\sum_{k=0}^{n}\gamma_{k+1}\mathbb{E}[h_k]) + \gamma_{k+1}\frac{c_0}{c_1}\big)\ ,\end{aligned} \tag{16}$$

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

$$\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}\big(1 - c_1 L(1+\sigma_1^2)\gamma_{k+1}\big)\mathbb{E}[h_k] \leq V_{0,n} + \sigma_0^2 L\sum_{k=0}^{n}\gamma_{k+1}^2 + \frac{c_0}{c_1}\sum_{k=0}^{n}\gamma_{k+1}\ . \tag{17}$$

Consequently, using (5) and noting that $\gamma_{n+1} \leq (2c_1 L(1+\sigma_1^2))^{-1}$, we obtain

$$\mathbb{E}[h_N] = \sum_{n'=0}^{n} \frac{\frac{\gamma_{n'+1}}{c_1}\big(1 - c_1 L(1+\sigma_1^2)\gamma_{n'+1}\big)\mathbb{E}[h_{n'}]}{\sum_{k=0}^{n}\frac{\gamma_{k+1}}{c_1}\big(1 - c_1 L(1+\sigma_1^2)\gamma_{k+1}\big)} \leq \frac{V_{0,n} + \sigma_0^2 L\sum_{k=0}^{n}\gamma_{k+1}^2}{\frac{1}{c_1}\sum_{k=0}^{n}\gamma_{k+1}/2} + 2c_0\ . \tag{18}$$

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

$$\textstyle\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[\|\boldsymbol{e}_{k+1}\|^2] \leq \sum_{k=0}^{n}\gamma_{k+1}^2\sigma^2\ . \tag{19}$$

Unlike in Theorem 1, the expected value of the inner product $\langle\nabla V(\boldsymbol{\eta}_k)\,|\,\boldsymbol{e}_{k+1}\rangle$ 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:*

$$\mathbb{E}\left[-\sum_{k=0}^{n}\gamma_{k+1}\langle\nabla V(\boldsymbol{\eta}_k)\,|\,\boldsymbol{e}_{k+1}\rangle\right] \leq C_h\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[\|h(\boldsymbol{\eta}_k)\|^2] + C_\gamma\sum_{k=0}^{n}\gamma_{k+1}^2 + C_{0,n}\ , \tag{20}$$

where C_h, C_γ and $C_{0,n}$ are defined in (12), (13), (14).

Finally, to prove the theorem, we combine Lemma 1, (19) and Lemma 2 to obtain:

$$\begin{aligned}&\sum_{k=0}^{n}\frac{\gamma_{k+1}}{c_1}\big(1 - c_1 L\gamma_{k+1}\big)\mathbb{E}[h_k] \\ &\leq V_{0,n} + C_{0,n} + \Big(\sigma^2 L + C_\gamma\Big)\sum_{k=0}^{n}\gamma_{k+1}^2 + C_h\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[h_k] + \frac{c_0}{c_1}\sum_{k=0}^{n}\gamma_{k+1}\ .\end{aligned} \tag{21}$$

Repeating a similar argument as in (18) using the distribution (10) shows the desired bound (11).

3. Applications

In this section, we present several applications pertaining to machine learning where the results in Section 2 apply 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 $f(x;\boldsymbol{\theta})$ where $\{f(x;\boldsymbol{\theta});\boldsymbol{\theta}\in\Theta\}$ is a parametric family of probability density functions and Θ is an open convex subset of $\mathbb{R}^d$. The observation $Y\in\mathsf{Y}$ is a deterministic function of X. We denote by $g(y;\boldsymbol{\theta})$ the (observed) likelihood function. The notations $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\theta}}\left[\cdot\,|\,Y\right]$ are used to denote the expectation and conditional expectation under the statistical model $\{f(x;\boldsymbol{\theta});\boldsymbol{\theta}\in\Theta\}$. We denote by π 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 $\{g(y;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$. 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 $S:\mathsf{X}\to\mathsf{S}$ be a measurable function. We assume that the complete data-likelihood function belongs to the curved exponential family

$$f(x;\boldsymbol{\theta}) = h(x)\exp\left(\langle S(x)\,|\,\phi(\boldsymbol{\theta})\rangle - \psi(\boldsymbol{\theta})\right)\ , \tag{22}$$

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

$$\overline{\boldsymbol{s}}(y;\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[S(X)\,|\,Y=y\right] \tag{23}$$

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

$$\ell(\boldsymbol{s};\boldsymbol{\theta}) := \psi(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta}) - \langle\boldsymbol{s}\,|\,\phi(\boldsymbol{\theta})\rangle\ , \tag{24}$$

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

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

$$\nabla\psi(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \nabla\mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) - \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top\boldsymbol{s} = \mathbf{0}\ . \tag{25}$$

In addition, for any $\boldsymbol{s}\in\mathsf{S}$, $\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$ is invertible and the map $\boldsymbol{s}\mapsto\overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is differentiable on S.

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

$$\hat{\boldsymbol{s}}_{n+1} = \hat{\boldsymbol{s}}_n + \gamma_{n+1}\big(\overline{\boldsymbol{s}}(Y_{n+1};\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{s}}_n\big),\ \ \hat{\boldsymbol{\theta}}_{n+1} = \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_{n+1})\ . \tag{26}$$

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 π and $g_{\boldsymbol{\theta}}$. Precisely, we set

$$V(\boldsymbol{s}) := \mathrm{KL}\left(\pi, g(\cdot;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\right) + \mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) \quad \mathrm{KL}\left(\pi, g(\cdot;\boldsymbol{\theta})\right) := \mathbb{E}_\pi\left[\log(\pi(Y))/g(Y;\theta)\right] \ . \tag{27}$$

We establish a few key results that relate the ro-EM method to an SA scheme seeking for a stationary point of $V(\boldsymbol{s})$. Denote by $\mathcal{F}_n$ the filtration generated by the random variables $\{\hat{s}_0, Y_k\}_{k\leq n}$. From (26) we can identify the drift term and its mean field respectively as

$$\begin{aligned} &H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) = \hat{\boldsymbol{s}}_n - \overline{\boldsymbol{s}}(Y_{n+1};\overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_n)) \ , \\ &h(\hat{\boldsymbol{s}}_n) = \mathbb{E}_\pi\left[H_{\hat{\boldsymbol{s}}_n}(Y_{n+1})|\mathcal{F}_n\right] = \hat{\boldsymbol{s}}_n - \mathbb{E}_\pi\left[\overline{\boldsymbol{s}}(Y_{n+1};\overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_n))\right] \ . \end{aligned} \tag{28}$$

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

Proposition 1 *Assume A8. Then*

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

Proposition 2 *Assume A8. Then, for $\boldsymbol{s} \in \mathsf{S}$,*

$$\nabla_{\boldsymbol{s}} V(\boldsymbol{s}) = \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})\Big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top h(\boldsymbol{s}) \ . \tag{29}$$

Proposition 1 relates the root(s) of the mean field $h(\boldsymbol{s})$ to the stationary condition of the regularized KL divergence. Together with an additional condition on the smallest eigenvalue of the Jacobian-Hessian-Jacobian product

$$\lambda_{\mathsf{min}}\big(\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s};\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top\big) \geq \upsilon > 0, \ \forall \ \boldsymbol{s} \in \mathsf{S} \ , \tag{30}$$

Proposition 2 shows that the mean field of the stochastic update in (28) satisfies A1 with $c_0 = 0$ and $c_1 = 1/\upsilon$. If we assume that the Lyapunov function in (27), and the stochastic update in (28) satisfy the assumptions in Case 1 [*i.e.,* A4], then these results show that within n iterations, the ro-EM method finds an $\mathcal{O}(\log n/\sqrt{n})$ stationary solution of the Lyapunov function. 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:

$$g(y;\boldsymbol{\theta}) \propto \Big(1 - \sum_{m=1}^{M-1}\omega_m\Big)\exp\Big(-\frac{(y-\mu_M)^2}{2}\Big) + \sum_{m=1}^{M-1}\omega_m\exp\Big(-\frac{(y-\mu_m)^2}{2}\Big) \ . \tag{31}$$

The parameters are denoted by $\boldsymbol{\theta} := (\omega_1, \ldots, \omega_{M-1}, \mu_1, \ldots, \mu_{M-1}, \mu_M) \in \mathcal{C}$ where the parameter set is defined as $\mathcal{C} = \Delta_{M-1} \times \mathbb{R}^M$ with $\Delta_{M-1} := \{(\omega_1, \cdots, \omega_{M-1}) \in \mathbb{R}^{M-1}, \omega_m \geq$

$0, \sum_{m=1}^{M-1} \omega_m \leq 1\}$. To apply the ro-EM method, we augment the nth data Y_n with the latent variable $Z_n \in \{1, \ldots, M\}$. Log likelihood of the complete data tuple is

$$\mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{1}_{\{z=M\}} \left[\log(1 - \sum_{m=1}^{M-1} \omega_m) - \frac{(y-\mu_M)^2}{2} \right] + \sum_{m=1}^{M-1} \mathbb{1}_{\{z=m\}} \left[\log(\omega_m) - \frac{(y-\mu_m)^2}{2} \right] . \tag{32}$$

The above can be written in the standard curved exponential family form (22). In particular, we partition the sufficient statistics as $S(\boldsymbol{x}) = (S^{(1)}(\boldsymbol{x})^\top, S^{(2)}(\boldsymbol{x})^\top, S^{(3)}(\boldsymbol{x}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, and partition $\phi(\boldsymbol{\theta}) = (\phi^{(1)}(\boldsymbol{\theta})^\top, \phi^{(2)}(\boldsymbol{\theta})^\top, \phi^{(3)}(\boldsymbol{\theta}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$. Using the fact that $\mathbb{1}_{\{z=M\}} = 1 - \sum_{m=1}^{M-1} \mathbb{1}_{\{z=m\}}$, (32) can be expressed in the standard form as (22) with

$$\begin{aligned} & s_m^{(1)} = \mathbb{1}_{\{z=m\}}, \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\} - \left\{ \log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2} \right\} , \\ & s_m^{(2)} = \mathbb{1}_{\{z=m\}} y, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m, \quad m = 1, \ldots, M-1, \quad s^{(3)} = y, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M , \end{aligned} \tag{33}$$

and $\psi(\boldsymbol{\theta}) = -\left\{ \log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2\sigma^2} \right\}$.

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

$$\widetilde{\omega}_m(Y_{n+1}; \hat{\boldsymbol{\theta}}_n) := \mathbb{E}_{\hat{\boldsymbol{\theta}}_n}[\mathbb{1}_{\{z=m\}} | Y = Y_{n+1}] = \frac{\hat{\omega}_{m,n} \exp(-\frac{1}{2}(Y_{n+1} - \hat{\mu}_{m,n})^2)}{\sum_{j=1}^{M} \hat{\omega}_{j,n} \exp(-\frac{1}{2}(Y_{n+1} - \hat{\mu}_{j,n})^2)} . \tag{34}$$

With the above notations, the E-step's update in (23) can be described with

$$\overline{s}(Y_{n+1}; \hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} \big(\widetilde{\omega}_1(Y_{n+1}; \hat{\boldsymbol{\theta}}_n), \ldots, \widetilde{\omega}_{M-1}(Y_{n+1}; \hat{\boldsymbol{\theta}}_n)\big)^\top \\ \big(Y_{n+1}\widetilde{\omega}_1(Y_{n+1}; \hat{\boldsymbol{\theta}}_n), \ldots, Y_{n+1}\widetilde{\omega}_{M-1}(Y_{n+1}; \hat{\boldsymbol{\theta}}_n)\big)^\top \\ Y_{n+1} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{s}}_n^{(1)} \\ \overline{\boldsymbol{s}}_n^{(2)} \\ \overline{s}_n^{(3)} \end{pmatrix} . \tag{35}$$

For the M-step, let $\epsilon > 0$ be a user designed parameter, we consider the following regularizer:

$$\mathrm{R}(\boldsymbol{\theta}) = \epsilon \textstyle\sum_{m=1}^{M} \left\{ \mu_m^2/2 - \log(\omega_m) \right\} - \epsilon \log\left(1 - \sum_{m=1}^{M-1} \omega_m\right) , \tag{36}$$

For any $\boldsymbol{s}$ with $\boldsymbol{s}^{(1)} \geq \boldsymbol{0}$, it can be shown that the regularized M-step in (26) evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1}\big(s_1^{(1)} + \epsilon, \ldots, s_{M-1}^{(1)} + \epsilon\big)^\top \\ \big((s_1^{(1)} + \epsilon)^{-1} s_1^{(2)}, \ldots, (s_{M-1}^{(1)} + \epsilon)^{-1} s_{M-1}^{(2)}\big)^\top \\ \big(1 - \sum_{m=1}^{M-1} s_m^{(1)} + \epsilon\big)^{-1}\big(s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}\big) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{37}$$

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 (26), (35), (37) yield a special case of an SA scheme on $\hat{\boldsymbol{s}}_n$ 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 $|Y_n| \leq \overline{Y}$ for any $n \geq 0$.*

The ro-EM method is initialized by setting $\hat{\boldsymbol{s}}_1 = (\mathbf{0}, \mathbf{0}, 0)^\top$ and begun with the M-step. Note that under A9, the sufficient statistics $\hat{\boldsymbol{s}}_n$ lie in the compact set $\mathsf{S} = \Delta_{M-1} \times [-\overline{Y}, \overline{Y}]^M$ 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* $\mathbb{E}[\|\overline{s}(Y_{n+1}; \hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{s}}_n\|^2 | \mathcal{F}_n] \leq 2M\overline{Y}^2$ *for all* $n \geq 0$.

Proposition 4 *Under A9 and the regularizer (36) set with* $\epsilon > 0$*, then for all* $(\boldsymbol{s}, \boldsymbol{s}') \in \mathsf{S}^2$*, there exists positive constants* υ, Υ, Ψ *such that:*

$$\langle \nabla V(\boldsymbol{s}) \,|\, h(\boldsymbol{s}) \rangle \geq \upsilon \|h(\boldsymbol{s})\|^2, \ \ \|\nabla V(\boldsymbol{s}) - \nabla V(\boldsymbol{s}')\| \leq \Psi \|\boldsymbol{s} - \boldsymbol{s}'\| \, . \tag{38}$$

The first part of Proposition 4 is a consequence of Proposition 2 and verifying the full rankness of the Jacobian-Hessian-Jacobian product in (29). 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 $c_0 = 0$, $c_1 = \upsilon^{-1}$], and A3 [with $L = \Psi$], A4 [with $\sigma_0^2 = 2M\overline{Y}^2$, $\sigma_1^2 = 0$] are satisfied. As such, applying Theorem 1 shows that

Corollary 1 *Under A9 and set* $\gamma_k = (2c_1 L(1 + \sigma_1^2)\sqrt{k})^{-1}$*, the ro-EM method for GMM (26), (35), (37) finds a stationary point of the regularized KL divergence (27) at the rate of* $\mathcal{O}(\log n / \sqrt{n})$*, where* n *is the number of iterations.*

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 $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|$, where $\boldsymbol{\theta}^\star$ is the optimal parameter. To our best knowledge, the rate results in Corollary 1 is unknown prior to this work.

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) $(\mathsf{S}, \mathsf{A}, \mathrm{R}, \mathrm{P})$, where S is a finite set of spaces (state-space), A is a finite set of action (action-space), $\mathrm{R} : \mathsf{S} \times \mathsf{A} \to [0, \mathrm{R}_{\max}]$ is a reward function and P is the transition model, *i.e.,* given an action $a \in \mathsf{A}$, $\mathrm{P}^a = \{\mathrm{P}^a_{s,s'}\}$ is a matrix, $\mathrm{P}^a_{s,s'}$ is the probability of transiting from the sth state to the s'th state upon taking action a. The agent's decision is characterized by a parametric family of policies $\{\boldsymbol{\Pi}_{\boldsymbol{\eta}}\}_{\boldsymbol{\eta} \in \mathcal{H}}$: $\Pi_{\boldsymbol{\eta}}(a; s)$ 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:

$$Q_{\boldsymbol{\eta}}((s, a); (s', a')) := \Pi_{\boldsymbol{\eta}}(a'; s') \, \mathrm{P}^a_{s,s'} \ , \tag{39}$$

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

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

$$J(\boldsymbol{\eta}) := \textstyle\sum_{s,a} \upsilon(s,a) \, \mathrm{R}(s,a) \; . \tag{40}$$

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

$$\nabla J(\boldsymbol{\eta}) = \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\eta}} \big[\, \mathrm{R}(S_T, A_T) \textstyle\sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\eta}}(A_{T-i}; S_{T-i}) \big] \; . \tag{41}$$

To approximate (41) 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

$$\widehat{\nabla}_T J(\boldsymbol{\eta}) := \mathrm{R}(S_T, A_T) \textstyle\sum_{i=0}^{T-1} \lambda^i \nabla \log \Pi_{\boldsymbol{\eta}}(A_{T-i}; S_{T-i}) \approx \nabla J(\boldsymbol{\eta}) \; , \tag{42}$$

where $(S_1, A_1, \ldots, S_T, A_T)$ is a realization of state-action sequence generated by the policy $\Pi_{\boldsymbol{\eta}}$. 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 direct policy gradient method (Baxter and Bartlett, 2001):

$$G_{n+1} = \lambda G_n + \nabla \log \Pi_{\boldsymbol{\eta}_n}(A_{n+1}; S_{n+1}) \; , \tag{43a}$$

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n + \gamma_{n+1} G_{n+1} \, \mathrm{R}(S_{n+1}, A_{n+1}) \; . \tag{43b}$$

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

$$\Pi_{\boldsymbol{\eta}}(a;s) = \big\{ \textstyle\sum_{a' \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \, | \, \boldsymbol{x}(s,a') - \boldsymbol{x}(s,a) \rangle \big) \big\}^{-1} \; , \tag{44}$$

where $\boldsymbol{x}(s,a) \in \mathbb{R}^d$ is a known feature vector. We make the following assumptions:

A10 *For all $s \in \mathsf{S}$, $a \in \mathsf{A}$, the feature vector $\boldsymbol{x}(s,a)$ and reward $\mathrm{R}(s,a)$ are bounded with $\|\boldsymbol{x}(s,a)\| \leq \bar{b}$, $|\,\mathrm{R}(s,a)| \leq \mathrm{R}_{\max}$.*

A11 *For all $\boldsymbol{\eta} \in \mathcal{H}$, the MC $\{(S_t, A_t)\}_{t \geq 1}$, as governed by the transition matrix $\boldsymbol{Q}_{\boldsymbol{\eta}}$ [cf. (39)], is uniformly geometrically ergodic: there exists $\rho \in [0,1)$, $K_R < \infty$ such that, for all $n \geq 0$,*

$$\|\boldsymbol{Q}_{\boldsymbol{\eta}}^n - \mathbf{1}\boldsymbol{\upsilon}_{\boldsymbol{\eta}}^\top\| \leq \rho^n K_R \; , \tag{45}$$

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

$$\|\boldsymbol{\upsilon}_{\boldsymbol{\eta}} - \boldsymbol{\upsilon}_{\boldsymbol{\eta}'}\| \leq L_Q \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \; , \quad \| \, \mathrm{J}_{\boldsymbol{\upsilon}_{\boldsymbol{\eta}}}^{\boldsymbol{\eta}}(\boldsymbol{\eta}) - \mathrm{J}_{\boldsymbol{\upsilon}_{\boldsymbol{\eta}}}^{\boldsymbol{\eta}}(\boldsymbol{\eta}') \| \leq L_\upsilon \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \; , \tag{46}$$

where $\mathrm{J}_{\boldsymbol{\upsilon}_{\boldsymbol{\eta}}}^{\boldsymbol{\eta}}(\boldsymbol{\eta})$ denotes the Jacobian of $\boldsymbol{\upsilon}_{\boldsymbol{\eta}}$ w.r.t. *$\boldsymbol{\eta}$.*

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 (45), the Lipschitz continuity conditions (46) can be implied using (Fort et al., 2011; Tadić and Doucet, 2017).

Our task is to verify that the policy gradient method (43) 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 $X_n = (S_n, A_n, G_n) \in \mathsf{X} := \mathsf{S} \times \mathsf{A} \times \mathbb{R}^d$, 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

$$H_{\boldsymbol{\eta}_n}(X_{n+1}) = G_{n+1}\,\mathrm{R}(S_{n+1}, A_{n+1}) \quad \text{with} \quad h(\boldsymbol{\eta}) = \lim_{T \to \infty} \mathbb{E}_{\tau_T \sim \Pi_{\boldsymbol{\eta}},\, S_1 \sim \overline{\Pi}_{\boldsymbol{\eta}}} \big[\widehat{\nabla}_T J(\boldsymbol{\eta}) \big] \;, \tag{47}$$

where $\widehat{\nabla}_T J(\boldsymbol{\eta})$ is defined in (42). Moreover, we let $P_{\boldsymbol{\eta}} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$ to be the Markov kernel associated with the MC $\{X_n\}_{n \geq 1}$. Observe that

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

$$\|\nabla \log \Pi_{\boldsymbol{\eta}}(a; s)\| \leq 2\overline{b}, \;\; \|\nabla \log \Pi_{\boldsymbol{\eta}}(a; s) - \nabla \log \Pi_{\boldsymbol{\eta}'}(a; s)\| \leq 8\overline{b}^2 \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \;. \tag{48}$$

Using the recursive update of (43a), we show that

$$\|G_n\| = \|\lambda G_{n-1} + \nabla \log \Pi_{\boldsymbol{\eta}}(A_n; S_n)\| \leq \lambda \|G_{n-1}\| + 2\overline{b} = \mathcal{O}(2\overline{b}\|G_0\|/(1-\lambda)) \;, \tag{49}$$

for any $n \geq 1$, which then implies that the stochastic update $H_{\boldsymbol{\eta}_n}(X_{n+1})$ in (43) is bounded since the reward is bounded using A10. The above proposition also implies that $h(\boldsymbol{\eta})$ is bounded for all $\boldsymbol{\eta} \in \mathcal{H}$. Therefore, the assumption A7 is satisfied.

Next, with a slight abuse of notation, we shall consider the compact state space $\mathsf{X} = \mathsf{S} \times \mathsf{A} \times \mathsf{G}$, with $\mathsf{G} = \{g \in \mathbb{R}^d : \|g\| \leq C_0 \overline{b}/(1-\lambda)\}$ and $C_0 \in [1, \infty)$, and analyze the policy gradient algorithm accordingly where $\{X_{n+1}\}_{n \geq 0}$ is in 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:*

$$\hat{H}_{\boldsymbol{\eta}}(x) = \textstyle\sum_{t=0}^{\infty} \left\{ P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}) \right\} \;, \tag{50}$$

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

$$\max\{\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x)\|, \|\hat{H}_{\boldsymbol{\eta}}(x)\|\} \leq L_{PH}^{(0)}, \;\; \left\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x)\right\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \;. \tag{51}$$

Proposition 7 *Under A10, A11, the gradient $\nabla J(\boldsymbol{\eta})$ is Υ-Lipschitz continuous, where we defined $\Upsilon := \mathrm{R}_{\max} |\mathcal{S}||\mathcal{A}|$. Moreover, for any $\boldsymbol{\eta} \in \mathcal{H}$ and let $\Gamma := 2\overline{b}\, \mathrm{R}_{\max}\, K_R \frac{1}{(1-\rho)^2}$, it holds that*

$$(1-\lambda)^2 \Gamma^2 + 2 \left\langle \nabla J(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle \geq \|h(\boldsymbol{\eta})\|^2, \;\; \|\nabla J(\boldsymbol{\eta})\| \leq \|h(\boldsymbol{\eta})\| + (1-\lambda)\Gamma \;. \tag{52}$$

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 shows that

Corollary 2 *Under A10, A11 and set $\gamma_k = (2c_1 L(1 + C_h)\sqrt{k})^{-1}$, the policy gradient algorithm (43) converges to an $\mathcal{O}((1-\lambda)^2\Gamma^2)$-quasi-stationary point for the average reward (40) at the rate of $\mathcal{O}(\log n/\sqrt{n})$, where n is the iteration number.*

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 $\mathcal{O}(\log(n)/\sqrt{n})$ 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.

References

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvari. Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*, 2018.

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive Algorithms and Stochastic Approximation*. 01 1990. ISBN 0-387-52894-6. doi: 10.1007/978-3-642-75894-2.

Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Olivier Cappé and Eric Moulines. On-line Expectation Maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. Stochastic Expectation Maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988, 2018.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R examples*. Chapman and Hall/CRC, 2014.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1466–1475, 2018.

Gersende Fort, Eric Moulines, and Pierre Priouret. Convergence of adaptive and interacting Markov chain monte carlo algorithms. *The Annals of Statistics*, 39(6):3262–3289, 2011.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, pages 703–710, 1994.

Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. 80:4026–4035, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/papini18a.html.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9918–9927. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8195-on-markov-chain-gradient-descent.pdf.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction, 2nd Edition*. MIT Press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

Vladislav B Tadić and Arnaud Doucet. Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304, 2017.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional em algorithm: Statistical optimization and asymptotic normality. In *Advances in neural information processing systems*, pages 2521–2529, 2015.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of Expectation Maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

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:*

$$\begin{aligned}&\textstyle\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}\big(1 - c_1 L \gamma_{k+1}\big) h_k \\ &\leq V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1}) + L \textstyle\sum_{k=0}^{n} \gamma_{k+1}^2 \|\boldsymbol{e}_{k+1}\|^2 + \sum_{k=0}^{n} \gamma_{k+1}\big(c_1^{-1} c_0 - \langle \nabla V(\boldsymbol{\eta}_k) \,|\, \boldsymbol{e}_{k+1} \rangle\,\big)\ .\end{aligned} \tag{53}$$

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

$$\begin{aligned}V(\boldsymbol{\eta}_{k+1}) &\leq V(\boldsymbol{\eta}_k) - \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, H_{\boldsymbol{\eta}_k}(X_{k+1}) \right\rangle + \frac{L\gamma_{k+1}^2}{2} \|H_{\boldsymbol{\eta}_k}(X_{k+1})\|^2 \\ &\leq V(\boldsymbol{\eta}_k) - \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, h(\boldsymbol{\eta}_k) + \boldsymbol{e}_{k+1} \right\rangle + L\gamma_{k+1}^2 \big(\|h(\boldsymbol{\eta}_k)\|^2 + \|\boldsymbol{e}_{k+1}\|^2\big)\ .\end{aligned} \tag{54}$$

The above implies that

$$\begin{aligned}\gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, h(\boldsymbol{\eta}_k) \right\rangle \leq V(\boldsymbol{\eta}_k) - V(\boldsymbol{\eta}_{k+1}) &- \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \boldsymbol{e}_{k+1} \right\rangle \\ &+ L\gamma_{k+1}^2 \big(\|h(\boldsymbol{\eta}_k)\|^2 + \|\boldsymbol{e}_{k+1}\|^2\big)\ .\end{aligned} \tag{55}$$

Using A1, $\langle \nabla V(\boldsymbol{\eta}_k) \,|\, h(\boldsymbol{\eta}_k) \rangle \geq \frac{1}{c_1}(h_k - c_0)$ and rearranging terms, we obtain

$$\begin{aligned}\frac{\gamma_{k+1}}{c_1}\big(1 - c_1 L \gamma_{k+1}\big) h_k \leq V(\boldsymbol{\eta}_k) - V(\boldsymbol{\eta}_{k+1}) &- \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \boldsymbol{e}_{k+1} \right\rangle \\ &+ L\gamma_{k+1}^2 \|\boldsymbol{e}_{k+1}\|^2 + \frac{c_0}{c_1}\gamma_{k+1}\ .\end{aligned} \tag{56}$$

Summing up both sides from $k = 0$ to $k = n$ gives the conclusion (15). ■

A.2. Proof of Lemma 2

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

$$\mathbb{E}\left[- \sum_{k=0}^{n} \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \boldsymbol{e}_{k+1} \right\rangle \right] \leq C_h \sum_{k=0}^{n} \gamma_{k+1}^2 \mathbb{E}[\|h(\boldsymbol{\eta}_k)\|^2] + C_\gamma \sum_{k=0}^{n} \gamma_{k+1}^2 + C_{0,n}\ , \tag{57}$$

where C_h, C_γ and $C_{0,n}$ are defined in (12), (13), (14).

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

$$\boldsymbol{e}_{n+1} = H_{\boldsymbol{\eta}_n}(X_{n+1}) - h(\boldsymbol{\eta}_n) = \hat{H}_{\boldsymbol{\eta}_n}(X_{n+1}) - P_{\boldsymbol{\eta}_n} \hat{H}_{\boldsymbol{\eta}_n}(X_{n+1})\ . \tag{58}$$

The inner product on the left hand side of (20) can thus be decomposed as

$$\mathbb{E}\left[- \sum_{k=0}^{n} \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \boldsymbol{e}_{k+1} \right\rangle \right] = \mathbb{E}[A_1 + A_2 + A_3 + A_4 + A_5]\ , \tag{59}$$

with

$$
\begin{aligned}
A_1 &:= -\sum_{k=1}^{n} \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \hat{H}_{\boldsymbol{\eta}_k}(X_{k+1}) - P_{\boldsymbol{\eta}_k} \hat{H}_{\boldsymbol{\eta}_k}(X_k) \right\rangle \ , \\
A_2 &:= -\sum_{k=1}^{n} \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, P_{\boldsymbol{\eta}_k} \hat{H}_{\boldsymbol{\eta}_k}(X_k) - P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k) \right\rangle \ , \\
A_3 &:= -\sum_{k=1}^{n} \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) - \nabla V(\boldsymbol{\eta}_{k-1}) \,|\, P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k) \right\rangle \ , \\
A_4 &:= -\sum_{k=1}^{n} \big(\gamma_{k+1} - \gamma_k\big) \left\langle \nabla V(\boldsymbol{\eta}_{k-1}) \,|\, P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k) \right\rangle \ , \\
A_5 &:= -\gamma_1 \left\langle \nabla V(\boldsymbol{\eta}_0) \,|\, \hat{H}_{\boldsymbol{\eta}_0}(X_1) \right\rangle + \gamma_{n+1} \left\langle \nabla V(\boldsymbol{\eta}_n) \,|\, P_{\boldsymbol{\eta}_n} \hat{H}_{\boldsymbol{\eta}_n}(X_{n+1}) \right\rangle \ .
\end{aligned}
$$

For A_1, we note that $\hat{H}_{\boldsymbol{\eta}_k}(X_{k+1}) - P_{\boldsymbol{\eta}_k} \hat{H}_{\boldsymbol{\eta}_k}(X_k)$ 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

$$
\begin{aligned}
A_2 &\leq L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1} \|\nabla V(\boldsymbol{\eta}_k)\| \|\boldsymbol{\eta}_k - \boldsymbol{\eta}_{k-1}\| \\
&= L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \|\nabla V(\boldsymbol{\eta}_k)\| \|H_{\boldsymbol{\eta}_{k-1}}(X_k)\| \\
&\overset{(a)}{\leq} L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \big(d_0 + d_1 \|h(\boldsymbol{\eta}_k)\|\big) \big(\|h(\boldsymbol{\eta}_{k-1})\| + \sigma\big) \\
&\overset{(b)}{\leq} L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \Big(d_0\sigma + d_0 \|h(\boldsymbol{\eta}_{k-1})\| + d_1\sigma \|h(\boldsymbol{\eta}_k)\| + d_1 \|h(\boldsymbol{\eta}_k)\| \|h(\boldsymbol{\eta}_{k-1})\|\Big) \ ,
\end{aligned}
\tag{60}
$$

where (a) is due to A2 on the norm of $\nabla V(\boldsymbol{\eta}_k)$ and A7 on the norm of $\boldsymbol{e}_k$, (b) is obtained by expanding the scalar product. Using the inequality $\|h(\boldsymbol{\eta}_n)\| \leq 1 + \|h(\boldsymbol{\eta}_n)\|^2$ and $2\|h(\boldsymbol{\eta}_k)\| \|h(\boldsymbol{\eta}_{k-1})\| \leq \|h(\boldsymbol{\eta}_k)\|^2 + \|h(\boldsymbol{\eta}_{k-1}\|^2$, we obtain:

$$
A_2 \leq L_{PH}^{(1)} \left((d_0 + d_0\sigma + d_1\sigma) \sum_{k=1}^{n} \gamma_k^2 + \big(d_0 + \frac{d_1}{2} + a d_1 \sigma + \frac{a d_1}{2}\big) \sum_{k=0}^{n} \gamma_{k+1}^2 \|h(\boldsymbol{\eta}_k)\|^2 \right) . \tag{61}
$$

For A_3, we obtain

$$
\begin{aligned}
A_3 &\overset{(a)}{\leq} L \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \|H_{\boldsymbol{\eta}_{k-1}}(X_k)\| \|P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k)\| \\
&\overset{(b)}{\leq} L L_{PH}^{(0)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \big(\|h(\boldsymbol{\eta}_{k-1})\| + \sigma\big) \\
&\leq L L_{PH}^{(0)} \left((1+\sigma) \sum_{k=1}^{n} \gamma_k^2 + \sum_{k=1}^{n} \gamma_k^2 \|h(\boldsymbol{\eta}_{k-1})\|^2 \right) \ ,
\end{aligned}
\tag{62}
$$

where (a) uses A3, (b) uses $H_{\boldsymbol{\eta}_{k-1}}(X_k) = h(\boldsymbol{\eta}_{k-1}) + \boldsymbol{e}_k$ and A6.

For A_4, we have

$$\begin{aligned} A_4 &\leq \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| \big(d_0 + d_1 \|h(\boldsymbol{\eta}_{k-1})\|\big) \|P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k)\| \\ &\overset{(a)}{\leq} L_{PH}^{(0)} \left((d_0+1) \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| + d_1 \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| \|h(\boldsymbol{\eta}_{k-1})\|^2 \right) \\ &\overset{(b)}{=} L_{PH}^{(0)} \left((d_0+1)\big(\gamma_1 - \gamma_{n+1}\big) + a' d_1 \sum_{k=1}^{n} \gamma_k^2 \|h(\boldsymbol{\eta}_{k-1})\|^2 \right) , \end{aligned} \tag{63}$$

where (a) is again an application of A6, 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

$$\begin{aligned} A_5 &\overset{(a)}{\leq} \gamma_1 \big(d_0 + d_1 \|h(\boldsymbol{\eta}_0)\|\big) L_{PH}^{(0)} + \gamma_{n+1} \big(d_0 + d_1 \|h(\boldsymbol{\eta}_n)\|\big) L_{PH}^{(0)} \\ &\overset{(b)}{\leq} L_{PH}^{(0)} \Big(d_0 \{\gamma_1 + \gamma_{n+1}\} + 2d_1 + d_1 \{\gamma_1^2 \|h(\eta_0)\|^2 + \gamma_{n+1}^2 \|h(\eta_n)\|^2\} \Big) \\ &\leq L_{PH}^{(0)} \Big(d_0 \{\gamma_1 + \gamma_{n+1}\} + 2d_1 + d_1 \sum_{k=0}^{n} \gamma_{k+1}^2 \|h(\boldsymbol{\eta}_k)\|^2 \Big) , \end{aligned} \tag{64}$$

where (a) is an application of A2 and A6, and (b) uses $a \leq 1 + a^2$. Gathering the relevant terms and taking expectations conclude the proof of this lemma. ■

Appendix B. Analysis of the ro-EM method

B.1. Proof of Proposition 1

Proposition *Assume A8. Then*

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

Proof We have

$$\nabla_{\boldsymbol{\theta}} \operatorname{KL}(\pi, g(\cdot;\boldsymbol{\theta})) = -\nabla_{\boldsymbol{\theta}} \mathbb{E}_\pi\big[\log g(Y;\boldsymbol{\theta})\big] = -\mathbb{E}_\pi\big[\nabla_{\boldsymbol{\theta}} \log g(Y;\boldsymbol{\theta})\big] , \tag{65}$$

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 $y \in \mathsf{Y}$ that

$$\nabla_{\boldsymbol{\theta}} \log g(y;\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) + J_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) \overline{s}(y;\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) + J_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}\big[S(\boldsymbol{X})|Y = y\big] . \tag{66}$$

Therefore, for any $\boldsymbol{s}$, it holds that

$$\begin{aligned} &\nabla_{\boldsymbol{\theta}} \operatorname{KL}\big(\pi, g(\cdot;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big) + \nabla_{\boldsymbol{\theta}} \operatorname{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) = \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \nabla_{\boldsymbol{\theta}} \operatorname{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) - J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) \mathbb{E}_\pi\big[\overline{s}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big] \\ &\overset{(a)}{=} J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) \Big(\boldsymbol{s} - \mathbb{E}_\pi\big[\overline{s}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big] \Big) \overset{(b)}{=} J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) h(\boldsymbol{s}) . \end{aligned} \tag{67}$$

where we have used the assumption A8 in (a) and the definition of $h(\boldsymbol{s})$ in (b). The conclusion follows directly from the identity (67) since $J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$ is full rank. ■

B.2. Proof of Proposition 2

Proposition *Assume A8. Then, for* $\boldsymbol{s} \in \mathsf{S}$,

$$\nabla_{\boldsymbol{s}} V(\boldsymbol{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})\Big)^{-1}\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top} h(\boldsymbol{s}) \ . \tag{68}$$

Proof Using chain rule and A8, we obtain

$$\begin{aligned} \nabla_{\boldsymbol{s}} V(\boldsymbol{s}) &= \mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s})^{\top}\Big(\nabla_{\boldsymbol{\theta}}\,\mathrm{KL}\left(\pi, g(\cdot;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\right) + \nabla_{\boldsymbol{\theta}}\,\mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big) \\ &= \mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s})^{\top}\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top} h(\boldsymbol{s}) \ , \end{aligned} \tag{69}$$

where the last equality uses the identity in (67). Consider the following vector map:

$$\boldsymbol{s} \to \nabla_{\boldsymbol{\theta}}\psi(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \nabla_{\boldsymbol{\theta}}\,\mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top}\boldsymbol{s} \ . \tag{70}$$

Taking the gradient of the above map *w.r.t.* $\boldsymbol{s}$ and note that the map is constant for all $\boldsymbol{s} \in \mathsf{S}$, we show that:

$$\boldsymbol{0} = -\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \Big(\underbrace{\nabla_{\boldsymbol{\theta}}^{2}\big(\psi(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta}) - \langle\phi(\boldsymbol{\theta})\,|\,\boldsymbol{s}\rangle\big)}_{=\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})}\big|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\boldsymbol{s})}\Big)\,\mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s}) \ . \tag{71}$$

This implies $\mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s}) = \big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s};\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1}\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$. Substituting into (69) yields the conclusion. ■

B.3. Proof of Proposition 3

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

Proof From (28), we note that the error term is given by

$$\boldsymbol{e}_{n+1} = H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) - h(\hat{\boldsymbol{s}}_n) = \begin{pmatrix} \mathbb{E}_{Y_{n+1}\sim\pi}[\overline{\boldsymbol{s}}_n^{(1)}|\mathcal{F}_n] - \overline{\boldsymbol{s}}_n^{(1)} \\ \mathbb{E}_{Y_{n+1}\sim\pi}[\overline{\boldsymbol{s}}_n^{(2)}|\mathcal{F}_n] - \overline{\boldsymbol{s}}_n^{(2)} \\ \mathbb{E}_{Y_{n+1}\sim\pi}[\overline{s}_n^{(3)}|\mathcal{F}_n] - \overline{s}_n^{(3)} \end{pmatrix} . \tag{72}$$

Obviously, it holds that $\mathbb{E}[\boldsymbol{e}_{n+1}|\mathcal{F}_n] = \boldsymbol{0}$. Furthermore, for all $m \in \{1,\ldots,M-1\}$, the mth element of the first block in $\boldsymbol{e}_{n+1}$ has a bounded conditional variance

$$\mathbb{E}\Big[\big|\mathbb{E}_{Y_{n+1}\sim\pi}[\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)] - \omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] \leq 1 \ . \tag{73}$$

For the second block in $\boldsymbol{e}_{n+1}$, the conditional variance of its mth element is

$$\begin{aligned} &\mathbb{E}\Big[\big|\mathbb{E}_{Y_{n+1}\sim\pi}[Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)] - Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] \\ &= \mathbb{E}\Big[\big|Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] - \big|\mathbb{E}_{Y_{n+1}\sim\pi}[Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)]\big|^2 \\ &\leq \mathbb{E}\Big[\big|Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] \leq \mathbb{E}\big[(Y_{n+1})^2\big] \leq \overline{Y}^2. \end{aligned} \tag{74}$$

Lastly, we also have $\mathbb{E}[|\mathbb{E}_{Y_{n+1}\sim\pi}[\overline{s}_n^{(3)}|\mathcal{F}_n] - \overline{s}_n^{(3)}|^2] \leq \overline{Y}^2$. Therefore, we conclude that $\mathbb{E}[\|\boldsymbol{e}_{n+1}\|^2|\mathcal{F}_n] \leq M - 1 + M\overline{Y}^2 < \infty$. ■

B.4. Proof of Proposition 4

Proposition *Under A9 and the regularizer (36) set with $\epsilon > 0$, then for all $(\boldsymbol{s}, \boldsymbol{s}') \in \mathsf{S}^2$, there exists positive constants υ, Υ, Ψ such that:*

$$\langle \nabla V(\boldsymbol{s}) \,|\, h(\boldsymbol{s}) \rangle \geq \upsilon \|h(\boldsymbol{s})\|^2, \;\; \|\nabla V(\boldsymbol{s})\| \leq \Upsilon \|h(\boldsymbol{s})\|, \;\; \|\nabla V(\boldsymbol{s}) - \nabla V(\boldsymbol{s}')\| \leq \Psi \|\boldsymbol{s} - \boldsymbol{s}'\|. \tag{75}$$

Proof We first check that A8 is satisfied under A9. In particular, one observes that when $\boldsymbol{s} \in \mathsf{S} = \Delta_{M-1} \times [-\overline{Y}, \overline{Y}]^M$, the $\textsf{M-step}$ update (37) is the unique solution satisfying the stationary condition of the minimization problem (26) and $\overline{\boldsymbol{\theta}}(\boldsymbol{s}) \in \mathcal{C}$.

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

$$\nabla V(\boldsymbol{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) \Big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s}; \boldsymbol{\theta}) \} \Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top h(\boldsymbol{s}) \; . \tag{76}$$

Using (33), we observe that for any given $\boldsymbol{\theta} \in \mathcal{C}$, the Jacobian of ϕ and the Hessian of $\ell(\boldsymbol{s}, \boldsymbol{\theta})$ are given by

$$\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{1-\sum_{m=1}^{M-1} \omega_m} \mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{\mathbf{1}}{\boldsymbol{\omega}}) & -\mathrm{Diag}(\boldsymbol{\mu}) & \mu_M \mathbf{1} \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{pmatrix} , \tag{77}$$

$$\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{\theta}) = \begin{pmatrix} \frac{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}{(1-\sum_{m=1}^{M-1} \omega_m)^2} \mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}{\boldsymbol{\omega}^2}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\boldsymbol{s}^{(1)} + \epsilon \mathbf{1}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 + \epsilon - \sum_{m=1}^{M-1} s_m^{(1)} \end{pmatrix} ,$$

where we have denoted $\frac{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}{\boldsymbol{\omega}^2}$ as the $(M-1)$-vector $\big(\frac{s_1^{(1)}+\epsilon}{\omega_1^2}, \ldots, \frac{s_{M-1}^{(1)}+\epsilon}{\omega_{M-1}^2}\big)$. Let us define $\boldsymbol{J}_{11}, \boldsymbol{H}_{11}$ as the top-left matrices in the above, evaluated at $\overline{\boldsymbol{\theta}}(\boldsymbol{s})$, as follows

$$\boldsymbol{J}_{11} := \frac{1}{1 - \frac{\mathbf{1}^\top(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})}{1+\epsilon M}} \mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{1+\epsilon M}{\boldsymbol{s}^{(1)} + \epsilon \mathbf{1}}) \tag{78}$$

$$\boldsymbol{H}_{11} := \frac{1 + \epsilon - \sum_{m=1}^{M-1} s_m^{(1)}}{(1 - \frac{\mathbf{1}^\top(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})}{1+\epsilon M})^2} \mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{(1+\epsilon M)^2}{\boldsymbol{s}^{(1)} + \epsilon \mathbf{1}}). \tag{79}$$

When $\epsilon > 0$, the above matrices, $\boldsymbol{J}_{11}$ and $\boldsymbol{H}_{11}$, are full rank and bounded if $\boldsymbol{s} \in \mathsf{S}$.

The matrix product $\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s}, \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top$ can hence be expressed as an outer product

$$\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s}, \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top = \boldsymbol{\mathcal{J}}(\boldsymbol{s}) \boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top \; , \tag{80}$$

with

$$\begin{aligned} \boldsymbol{\mathcal{J}}(\boldsymbol{s}) &:= \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) \begin{pmatrix} \boldsymbol{H}_{11}^{-\frac{1}{2}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{J}_{11}\boldsymbol{H}_{11}^{-\frac{1}{2}} & -\mathrm{Diag}(\frac{\boldsymbol{s}^{(2)}}{(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})^{\frac{3}{2}}}) & \frac{s^{(3)}-\mathbf{1}^\top\boldsymbol{s}^{(2)}}{(1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)})^{\frac{3}{2}}} \mathbf{1} \\ \mathbf{0} & \mathrm{Diag}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}} \end{pmatrix} . \end{aligned} \tag{81}$$

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

$$\langle \nabla V(\boldsymbol{s}) \,|\, h(\boldsymbol{s}) \rangle = \Big\langle \mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top h(\boldsymbol{s}) \,|\, h(\boldsymbol{s}) \Big\rangle \geq \upsilon \| h(\boldsymbol{s}) \|^2 \, . \tag{82}$$

The second part in (38) can be verified by observing that $\mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}^{\boldsymbol{\theta}}_{\ell}(\boldsymbol{s};\boldsymbol{\theta}) \} \Big)^{-1} \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top$ is bounded due to A9.

For the third part in (38), again from (76) we obtain:

$$\nabla V(\boldsymbol{s}) = \mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top h(\boldsymbol{s}) \, . \tag{83}$$

From (81), it can be seen that $\mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top$ is Lipschitz continuous in $\boldsymbol{s}$ and bounded, *i.e.,* there exists constants $L_J, C_J < \infty$ such that

$$\| \mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top - \mathcal{J}(\boldsymbol{s}') \mathcal{J}(\boldsymbol{s}')^\top \| \leq L_J \| \boldsymbol{s} - \boldsymbol{s}' \|, ~~ \| \mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top \| \leq C_J, \, \forall \, \boldsymbol{s}, \boldsymbol{s}' \in \mathsf{S} \, . \tag{84}$$

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

$$\begin{aligned} \| h(\boldsymbol{s}) - h(\boldsymbol{s}') \| &= \| \boldsymbol{s} - \boldsymbol{s}' + \mathbb{E}_{Y \sim \pi} \big[\overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}')) - \overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) \big] \| \\ &\overset{(a)}{\leq} \| \boldsymbol{s} - \boldsymbol{s}' \| + \mathbb{E}_{Y \sim \pi} \big[\| \overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}')) - \overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) \| \big] \, , \end{aligned} \tag{85}$$

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

$$\overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}')) - \overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) = \left(\begin{array}{c} \widetilde{\boldsymbol{\omega}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})') - \widetilde{\boldsymbol{\omega}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) \\ Y \big(\widetilde{\boldsymbol{\omega}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})') - \widetilde{\boldsymbol{\omega}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) \big) \\ 0 \end{array} \right) \, , \tag{86}$$

where $\widetilde{\boldsymbol{\omega}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))$ is a collection of the $M-1$ terms $\widetilde{\omega}_m(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))$, $m = 1, \ldots, M-1$ [cf. (34)]. Observe that

$$\widetilde{\omega}_m(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) = \frac{ \frac{s_m^{(1)} + \epsilon}{1 + \epsilon M} \exp(- \frac{1}{2} (Y - \frac{s_m^{(2)}}{s_m^{(1)} + \epsilon})^2) }{ \sum_{j=1}^M \frac{s_j^{(1)} + \epsilon}{1 + \epsilon M} \exp(- \frac{1}{2} (Y - \frac{s_j^{(2)}}{s_j^{(1)} + \epsilon})^2) } \, . \tag{87}$$

Under A9 and the condition that $\boldsymbol{s} \in \mathsf{S}$, *i.e.,* a compact set, there exists $L_\omega < \infty$ such that

$$| \widetilde{\omega}_m(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) - \widetilde{\omega}_m(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}')) |^2 \leq L_\omega^2 \| \boldsymbol{s} - \boldsymbol{s}' \|^2 \, , \tag{88}$$

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

$$\| \overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}')) - \overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s})) \| \leq (M-1)(1 + \overline{Y}) L_\omega \| \boldsymbol{s} - \boldsymbol{s}' \| \, , \tag{89}$$

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

$$\begin{aligned} & \| \nabla V(\boldsymbol{s}) - \nabla V(\boldsymbol{s}') \| = \| \mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top h(\boldsymbol{s}) - \mathcal{J}(\boldsymbol{s}') \mathcal{J}(\boldsymbol{s}')^\top h(\boldsymbol{s}') \| \\ & = \| \mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top (h(\boldsymbol{s}) - h(\boldsymbol{s}')) + \big(\mathcal{J}(\boldsymbol{s}) \mathcal{J}(\boldsymbol{s})^\top - \mathcal{J}(\boldsymbol{s}') \mathcal{J}(\boldsymbol{s}')^\top \big) h(\boldsymbol{s}') \| \\ & \leq \big(L_h C_J + L_J C_h \big) \| \boldsymbol{s} - \boldsymbol{s}' \|, \end{aligned} \tag{90}$$

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 leads to the convergence of the policy gradient algorithm analyzed in Section 3.2.

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

$$\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\eta}}(s,a) := \frac{\partial \log \Pi(a;s,\boldsymbol{\eta})}{\partial \eta_i}, \quad r(s,a) := \mathcal{R}(s,a). \tag{91}$$

Using the above notations, the mean field in (47) can be evaluated as

$$h(\boldsymbol{\eta}) = \sum_{t=0}^{\infty} \sum_{(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}} \lambda^t \mathcal{R}(s',a') \tilde{Q}_{\boldsymbol{\eta}}^t((s,a);(s',a')) \nabla \log \Pi(a;s,\boldsymbol{\eta}) \upsilon_{\boldsymbol{\eta}}(s,a). \tag{92}$$

In particular, its ith element can be expressed as

$$h_i(\boldsymbol{\eta}) = \sum_{t=0}^{\infty} \lambda^t \boldsymbol{\upsilon}_{\boldsymbol{\eta}}^\top \mathrm{Diag}(\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^t \boldsymbol{r} \ . \tag{93}$$

We also define the difference between $h(\boldsymbol{\eta})$ and $\nabla J(\boldsymbol{\eta})$ as

$$\Delta(\boldsymbol{\eta}) := h(\boldsymbol{\eta}) - \nabla J(\boldsymbol{\eta}). \tag{94}$$

C.1. Useful Lemmas

Lemma 3 *Let A10, A11 hold. For any $(\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{H}^2$ and $t \geq 0$, one has*

$$\|\boldsymbol{Q}_{\boldsymbol{\eta}}^t - \boldsymbol{Q}_{\boldsymbol{\eta}'}^t\| \leq C_1 \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|, \quad \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^t - \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^t\| \leq C_1 \big(t\rho^t\big) \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \ , \tag{95}$$

where we have set $C_1 := \rho K_R^2 \big(2\bar{b} + L_Q\big) + L_Q$ in the above.

Proof For part 1), we observe that each entry of $\boldsymbol{Q}_{\boldsymbol{\eta}}$ is given by [cf. (39)]:

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

which is Lipschitz continuous *w.r.t.* $\boldsymbol{\eta}$ since

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

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

$$|Q_{\boldsymbol{\eta}}((s,a);(s',a')) - Q_{\boldsymbol{\eta}'}((s,a);(s',a'))| \leq 2\bar{b} |P_{s,s'}^a| \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \ . \tag{96}$$

Since $|P_{s,s'}^a| \leq 1$ for any s, s', a, we have $\|\boldsymbol{Q}_{\boldsymbol{\eta}} - \boldsymbol{Q}_{\boldsymbol{\eta}'}\| \leq 2\bar{b}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$.

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

$$\begin{aligned} \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^{t+1} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^{t+1} &= \sum_{\tau=0}^{t} \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^{\tau} \big(\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'} \big) \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^{t-\tau} \\ &= \sum_{\tau=0}^{t} \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^{\tau} \big(\boldsymbol{Q}_{\boldsymbol{\eta}} - \boldsymbol{Q}_{\boldsymbol{\eta}'} - \mathbf{1}(\boldsymbol{\upsilon}_{\boldsymbol{\eta}} - \boldsymbol{\upsilon}_{\boldsymbol{\eta}'})^{\top} \big) \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^{t-\tau} \; . \end{aligned} \tag{97}$$

As such,

$$\begin{aligned} \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^{t+1} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^{t+1}\| &\leq \sum_{\tau=0}^{t} \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^{\tau}\| \big\| \boldsymbol{Q}_{\boldsymbol{\eta}} - \boldsymbol{Q}_{\boldsymbol{\eta}'} - \mathbf{1}(\boldsymbol{\upsilon}_{\boldsymbol{\eta}} - \boldsymbol{\upsilon}_{\boldsymbol{\eta}'})^{\top} \big\| \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^{t-\tau}\| \\ &\leq K_R^2 \sum_{\tau=0}^{t} \rho^{\tau} \rho^{t-\tau} \big(\|\boldsymbol{Q}_{\boldsymbol{\eta}} - \boldsymbol{Q}_{\boldsymbol{\eta}'}\| + \|\boldsymbol{\upsilon}_{\boldsymbol{\eta}} - \boldsymbol{\upsilon}_{\boldsymbol{\eta}'}\| \big) \\ &\leq K_R^2 \big(2\bar{b} + L_Q \big) \big(t\rho^t \big) \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \; . \end{aligned} \tag{98}$$

Consequently,

$$\begin{aligned} \|\boldsymbol{Q}_{\boldsymbol{\eta}}^{t+1} - \boldsymbol{Q}_{\boldsymbol{\eta}'}^{t+1}\| &\leq \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^{t+1} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}'}^{t+1}\| + \|\boldsymbol{\upsilon}_{\boldsymbol{\eta}} - \boldsymbol{\upsilon}_{\boldsymbol{\eta}'}\| \\ &\leq \big(K_R^2 \big(t\rho^t \big) \big(2\bar{b} + L_Q \big) + L_Q \big) \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \; . \end{aligned} \tag{99}$$

Setting $C_1 = \rho K_R^2 \big(2\bar{b} + L_Q \big) + L_Q$ completes the proof. ■

Lemma 4 *Let A10, A11 hold. The following statements are true:*

1. *The average reward $J(\boldsymbol{\eta})$ is differentiable and for any $(\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{H}^2$, one has*

$$\|\nabla J(\boldsymbol{\eta}) - \nabla J(\boldsymbol{\eta}')\| \leq \mathrm{R}_{\max} \, |\mathcal{S}||\mathcal{A}| \, L_{\upsilon} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \; . \tag{100}$$

2. *For any $\boldsymbol{\eta} \in \mathcal{H}$, one has*

$$\|\Delta(\boldsymbol{\eta})\| \leq 2\bar{b} \, \mathrm{R}_{\max} \, K_R \frac{1-\lambda}{(1-\rho)^2} \; . \tag{101}$$

Proof For part 1), we observe that

$$J(\boldsymbol{\eta}) = \mathbb{E}_{(S,A) \sim \boldsymbol{\upsilon}_{\boldsymbol{\eta}}} \big[\mathcal{R}(S,A) \big] = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \upsilon_{\boldsymbol{\eta}}(s,a) \mathcal{R}(s,a) \; . \tag{102}$$

It follows from the Lipschitz continuity of $\mathrm{J}_{\boldsymbol{\upsilon}_{\eta}}^{\boldsymbol{\eta}}(\boldsymbol{\eta})$ [cf. A11] that

$$\begin{aligned} \|\nabla J(\boldsymbol{\eta}) - \nabla J(\boldsymbol{\eta}')\| &\leq \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\mathcal{R}(s,a)| \|\nabla \upsilon_{\boldsymbol{\eta}}(s,a) - \nabla \upsilon_{\boldsymbol{\eta}'}(s,a)\| \\ &\leq \mathrm{R}_{\max} \, |\mathcal{S}||\mathcal{A}| \, L_{\upsilon} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \; . \end{aligned} \tag{103}$$

The above verifies (100).

For part 2), we define

$$J_T(\boldsymbol{\eta},(s,a)) := \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s',a') Q_{\boldsymbol{\eta}}^T((s,a);(s',a')) \ , \tag{104}$$

$$g(\boldsymbol{\eta}) := \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s,a)\tilde{Q}_{\boldsymbol{\eta}}^t((s,a);(s',a'))\nabla \log \Pi(a;s,\boldsymbol{\eta})\upsilon_{\boldsymbol{\eta}}(s,a) \ . \tag{105}$$

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

$$\begin{aligned} &\Delta(\boldsymbol{\eta}) = h(\boldsymbol{\eta}) - g(\boldsymbol{\eta}) \\ &= \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} (\lambda^t - 1)\mathcal{R}(s,a)\tilde{Q}_{\boldsymbol{\eta}}^t((s,a);(s',a'))\nabla \log \Pi(a;s,\boldsymbol{\eta})\upsilon_{\boldsymbol{\eta}}(s,a) \ . \end{aligned} \tag{106}$$

and in particular, the ith element is given by

$$\Delta_i(\boldsymbol{\eta}) = \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} (\lambda^t - 1)\boldsymbol{\upsilon}_{\boldsymbol{\eta}}^\top \mathrm{Diag}(\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\eta}})\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^t \boldsymbol{r} \ , \tag{107}$$

which can be bounded as

$$\begin{aligned} |\Delta_i(\boldsymbol{\eta})| &\leq \sum_{t=0}^{\infty}(1-\lambda^t)\|\boldsymbol{\upsilon}_{\boldsymbol{\eta}}\|\|\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\eta}}\|_\infty \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\eta}}^t\|\|\boldsymbol{r}\| \\ &\overset{(a)}{\leq} 2\bar{b}\,\mathrm{R}_{\max} K_R \sum_{t=0}^{\infty}(1-\lambda^t)\rho^t \leq 2\bar{b}\,\mathrm{R}_{\max} K_R \frac{1-\lambda}{(1-\rho)^2} \ , \end{aligned} \tag{108}$$

where (a) uses A11, A10, and Proposition 5. The above implies that $\|\Delta(\boldsymbol{\eta})\| \leq 2\bar{b}\,\mathrm{R}_{\max} K_R \frac{1-\lambda}{(1-\rho)^2}$.
■

Lemma 5 *Let A10, A11 hold. Denote the joint state x as $x = (s,a,g) \in \mathcal{S}\times\mathcal{A}\times\mathbb{R}^d$. There exists $\delta\in[0,1)$, $C_2\in[1,\infty)$ such that for any $t\geq 0$,*

$$\begin{aligned} &\|P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\| \leq C_2 t \delta^t (1+\|g\|) \ , \\ &\Big\| \big(P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\big) - \big(P_{\boldsymbol{\eta}'}^t H_{\boldsymbol{\eta}'}(x) - h(\boldsymbol{\eta}')\big)\Big\| \leq C_2 t \delta^t \|\boldsymbol{\eta}-\boldsymbol{\eta}'\|(1+\|g\|) \ . \end{aligned} \tag{109}$$

Proof Denoting the joint state as $x=(s,a,g)$, we observe that

$$\begin{aligned} &P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) = \mathbb{E}_{\Pi_{\boldsymbol{\eta}}}\big[\mathcal{R}(S_t,A_t)G_t \,|\, (S_0,A_0)=(s,a), G_0 = g\big] \\ &= \mathbb{E}_{\Pi_{\boldsymbol{\eta}}}\Bigg[\mathcal{R}(S_t,A_t)\Big(\lambda^t g + \sum_{i=1}^{t-1}\lambda^i \nabla\log\Pi(A_i;S_i,\boldsymbol{\eta})\Big) \,\Big|\, (S_0,A_0)=(s,a)\Bigg] \\ &= \sum_{i=0}^{t-1} \sum_{(s',a'),(s'',a'')\in\mathcal{S}\times\mathcal{A}} \lambda^i \mathcal{R}(s'',a'')Q_{\boldsymbol{\eta}}^i((s',a');(s'',a''))\nabla\log\Pi(a';s',\boldsymbol{\eta})Q_{\boldsymbol{\eta}}^{t-i}((s,a);(s',a')) \\ &\quad + \lambda^t g \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s',a')Q_{\boldsymbol{\eta}}^t((s,a);(s',a')) \ . \end{aligned}$$

The jth element of the above is thus given by

$$\big[P^t_{\boldsymbol{\eta}} H_{\boldsymbol{\eta}}(x)\big]_j = \sum_{i=0}^{t-i} \lambda^i \boldsymbol{e}^\top_{(s,a)} \boldsymbol{Q}^{t-i}_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} + \lambda^t g_j \mathbf{1}^\top \boldsymbol{Q}^t_{\boldsymbol{\eta}} \boldsymbol{r} \ , \tag{110}$$

where g_j is the jth element of g and $\boldsymbol{e}_{(s,a)}$ is the (s,a)th coordinate vector. Moreover, we recall that

$$h_j(\boldsymbol{\eta}) = \sum_{t=0}^{\infty} \lambda^t \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \tilde{\boldsymbol{Q}}^t_{\boldsymbol{\eta}} \boldsymbol{r} \ . \tag{111}$$

Note that

$$\begin{aligned} \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \mathbf{1} &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} v_{\boldsymbol{\eta}}(s,a) \nabla_j \log \Pi(a;s,\boldsymbol{\eta}) \\ &= \sum_{s\in\mathcal{S}} \Big(\sum_{a\in\mathcal{A}} \underbrace{\Pi(a;s,\boldsymbol{\eta}) \nabla_j \log \Pi(a;s,\boldsymbol{\eta})}_{=\nabla_j \Pi(a;s,\boldsymbol{\eta})} \Big) \overline{\Pi}_{\boldsymbol{\eta}}(s) = 0 \ . \end{aligned} \tag{112}$$

where we recalled that $\overline{\Pi}_{\boldsymbol{\eta}}(s)$ is the stationary distribution for the MDP on the state. Using the decomposition $\tilde{\boldsymbol{Q}}^t_{\boldsymbol{\eta}} = \boldsymbol{Q}^t_{\boldsymbol{\eta}} - \mathbf{1}\boldsymbol{v}^\top_{\boldsymbol{\eta}}$, we observe

$$\begin{aligned} h_j(\boldsymbol{\eta}) &= \sum_{i=0}^{t-1} \lambda^i \Big\{ \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} - \underbrace{\boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \mathbf{1}}_{=0} \boldsymbol{v}^\top_{\boldsymbol{\eta}} \boldsymbol{r} \Big\} + \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \tilde{\boldsymbol{Q}}^i_{\boldsymbol{\eta}} \boldsymbol{r} \\ &= \sum_{i=0}^{t-1} \lambda^i \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} + \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \tilde{\boldsymbol{Q}}^i_{\boldsymbol{\eta}} \boldsymbol{r} \ . \end{aligned}$$

Therefore,

$$\begin{aligned} &\big[P^t_{\boldsymbol{\eta}} H_{\boldsymbol{\eta}}(x)\big]_j - h_j(\boldsymbol{\eta}) \\ &= \sum_{i=0}^{t-1} \lambda^i \Big\{ \boldsymbol{e}^\top_{(s,a)} (\tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\eta}} + \mathbf{1}\boldsymbol{v}_{\boldsymbol{\eta}}) \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} - \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} \Big\} \\ &\quad + \lambda^t g_j \mathbf{1}^\top \boldsymbol{Q}^t_{\boldsymbol{\eta}} \boldsymbol{r} - \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \tilde{\boldsymbol{Q}}^i_{\boldsymbol{\eta}} \boldsymbol{r} \\ &= \sum_{i=0}^{t-1} \lambda^i \boldsymbol{e}^\top_{(s,a)} \tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} + \lambda^t g_j \mathbf{1}^\top \boldsymbol{Q}^t_{\boldsymbol{\eta}} \boldsymbol{r} - \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}^\top_{\boldsymbol{\eta}} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}) \tilde{\boldsymbol{Q}}^i_{\boldsymbol{\eta}} \boldsymbol{r} \ . \end{aligned} \tag{113}$$

Consequently, we obtain the upper bound as

$$\begin{aligned} \big| \big[P^t_{\boldsymbol{\eta}} H_{\boldsymbol{\eta}}(x)\big]_j - h_j(\boldsymbol{\eta}) \big| \leq & \sum_{i=0}^{t-1} \lambda^i \| \tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\eta}} \| \| \nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}} \|_\infty \| \boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r} \| + \lambda^t |g_j| \| \boldsymbol{Q}^t_{\boldsymbol{\eta}} \boldsymbol{r} \| \\ & + \sum_{i=t}^{\infty} \lambda^i \| \boldsymbol{v}_{\boldsymbol{\eta}} \| \| \nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}} \|_\infty \| \tilde{\boldsymbol{Q}}^i_{\boldsymbol{\eta}} \boldsymbol{r} \| \ . \end{aligned} \tag{114}$$

Using A10, A11 and notice that $\|\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\eta}}\|_\infty \le 2\bar{b}$, $\|\boldsymbol{Q}^i_{\boldsymbol{\eta}} \boldsymbol{r}\| \le \overline{R}$, $\|\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\eta}}\boldsymbol{r}\| \le \overline{R} K_R \sqrt{|\mathcal{S}||\mathcal{A}|}\rho^i$, we obtain

$$\big|\big[P^t_{\boldsymbol\eta} H_{\boldsymbol\eta}(x)\big]_j - h_j(\boldsymbol\eta)\big| \le 2\bar b \overline{R} K_R \sum_{i=0}^{t-1}\lambda^i\rho^{t-i} + \lambda^t |g_j|\overline{R} + 2\bar b\overline{R}K_R\sqrt{|\mathcal S||\mathcal A|}\sum_{i=t}^\infty \lambda^i\rho^i \,. \tag{115}$$

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[1]

$$\big|\big[P^t_{\boldsymbol\eta} H_{\boldsymbol\eta}(x)\big]_j - h_j(\boldsymbol\eta)\big| \le C_2'\big(t\delta^t\big)\big(1+\|g\|\big) \,, \tag{116}$$

which naturally implies the first equation in (109).

For the second equation in (109),

$$\begin{aligned}
&\big[P^t_{\boldsymbol\eta} H_{\boldsymbol\eta}(x)\big]_j - h_j(\boldsymbol\eta) - \Big\{\big[P^t_{\boldsymbol\eta'} H_{\boldsymbol\eta'}(x)\big]_j - h_j(\boldsymbol\eta')\Big\} \\
&= \sum_{i=0}^{t-1}\lambda^i \boldsymbol{e}^\top_{(s,a)}\{\tilde{\boldsymbol Q}^{t-i}_{\boldsymbol\eta}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta})\boldsymbol Q^i_{\boldsymbol\eta} - \tilde{\boldsymbol Q}^{t-i}_{\boldsymbol\eta'}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta'})\boldsymbol Q^i_{\boldsymbol\eta'}\}\boldsymbol r \\
&\quad + \lambda^t g_j \mathbf{1}^\top\big(\boldsymbol Q^t_{\boldsymbol\eta} - \boldsymbol Q^t_{\boldsymbol\eta'}\big)\boldsymbol r + \sum_{i=t}^{\infty}\lambda^i\{\boldsymbol v^\top_{\boldsymbol\eta'}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta'})\tilde{\boldsymbol Q}^i_{\boldsymbol\eta'} - \boldsymbol v^\top_{\boldsymbol\eta}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta})\tilde{\boldsymbol Q}^i_{\boldsymbol\eta}\}\boldsymbol r \,.
\end{aligned} \tag{117}$$

This leads to the upper bound:

$$\begin{aligned}
&\Big|\big[P^t_{\boldsymbol\eta} H_{\boldsymbol\eta}(x)\big]_j - h_j(\boldsymbol\eta) - \Big\{\big[P^t_{\boldsymbol\eta'} H_{\boldsymbol\eta'}(x)\big]_j - h_j(\boldsymbol\eta')\Big\}\Big| \\
&\le \sqrt{|\mathcal S||\mathcal A|}\overline{R}\sum_{i=0}^{t-i}\lambda^i\big\|\tilde{\boldsymbol Q}^{t-i}_{\boldsymbol\eta}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta})\boldsymbol Q^i_{\boldsymbol\eta} - \tilde{\boldsymbol Q}^{t-i}_{\boldsymbol\eta'}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta'})\boldsymbol Q^i_{\boldsymbol\eta'}\big\| \\
&\quad + \lambda^t|\mathcal S||\mathcal A|\|\boldsymbol Q^t_{\boldsymbol\eta} - \boldsymbol Q^t_{\boldsymbol\eta'}\| + \sqrt{|\mathcal S||\mathcal A|}\overline{R}\sum_{i=t}^{\infty}\lambda^i\|\boldsymbol v^\top_{\boldsymbol\eta'}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta'})\tilde{\boldsymbol Q}^i_{\boldsymbol\eta'} - \boldsymbol v^\top_{\boldsymbol\eta}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta})\tilde{\boldsymbol Q}^i_{\boldsymbol\eta}\| \,.
\end{aligned} \tag{118}$$

Using the boundedness and Lipschitz continuity of $\nabla_j\boldsymbol\Pi_{\boldsymbol\eta}$, $\boldsymbol v_{\boldsymbol\eta}$, $\boldsymbol Q^t_{\boldsymbol\eta}$, $\tilde{\boldsymbol Q}^t_{\boldsymbol\eta}$ [cf. Lemma 3], let $C_{2,1}, C_{2,2} \in [1,\infty)$, the norms in the above can be bounded as

$$\begin{aligned}
&\big\|\tilde{\boldsymbol Q}^{t-i}_{\boldsymbol\eta}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta})\boldsymbol Q^i_{\boldsymbol\eta} - \tilde{\boldsymbol Q}^{t-i}_{\boldsymbol\eta'}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta'})\boldsymbol Q^i_{\boldsymbol\eta'}\big\| \le C_{2,1}\big((t-i)\rho^{t-i}\big)\|\boldsymbol\eta - \boldsymbol\eta'\| \\
&\big\|\boldsymbol v^\top_{\boldsymbol\eta'}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta'})\tilde{\boldsymbol Q}^i_{\boldsymbol\eta'} - \boldsymbol v^\top_{\boldsymbol\eta}\mathrm{Diag}(\nabla_j\boldsymbol\Pi_{\boldsymbol\eta})\tilde{\boldsymbol Q}^i_{\boldsymbol\eta}\big\| \le C_{2,2}\big(i\rho^i\big)\|\boldsymbol\eta - \boldsymbol\eta'\| \\
&\big\|\boldsymbol Q^t_{\boldsymbol\eta} - \boldsymbol Q^t_{\boldsymbol\eta'}\big\| \le C_1\|\boldsymbol\eta - \boldsymbol\eta'\| \,.
\end{aligned} \tag{119}$$

The above shows that the three terms in the right hand side of (118) are proportional to $(1+\|g\|)\|\boldsymbol\eta - \boldsymbol\eta'\|$ and decay geometrically with t. This implies there exists $C_2'' \in [1,\infty)$, $\delta \in [0,1)$ such that

$$\Big\|P^t_{\boldsymbol\eta} H_{\boldsymbol\eta}(x) - h(\boldsymbol\eta) - \Big\{P^t_{\boldsymbol\eta'} H_{\boldsymbol\eta'}(x) - h(\boldsymbol\eta')\Big\}\Big\| \le C_2''\big(t\delta^t\big)(1+\|g\|)\|\boldsymbol\eta - \boldsymbol\eta'\| \,. \tag{120}$$

Setting $C_2 = \max\{C_2', C_2''\}$ concludes the proof of the current lemma. ■

[1] Note that an exact characterization for C_2' is also possible.

C.2. Proof of Proposition 5

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

$$\|\nabla \log \Pi_{\boldsymbol{\eta}}(a;s)\| \leq 2\bar{b}, \quad \|\nabla \log \Pi_{\boldsymbol{\eta}}(a;s) - \nabla \log \Pi_{\boldsymbol{\eta}'}(a;s)\| \leq 8\bar{b}^2 \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \,. \tag{121}$$

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

$$\nabla \log \Pi_{\boldsymbol{\eta}}(a;s) = \frac{1}{\sum_{a' \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle\big)} \sum_{b \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle\big) \Delta \boldsymbol{x}(a,b) \,. \tag{122}$$

Observe that

$$\|\nabla \log \Pi_{\boldsymbol{\eta}}(a;s)\| \leq \max_{a,b \in \mathsf{A}} \|\boldsymbol{x}(s,a) - \boldsymbol{x}(s,b)\| \leq 2\bar{b} \,. \tag{123}$$

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

$$\begin{aligned}
&\nabla^2 \log \Pi_{\boldsymbol{\eta}}(a;s) = \\
&\frac{1}{\sum_{a' \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle\big)} \sum_{b \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle\big) \Delta \boldsymbol{x}(a,b) \Delta \boldsymbol{x}(b,a)^\top - \\
&\Big(\sum_{b \in \mathsf{A}} \frac{\exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle\big)}{\sum_{a' \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle\big)} \Delta \boldsymbol{x}(a,b)\Big) \Big(\frac{\exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle\big)}{\sum_{a' \in \mathsf{A}} \exp\big(\langle \boldsymbol{\eta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle\big)} \Delta \boldsymbol{x}(a,b)\Big)^\top \,.
\end{aligned} \tag{124}$$

It can be checked that

$$\|\nabla^2 \log \Pi_{\boldsymbol{\eta}}(a;s)\| \leq \max_{a,b \in \mathsf{A}} \big\|\Delta \boldsymbol{x}(a,b) \Delta \boldsymbol{x}(b,a)^\top\big\| + \big(\max_{a,b \in \mathsf{A}} \|\Delta \boldsymbol{x}(a,b)\|\big)^2 \leq 8\bar{b}^2 \,. \tag{125}$$

This implies smoothness condition in (48). ■

C.3. Proof of Proposition 6

Proposition *Under A10, A11, the function*

$$\hat{H}_{\boldsymbol{\eta}}(x) = \textstyle\sum_{t=0}^{\infty} \big\{P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\big\} \,, \tag{126}$$

is well defined and satisfies the Poisson equation (7). For all $x \in \mathsf{X}$, $(\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{H}^2$, *there exists constants* $L_{PH}^{(0)}$, $L_{PH}^{(1)}$ *such that*

$$\max\{\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x)\|, \|\hat{H}_{\boldsymbol{\eta}}(x)\|\} \leq L_{PH}^{(0)}, \quad \left\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x)\right\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \,. \tag{127}$$

Proof From Lemma 5, there exists $C_2 \in [1, \infty)$, $\delta \in [0,1)$ such that

$$\|P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\| \leq C_2 t \delta^t (1 + \|g\|), \; \forall\, t \geq 1, \; \forall\, x \in \mathsf{X} \,, \tag{128}$$

It follows that the solution to the Poisson equation $\hat{H}_{\boldsymbol{\eta}}(x)$ in (50) is well defined.

Moreover, it satisfies (7) and

$$\max\{\|\hat{H}_{\boldsymbol{\eta}}(x)\|, \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x)\|\} \leq L_{PH}^{(0)} \, , \tag{129}$$

for some $L_{PH}^{(0)} < \infty$ (note that g is bounded as specified by the state space X). As such, the first equation in (51) of the proposition is proven. Finally, applying the definition of $\hat{H}_{\boldsymbol{\eta}}(x)$ shows that

$$P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}'}(x) = \sum_{t=1}^{\infty} \Big\{ \big(P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\big) - \big(P_{\boldsymbol{\eta}'}^t H_{\boldsymbol{\eta}'}(x) - h(\boldsymbol{\eta}')\big) \Big\} \, . \tag{130}$$

Using Lemma 5, this implies

$$\begin{aligned} \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}'}(x)\| &\leq \sum_{t=1}^{\infty} \Big\| \big(P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\big) - \big(P_{\boldsymbol{\eta}'}^t H_{\boldsymbol{\eta}'}(x) - h(\boldsymbol{\eta}')\big) \Big\| \\ &\leq \sum_{t=1}^{\infty} \Big\{ C_2 \big(t\delta^t\big) \big(1 + \|g\|\big) \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \Big\} \, . \end{aligned} \tag{131}$$

As such, there exists $L_{PH}^{(1)} \in [1, \infty)$ such that

$$\|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}'}(x)\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \, , \tag{132}$$

for all $x \in \mathsf{X}$. This proves the second equation in (51) of the proposition. ■

C.4. Proof of Proposition 7

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

$$(1-\lambda)^2 \Gamma^2 + 2 \left\langle \nabla J(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle \geq \|h(\boldsymbol{\eta})\|^2, \; \|\nabla J(\boldsymbol{\eta})\| \leq \|h(\boldsymbol{\eta})\| + (1-\lambda)\Gamma \, , \tag{133}$$

where $\Gamma := 2\bar{b}\, \mathrm{R}_{\max}\, K_R \frac{1}{(1-\rho)^2}$.

Proof The first statement is a direct application of part 1) in Lemma 4 which holds under A10, A11. To prove the second statement, let us define the error vector as

$$\Delta(\boldsymbol{\eta}) := h(\boldsymbol{\eta}) - \nabla J(\boldsymbol{\eta}) \tag{134}$$

Applying Lemma 4 shows that $\sup_{\boldsymbol{\eta} \in \mathcal{H}} \|\Delta(\boldsymbol{\eta})\|^2 \leq \Gamma^2 (1-\lambda)^2$. We observe that

$$\begin{aligned} \left\langle \nabla J(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle = \left\langle h(\boldsymbol{\eta}) - \Delta(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle &= \|h(\boldsymbol{\eta})\|^2 - \left\langle \Delta(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle \\ &\geq \|h(\boldsymbol{\eta})\|^2 - \frac{1}{2}\big(\|h(\boldsymbol{\eta})\|^2 + \|\Delta(\boldsymbol{\eta})\|^2\big) \, . \end{aligned} \tag{135}$$

This implies

$$\frac{\Gamma^2}{2}(1-\lambda)^2 + \left\langle \nabla J(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle \geq \frac{1}{2} \|h(\boldsymbol{\eta})\|^2 \, . \tag{136}$$

Furthermore, it is straightforward to show that

$$\|\nabla J(\boldsymbol{\eta})\| \leq \|h(\boldsymbol{\eta})\| + \|\Delta(\boldsymbol{\eta})\| \leq \|h(\boldsymbol{\eta})\| + \Gamma(1-\lambda) \, , \tag{137}$$

which concludes the proof. ■

Appendix D. Existence and regularity of the solutions of Poisson equations

Consider the following assumptions:

A12 *For any* $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathbb{R}^d$*, we have* $\sup_{x \in \mathsf{X}} \|P_{\boldsymbol{\eta}}(x, \cdot) - P_{\boldsymbol{\eta}'}(x, \cdot)\|_{\mathrm{TV}} \leq L_P \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$.

A13 *For any* $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathbb{R}^d$*, we have* $\sup_{x \in \mathsf{X}} \|H_{\boldsymbol{\eta}}(x) - H_{\boldsymbol{\eta}'}(x)\| \leq L_H \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$.

A14 *There exists* $\rho < 1$*,* $K_P < \infty$ *such that*

$$\sup_{\boldsymbol{\eta} \in \mathbb{R}^d, x \in \mathsf{X}} \|P_{\boldsymbol{\eta}}^n(x, \cdot) - \pi_{\boldsymbol{\eta}}(\cdot)\|_{\mathrm{TV}} \leq \rho^n K_P, \tag{138}$$

Lemma 6 *Assume A12–14. Then, for any* $\boldsymbol{\eta} \in \mathcal{H}$ *and* $x \in \mathsf{X}$*,*

$$\|\hat{H}_{\boldsymbol{\eta}}(x)\| \leq \frac{\sigma K_P}{1 - \rho}\,, \tag{139}$$

$$\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x)\| \leq \frac{\sigma \rho K_P}{1 - \rho}\,. \tag{140}$$

Moreover, for $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathcal{H}$ *and* $x \in \mathsf{X}$*,*

$$\left\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x)\right\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|\,, \tag{141}$$

where

$$L_{PH}^{(1)} = \frac{K_P^2 \sigma L_P}{(1 - \rho)^2} (2 + K_P) + \frac{K_P}{1 - \rho} L_H\,. \tag{142}$$

Proof Note that, under A14,

$$\begin{aligned} &\sum_{i=0}^{\infty} \left\| P_{\boldsymbol{\eta}}^i (H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})) - \pi_{\boldsymbol{\eta}}\big(H_{\boldsymbol{\eta}}(\cdot) - h(\boldsymbol{\eta})\big) \right\| \\ &\leq \|H_{\boldsymbol{\eta}}(\cdot) - h(\boldsymbol{\eta})\|_{\infty}\, K_P \sum_{i=0}^{\infty} \rho^i \leq \frac{\sigma K_P}{1 - \rho}\,. \end{aligned} \tag{143}$$

Therefore, for all $\boldsymbol{\eta} \in \mathcal{H}$ and $x \in \mathsf{X}$, the series

$$\sum_{i=0}^{\infty} P_{\boldsymbol{\eta}}^i (H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})) - \pi_{\boldsymbol{\eta}}\big(H_{\boldsymbol{\eta}}(\cdot) - h(\boldsymbol{\eta})\big) \tag{144}$$

is uniformly converging and is a solution of the Poisson equation (7). In addition, (139) and (140) follow directly from (143). Under A14, applying a simple modification[2] of (Fort et al., 2011, Lemma 4.2, 1st statement) shows[3] that for any $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathcal{H}$, we have

$$\|\pi_{\boldsymbol{\eta}} - \pi_{\boldsymbol{\eta}'}\|_{\mathrm{TV}} \leq \frac{K_P(1 + K_P)}{1 - \rho} \sup_{x \in \mathsf{X}} \|P_{\boldsymbol{\eta}}(x, \cdot) - P_{\boldsymbol{\eta}'}(x, \cdot)\|_{\mathrm{TV}}\,. \tag{145}$$

[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.

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

$$\begin{aligned}
&\left\| P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x) \right\| \\
&\leq \frac{K_P^2}{(1-\rho)^2} \Big(\sup_{\boldsymbol{\eta} \in \mathcal{H}, x \in \mathsf{X}} \| H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}) \| \Big) \Big(\sup_{x \in \mathsf{X}} \| P_{\boldsymbol{\eta}}(x, \cdot) - P_{\boldsymbol{\eta}'}(x, \cdot) \|_{\mathrm{TV}} \Big) \\
&\quad + \frac{K_P}{1-\rho} \Big(\sup_{\boldsymbol{\eta} \in \mathcal{H}, x \in \mathsf{X}} \| H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}) \| \Big) \| \pi_{\boldsymbol{\eta}} - \pi_{\boldsymbol{\eta}'} \|_{\mathrm{TV}} + \frac{K_P}{1-\rho} \sup_{x \in \mathsf{X}} \| H_{\boldsymbol{\eta}}(x) - H_{\boldsymbol{\eta}'}(x) \| \\
&\leq \left(\frac{K_P^2 \sigma L_P}{(1-\rho)^2} (2 + K_P) + \frac{K_P}{1-\rho} L_H \right) \| \boldsymbol{\eta} - \boldsymbol{\eta}' \| = L_{PH}^{(1)} \| \boldsymbol{\eta} - \boldsymbol{\eta}' \| \ ,
\end{aligned} \tag{146}$$

where the last inequality is due to A12, A13, A7 and (145). ■