Convergence Analysis of the Approximate Newton Method for Markov Decision Processes

Thomas Furmston
Department of Computer Science
University College London
T.Furmston@cs.ucl.ac.uk

Guy Lever
Department of Computer Science
University College London
G.Lever@cs.ucl.ac.uk

Abstract

Recently two approximate Newton methods were proposed for the optimisation of Markov Decision Processes. While these methods were shown to have desirable properties, such as a guarantee that the preconditioner is negative-semidefinite when the policy is log-concave w.r.t. the policy parameters, and were demonstrated to have strong empirical performance in challenging domains, such as the game of Tetris, no convergence analysis was provided. The purpose of this paper is to provide such an analysis. We start by providing a detailed analysis of the Hessian of a Markov Decision Process, which is formed of a negative-semidefinite component, a positive-semidefinite component and a remainder term. The first part of our analysis details how the negative-semidefinite and positive-semidefinite components relate to each other, and how these two terms contribute to the Hessian. The next part of our analysis shows that under certain conditions, relating to the richness of the policy class, the remainder term in the Hessian vanishes in the vicinity of a local optimum. Finally, we bound the behaviour of this remainder term in terms of the mixing time of the Markov chain induced by the policy parameters, where this part of the analysis is applicable over the entire parameter space. Given this analysis of the Hessian we then provide our local convergence analysis of the approximate Newton framework.

1 Markov Decision Processes

Markov Decision Processes (MDPs) are the most commonly used model for the description of sequential decision making processes in a fully observable environment, see e.g. [2]. An MDP is described by the tuple \( \{S,A,H,p_1,p,\pi,R\} \), where \( S \) and \( A \) are sets, known respectively as the state space and action space, \( H \in \mathbb{N} \) is the planning horizon, which can be either finite or infinite, and \( \{p_1,p,\pi,R\} \) are functions that are referred as the initial state distribution, transition dynamics, policy (or controller) and the reward function. In general the state and action spaces can be arbitrary sets, but we restrict our attention to discrete sets. We use boldface notation to represent a vector and also use the notation \( z = (s,a) \) to denote a state-action pair. Given an MDP, the trajectory of the agent is determined by the following recursive procedure: given the agent’s state, \( s_t \), at a given time-point, \( t \in \mathbb{N}_H \), an action is selected according to the policy, \( a_t \sim \pi(\cdot|s_t) \); the agent will then transition to a new state according to the transition dynamics, \( s_{t+1} \sim p(\cdot|a_t,s_t) \); this process is iterated sequentially through all of the time-points in the planning horizon, where the state of the initial time-point is determined by the initial state distribution \( s_1 \sim p_1(\cdot) \). At each time-point the agent receives a scalar reward that is determined by the reward function, which is a function of the current state-action pair. Typically the reward function is assumed to be bounded, but, as the objective is linear in the reward function, we assume w.l.o.g that it is non-negative.

The most widely used objective in the MDP framework is to maximise the total expected reward of the agent over the course of the planning horizon. This objective can take various forms, including
an infinite planning horizon, with either discounted or average rewards, or a finite planning horizon. Theoretical contributions of this paper are applicable to all three frameworks, but for notational ease and for reasons of space we concern ourselves with the infinite horizon framework with discounted rewards. In this framework the boundedness of the objective function is ensured by the introduction of a discount factor, \( \gamma \in [0, 1) \), which scales the rewards of the various time-points in a geometric manner. Writing the objective function and trajectory distribution directly in terms of the parameter vector, \( w \in \mathcal{W} \), then we have

\[
U(w) = \sum_{t=1}^{\infty} \mathbb{E}_{p_t(a; s; w)} \left[ \gamma^{t-1} R(a, s) \right],
\]

where we have denoted the parameter space by \( \mathcal{W} \subset \mathbb{R}^{nw} \) and have used the notation \( p_t(a, s; w) \) to represent the state-action occupancy marginal, \( p(s_t = s, a_t = a; w) \), of the joint state-action trajectory distribution, which, for any \( H \in \mathbb{N} \), is given by

\[
p(a_1:H, s_1:H; w) = \pi(a_H|s_H; w) \left\{ \prod_{t=1}^{H-1} p(s_{t+1}|a_t, s_t)\pi(a_t|s_t; w) \right\} p_1(s_1).
\]

Note that the policy is now written in terms of its parametric representation, \( \pi(a|s; w) \).

Recently two approximate Newton methods were proposed for the optimisation of Markov Decision Processes [5]. These methods have various properties that make them attractive in practice, such as affine (or scale) invariance and a guarantee that the preconditioner will be negative-semidefinite when the policy is log-concave in the policy parameters. While these methods have been demonstrated to have strong empirical performance in challenging domains, such as the game of Tetris, no convergence analysis was provided in [5]. The purpose of this paper is to provide such an analysis, where we begin in sections (3 & 4) by giving a detailed analysis of the structure of the Hessian of a Markov Decision Processes. Firstly, in section (3) we analyse the negative-definite and positive-definite terms of the Hessian, showing how these two terms relate to each other. In section (4) we demonstrate that under certain conditions the remaining term in the Hessian vanishes in the vicinity of a local optimum, where these conditions express the expressive power of the policy class in relation to the difficulty of the MDP. Finally, in section (5) we use our analysis of the Hessian to provide a detailed local convergence analysis of the approximate Newton method.

2 Approximate Newton Method

In gradient-based optimisation of Markov Decision Processes the update of the policy parameters take the form

\[
w^{\text{new}} = w + \alpha M(w) \nabla_w U(w),
\]

where \( \alpha \in \mathbb{R}^+ \) is the step-size parameter, and \( M(w) \) is some preconditioning matrix that possibly depends on \( w \). Provided that \( M(w) \) is positive-definite, and the step-size is sufficiently small, such an update will increase the total expected reward. Iterating this process will lead to a sequence of iterates that converge to a local optimum of (1), provided the step-size sequence is appropriately selected.

There are various choices for \( M(w) \) in (3), including steepest gradient ascent [14], where \( M(w) = I \), and natural gradient ascent [9], where \( M(w) = G^{-1}(w) \), where \( G(w) \) is the Fisher information matrix of the trajectory distribution. All of these choices have their own advantages and disadvantages, but in this paper our focus is exclusively on the full and diagonal approximate Newton methods [5], where, respectively, we have \( M(w) = -\mathcal{H}^{-1}_2(w) \) and \( M(w) = -D^{-1}_2(w) \), where \( \mathcal{H}_2(w) \) is the term from the Hessian given in (2) and \( D_2(w) \) is the diagonal matrix formed from the main diagonal of \( \mathcal{H}_2(w) \). It is convenient for later reference to note the form of the gradient, \( \nabla_w U(w) \), and the Hessian, \( \mathcal{H}(w) = \nabla_w \nabla_w U(w) \). Firstly, the gradient can be written in the following form

\[
\nabla_w U(w) = \sum_{z \in Z} p_z(z; w) Q(z; w) \nabla_w \log \pi(a|s; w).
\]

The term \( p_z(z; w) \) is a geometric weighted average of state-action occupancy marginals given by

\[
p_z(z; w) = \sum_{t=1}^{\infty} \gamma^{t-1} p_t(z; w),
\]

where we have denoted the parameter space by \( \mathcal{W} \subset \mathbb{R}^{nw} \) and have used the notation \( p_t(a, s; w) \) to represent the state-action occupancy marginal, \( p(s_t = s, a_t = a; w) \), of the joint state-action trajectory distribution, which, for any \( H \in \mathbb{N} \), is given by

\[
p(a_1:H, s_1:H; w) = \pi(a_H|s_H; w) \left\{ \prod_{t=1}^{H-1} p(s_{t+1}|a_t, s_t)\pi(a_t|s_t; w) \right\} p_1(s_1).
\]

Note that the policy is now written in terms of its parametric representation, \( \pi(a|s; w) \).

Recently two approximate Newton methods were proposed for the optimisation of Markov Decision Processes [5]. These methods have various properties that make them attractive in practice, such as affine (or scale) invariance and a guarantee that the preconditioner will be negative-semidefinite when the policy is log-concave in the policy parameters. While these methods have been demonstrated to have strong empirical performance in challenging domains, such as the game of Tetris, no convergence analysis was provided in [5]. The purpose of this paper is to provide such an analysis, where we begin in sections (3 & 4) by giving a detailed analysis of the structure of the Hessian of a Markov Decision Processes. Firstly, in section (3) we analyse the negative-definite and positive-definite terms of the Hessian, showing how these two terms relate to each other. In section (4) we demonstrate that under certain conditions the remaining term in the Hessian vanishes in the vicinity of a local optimum, where these conditions express the expressive power of the policy class in relation to the difficulty of the MDP. Finally, in section (5) we use our analysis of the Hessian to provide a detailed local convergence analysis of the approximate Newton method.
while the term $Q(z; w)$ is referred to as the state-action value function and is equal to the total expected future reward from the current time-point onwards, given the current state-action pair, $z$, and parameter vector, $w$, i.e.

$$Q(z; w) = \sum_{t=1}^{\infty} \mathbb{E}_{p_{\pi}(z'; w)} \left[ \gamma^{t-1} R(z') \right] \Big| z_1 = z.$$ 

This is a standard result and due to reasons of space we have omitted the details, but see e.g. [14] or section 7.1 of the supplementary material for more details. Similarly, the Hessian can be shown to take the form

$$\mathcal{H}(w) = \mathcal{H}_{11}(w) + \mathcal{H}_{12}(w) + \mathcal{H}_{12}^T(w),$$

where

$$\mathcal{H}_{12}(w) = \sum_{z \in Z} p_{\pi}(z; w) Q(z; w) \nabla_w \nabla^T_w \log \pi(a|s; w),$$

$$\mathcal{H}_{11}(w) = \sum_{z \in Z} p_{\pi}(z; w) Q(z; w) \nabla_w \log \pi(a|s; w) \nabla^T_w \log \pi(a|s; w),$$

$$\mathcal{H}_{12}(w) = \sum_{z \in Z} p_{\pi}(z; w) \nabla_w \log \pi(a|s; w) \nabla^T_w Q(z; w).$$

It will be convenient in sections of the paper to use the notation $\mathcal{H}_1(w) = \mathcal{H}_{11}(w) + \mathcal{H}_{12}(w)$, so that $\mathcal{H}_1(w)$ represents the portion of the Hessian not included in $\mathcal{H}_2(w)$. We have omitted the details of the derivation, but these can be found in section 7.2 of the supplementary material. Alternative derivations of the Hessian can be found in [1][9]. See also equation (6) of [9] for the Hessian in the average reward framework. Finally, an alternative formulation of the Hessian is provided in [5].

In approximate Newton methods, such as the Gauss-Newton method, it is of interest to understand the behaviour of the approximate Hessian and its relation to the Hessian, both in the vicinity of a local optimum and over the whole parameter space. For instance, it is of interest to understand the conditions under which the approximate Hessian converges to the true Hessian as the iterates converge to a local optimum. This is a desirable quality as it results in a quadratic rate of convergence, see e.g. [12] for more details. For instance, the approximate Hessian used in the Gauss-Newton method converges to the true Hessian in the case where the residuals (of the non-linear least squares problem) tend to zero. Before we proceed to the local convergence analysis of the approximate Newton method we first provide an analysis of the Hessian in sections 3 & 4, where we shall consider the $\mathcal{H}_{11}(w) + \mathcal{H}_2(w)$ term in section 3 and the $\mathcal{H}_{12} + \mathcal{H}_{12}^T$ term in section 4.

## 3 Analysis of the $\mathcal{H}_{11} + \mathcal{H}_2$ term

It can be seen from (6) that $\mathcal{H}_{11}(w)$ is positive-semidefinite, while $\mathcal{H}_2(w)$ is negative-semidefinite under the condition that the policy is log-concave in the policy parameters. In this section we shall consider the term $\mathcal{H}_{11}(w) + \mathcal{H}_2(w)$ that appears in the Hessian and how these two matrices relate to each other.

We start by writing $\mathcal{H}_2(w)$ and $\mathcal{H}_{11}(w)$ in terms of the advantage function, $A(z; w) = Q(z; w) - V(s; w)$, and the value function, $V(s; w) = \sum_{a \in A} \pi(a|s; w) Q(z; w)$. In particular, given that $Q(z; w) = A(z; w) + V(s; w)$ we have that

$$\mathcal{H}_{11}(w) = A_{11}(w) + V_{11}(w), \quad \mathcal{H}_2(w) = A_2(w) + V_2(w),$$

where the matrix $A_{11}(w)$ is defined as follows

$$A_{11}(w) = \sum_{z \in Z} p_{\pi}(z; w) A(z; w) \nabla_w \log \pi(a|s; w) \nabla^T_w \log \pi(a|s; w),$$

and the matrices $V_{11}(w)$, $A_2(w)$ and $V_2(w)$ are defined in a similar manner. As the value function doesn’t depend on the action the matrix, $V_{11}(w)$, can be written in the equivalent form

$$V_{11}(w) = \sum_{s \in S} p_{\pi}(s; w) V(s; w) \sum_{a \in A} \pi(a|s; w) \nabla_w \log \pi(a|s; w) \nabla^T_w \log \pi(a|s; w).$$
where a similar relation holds for $V_2(w)$. Furthermore, as the value function is non-negative the matrix $V_1(w)$ is positive-semidefinite, while the matrix $V_2(w)$ is negative-semidefinite when the policy is log-concave in the policy parameters. As the advantage function can take on negative values the same is not true of either $A_1(w)$ or $A_2(w)$. Finally, when the magnitude of the value function is large in comparison to that of the advantage function it can be expected that the norm of $A_1(w)$ and $A_2(w)$ will be small in comparison to that of $V_1(w)$ and $V_2(w)$ respectively. This is quite a natural condition in practice as the value function is non-negative, while the advantage function (evaluated at any given state) is zero mean with respect to the policy, i.e. $\sum_{a \in A} \pi(a|s; w)A(z; w) = 0$, for all $s \in S$.

It can be seen from (6) that $all$ $s$ at any given state) is zero mean with respect to the policy, where a similar relation holds for (7) we have split this weighting into the contribution from each state, w.r.t. $H$ in terms of (4) we have split this weighting into the contribution from each state, i.e. the value function, and the contribution from the relative advantage of each action in a given state, i.e. the advantage function. In terms of $H_{11}(w) + H_{12}(w)$ we have that

$$H_{11}(w) + H_{12}(w) = A_{11}(w) + V_1(w) + A_2(w) + V_2(w).$$

(9)

The purpose of writing $H_{11}(w) + H_{12}(w)$ in this form is that the contributions from the value function in these two terms cancel. In other words $V_1(w) = -V_2(w)$ and we have

$$H_{11}(w) + H_{12}(w) = A_2(w) + A_{11}(w).$$

(10)

Due to reasons of space we give a proof of the relation $V_1(w) = -V_2(w)$ in section[7.3] of the supplementary material.

There is a further simplification that often occurs in practice and should be noted. In particular, when the Hessian of the log-policy does not depend on the action, i.e. the curvature of the log-policy is constant for each state when viewed as a function of the parameter vector, then we have that $A_2(w) = 0$. Again, due to reasons of space we give the proof of this relation in section[7.3] of the supplementary material. In terms of (10) this condition gives

$$H_{11}(w) + H_{12}(w) = A_{11}(w).$$

(11)

A commonly used class of policies where this property holds is $\pi(a|s; w) \propto \exp(w^\top \phi(a, s))$, where $\phi(a, s)$ is some vector of features that depends on the state-action pair, $(a, s) \in A \times S$. Under this class of policy we have $\nabla_w \nabla_w^\top \log \pi(a|s; w) = -\text{Cov}_{a' \sim \pi_\phi(s; w)}(\phi(a', s), \phi(a', s))$, which doesn’t depend on $a$. To ease the notation we shall assume that $H_{11}(w) + H_{12}(w)$ has the form (11) for the remainder of the paper, where similar results hold when this property do not hold. Also note that the relation, $A_2(w) = 0$, shows that it is no longer necessary that the policy is log-concave in the policy parameters to ensure $H_2(w)$ is negative-semidefinite. Indeed, in this case we have $H_2(w) = V_2(w) = -V_1(w)$.

4 Analysis of the $H_{12} + H_{12}^\top$ term

Up until now we have focused our attention on the terms $H_2$, $H_{11}$ and $A_{11}$. In this section we shall analyse on the term, $H_{12} + H_{12}^\top$. Since $H_{12} + H_{12}^\top$ is omitted from the approximate Newton preconditioner, and cannot be guaranteed to be either positive or negative semi-definite, it is useful to understand the effect of omitting this term. In particular, we show that this term will be negligible in certain reasonable situations: either in the vicinity of a local optimum, or over the entire parameter space.

4.1 Analysis in Vicinity of a Local Optimum

In this section we shall consider the conditions under which the term $H_{12} + H_{12}^\top$ vanishes at a local optimum. We start by noting that

$$H_{12}(w) = \sum_{z \in Z} p_z(z; w)\nabla_w \log \pi(a|s; w)\nabla_w^\top \left( R(z) + \gamma \sum_{s'} p(s'|a, s)V(s'; w) \right)$$

$$= \gamma \sum_{z \in Z} p_z(z; w)\nabla_w \log \pi(a|s; w) \sum_{s'} p(s'|a, s)\nabla_w^\top V(s'; w)$$
so that \( \nabla^\top_w V(s'; w) = 0 \) for all \( s' \in \mathcal{S} \) then \( H_{12}(w) + H_{12}'(w) = 0 \). It is sufficient therefore to require that \( \nabla^\top_{w^*} V(s; w) = 0 \), for all \( s \in \mathcal{S} \), at a local optimum \( w^* \in \mathcal{W} \). We therefore consider the situations in which this occurs. We start by introducing the notion of a non-decreasing policy class.

**Definition 1.** A policy parameterisation is said to be non-decreasing w.r.t. a Markov Decision Process if whenever there exists a search direction, \( \eta \in \mathcal{W} \), and step-size, \( \alpha \in \mathbb{R}^+ \), such that

\[
V(s; w + \alpha \eta) \geq V(s; w),
\]

for some state, \( s \in \mathcal{S} \), and some parameter vector, \( w \in \mathcal{W} \), then there exists a projection\(^1\) of \( \eta \), denoted by \( \eta_P = \mathcal{P}(\eta) \), such that

\[
V(s; w + \alpha \eta_P) = V(s; w),
\]

for all \( s \in \mathcal{S} \), where the inequality in (13) is strict for at least one state, \( \bar{s} \in \mathcal{S} \). Furthermore, for any state, \( s \in \mathcal{S} \), such that (13) forms an equality then we have that

\[
\pi(a|s; w + \alpha \eta_P) = \pi(a|s; w), \quad \forall a \in \mathcal{A}.
\]

We now show that the property of a non-decreasing policy parameterisation is sufficient to ensure that \( \nabla_{w^*} V(s; w^*) = 0 \), for all \( s \in \mathcal{S} \), at a local optimum \( w^* \in \mathcal{W} \). This is summarised in the following lemma.

**Lemma 1.** Suppose that \( w^* \in \mathcal{W} \) is a local optimum of the differentiable objective function, \( U(w) = \mathbb{E}_{\pi(s)}[V(s; w)] \). Suppose that there is an open neighbourhood of \( w^* \), denoted by \( \mathcal{U} \), such that for all \( w \in \mathcal{U} \) the Markov chain induced by \( w \) is ergodic. Suppose that the policy parameterisation is non-decreasing w.r.t. the given Markov Decision Process. Then \( w^* \) is a stationary point of \( V(s; w) \) for all \( s \in \mathcal{S} \).

**Proof.** See section (7.4) of the supplementary material \( \square \)

We now show that tabular policies are non-decreasing, regardless of the given Markov Decision Process. By tabular policies we mean that for each state, \( s \in \mathcal{S} \), the conditional distribution, \( \pi(a|s; w_s) \), is parameterised by a separate parameter vector, \( w_s \in \mathbb{R}^{n_s} \).

**Lemma 2.** Suppose that a given a Markov Decision Process has a tabular policy parameterisation, then the policy parameterisation is non-decreasing.

**Proof.** See section (7.4) of the supplementary material \( \square \)

An immediate corollary is that for tabular policies inducing ergodic Markov chains, local optima of the objective function are also stationary points of the value functions \( V(s; w) \) for all states. This is summarised in the following theorem.

**Theorem 1.** Let \( w^* \in \mathcal{W} \) be a local optimum, in a tabular policy parameterization, of the differentiable objective \( U(w) = \mathbb{E}_{\pi(s)}[V(s; w)] \). Suppose that there is an open neighbourhood of \( w^* \), denoted by \( \mathcal{U} \), such that for all \( w \in \mathcal{U} \) the Markov chain induced by \( w \) is ergodic, and that \( V(s; w) \) is differentiable for all states \( s \). Then \( H_{12}(w^*) = H_{12}'(w^*) = 0 \).

A final point to note is that when we have the additional condition that the gradient of the value function is continuous in \( w \) (at \( w = w^* \)) then \( H_{12}(w) + H_{12}'(w) \to 0 \) as \( w \to w^* \). This condition will be satisfied if, for example, the policy is continuously differentiable w.r.t. the policy parameters.

### 4.1.1 Examples

We now consider two simple maze navigation MDPs to illustrate this theory, where these MDPs are displayed in Figure (1). Walls of the maze are solid lines, while the dotted lines indicate state boundaries and are passable. The agent begins in state marked ‘S’ and receives a reward at the goal ‘G’, and is then reset to the start state. With four possible actions (up, down, left, right), the optimal policy is to move, with probability one, in the direction of the arrows. We consider the

\(^1\)This is meant informally, \( \eta_P \) is not necessarily a projection in a formal sense.
Due to the relation, case, using the notation \( \hat{\gamma} \). Indeed, consider the extreme case where \( p \) given by \( z \) next state distribution of the transition dynamics conditioned on the current state-action pair being \( p \) state distribution is now given by \( p \). where \( \gamma \) of the terms in \( \mathcal{H}_{12} \rangle \) of the supplementary material. In (14) we have used the following notation \( \mathcal{H} \). In this section we perform some further analyses of \( \mathcal{H} \). In accordance with the theory the term \( \mathcal{H}_{12} + \mathcal{H}_{12}^\top \) vanishes at the optimal policy. (b) Under the feature representation states 4, 5 and 6 receive the same feature, but now the optimal policy differs among these states. The policy class is not rich enough for this MDP and the term \( \mathcal{H}_{12} + \mathcal{H}_{12}^\top \) does not vanish at the optimal policy.

4.2 Analysis over the Entire Parameter Space

In this section we perform some further analyses of \( \mathcal{H}_{12} + \mathcal{H}_{12}^\top \), both in terms of the discount factor and the mixing time of the Markov chain induced by the policy parameters. Unlike section 4.1, the results of this section are applicable over the entire parameter space.

To aid the analyses it is helpful to rewrite \( \mathcal{H}_{12} \) into the following form

\[
\mathcal{H}_{12}(\mathbf{w}) = \gamma \sum_{z \in \mathcal{Z}} p_z(z; \mathbf{w}) \nabla_w \log \pi(a|s; \mathbf{w}) \sum_{z' \in \mathcal{Z}} p_{\gamma, z}(z'; \mathbf{w}) \nabla_w \log \pi(a'|s'; \mathbf{w}) Q(z'; \mathbf{w}),
\]

Due to reasons of space we have omitted the details of these derivations, but these can be found in section 7.2 of the supplementary material. In (14) we have used the following notation

\[
p_{\gamma, z}(z'; \mathbf{w}) = \sum_{i=1}^{\infty} \gamma^{i-1} p_{i, z}(z'; \mathbf{w}),
\]

where \( p_{i, z}(z'; \mathbf{w}) \) is the state-action occupancy marginal of the \( i \)-th time-point, given that the initial state distribution is now given by \( p(\cdot|z) \). In other words, the initial state distribution is given by the next state distribution of the transition dynamics conditioned on the current state-action pair being given by \( z \). Note that the difference between [5] & [15] is that the initial state distribution of [5] is given by the initial state distribution of the original Markov Decision Process, while the initial state distribution of [15] is given by the next state distribution of the transition dynamics.

The analysis of this section is based on two observations. The first observation is that the magnitude of the terms in \( \mathcal{H}_{12}(\mathbf{w}) \) is related to how much the term \( p_{\gamma, z}(\cdot; \mathbf{w}) \) depends upon \( z \), for each \( z \in \mathcal{Z} \). Indeed, consider the extreme case where \( p_{\gamma, z}(\cdot; \mathbf{w}) \) is independent of \( z \), for each \( z \in \mathcal{Z} \). In this case, using the notation \( \bar{p}_{\gamma}(\cdot; \mathbf{w}) \equiv p_{\gamma, z}(\cdot; \mathbf{w}) \), for each \( z \in \mathcal{Z} \), we have

\[
\mathcal{H}_{12}(\mathbf{w}) = \mathbb{E}_{\bar{p}_{\gamma}(z; \mathbf{w})} \left[ \nabla_w \log \pi(a|s; \mathbf{w}) \right] \mathbb{E}_{\bar{p}_{\gamma}(z'; \mathbf{w})} \left[ \nabla_w \log \pi(a'|s'; \mathbf{w}) Q(z'; \mathbf{w}) \right].
\]

Due to the relation, \( \mathbb{E}_{\bar{p}_{\gamma}(z; \mathbf{w})} \left[ \nabla_w \log \pi(a|s; \mathbf{w}) \right] = 0 \), it follows that in this case we have \( \mathcal{H}_{12}(\mathbf{w}) = 0 \). This is an extreme example, but one can expect that the magnitude of \( \mathcal{H}_{12}(\mathbf{w}) \) will be small when
the dependence of $p_{\gamma,z}(\cdot;w)$ upon $z$ is small, for each $z \in Z$. The second observation is that, for each $z \in Z$, the dependence of $p_{\gamma,z}(\cdot;w)$ upon $z$ is directly related to both the mixing time of the Markov chain, induced by $w$, and the discount factor. We now formalise these observations to obtain a bound on the terms of $\mathcal{H}_{12}(w)$ in terms of the discount factor and the mixing time.

**Lemma 3.** Given $w \in W$, suppose that the Markov chain induced by $w$ is ergodic. Denote the second largest eigenvalue (in terms of absolute value) of the state-action transition matrix by $\lambda_2$. Under these conditions and given some matrix norm, $\| \cdot \|$, the matrix $\mathcal{H}_{12}(w)$ satisfies the bound

$$
\|\mathcal{H}_{12}(w)\| \leq \frac{\eta |\lambda_2|}{(1 - \gamma)^2(1 - \gamma|\lambda_2|)},
$$

for some positive constant, $\eta > 0$. In particular, we have that $\|\mathcal{H}_{12}(w)\| \to 0$ as either $\gamma \to 0$ or $|\lambda_2| \to 0$.

## 5 Convergence Analysis

In this section we perform our local convergence analysis of the approximate Newton framework. The first contribution of this section is to provide the local convergence analysis of the EM-algorithm, where we shall relate the convergence behaviour of the EM-algorithm as properties of this algorithm. We shall then consider the local convergence analysis of the approximate Newton method with a fixed step-size of one. See theorem 1 of [5] for more details.

It was shown in [5] that the EM-algorithm is the same, up to first order, as applying the full approximate Newton method. See section(10.1.6) of [13] for more details.

A formal proof of Ostrowski’s theorem can be found in [13]. Some additional results that we require are that root super-linear convergence is obtained when $\rho(\nabla G(w^*)) = 0$, while quotient super-linear convergence is obtained when $\nabla G(w^*) = 0$. See section(10.1.6) of [13] for more details.

It was shown in [5] that the EM-algorithm is the same, up to first order, as applying the full approximate Newton method with a fixed step-size of one. See theorem 1 of [5] for more details. For this reason we first categorise the convergence behaviour of the EM-algorithm when applied to Markov Decision Processes. This is, to our knowledge, the first formal derivation of the convergence properties for this application of the EM-algorithm.

**Lemma 4.** Suppose that the sequence, $\{w_k\}_{k \in \mathbb{N}}$, is generated by an application of the EM-algorithm, where the sequence converges to $w^*$. Denoting the update operation of the EM-algorithm by $G_{\text{EM}}$, so that $w_{k+1} = G_{\text{EM}}(w_k)$, then

$$
\nabla G_{\text{EM}}(w^*) = -\mathcal{H}_2(w^*)^{-1}\mathcal{H}_1(w^*). \tag{16}
$$

When the policy parameterisation is non-decreasing w.r.t. the given Markov Decision Process this simplifies to $\nabla G_{\text{EM}}(w^*) = I - \mathcal{H}_2(w^*)^{-1}\mathcal{H}_1(w^*)$. When the Hessian, $\mathcal{H}(w^*)$, is negative-definite then $\rho(\nabla G_{\text{EM}}(w^*)) < 1$ and $w^*$ is a local point of an attraction for the EM-algorithm. When $\mathcal{H}(w^*)$ is negative-semidefinite and $\mathcal{H}_2(w^*)$ is negative-definite then there is a ridge in the objective function going through $w^*$.

**Proof.** See section[7,5] of the supplementary material. □
We now consider the rate of convergence of the EM-algorithm, where the rate of convergence of this algorithm is given by \( \varrho = \rho(\nabla G_{EM}(w^*)) \). As was noted in the proof of lemma 4 the rate of convergence is given by the maximal value of the relative Rayleigh quotient, i.e.

\[
\varrho = \max_{x \in \mathbb{R}^n} R(x) = \max_{x \in \mathbb{R}^n} \frac{x^\top H_1(w^*)x}{x^\top H_2(w^*)x} = 1 - \min_{x \in \mathbb{R}^n} \frac{x^\top H(w^*)x}{x^\top H_2(w^*)x} = 1 - \lambda_{em}, \tag{17}
\]

where \( \lambda_{em} \) is the minimum of the ratio of quadratic forms, \( x^\top H(w^*)x / x^\top H_2(w^*)x \). In the case where the policy parameterisation is non-decreasing w.r.t. the given MDP then \( H(w^*) = A_{11}(w^*) \), so that \( \lambda_{em} \) is the minimum of the ratio of quadratic forms, \( x^\top A_{11}(w^*)x / x^\top H_2(w^*)x \). In either case \( \lambda_{em} \in (0, 1] \). It can be seen through (17) that the rate of convergence of the EM-algorithm can be related to the quality with which \( H_2(w^*) \) approximates \( H(w^*) \). For instance, if \( H_2(w^*) \approx H(w^*) \) then \( \lambda_{em} \) will be close to one and \( \varrho \) will be close to zero, resulting in fast convergence.

By contrast, if \( H_2(w^*) \) is a poor approximation to \( H(w^*) \), then the contribution from \( H_1(w^*) \) will be more significant. In this case \( \lambda_{em} \) will be closer to zero, \( \varrho \) will be larger and so the rate of convergence will be slower.

Having considered the local convergence analysis of the EM-algorithm we now consider the local convergence analysis of the approximate Newton framework. We shall focus on the full approximate Newton method, where the analysis of the diagonal approximate Newton method follows similarly.

We denote the parameter update function of the full approximate Newton method by \( G_{FAN} \), where we have \( G_{FAN}(w) = w - \alpha H_2^{-1}(w)\nabla U(w) \), where \( \alpha \in \mathbb{R}^+ \) is the step-size. The mapping \( G_{FAN}(w) \) depends on the step-size, but for notational reasons we omit this from the notation. We summarise our local convergence results for the full approximate Newton method in the following lemma.

**Lemma 5.** Suppose that the sequence, \( \{w_k\}_{k \in \mathbb{N}} \), is generated by an application of the full approximate Newton method using a fixed step-size of \( \alpha \in \mathbb{R}^+ \), where the sequence converges to \( w^* \), then

\[
\nabla G_{FAN}(w^*) = I - \alpha H_2^{-1}(w^*)H(w^*) = (1 - \alpha)I - \alpha H_2^{-1}(w^*)A_{11}(w^*), \tag{18}
\]

When the policy parameterisation is non-decreasing w.r.t. the given Markov Decision Process this simplifies to \( \nabla G_{FAN}(w^*) = I - \alpha H_2^{-1}(w^*)A_{11}(w^*) \). When a fixed step-size of one is used in the full approximate Newton method then \( \nabla G_{FAN}(w^*) = \nabla G_{EM}(w^*) \), so that the rate of convergence is the same as the rate of convergence of the EM-algorithm. The optimal step-size of the full approximate Newton method is given by \( \alpha = 2(2 - (\lambda_{min} + \lambda_{max}))^{-1} \), where \( \lambda_{min} \) and \( \lambda_{max} \) are respectively the minimal and maximal eigenvalues of \( \nabla G_{EM}(w^*) \), and the point \( w^* \) remains a point of attraction of the full approximate Newton method when using this step-size.

**Proof.** See section 7.5 of the supplementary material.

The optimal step-size in lemma 5 will not be known in practice and instead some type of line search [12] or adaptive step-size procedure [6] should be considered.

## 6 Summary

In this paper we have provided a local convergence analysis of the approximate Newton framework for Markov Decision Processes, while also providing a detailed analysis of the Hessian. Our analysis suggests that while \( H_2 \nrightarrow H \) as \( w \rightarrow w^* \), so that in general neither super-linear nor quadratic convergence will be obtained by the approximate Newton framework, the Hessian can have a particularly simple form in the vicinity of a local optimum. In particular, when the policy parameterisation is non-decreasing w.r.t. the given MDP then we have that \( H \rightarrow A_{11} \) as \( w \rightarrow w^* \). This is an interesting point because it means that a quotient super-linear, or possibly quadratic, rate of convergence could be obtained without needing to calculate all the terms in the Hessian. Unfortunately, the matrix \( A_{11} \) is not guaranteed to be negative-definite and so using this matrix as a preconditioner may not be beneficial when not in the vicinity of a local optimum.
References

[1] J. Baxter and P. Bartlett. Infinite Horizon Policy Gradient Estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, second edition, 2000.

[3] G. Casella and R. Berger. *Statistical Inference*. Dunbury, 2001.

[4] P. Dayan and G. E. Hinton. Using Expectation-Maximization for Reinforcement Learning. *Neural Computation*, 9:271–278, 1997.

[5] T. Furmston and D. Barber. A Unifying Perspective of Parametric Policy Search Methods for Markov Decision Processes. *NIPS*, 22:1523–1530, 2012.

[6] A. George and W. Powell. Adaptive Step-Sizes for Recursive Estimation with Applications in Approximate Dynamic Programming. *Machine Learning*, 65:167–198, 2006.

[7] M. R. Hestenes. *Optimisation Theory : The Finite Dimensional Case*. John Wiley & Sons, first edition, 1975.

[8] S. Kakade. Optimizing Average Reward Using Discounted Rewards. *COLT*, 14:605–615, 2001.

[9] S. Kakade. A Natural Policy Gradient. *NIPS*, 14:1531–1538, 2002.

[10] J. Kober and J. Peters. Policy Search for Motor Primitives in Robotics. *Machine Learning*, 84(1-2):171–203, 2011.

[11] K. Lange. A Gradient Algorithm that is Locally Equivalent to the EM-algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):425–437, 1995.

[12] J. Nocedal and S. Wright. *Numerical Optimisation*. Springer, 2006.

[13] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, first edition, 1970.

[14] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. *NIPS*, 13:1057–1063, 2000.

[15] M. Toussaint, A. Storkey, and S. Harmeling. *Bayesian Time Series Models*, chapter Expectation-Maximization Methods for Solving (PO)MDPs and Optimal Control Problems. Cambridge University Press, 2011. In press. See userpage.fu-berlin.de/~mtoussai.
7 Supplementary Material

7.1 Gradient Derivation

For ease of reference in this section we give a brief outline of the derivation for the derivative of (1). As in the rest of the paper we focus on the case of an infinite planning horizon with discounted rewards, where other frameworks follow similarly. To this end it is helpful to introduce the following notation

\[ p_\gamma(z; w) = \sum_{t=1}^{\infty} \gamma^{t-1} p_t(z; w), \]

Note that in terms of (19) the objective function takes the form,

\[ U(w) = \sum_{z \in Z} p_\gamma(z; w) R(z), \]

which means that the gradient can be written as follows

\[ \nabla_w U(w) = \sum_{z \in Z} \left( \nabla_w p_\gamma(z; w) R(z) + p_\gamma(z; w) \nabla_w R(z) \right). \]

Due to the fact that the reward function is independent of the policy parameters this simplifies further to

\[ \nabla_w U(w) = \sum_{z \in Z} \nabla_w p_\gamma(z; w) R(z). \]

Hence, to calculate the gradient of the MDP objective it is sufficient to find the derivative of (19), which we now detail. The first point to note is that, for each \( z' \in Z \), the term \( p_\gamma(z'; w) \) can be written in the form

\[ p_\gamma(z'; w) = \gamma \sum_{z \in Z} P(z'|z; w) p_\gamma(z; w) + p_1(z'; w). \]

Taking the derivative of \( p_\gamma(z'; w) \) gives

\[ \nabla_w p_\gamma(z'; w) = \gamma \sum_{z \in Z} \left( \nabla_w P(z'|z; w) p_\gamma(z; w) + \gamma P(z'|z; w) \nabla_w p_\gamma(z; w) \right) + \nabla_w p_1(z'; w). \]

Assuming that the components of \( \nabla_w \log \pi(a|s; w) \) are uniformly bounded, for each \( a, s \in A \times S \), then, using the log-trick, we have

\[ \nabla_w P(z'|z; w) = P(z'|z; w) \nabla_w \log P(z'|z; w), \]

it can be shown that

\[ \nabla_w p_\gamma(z'; w) = \gamma \sum_{z \in Z} P(z'|z; w) \nabla_w p_\gamma(z; w) + p_\gamma(z'; w) \nabla_w \log \pi(a'|s'; w). \]

Given the current parameter vector, \( w \in \mathcal{W} \), the reward function can be written through the Bellman equation in the following manner

\[ R(z) = Q(z; w) - \gamma \sum_{z' \in Z} P(z'|z; w) Q(z'; w). \]

Using this form of the reward function the gradient can be written in the form

\[ \nabla_w U(w) = \sum_{z \in Z} \nabla_w p_\gamma(z; w) \left( Q(z; w) - \gamma \sum_{z' \in Z} P(z'|z; w) Q(z'; w) \right). \]

Using (21) we have that

\[ \gamma \sum_{z \in Z} \nabla_w p_\gamma(z; w) \sum_{z' \in Z} P(z'|z; w) Q(z'; w) = \sum_{z' \in Z} \left( \nabla_w p_\gamma(z'; w) \right. \]

\[ \left. - p_\gamma(z'; w) \nabla_w \log \pi(a'|s'; w) \right) Q(z'; w), \]

so that the gradient takes the form

\[ \nabla_w U(w) = \sum_{z \in Z} p_\gamma(z; w) \nabla_w \log \pi(z; w) Q(z; w), \]

which completes the derivation.
7.2 Hessian Derivation

For ease of reference in this section we give a brief outline of the derivation for the Hessian of (1). As in the rest of the paper we focus on the case of an infinite planning horizon with discounted rewards, where other frameworks follow similarly. Using the same technique as in section (7.1) the Hessian can be shown to take the form

$$\nabla_w \nabla^T_w U(w) = \sum_{z \in Z} \nabla_w \nabla^T_w p_\gamma(z; w) R(z).$$

Hence, to calculate the Hessian of (1) it is necessary to calculate the Hessian of $p_\gamma(z'; w)$, for each $z' \in Z$, which we now detail. Using (21) we have

$$\nabla_w \nabla^T_w p_\gamma(z'; w) = \gamma \sum_{z \in Z} P(z'|z; w) \nabla_w \nabla^T_w p_\gamma(z; w) + \gamma \sum_{z \in Z} \nabla_w P(z'|z; w) \nabla_w^T p_\gamma(z; w)$$

$$+ p_\gamma(z'; w) \nabla_w \nabla^T_w \log \pi(a'|s'; w) + \nabla_w p_\gamma(z'; w) \nabla^T_w \log \pi(a'|s'; w).$$

Using the relation (22) on the last term in $\nabla_w \nabla^T_w p_\gamma(z'; w)$ gives

$$\nabla_w \nabla^T_w p_\gamma(z'; w) = \gamma \sum_{z \in Z} P(z'|z; w) \nabla_w \nabla^T_w p_\gamma(z; w)$$

$$+ \gamma \sum_{z \in Z} \left( \nabla_w P(z'|z; w) \nabla^T_w p_\gamma(z; w) + \nabla_w p_\gamma(z; w) \nabla^T_w P(z'|z; w) \right)$$

$$+ p_\gamma(z'; w) \left( \nabla_w \nabla^T_w \log \pi(a'|s'; w) + \nabla_w \log \pi(a'|s'; w) \nabla^T_w \log \pi(a'|s'; w) \right).$$

Again using the relation (22) we have that the Hessian of (1) takes the form

$$\nabla_w \nabla^T_w U(w) = \sum_{z \in Z} \nabla_w \nabla^T_w p_\gamma(z; w) \left( Q(z; w) - \gamma \sum_{z' \in Z} P(z'|z; w) Q(z'; w) \right).$$

Using the form for the Hessian of $p_\gamma(z; w)$ we have that

$$\gamma \sum_{z \in Z} \sum_{z' \in Z} \nabla_w \nabla^T_w p_\gamma(z; w) P(z'|z; w) Q(z'; w) = \sum_{z \in Z} \nabla_w \nabla^T_w p_\gamma(z; w) Q(z'; w)$$

$$- \sum_{z' \in Z} p_\gamma(z'; w) \left( \nabla_w \nabla^T_w \log \pi(a'|s'; w) + \nabla_w \log \pi(a'|s'; w) \nabla^T_w \log \pi(a'|s'; w) \right) Q(z'; w)$$

$$- \gamma \sum_{z \in Z} \sum_{z' \in Z} \left( \nabla_w P(z'|z; w) \nabla^T_w p_\gamma(z; w) + \nabla_w p_\gamma(z; w) \nabla^T_w P(z'|z; w) \right) Q(z'; w),$$

so the Hessian of (1) simplifies to

$$\nabla_w \nabla^T_w U(w) = H_{11}(w) + H_{12}(w) + H_{11}^T(w),$$

where

$$H_{11}(w) = \sum_{z \in Z} p_\gamma(z; w) \nabla_w \log \pi(a|s; w) \nabla^T_w \log \pi(a|s; w) Q(z; w),$$

$$H_{12}(w) = \sum_{z \in Z} p_\gamma(z; w) \nabla_w \nabla^T_w \log \pi(a|s; w) Q(z; w),$$

$$H_{11}^T(w) = \sum_{z \in Z} \sum_{z' \in Z} \nabla_w p_\gamma(z; w) \nabla^T_w P(z'|z; w) Q(z'; w).$$

The term $H_{12}(w)$ involves terms of the form $\nabla_w p_\gamma(z; w), z \in Z$. In order to remove these terms we use the fact that the derivative of the state-action value function can be written in the form

$$\nabla_w Q(z; w) = \gamma \sum_{z' \in Z} \left( \nabla_w P(z'|z; w) Q(z'; w) + P(z'|z; w) \nabla_w Q(z'; w) \right).$$
Using (23) the matrix \( H_{12}(w) \) can be written in the equivalent form
\[
H_{12}(w) = \sum_{z \in Z} \nabla_w p_{\gamma}(z; w) \left( \nabla^T_w Q(z; w) - \gamma \sum_{z' \in Z} P(z'|z; w) \nabla^T_w Q(z'; w) \right).
\]

Using the fact that
\[
\gamma \sum_{z \in Z} \sum_{z' \in Z} \nabla_w p_{\gamma}(z; w) P(z'|z; w) Q(z'; w) = \sum_{z \in Z} \left( \nabla_w p_{\gamma}(z; w) - p_{\gamma}(z; w) \nabla_w \log \pi(a|s; w) \right) Q(z; w),
\]
gives
\[
H_{12}(w) = \sum_{z \in Z} p_{\gamma}(z; w) \nabla_w \log \pi(a|s; w) \nabla^T_w Q(z; w).
\]

We now derive the form of the term \( H_{12}(w) \) given in (14). This is done by rewriting the gradient of the state-action value function in an equivalent form. In particular, given that the state-action value function takes the form
\[
Q(z; w) = R(z) + \gamma \sum_{z' \in Z} p(s'|z) \pi(a'|s'; w) Q(z'; w)
\]
and
\[
Q(z; w) = \sum_{t=1}^{\infty} E_{p_{\gamma}(z'; w)} \left[ \gamma^{t-1} R(z') \right] \big| z_1 = z,
\]
then we have
\[
Q(z; w) = R(z) + \gamma U_z(w),
\]
where \( U_z(w) \) is the objective function of the Markov Decision Process that is identical to the original Markov decision process, but where the initial state distribution given by \( p(\cdot|z) \). This means that the gradient of the state-action value function takes the form
\[
\nabla_w Q(z; w) = \gamma \nabla_w U_z(w),
\]
\[
= \gamma \sum_{z' \in Z} p_{\gamma}(z'|w) \nabla^T_w \log \pi(a'|s'; w) Q(z'; w),
\]
where the second line follow through an application of the policy gradient theorem. The form of \( H_{12}(w) \) given in (14) now follows.

### 7.3 Analysis of the \( H_{11} + H_{2} \) term

In this section we prove the two relations that were stated without proof in section (3), namely that \( V_{11}(w) = -V_2(w) \) and \( A_2(w) = 0 \).

We first show the relation, \( V_{11}(w) = -V_2(w) \). Assuming the policy satisfies the Fisher regularity conditions, see e.g. [3], we have
\[
\sum_{a \in A} \pi(a|s; w) \nabla_w \log \pi(a|s; w) \nabla^T_w \log \pi(a|s; w) = -\sum_{a \in A} \pi(a|s; w) \nabla_w \nabla^T_w \log \pi(a|s; w),
\]
for each \( s \in S \) and \( w \in W \). This means that the matrix \( V_{11}(w) \) can be written in the form
\[
V_{11}(w) = \sum_{s \in S} p_{\gamma}(s; w) V(s; w) \sum_{a \in A} \pi(a|s; w) \nabla_w \log \pi(a|s; w) \nabla^T_w \log \pi(a|s; w),
\]
\[
= -\sum_{s \in S} p_{\gamma}(s; w) V(s; w) \sum_{a \in A} \pi(a|s; w) \nabla_w \nabla^T_w \log \pi(a|s; w) = -V_2(w).
\]

This shows that \( V_{11}(w) = -V_2(w) \).

To show the second relation, \( A_2(w) = 0 \), we assume that the Hessian of the log-policy is independent of the action. In particular, denoting the Hessian of the log-policy by \( \nabla_w \nabla^T_w \log \pi(s; w) \), \( s \in S \), we have
\[
A_2(w) = \sum_{z \in Z} p_{\gamma}(z; w) A(z; w) \nabla_w \nabla^T_w \log \pi(s; w),
\]
\[
= \sum_{s \in S} p_{\gamma}(s; w) \nabla_w \nabla^T_w \log \pi(s; w) \sum_{a \in A} \pi(a|s; w) A(z; w).
\]
The relation \( A_2(w) = 0 \) now follows because \( \sum_{a \in A} \pi(a|s; w) A(z; w) = 0 \), for all \( s \in S \).
Lemma 7.4 Analysis in Vicinity of a Local Optimum

Suppose that $w^* \in \mathcal{W}$ is a local optimum of the differentiable objective function, $U(w) = \mathbb{E}_{p_1(s)} [V(s;w)]$. Suppose that there is an open neighbourhood of $w^*$, denoted by $\mathcal{U}$, such that for all $w \in \mathcal{U}$ the Markov chain induced by $w$ is ergodic. Suppose that the policy parameterisation is non-decreasing w.r.t. the given Markov Decision Process. Then $w^*$ is a stationary point of $V(s;w)$ for all $s \in \mathcal{S}$.

\textit{Proof.} In order to obtain a contradiction suppose that $w^*$ is not a stationary point of $V(s;w)$, for each $s \in \mathcal{S}$. This means that there exists a state, $\bar{s} \in \mathcal{S}$, and a vector in the parameter space, $\eta \in \mathcal{W}$, such that $\eta$ is a strict ascent direction of $V(\bar{s};w)$ at $w^*$. As the policy parameterisation is non-decreasing there exists a projection mapping, $P$, and a sufficiently small step-size, $\alpha \in \mathbb{R}^+$, such that $w^* + \alpha \eta \in \mathcal{U}$ and

$$V(s;w^* + \alpha \eta) \geq V(s;w^*), \tag{24}$$

for all $s \in \mathcal{S}$, where there exists $\bar{s} \in \mathcal{S}$ such that $V(\bar{s};w^* + \alpha \eta) > V(\bar{s};w^*)$.

In order to obtain a contradiction we will show that there is no $s \in \mathcal{S}$ such that (24) holds with an equality. Given this property a contradiction is obtained because it follows that

$$U(w^* + \alpha \eta) = \mathbb{E}_{p_1(s)} [V(s;w^* + \alpha \eta)],$$

$$> \mathbb{E}_{p_1(s)} [V(s;w^*)] = U(w^*),$$

and

$$S_\geq = \{s \in \mathcal{S} | V(s;w^* + \alpha \eta) = V(s;w^*)\},$$

$$S_\leq = \{s \in \mathcal{S} | V(s;w^* + \alpha \eta) > V(s;w^*)\},$$

we wish to show that $S_\leq = \emptyset$. In particular, for a contradiction, suppose that $S_\leq \neq \emptyset$. This means, given the ergodicity of the Markov chain induced by $w^*$ and the fact that $S_\geq \neq \emptyset$, that there exists $s \in S_\leq$ and $s' \in S_\geq$ such that

$$p(s'|s;w^*) = \sum_{a \in \mathcal{A}} p(s'|s,a) \pi(a|s;w^*) > 0.$$

As $s \in S_\leq$ we have that

$$p(s'|s;w^*) = p(s'|s;w^* + \alpha \eta).$$

We now consider the form $V(s;w^*)$. In particular, we have

$$V(s;w^*) = \sum_{a \in \mathcal{A}} \pi(a|s;w^*) R(a,s) + \gamma \sum_{s_{\text{next}} \in \mathcal{S}} p(s_{\text{next}}|s;w^*) V(s_{\text{next}};w^*),$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s;w^* + \alpha \eta) R(a,s) + \gamma \sum_{s_{\text{next}} \in \mathcal{S}} p(s_{\text{next}}|s;w^* + \alpha \eta) V(s_{\text{next}};w^*),$$

$$< \sum_{a \in \mathcal{A}} \pi(a|s;w^* + \alpha \eta) R(a,s) + \gamma \sum_{s_{\text{next}} \in \mathcal{S}} p(s_{\text{next}}|s;w^* + \alpha \eta) V(s_{\text{next}};w^* + \alpha \eta),$$

$$= V(s;w^* + \alpha \eta),$$

where the inequality follows from the fact that $p(s'|s;w^* + \alpha \eta) > 0$, for some $s' \in S_\geq$. This is a contradiction of the fact that $s \in S_\leq$, so it follows that $S_\leq = \emptyset$ and for all $s \in \mathcal{S}$ we have

$$V(s;w^*) < V(s;w^* + \alpha \eta),$$

which completes the proof. \qed
We now prove a technical lemma about the gradient of the value function in the case of a tabular policy. As we are considering a tabular policy we have a separate parameter vector \( w_s \) for each state \( s \) and denote the parameter vector of the entire policy by \( w \), where this is given by the concatenation of the parameter vectors of the various states. The dimension of \( w \) is given by \( n = \sum_{s \in \mathcal{S}} n_s \). For each state, \( s \in \mathcal{S} \), we also introduce the projection mapping, \( P_s \), that maps the elements corresponding to state \( s \) to themselves, while all other elements are set to zero. As before we denote the resulting projection as \( w_{P_s} = P_s(w) \). In order to show that tabular policies are non-decreasing we start by relating the gradient of \( V(s; w) \) to the gradient of \( V(\hat{s}; w) \), where the gradient is taken \( w.r.t. \) the policy parameters of state \( \hat{s} \), while the policy parameters of the remaining states are held fixed.

**Lemma 6.** Suppose we are given a Markov Decision Process with a tabular policy such that \( V(s; w) \) is differentiable for each \( s \in \mathcal{S} \). Given \( \bar{s}, \hat{s} \in \mathcal{S} \), such that \( \bar{s} \neq \hat{s} \), then we have that

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = P_{\bar{s}} \nabla_{w_{\hat{s}}} V(\hat{s}; w),
\]

where we have the notation \( \nabla_{w_{\bar{s}}} V(\bar{s}; w) \) to denote the gradient of the value function \( w.r.t. \) the policy parameter of state \( \bar{s} \), while the policy parameters of all other states are held fixed. The term \( P_{\bar{s}} \) in (25) is given by

\[
p_{\text{hit}} = \sum_{t=2}^{\infty} \gamma^{t-1} p(s_t = \bar{s} | s_1 = \hat{s}, s_{1:t-1} \neq \bar{s}; w).
\]

Furthermore, when Markov chain induced by the policy parameters is ergodic then \( p_{\text{hit}} > 0 \).

**Proof.** Given the equality

\[
V(s; w) = \sum_{a \in A} \pi(a | s; w) Q(s, a; w),
\]

we have that

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = \sum_{a \in A} \left( \nabla_{w_{\bar{s}}} \pi(a | \bar{s}; w) Q(\bar{s}, a; w) + \pi(a | \bar{s}; w) \nabla_{w_{\bar{s}}} Q(\bar{s}, a; w) \right).
\]

As the policy is tabular and \( \bar{s} \neq \hat{s} \) we have that \( \nabla_{w_{\bar{s}}} \pi(a | \bar{s}; w) = 0 \), so that this simplifies to

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = \sum_{a \in A} \pi(a | \bar{s}; w) \nabla_{w_{\hat{s}}} Q(\hat{s}, a; w).
\]

Using the fact that

\[
Q(s, a; w) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) V(s'; w),
\]

means that we have

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = \gamma \sum_{s' \in \mathcal{S}} p(s' | \bar{s}; w) \nabla_{w_{\hat{s}}} V(s'; w).
\]

Splitting the summation in (26) between the state \( \bar{s} \) and all other states gives

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = \gamma p(\bar{s} | \bar{s}; w) \nabla_{w_{\bar{s}}} V(\bar{s}; w) + \gamma \sum_{s' \in \mathcal{S} \setminus \{ \bar{s} \}} p(s' | \bar{s}; w) \nabla_{w_{\bar{s}}} V(s'; w).
\]

Applying equation (27) recursively gives

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = \sum_{t=2}^{\infty} \gamma^{t-1} p(s_t = \bar{s} | s_1 = \hat{s}, s_{1:t-1} \neq \bar{s}; w) \nabla_{w_{\bar{s}}} V(\bar{s}; w).
\]

Defining

\[
p_{\text{hit}} = \sum_{t=2}^{\infty} \gamma^{t-1} p(s_t = \bar{s} | s_1 = \hat{s}, s_{1:t-1} \neq \bar{s}; w),
\]

then we have

\[
\nabla_{w_{\bar{s}}} V(\bar{s}; w) = p_{\text{hit}} \nabla_{w_{\hat{s}}} V(\hat{s}; w),
\]

which completes the proof. The probability, \( p(s_t = \bar{s} | s_1 = \hat{s}, s_{1:t-1} \neq \bar{s}; w) \), is equivalent to the probability of the first hitting time (of hitting state \( \bar{s} \) when starting in state \( \hat{s} \)) is equal to \( t \). The strict inequality, \( p_{\text{hit}} > 0 \), follows from the ergodicity of the Markov chain induced by \( w \).

\[\square\]
Lemma. Suppose that a given a Markov Decision Process has a tabular policy parameterisation, then the policy parameterisation is non-decreasing.

Proof. Suppose we are given parameter vector, \( w \in W \), a search direction, \( \eta \), and a step-size, \( \alpha \in \mathbb{R}^+ \), such that
\[
V(\hat{s}; w + \alpha\eta) > V(\hat{s}; w),
\]
for some \( \hat{s} \in S \). As \( \eta \) is a strict ascent direction we have the following inequality
\[
\eta^\top \nabla_w V(\hat{s}; w) > 0,
\]
which means that, using the equality \( \eta = \sum_{s \in S} \eta_{p_s} \), we have
\[
\eta^\top \nabla_w V(\hat{s}; w) = \sum_{s \in S} \eta_{p_s}^\top \nabla_{w_s} V(\hat{s}; w) > 0.
\]
Hence there exists \( \bar{s} \in S \) such that
\[
\eta_{p_{\bar{s}}}^\top \nabla_{w_{\bar{s}}} V(\bar{s}; w) > 0. \tag{30}
\]
We have that either \( \bar{s} = \bar{s} \), in which case we have \( \eta_{p_{\bar{s}}}^\top \nabla_{w_{\bar{s}}} V(\bar{s}; w) > 0 \), or that \( \bar{s} \neq \bar{s} \) in which case we have, by Lemma[6], that \( \nabla_{w_{\bar{s}}} V(\bar{s}; w) = p_{hit} \nabla_{w_{\bar{s}}} V(\bar{s}; w) \). As \( p_{hit} \geq 0 \) we again have that \( \eta_{p_{\bar{s}}}^\top \nabla_{w_{\bar{s}}} V(\bar{s}; w) > 0 \). In either case we have that \( \eta_{p_s} \) is a strict ascent direction for the value function evaluated at \( \bar{s} \).

We now show that the projection \( \eta_{p_s} \) satisfies the properties of a non-decreasing policy. Firstly, we have at least one state where the inequality in \([13]\) is strict for a sufficiently small step-size, namely the state \( \bar{s} \). Additionally, it can be seen from Lemma[6] that for all other states the inequality holds in \([13]\), again for a sufficiently small step-size, where the inequality will be strict if \( p_{hit} > 0 \) and an equality if \( p_{hit} = 0 \). Furthermore, as the policy is tabular we have that for all \( s \in S \), such that \( s \neq \bar{s} \), and for all \( a \in \mathcal{A} \) the following equality holds
\[
\pi(a|s; w + \alpha\eta_{p_s}) = \pi(a|s; w).
\]
Hence the policy parameterisation is non-decreasing.

7.5 Convergence Analysis

We start this section with a technical lemma that will be useful in this section. In particular, we show that when \( \mathcal{H}(w^*) \) is negative-definite then \( \mathcal{H}_2(w^*) \) is also negative-definite. We actually show a stronger results, where we show that this matrix is negative-definite in an open neighbourhood of \( w^* \). Note that this result doesn’t require the policy to be log-concave in the policy parameters.

Lemma 7. Suppose that the point \( w^* \in W \) is a local optimum of \( U(w) \) such that \( \mathcal{H}(w^*) \) is negative-definite. Given any sequence, \( \{w_k\}_{k \in \mathbb{N}} \), that converges to \( w^* \), then \( \exists K \in \mathbb{N} \) such that \( \forall k \geq K \) the matrix \( \mathcal{H}_2(w_k) \) is negative-definite.

Proof. First note that the matrix \( \mathcal{H}_2(w) \) can be written in the form
\[
\mathcal{H}_2(w) = \mathcal{H}(w) - (\mathcal{H}_{11}(w) + \mathcal{H}_{12}(w) + \mathcal{H}_{12}^\top(w)).
\]
Furthermore, for any \( w \in W \) the matrices \( \mathcal{H}_{11}(w) + \mathcal{H}_{12}(w) + \mathcal{H}_{12}^\top(w) \) is positive-semidefinite. As \( \mathcal{H}(w^*) \) is negative-definite we have that
\[
\mathcal{H}_2(w^*) = \mathcal{H}(w^*) - (\mathcal{H}_{11}(w^*) + \mathcal{H}_{12}(w^*) + \mathcal{H}_{12}^\top(w^*)),
\]
is negative-definite. The entries of \( \mathcal{H}_2(w) \) depend continuously on \( w \in W \), so that, defining the characteristic polynomials
\[
p(\lambda) = \det(\lambda I - \mathcal{H}(w^*)), \quad p_k(\lambda) = \det(\lambda I - \mathcal{H}(w_k)), \quad k \in \mathbb{N},
\]
we have that \( \lim_{k \to \infty} p_k(\lambda) = p(\lambda) \). As the zeros of a polynomial depend continuously upon its coefficients and \( \mathcal{H}_2(w^*) \) is negative-definite, so that the zeros of \( p(\lambda) \) are strictly negative and bounded away from zero, it follows that \( \exists K \in \mathbb{N} \) such that for all \( k \geq K \) the zeros of \( p_k(\lambda) \) are strictly negative. Hence, for all \( k \geq K \) the matrix \( \mathcal{H}_2(w_k) \) is guaranteed to be negative-definite.
Lemma 4. Suppose that the sequence, \( \{w_k\}_{k \in \mathbb{N}} \), is generated by an application of the EM-algorithm, where the sequence converges to \( w^* \). Denoting the update operation of the EM-algorithm by \( G_{\text{EM}} \), so that \( w_{k+1} = G_{\text{EM}}(w_k) \), then

\[
\nabla G_{\text{EM}}(w^*) = -\mathcal{H}_2(w^*)^{-1}\mathcal{H}_1(w^*). 
\]

When the policy parameterisation is non-decreasing w.r.t. the given Markov Decision Process this simplifies to \( \nabla G_{\text{EM}}(w^*) = -\mathcal{H}_2(w^*)^{-1}\mathcal{H}_1(w^*) \). When the Hessian, \( \mathcal{H}(w^*) \), is negative-definite then \( \rho(\nabla G_{\text{EM}}(w^*)) < 1 \) and \( w^* \) is a local point of attraction for the EM-algorithm.

Proof. In the EM-algorithm the update of the policy parameters takes the form

\[
G_{\text{EM}}(w_k) = \arg\max_{w \in W} Q(w, w_k),
\]

where the function \( Q(w, w') \) is given by

\[
Q(w, w') = \mathbb{E}_{\pi(z; w')Q(z; w')} \left[ \log \pi(a|s; w) \right],
\]

and we have used the notation \( \mathbb{E}_{\pi(z; w')Q(z; w')} [f(z)] \) to denote the summation of \( f(w,r) \) the non-negative function, \( \pi(z; w')Q(z; w') \). More details of the EM-algorithm for Markov Decision Processes can be found in [4, 5, 10, 15]. Note that \( Q \) is a two parameter function, where the first parameter occurs inside the expectation, while the second parameter parameterises the function w.r.t. which the expectation is taken. Also note that \( Q(w, w') \) satisfies the following identities

\[
\nabla^10 Q(w, w') = \mathbb{E}_{\pi(z; w')Q(z; w')} \left[ \nabla_w \log \pi(a|s; w) \right],
\]

\[
\nabla^20 Q(w, w') = \mathbb{E}_{\pi(z; w')Q(z; w')} \left[ \nabla_w \nabla_w^\top \log \pi(a|s; w) \right],
\]

\[
\nabla^11 Q(w, w') = \sum_{z \in Z} \nabla'_w \left( \pi(z; w')Q(z; w') \right) \nabla^w_\top \log \pi(a|s; w).
\]

Note that when we set \( w = w' \) in the first two of these terms we have \( \nabla^10 Q(w, w) = \nabla_w U(w) \), \( \nabla^20 Q(w, w) = \mathcal{H}_2(w) \). A key identity that we need for the proof is that \( \nabla^{11} Q(w, w) = \mathcal{H}_1(w) \). This follows from the observation that \( \nabla_w U(w) = \nabla^10 Q(w, w) \), so that

\[
\nabla_w \nabla^w_\top U(w) = \nabla_w \left( \nabla^10 Q(w, w) \right) = \nabla^20 Q(w, w) + \nabla^{11} Q(w, w),
\]

so that

\[
\mathcal{H}_1(w) = \mathcal{H}(w) - \mathcal{H}_2(w) = \nabla^20 Q(w, w) + \nabla^{11} Q(w, w) - \nabla^20 Q(w, w),
\]

as claimed.

Now, to calculate the matrix \( \nabla G_{\text{EM}}(w^*) \) we perform a Taylor series expansion of \( \nabla^10 Q(w_{k+1}, w_k) \) in both parameters around the point \((w^*, w^*)\), which gives

\[
\nabla^10 Q(w_{k+1}, w_k) = \nabla^10 Q(w^*, w^*) + \nabla^11 Q(w^*, w^*)(w_{k+1} - w^*)
\]

\[
+ \nabla^20 Q(w^*, w^*)(w_k - w^*) + \ldots.
\]

As \( w^* \) is a local optimum of \( U(w) \) we have that \( \nabla^10 Q(w^*, w^*) = 0 \). Furthermore, as the sequence \( \{w_k\}_{k \in \mathbb{N}} \) was generated by the EM-algorithm, we have, for each \( k \in \mathbb{N} \), that \( w_{k+1} = \arg\max_{w \in W} Q(w, w_k) \), which implies that \( \nabla^10 Q(w_{k+1}, w_k) = 0 \). Finally, as \( \nabla^11 Q(w^*, w^*) = \mathcal{H}_2(w^*) \) and \( \nabla^20 Q(w^*, w^*) = \mathcal{H}_1(w^*) \) we have

\[
0 = \mathcal{H}_2(w^*)(w_{k+1} - w^*) + \mathcal{H}_1(w^*)(w_k - w^*) + \ldots.
\]

Using the fact that \( w_{k+1} = G_{\text{EM}}(w_k) \) and \( w^* = G_{\text{EM}}(w^*) \), taking the limit \( k \to \infty \) gives

\[
0 = \mathcal{H}_2(w^*) \nabla_w G_{\text{EM}}(w^*) + \mathcal{H}_1(w^*),
\]

16
so that
\[ \nabla_w G_{\text{EM}}(w^*) = -\mathcal{H}_2^{-1}(w^*)\mathcal{H}_1(w^*). \]
In the case where the policy parameterisation is non-decreasing w.r.t. the given MDP then we have \( \mathcal{H}_1(w^*) + \mathcal{H}_2(w^*)^\top = 0 \) so that \( \mathcal{H}_1(w^*) = \mathcal{H}_1(w^*). \) Hence in this case we have that
\[ \nabla_w G_{\text{EM}}(w^*) = I - \mathcal{H}_2^{-1}(w^*)\mathcal{A}_{11}(w^*), \]
where we have used the fact that \( \mathcal{H}(w^*) + \mathcal{H}_1(w^*) = \mathcal{A}_{11}(w^*). \)

We now show that when \( \mathcal{H}(w^*) \) is negative-definite that \( w^* \) is a point of attraction for the EM-algorithm, in particular that the eigenvalues of \( \nabla G_{\text{EM}}(w^*) \) are all contained in the interval \([0, 1] \).

To do so we use the notion of relative eigenvalues, see e.g. [7]. A scalar, \( \lambda \in \mathbb{R} \), is an eigenvalue of a symmetric matrix, \( A \), relative to a positive-definite matrix, \( B \), if there exists a non-zero vector, \( x \), such that \( Ax = \lambda Bx \). Now suppose that \( x \in \mathbb{R}^{nw} \) is an eigenvector of \( \nabla G_{\text{EM}}(w^*) \), with corresponding eigenvalue, \( \lambda \in \mathbb{R} \). Given [7] this means that \( x \) satisfies the relation, \( \mathcal{H}_1(w^*)x = -\lambda \mathcal{H}_2(w^*)x \), so that \( x \) is an eigenvector of \( \mathcal{H}_1(w^*) \), relative to \( -\mathcal{H}_2(w^*) \), where we know that \( -\mathcal{H}_2(w^*) \) is positive-definite from lemma 7. This means that the eigenvalues of \( \nabla G_{\text{EM}}(w^*) \) are in one-to-one correspondence with the eigenvalues of \( \mathcal{H}_1(w^*) \), relative to \( -\mathcal{H}_2(w^*) \). An important property about relative eigenvalues is that the eigenvalues of \( A \), relative to \( B \), are contained in the image of the relative Rayleigh quotient,
\[ R(x) = \frac{x^\top Ax}{x^\top Bx}. \]

Furthermore, the maximal and minimal relative eigenvalues are given by the maximal and minimal values of the relative Rayleigh quotient. See [7] for more details. This means that the eigenvalues of \( \mathcal{H}_1(w^*) \), relative to \( -\mathcal{H}_2(w^*) \), are contained in the image of
\[ R(x) = \frac{x^\top \mathcal{H}_1(w^*)x}{-x^\top \mathcal{H}_2(w^*)x} = \frac{x^\top (\mathcal{H}(w^*) - \mathcal{H}_2(w^*))x}{-x^\top \mathcal{H}_2(w^*)x}, \]
(32)
where we have used the relation \( \mathcal{H}_1(w^*) = \mathcal{H}(w^*) - \mathcal{H}_2(w^*) \). As \( -\mathcal{H}_2(w^*) \) is positive-definite, \( \mathcal{H}(w^*) \) is negative-definite and \( \mathcal{H}_1(w^*) \) is positive-semidefinite it can be seen that \( R(x) \in [0, 1] \), for all \( x \in \mathbb{R}^{nw} \). Hence, all of the eigenvalues of \( \nabla G_{\text{EM}}(w^*) \) are contained in the interval \([0, 1] \), so that by Ostrowski’s theorem \( w^* \) is a point of attraction. In the case where \( \mathcal{H}_2(w^*) \) is negative-definite, but \( \mathcal{H}(w^*) \) is only negative-semidefinite, it can be seen from the relation
\[ \mathcal{H}(w^*) = \mathcal{H}_2(w^*)(I - \nabla{\text{FAN}}(w^*)), \]
that \( \nabla G_{\text{EM}}(w^*) \) has eigenvalues equal to one, which correspond to ridges in the objective function.

**Lemma 5.** Suppose that the sequence, \( \{w_k\}_{k \in \mathbb{N}} \), is generated by an application of the full approximate Newton method using a fixed step-size of \( \alpha \in \mathbb{R}^+ \), where the sequence converges to \( w^* \). Denoting the update operation of the full approximate Newton method by \( G_{\text{FAN}} \), so that \( w_{k+1} = w_k - \alpha \mathcal{H}_2^{-1}(w_k)\nabla U(w_k) \), then
\[ \nabla G_{\text{FAN}}(w^*) = I - \alpha \mathcal{H}_2^{-1}(w^*)\mathcal{H}(w^*) = (1 - \alpha)I - \alpha \mathcal{H}_2^{-1}(w^*)\mathcal{H}_1(w^*), \]

When the policy parameterisation is non-decreasing w.r.t. the given Markov Decision Process this simplifies to \( \nabla G_{\text{FAN}}(w^*) = I - \alpha \mathcal{H}_2^{-1}(w^*)\mathcal{A}_{11}(w^*) \). When a fixed step-size of one is used in the full approximate Newton method then \( \nabla G_{\text{FAN}}(w^*) = \nabla G_{\text{EM}}(w^*) \), so that the rate of convergence is the same as the rate of convergence of the EM-algorithm. The optimal step-size of the full approximate Newton method is given by
\[ \hat{\alpha} = \frac{2}{\lambda_{\min} + \lambda_{\max}}, \]
where \( \lambda_{\min} \) and \( \lambda_{\max} \) are respectively the minimal and maximal eigenvalues of \( \nabla G_{\text{EM}}(w^*) \), and the point \( w^* \) remains a point of attraction of the full approximate Newton method when using this step-size.

**Proof.** In order to apply Ostrowski’s theorem to the approximate Newton framework it is necessary to show i) that \( G_{\text{FAN}}(w^*) \) is Frechet differentiable and ii) that the spectral radius of \( \nabla G_{\text{FAN}}(w^*) \) is
strictly less than one. Assuming, for the moment, that \( G_{\text{FAN}}(w^*) \) is Fréchet differentiable, then we have
\[
\nabla G_{\text{FAN}}(w^*) = I - \alpha \nabla H_2^{-1}(w^*) \nabla \top U(w^*) - \alpha H_2^{-1}(w^*) \nabla \top U(w^*).
\]
Using the fact that \( \nabla U(w^*) = 0 \) and \( \nabla \top U(w) = H_1(w) + H_2(w), \forall w \in \mathcal{W} \), this simplifies to
\[
\nabla G_{\text{FAN}}(w^*) = H_2^{-1}(w^*) \left( H_2(w^*) - \alpha \nabla \top U(w^*) \right),
\]
\[
= H_2^{-1}(w^*) \left( (1 - \alpha)H_2(w^*) - \alpha H_1(w^*) \right),
\]
\[
= (1 - \alpha)I - \alpha H_2^{-1}(w^*) H_1(w^*). \tag{33}
\]
A more formal proof of \((33)\), requiring no assumptions on the differentiability of \( H_2(w) \), can be made using the methods described in section(10.2.1) of \cite{13}. When the policy parameterisation is non-decreasing \( \text{w.r.t.} \) the given MDP then \( H_{12}(w^*) + H_{12}(w^*)^\top = 0 \), so that \( H_1(w^*) = H_{11}(w^*) \).
It then follows that \( \nabla G_{\text{FAN}}(w^*) = I - \alpha H_2^{-1}(w^*) \Lambda_11(w^*) \), where we have used the fact that \( \nabla \Lambda(w^*) = \Lambda_{11}(w^*) + H_2(w^*) = A_{11}(w^*) \).

We now relate the rate of convergence of the Full approximate Newton method to the rate of convergence of the EM-algorithm, where we use analogous arguments to those presented in \cite{11}. Firstly, note that \( \nabla G_{\text{FAN}}(w^*) = (1 - \alpha)I + \alpha \nabla G_{\text{EM}}(w^*) \), so that when \( \alpha = 1 \) the two methods have the same rate of convergence. Furthermore, an eigenvector of \( \nabla G_{\text{EM}}(w^*) \) is also an eigenvector of \( \nabla G_{\text{FAN}}(w^*) \), but with an eigenvalue of \( \lambda_{\alpha} = 1 - \alpha + \alpha \lambda \) instead of \( \lambda \). As all the eigenvalues of \( \nabla G_{\text{EM}}(w^*) \) are contained in the interval \((0, 1)\) then the spectral radius of \( \nabla G_{\text{FAN}}(w^*) \) is given as
\[
\rho(\nabla G_{\text{FAN}}(w^*)) = \max(|1 - \alpha + \alpha \lambda_{\text{min}}|, |1 - \alpha + \alpha \lambda_{\text{max}}|), \tag{34}
\]
where \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) are respectively the minimal and maximal eigenvalues of \( \nabla G_{\text{EM}}(w^*) \). An optimal rate of convergence is obtained for the full approximate Newton method by minimising \((34)\) \( \text{w.r.t.} \) the step-size. This minimum is obtained when \( 1 - \alpha + \alpha \lambda_{\text{min}} \) and \( 1 - \alpha + \alpha \lambda_{\text{max}} \) are of equal magnitude but of opposite sign, where the resulting step-size is given by
\[
\hat{\alpha} = \frac{2}{2 - (\lambda_{\text{min}} + \lambda_{\text{max}})}.
\]
The spectral radius of \( \nabla G_{\text{FAN}}(w^*) \) that results from this step-size is \((\lambda_{\text{max}} - \lambda_{\text{min}})/(2 - (\lambda_{\text{min}} + \lambda_{\text{max}}))\), which is less than one so that \( w^* \) is still a point of attraction when using this step-size. \( \square \)

### 7.6 Analysis over the Entire Parameter Space

Throughout this section we shall assume that the stat-action space of the MDP is finite, where \( |S| \times |A| = N \). Given a parameter vector, \( w \in \mathcal{W} \), we suppose that the Markov chain induced by \( w \) is ergodic. We denote the state-action transition matrix, induced by \( w \), by \( P(w) \), and when the context is clear we shall use the notation \( P \). Given \( t \in \mathbb{N} \), we denote the state-action occupancy marginal at the \( t \)th-point time by \( p_t \), so that \( p_t P = p_{t+1} \). We shall assume that the \( P \) has \( N \) linear independent eigenvectors.\footnote{This assumption is not necessary for the results in this section. We make these assumptions as they make the form of the bounds more readable.} We denote the eigenvalues of \( P \) by \( \{\lambda_n\}_{n=1}^N \), where the indices of the eigenvalues satisfy the ordering \(|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_N|\), so that, under the assumption of ergodicity, \( \lambda_1 = 1 \) and \(|\lambda_2| < 1 \). As \( P \) has \( N \) linearly independent eigenvectors it is possible to decompose \( P \) into the form \( P = SAS^{-1} \), where \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) \). The columns of \( S \) are formed from a set of (linearly independent) right eigenvectors of \( P \), where the ordering of these eigenvectors respects the ordering of the eigenvalues in \( \Lambda \). Likewise, the rows of \( S^{-1} \) are left eigenvectors of \( P \), where, again, the ordering of these eigenvectors respects the ordering of the eigenvalues in \( \Lambda \). We denote these sets of right and left eigenvectors by \( X = \{x_1, \ldots, x_N\} \) and \( Y = \{y_1, \ldots, y_N\} \), respectively. Under the conditions specified it is possible to set \( y_1 = \mu \), where \( \mu \) is the stationary distribution of the Markov chain induced by the policy parameters, and \( x_1 = e \), where \( e = (1, \ldots, 1)^\top \).
A final assumption is that the derivative of the log-policy, \( w.r.t. \) to the various components of the parameter vector, is uniformly bounded. In other words, there exists \( M \in \mathbb{R}_+^\times \) such that
\[
\left| \frac{d}{dw_i} \log \pi(a|s; w) \right| < M,
\]
for all \((a, s) \in A \times S, w \in \mathcal{W}\).

**Lemma 3.** Given \( w \in \mathcal{W} \), suppose that the Markov chain induced by \( w \) is ergodic. Denote the second largest eigenvalue (in terms of absolute value) of the state-action transition matrix by \( \lambda_2 \). Under these conditions and given some matrix norm, \( \| \cdot \| \), the matrix \( \mathcal{H}_{12}(w) \) satisfies the bound
\[
\| \mathcal{H}_{12}(w) \| \leq \frac{\eta \gamma |\lambda_2|}{(1 - \gamma)^2 (1 - \gamma |\lambda_2|)},
\]
for some positive constant, \( \eta > 0 \).

**Proof.** It is useful to write certain terms in vector notation. In particular, we use the notation \( p_{\gamma|z} \) to denote \( p_\gamma(z'|z; w), z, z' \in Z \). We also introduce the notation
\[
\mathbf{p}_{\gamma, \mu} = \sum_{t=1}^{\infty} \gamma^{t-1} \mu,
\]
where \( \mu \) is the stationary distribution of the Markov chain, induced by the policy parameters. Note that, as \( \mu \) is independent of time, we have \( \mathbf{p}_{\gamma, \mu} = (1-\gamma)^{-1} \mu \).

The matrix \( \mathcal{H}_{12}(w) \) can be written as follows
\[
\mathcal{H}_{12}(w) = \Pi J \Pi^\top
\]
where \( \Pi \) is an \( n_w \times N \) matrix, where the \((i, j)\)th element of \( \Pi \) is given by
\[
\Pi_{ij} = \frac{d}{dw_i} \log \pi(z = j; w),
\]
and \( J \) is an \( N \times N \) matrix, where the \((i, j)\)th element of \( J \) is given by
\[
J_{ij} = p_\gamma(z = i; w) p_\gamma(z = j|z = i; w) Q(z = j; w).
\]
Likewise, if we define \( I = \mathbf{p}_{\gamma, \mu} \) then, using the same argument as in section 4.2, it can be shown that \( 0 = \Pi J \Pi^\top \). This means that \( \mathcal{H}_{12}(w) \) can be written in the equivalent form
\[
\mathcal{H}_{12}(w) = \Pi (J - I) \Pi^\top.
\]
This is the form of \( \mathcal{H}_{12}(w) \) that we will use to obtain our bound in (35). Additionally, we use the fact that for any matrix norm, \( \| \cdot \| \), there exists a positive constant \( \eta \in \mathbb{R}_+^\times \) such that
\[
\| \mathcal{H}_{12}(w) \| \leq \eta \max_{i,j} |\mathcal{H}_{12}^{ij}(w)|,
\]
where \( \mathcal{H}_{12}^{ij}(w) \) denotes the \((i, j)\)th element of \( \mathcal{H}_{12}(w) \). Hence, to obtain the bound (35) it suffices to obtain an appropriate bound on \( \max_{i,j} |\mathcal{H}_{12}^{ij}(w)| \).

Denoting the \( i \)th row of \( \Pi \) by \( \Pi_{i,:} \), then the \((i, j)\)th element of \( \mathcal{H}_{12}(w) \) can be written as follows
\[
\mathcal{H}_{12}^{ij}(w) = \Pi_{i,:} (J - I) \Pi_{j,:}^\top.
\]
The Cauchy-Schwarz inequality now gives
\[
|\mathcal{H}_{12}^{ij}(w)| \leq \| \Pi_{i,:} \| \| (J - I) \Pi_{j,:}^\top \|
\]
where \( \| \cdot \| \) is some given vector-norm. Considering a vector induced matrix-norm, this gives
\[
|\mathcal{H}_{12}^{ij}(w)| \leq \| \Pi_{i,:} \| \| \Pi_{j,:} \| \| J - I \|.
\]
As the eigenvalues of $P$, now the term $x$ As
by the equivalence of norms, see e.g. section(2.2.1) of [13], and the uniform boundedness of the derivative of the log-policy, there exists a constant $\eta_1 \in \mathbb{R}^+$, such that
\[
|H_{ij}^2(w)| \leq \eta_1 M^2 \|J - I\|.
\]
(36)

As $J - I$ is a square matrix there exists a constant, $\eta_2 \in \mathbb{R}^+$, such that
\[
\|J - I\| \leq \eta_2 \max_{i \in \mathbb{N}} \|(J - I)_{i,\cdot}\|,
\]
where $(J - I)_{i,\cdot}$ is the $i^{th}$ row of $J - I$, see exercise 2.2.2 of [13]. Using this bound in (36) gives
\[
|H_{ij}^2(w)| \leq \eta M^2 \max_{i \in \mathbb{N}} \|(J - I)_{i,\cdot}\|,
\]
(37)
for the strictly positive constant, $\eta = \eta_1 \eta_2$.

Note that all of the components of the $i^{th}$ row of $J - I$ contain a multiplicative factor of $p_\gamma(z = i; w)$, which can be bounded by $(1 - \gamma)^{-1}$. Furthermore, the state-action value function can be bounded by $R_{\text{max}}/(1 - \gamma)$, where $R_{\text{max}}$ is the maximum value of the reward function. After a simple manipulation using the supremum-norm, we can use (37) to obtain the bound
\[
|H_{ij}^2(w)| \leq \eta R_{\text{max}} M^2 \max_{i \in \mathbb{N}} \|p_\gamma,\cdot - p_\gamma,\cdot\|.
\]
(38)

To complete the proof we obtain a bound on the term $\|p_\gamma,\cdot - p_\gamma,\cdot\|$, for each $i \in \mathbb{N}$. To do so, we write these two vectors in terms of the eigendecomposition of $P$. In particular, we have
\[
p_\gamma,\cdot = \sum_{t=1}^\infty \gamma^{-t} p_{i,\cdot} S \sum_{t=1}^\infty \gamma^{-t} A^{t-1} S^{-1}.
\]
As the eigenvalues of $P$ are all contained in the unit circle and $\gamma \in [0, 1)$, we have
\[
p_\gamma,\cdot = \sum_{t=1}^\infty \gamma^{-t} p_{i,\cdot} S \sum_{t=1}^\infty \gamma^{-t} A^{t-1} S^{-1}.
\]
As $x_1 = (1, \ldots, 1)^T$ and $y_1 = \mu$, we have that
\[
p_\gamma,\cdot = \frac{1}{1 - \gamma} \mu + p_{1,\cdot} S \sum_{t=1}^\infty \gamma^{-t} A^{t-1} S^{-1}.
\]
This means that $(p_\gamma,\cdot - p_\gamma,\cdot)^T$ is given by
\[
(p_\gamma,\cdot - p_\gamma,\cdot)^T = p_{i,\cdot} S \sum_{t=1}^\infty \gamma^{-t} \mu S^{-1}.
\]
Now the term $p_{1,\cdot}$ is nothing but the next state-action distribution, given that the current state-action pair is $z = i$. In other words, $p_{1,\cdot}$ is the row of $P$ corresponding to state-action pair $z = i$. This means that $(p_\gamma,\cdot - p_\gamma,\cdot)^T$ takes the form
\[
(p_\gamma,\cdot - p_\gamma,\cdot)^T = S_{i,\cdot} \sum_{t=1}^\infty \gamma^{-t} \mu S^{-1},
\]
where $P_{1,\cdot}$ and $S_{i,\cdot}$ denote the $i^{th}$ row of $P$ and $S$, respectively.
Using the Cauchy-Schwarz inequality we have
\[
\|p_{\gamma, z=i} - p_\gamma^i\| \leq \left\| S_i \right\| \left\| \text{diag} \left( 0, \frac{\lambda_2}{1 - \gamma \lambda_2}, \ldots, \frac{\lambda_N}{1 - \gamma \lambda_N} \right) \right\| \left\| S^{-1} \right\|,
\]
for any vector-norm, and corresponding induced matrix-norm, \(\| \cdot \|\). A further application of the inequality, \(\|Ax\| \leq \|A\|\|x\|\), gives
\[
\|p_{\gamma, z=i} - p_\gamma^i\| \leq \left\| S_i \right\| \left\| S^{-1} \right\| \left\| \text{diag} \left( 0, \frac{\lambda_2}{1 - \gamma \lambda_2}, \ldots, \frac{\lambda_N}{1 - \gamma \lambda_N} \right) \right\|.
\]
If we now consider the 2-norm, then the induced matrix-norm satisfies the identity \(\|A\| = \sqrt{\lambda}\), where \(\lambda\) is the largest eigenvalue of \(A^T A\). In this case we have
\[
\|p_{\gamma, z=i} - p_\gamma^i\| \leq \frac{|\lambda_2|}{1 - \gamma |\lambda_2|} \left\| S_i \right\| \left\| S^{-1} \right\|. \tag{38}
\]
Using this bound for \(\|p_{\gamma, z=i} - p_\gamma^i\|, i \in \mathbb{N}_N\), in (38) gives
\[
|\mathcal{H}_{ij}^i(w)| \leq \frac{\eta \|\lambda_2\|}{(1 - \gamma)^2 (1 - \gamma |\lambda_2|)}
\]
for some positive constant, \(\eta \in \mathbb{R}^+\). This completes the proof.

\[\square\]