Supplementary material for "Temporal Difference Based Actor Critic Algorithms - Convergence and Neural Implementation"

A Proof of Theorem 2.4

The following theorem was proved in [3, 4, 6]. It relates the gradient of the average reward per stage to the differential value function. We present the proof here, which will be used in the sequel.

Theorem A.1. *The gradient of the average reward per stage can be expressed by*

$$\nabla\eta(\theta) = \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y)\psi(x,u|\theta)h(y|\theta). \tag{A.1}$$

Proof. We begin with Poisson's equation (2) in vector form

$$h(\theta) = r - e\eta(\theta) + P(\theta)h(\theta),$$

where $h(\theta) = [h(x|\theta]_{x\in\mathcal{X}}$ and e is a column vector of 1's. Taking the derivative with respect to θ and rearranging yields

$$e\nabla\eta(\theta) = -\nabla h(\theta) + \nabla P(\theta)h(\theta) + P(\theta)\nabla h(\theta).$$

Multiplying the left hand side of the last equation by the stationary distribution $\pi(\theta)'$ yields

$$\begin{aligned}\nabla\eta(\theta) &= -\pi(\theta)'\nabla h(\theta) + \pi(\theta)'\nabla P(\theta)h(\theta) + \pi(\theta)'P(\theta)\nabla h(\theta)\\ &= -\pi(\theta)'\nabla h(\theta) + \pi(\theta)'\nabla P(\theta)h(\theta) + \pi(\theta)'\nabla h(\theta)\\ &= \pi(\theta)'\nabla P(\theta)h(\theta).\end{aligned}$$

Expressing the result explicitly we obtain

$$\begin{aligned}\nabla\eta(\theta) &= \sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)h(y|\theta)\\ &= \sum_{x,y\in\mathcal{X}} P(x)\nabla\left\{\sum_u \left(P(y|x,u)\mu(u|x,\theta)\right)\right\}h(y|\theta)\\ &= \sum_{x,y\in\mathcal{X}} P(x)\sum_u \left(P(y|x,u)\nabla\mu(u|x,\theta)\right)h(y|\theta)\\ &= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(y|x,u)P(x)\nabla\mu(u|x,\theta)h(y|\theta)\\ &= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(y|x,u)\mu(u|x,\theta)P(x)\frac{\nabla\mu(u|x,\theta)}{\mu(u|x,\theta)}h(y|\theta)\\ &= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y)\psi(x,u|\theta)h(y|\theta).\end{aligned} \tag{A.2}$$

□

Now we can prove Theorem 2.4. We start with the first line in (A.2).

$$\begin{aligned}\nabla\eta(\theta) &= \sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)h(y|\theta)\\ &= \sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)\left(h(y|\theta) - h(x|\theta) + r(x) - \eta(\theta) + f(x)\right)\\ &- \sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)\left(-h(x|\theta) + r(x) - \eta(\theta) + f(x)\right)\\ &= \sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)\left(d(x,y) + f(x)\right)\\ &- \sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)\left(-h(x|\theta) + r(x) - \eta(\theta) + f(x)\right).\end{aligned}$$

Next, we show that the second term equals 0. We define $F(x,\theta) \triangleq -h(x|\theta) + r(x) - \eta(\theta) + f(x)$ and obtain

$$\begin{aligned}\sum_{x,y\in\mathcal{X}} P(x)\nabla P(y|x,\theta)F(x,\theta) &= \sum_x P(x)F(x,\theta)\sum_y \nabla P(y|x,\theta)\\ &= \sum_x P(x)F(x,\theta)\nabla\sum_y P(y|x,\theta)\\ &= \sum_x P(x)F(x,\theta)\nabla 1\\ &= 0.\end{aligned}$$

Following the same steps as in the proof of Theorem A.1 we have

$$\nabla\eta(\theta) = \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y)\psi(x,u|\theta)\left(d(x,y) + f(x)\right).$$

B Proof of Theorem 2.5

We introduce the following assumption, which adds constraints to the iterations for θ and $\tilde{h}$, and will be used in the sequel to prove Theorem 2.5. This assumption may seem restrictive at first but in practice it is not. The reason is that we usually assume the bounds of the constraints to be large enough so the iterates practically do not reach those bounds. For example, if θ represents synaptic weights of a neural network, above a certain value the synapses saturate. Thus we will choose the bound for θ to be above its saturation value.

Assumption B.1. *We denote by $\{\theta_i\}_{i=1}^K$ the components of θ, and choose positive constants B_θ and $B_{\tilde{h}}$. We define the set $H \subset \mathbb{R}^K \times \mathbb{R}^{|\mathcal{X}|}$ to be*

$$H \triangleq \left\{(\theta_i, \tilde{h}(x))\Big| -B_\theta \le \theta_i \le B_\theta, \quad 1 \le i \le K, \quad -B_{\tilde{h}} \le \tilde{h}(x) \le B_{\tilde{h}}, \quad \forall x \in \mathcal{X}\right\}.$$

Let Π_H be an operator which projects $(\theta, \tilde{h}(x))$ onto H.

For future purposes, we express Algorithm 1 in a different way. Define the augmented parameter vector y_m by

$$y_m \triangleq \begin{pmatrix}\theta_m\\ \tilde{h}_m\\ \tilde{\eta}_m\end{pmatrix}, \quad \theta_m \in \mathbb{R}^K, \tilde{h}_m \in \mathbb{R}^{|\mathcal{X}|}, \tilde{\eta}_m \in \mathbb{R}.$$

The algorithm ignoring the constraints of Assumption B.1, takes the form

$$y_{m+1} = y_m + \gamma_m V_m(y_m) \quad \text{(unconstrained iteration)}, \tag{B.1}$$

where the components of $V_m(y_m)$ are determined according Algorithm 1. Including the constraints we can write the iterates of Algorithm 1 as

$$y_{m+1} = \Pi_H[y_m + \gamma_m V_m(y_m)] \quad \text{(constrained iteration)}. \tag{B.2}$$

We note that we can write the projected equation (B.2) as

$$y_{m+1} = y_m + \gamma_m V_m(y_m) + \gamma_m Z_m(y_m), \tag{B.3}$$

where $Z_m(y_m)$ is a projection term. Define $\mathcal{F}_m$ to be the σ-algebra generated by x_n, $0 \leq n \leq t_m$ namely, $\mathcal{F}_m = \sigma\{x_0, x_1, ..., x_{t_m}\}$, representing the history of the algorithm up to the time t_m. Set

$$v(y_m) \triangleq \mathrm{E}\left[V_m(y_m)|\mathcal{F}_m\right]$$

to be the average change of the algorithm during the m-th cycle, where the m-th cycle is defined to consist of the times between t_m and t_{m+1}. We can then rewrite (B.2) as

$$y_{m+1} = \Pi_H[y_m + \gamma_m v(y_m) + \varepsilon_m], \text{ where } \varepsilon_m = \gamma_m(V_m(y_m) - v(y_m)).$$

We note that the vector $v(y_m)$ is the deterministic part of the iterate while the vector ε_m is the stochastic part.

In order to prove Theorem 2.5, we use techniques from the theory of constrained stochastic approximation [5], in particular, Theorem 5.2.1 in [5] adapted to our purposes.

Theorem B.1. *Consider an iterate scheme as in* (B.3)*, and assume the following:*

1. $\sup_m \mathrm{E}[|V_m|^2] < \infty$,
2. *There exists a measurable function* $v(y)$ *such that* $\mathrm{E}[V_m|y_0, V_i, i < m] = v(y_m)$,
3. *The function* $v(y)$ *is continuous,*
4. *The sequence* γ_m *satisfies* $\sum \gamma_m = \infty$ *and* $\sum \gamma_m^2 < \infty$.

We define $\{y^n(\cdot)\}$ *to be a set of continuous time functions, which are shifted functions of the linear interpolations of the* y_m[1]*. Then, there is a set* N *of probability zero such that for* $\omega \notin N$*, the set of functions* $\{y^n(\omega, \cdot)\}$ *is equicontinuous. Let* $y(\omega, \cdot)$ *denote the limit of some convergent sub sequence. Then this pair satisfies the projected ODE*

$$\dot{y} = \Pi_H[v(y)].$$

Thus, we need to prove that the assumptions of Theorem B.1 are valid. We devote the rest of this section to this purpose. Sub-vectors of $v(y_m)$ will be denoted by $v(\theta_m)$, $v(\tilde{h}_m)$, and $v(\tilde{\eta}_m)$. We examine the components of $v(y_m)$. Define $T_m = t_{m+1} - t_m$, representing the time between two consecutive

[1] Define $\hat{t}_0 = 0$ and $\hat{t}_m = \sum_{i=0}^{m-1} \gamma_i$. Define the *continuous time interpolation* $y^0(\cdot)$ on $(-\infty, \infty)$ by $y^0(t) = y_0$ for $t < 0$ and $y^0(t) = y_m$ for $\hat{t}_{m-1} < t \leq \hat{t}_m$. Define the sequence of shifted process by $y^m(t) \triangleq y^0(t + \hat{t}_m)$.

hitting times of the recurrent state x^*. Therefore, we can write the actor iterate θ_m as

$$\begin{aligned}
v(\theta_m) &= \mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} \tilde{d}(x_{n+1},x_n)\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right] \\
&= \mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} d(x_{n+1},x_n)\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right] \\
&+ \mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} (\tilde{d}(x_{n+1},x_n)-d(x_{n+1},x_n))\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right] \\
&= \mathrm{E}_{\theta_m}[T_m]\nabla\eta(\theta_m) + (\eta(\theta_m)-\tilde{\eta}_m)\mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} \psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right] \\
&+ \overbrace{\mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} \left(\tilde{h}_m(x_{n+1})-h(x_{n+1}|\theta_m)\right)\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right]}^{A_1} \\
&+ \overbrace{\mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} \left(h(x_n|\theta_m)-\tilde{h}_m(x_n)\right)\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right]}^{A_2},
\end{aligned} \tag{B.4}$$

where for the last equality we used Theorem 2.4 and (3). Next, we express A_1 and A_2 as functions of the true differential value, $h(x|\theta)$, and the estimated differential value, $\tilde{h}_m(x)$. For A_1 we have

$$\begin{aligned}
A_1 &= \mathrm{E}\left[\sum_{x\in\mathcal{X}}\sum_{n=t_m}^{t_{m+1}-1} (1\{x_{n+1}=x\}h(x|\theta_m)-\tilde{h}_m(x))\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right] \\
&= \sum_{x\in\mathcal{X}}\left(h(x|\theta_m)-\tilde{h}_m(x)\right)\mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} (1\{x_{n+1}=x\}\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right],
\end{aligned}$$

and for A_2 we have

$$\begin{aligned}
A_2 &= \mathrm{E}\left[\sum_{x\in\mathcal{X}}\sum_{n=t_m}^{t_{m+1}-1} (1\{x_n=x\}h(x|\theta_m)-\tilde{h}_m(x))\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right] \\
&= \sum_{x\in\mathcal{X}}\left(h(x|\theta_m)-\tilde{h}_m(x)\right)\mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} (1\{x_n=x\}\psi(x_n,u_n|\theta_m)\Big|\mathcal{F}_m\right].
\end{aligned}$$

where $1\{A\}$ is an indicator function which equals 1 if A is true, and 0 otherwise. Thus, we have

$$A_1+A_2 = \sum_{x\in\mathcal{X}} D^{(x)}(\theta)(h(x|\theta)-\tilde{h}(x)),$$

where $D^{(x)}$ was defined in Theorem 2.5. Summarizing the above we get

$$v(\theta_m) = \mathrm{E}_{\theta_m}[T_m]\nabla\eta(\theta_m) + C(\theta)(\eta(\theta_m)-\tilde{\eta}_m) + \sum_{x\in\mathcal{X}} D^{(x)}(\theta)(h(x|\theta)-\tilde{h}(x)).$$

For the critic we have for all $x \in \mathcal{X}$

$$\begin{aligned} v(\tilde{h}_m(x)) &= \mathrm{E}\left[\Gamma_h \sum_{n=N(x)}^{t_{m+1}-1} \tilde{d}(x_{n+1}, x_n) \Big| \mathcal{F}_m\right] \\ &= \mathrm{E}\left[\Gamma_h \left(\sum_{n=N(x)}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m) - \tilde{h}_m(x)\right) \Big| \mathcal{F}_m\right] \\ &= \mathrm{E}\left[\Gamma_h \left(\sum_{n=N(x)}^{t_{m+1}-1} (r(x_n) - \eta(\theta_m)) - \tilde{h}_m(x)\right) \Big| \mathcal{F}_m\right] \\ &+ \mathrm{E}\left[\Gamma_h \sum_{n=N(x)}^{t_{m+1}-1} (\eta(\theta_m) - \tilde{\eta}_m) \Big| \mathcal{F}_m\right] \\ &= \Gamma_h \Big(h(x|\theta_m) - \tilde{h}_m(x)\Big) + \Gamma_h \mathrm{E}_{\theta_m}[T_m](\eta(\theta_m) - \tilde{\eta}_m), \end{aligned} \tag{B.5}$$

and

$$\begin{aligned} v_m(\tilde{\eta}_m) &= \Gamma_\eta \mathrm{E}\left[\sum_{n=t_m}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m) | \mathcal{F}_m\right] \\ &= \Gamma_\eta \mathrm{E}_{\theta_m}[T_m](\eta(\theta_m) - \tilde{\eta}_m). \end{aligned} \tag{B.6}$$

The following lemma will establish the boundedness of the first two moments of T_m.

Lemma B.2. *The first two moments of the random times $\{T_m\}$ are bounded by a constant B_T, for all $\theta \in \mathbb{R}^K$ and for all m, $1 \le m < \infty$.*

Proof. The boundedness from below is trivial since the random times are positive. According to Assumption 2.1(i) and Lemma 2.1, each Markov chain in $\bar{\mathcal{P}}$ is a periodic and recurrent. Thus, we can show that for each $\theta \in \mathbb{R}^K$ there exist a constant $B_T(\theta)$, $0 < B_T(\theta) < 1$, where [2]

$$P(T_m = k|\theta_m) \le B_T^k(\theta_m), \quad 1 \le m < \infty, \quad 1 \le k < \infty. \tag{B.7}$$

Therefore,

$$\mathrm{E}_{\theta_m}[T_m] = \sum_{k=1}^{\infty} kP(T_m = k|\theta_m) \le \sum_{k=1}^{\infty} kB_T^k(\theta_m) \le B_{T_1}(\theta_m) < \infty,$$

and

$$\mathrm{E}_{\theta_m}[T_m^2] = \sum_{k=1}^{\infty} k^2 P(T_m = k|\theta_m) \le \sum_{k=1}^{\infty} k^2 B_T^k(\theta_m) \le B_{T_2}(\theta_m) < \infty.$$

Moreover, since the set $\bar{\mathcal{P}}$ is closed, and by Assumption 2.1 the above hold for the closure of $\bar{\mathcal{P}}$ as well, there exists a constant B_T satisfying $B_T = \max\{\sup_\theta B_{T_1}(\theta), \sup_\theta B_{T_2}(\theta)\} < \infty$. □

The following lemma establishes summation and convergence properties of the the random variable $\gamma_m T_m$. These properties will be used later in order to show that the iterate $\tilde{\eta}_m$ is bounded.

Lemma B.3. *Given the result of Lemma B.2, and that $\{\gamma_m\}_{m=1}^{\infty}$ satisfies the assumptions of Algorithm 1, we have:*

[2] An MC is periodic if the greatest common divisor of the set $\{n_x = \min\{n|P_{xx}^{(n)} > 0\}|1 \le x \le |\mathcal{X}|\}$ is larger than 1, where $P_{xx}^{(n)}$ is the probability of starting from state x and returning to it in n steps. According to Assumption 2.1(i) and Lemma 2.1 the chain is recurrent and aperiodic, thus, $P_{xy}(\theta) < 1$, for all $\theta \in \mathbb{R}^K$, $x, y \in \mathcal{X}$ (otherwise the chain is neither aperiodic nor recurrent). A discussion of this property is found in [2], Section 2.4.2.

1. The second moments of the random times T_m satisfies

$$\mathrm{E}\left[\sum_{m=1}^{\infty}\gamma_m^2 T_m^2\right] < \infty,$$

2. $\lim_{m\to\infty}\gamma_m T_m = 0$, w.p. 1.

Proof. For the first part of the lemma we have

$$\begin{aligned}
\mathrm{E}\left[\sum_{m=1}^{\infty}\gamma_m^2 T_m^2\right] &= \lim_{k\to\infty}\mathrm{E}\left[\sum_{m=1}^{k}\gamma_m^2 T_m^2\right] \\
&= \lim_{k\to\infty}\sum_{m=1}^{k}\gamma_m^2\mathrm{E}\left[T_m^2\right] \\
&\leq B_T\sum_{m=1}^{\infty}\gamma_m^2 \\
&< \infty,
\end{aligned}$$

where in the first equality we used the Monotone Convergence Theorem. Thus, the first part of the lemma is established. We prove the second part by contradiction. Let us assume that $\mathrm{limsup}_{m\to\infty}\gamma_m T_m = \epsilon^+$, and $\mathrm{liminf}_{m\to\infty}\gamma_m T_m = \epsilon^-$. Choose $\epsilon^2 = 1/2\cdot\max\{|\epsilon^+|^2, |\epsilon^-|^2\}$. Thus, $\gamma_m^2 T_m^2$ is greater than ϵ^2 infinitely often, which yields

$$\begin{aligned}
\sum_{m=1}^{\infty}\gamma_m^2 T_m^2 &\geq \sum_{m=1}^{\infty}\gamma_m^2 T_m^2 1\left\{\gamma_m^2 T_m^2 > \epsilon^2/2\right\} \\
&\geq \epsilon^2\sum_{m=1}^{\infty}1\left\{\gamma_m^2 T_m^2 > \epsilon^2/2\right\} \\
&= \infty,
\end{aligned}$$

yielding a contradiction. □

The next lemma shows that the iterates of $\tilde{\eta}_m$ are bounded.

Lemma B.4. *The sequence $\tilde{\eta}_m$ is bounded w.p. 1.*

Proof. Using lemma B.3 we can choose M such that $\gamma_m T_m < \epsilon < 1$ for all $m > M$. Using Assumption 2.1(ii) for the boundedness of the rewards, we have

$$\begin{aligned}
\tilde{\eta}_{m+1} &= (1-\gamma_m T_m)\tilde{\eta}_m + \Gamma_\eta\gamma_m\sum_{n=t_m}^{t_{m+1}-1} r(x_n) \\
&\leq (1-\gamma_m T_m)\tilde{\eta}_m + \Gamma_\eta\gamma_m T_m B_r \\
&\leq \begin{cases} \tilde{\eta}_m & \text{if } \tilde{\eta}_m > B_r\Gamma_\eta, \\ B_r\Gamma_\eta & \text{if } \tilde{\eta}_m \leq B_r\Gamma_\eta, \end{cases} \\
&= \max\{\tilde{\eta}_m, B_r\Gamma_\eta\},
\end{aligned} \tag{B.8}$$

which means that each iterate is bounded above by the previous iterate or by a constant. Using similar arguments we can prove that $\tilde{\eta}_m$ is bounded below. □

Next, we prove the first assumption of Theorem B.1.

Lemma B.5. *The vector* $\mathrm{E}\left[|V_m(y_m)|^2\right]$ *is bounded w.p. 1.*

Proof. We prove the boundedness of the three parts of the vector $V_m(y_m)$. The mean $\mathrm{E}\left[|\tilde{\eta}|^2\right] < \infty$ since by Lemma B.4 the iterate $\tilde{\eta}_m$ is bounded. Next, we look at the mean of the squared iterate $\tilde{h}_m(x)$. For all $x \in \mathcal{X}$,

$$
\begin{aligned}
\mathrm{E}\left[|\tilde{h}_{m+1}(x)|^2\right] &= \mathrm{E}\left[\left(\tilde{h}_m + \gamma_m \Gamma_h \left(\sum_{n=N_m(x)}^{t_{m+1}-1} \tilde{d}(x_n, x_{n+1})\right)\right)^2\right] \\
&\overset{(i)}{=} \mathrm{E}\left[\left((1-\gamma_m)\tilde{h}_m(x) + \gamma_m \Gamma_h \left(\sum_{n=N_m(x)}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m)\right)\right)^2\right] \\
&\leq (1-\gamma_m)^2 \mathrm{E}\left[\tilde{h}_m^2(x)\right] + \gamma_m^2 \Gamma_h^2 \mathrm{E}\left[\left(\sum_{n=N_m(x)}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m)\right)^2\right] \\
&\quad + 2(1-\gamma_m)\gamma_m \mathrm{E}\left[\left|\tilde{h}_m(x)\right| \left|\Gamma_h \sum_{n=N_(x)}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m)\right|\right] \\
&\overset{(ii)}{\leq} (1-\gamma_m)^2 \mathrm{E}\left[\tilde{h}_m^2(x)\right] + \gamma_m^2 \Gamma_h^2 \mathrm{E}\left[\left(\sum_{n=N_m(x)}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m)\right)^2\right] \\
&\quad + 2(1-\gamma_m)\gamma_m \sqrt{\mathrm{E}\left[\tilde{h}_m^2(x)\right] \mathrm{E}\left[\left(\Gamma_h \sum_{n=N_(x)}^{t_{m+1}-1} (r(x_n) - \tilde{\eta}_m)\right)^2\right]} \\
&\leq \left((1-\gamma_m)\sqrt{\mathrm{E}\left[\tilde{h}_m^2(x)\right]} + \gamma_m \sqrt{\Gamma_h^2 B_T (B_r + B_{\tilde{\eta}})^2}\right)^2 \\
&\leq \left(\max\left\{1, \sqrt{\mathrm{E}\left[\tilde{h}_m^2(x)\right]}, \sqrt{\Gamma_h^2 B_T (B_r + B_{\tilde{\eta}})^2}\right\}\right)^2 \\
&\leq \max\left\{1, \mathrm{E}\left[\tilde{h}_m^2(x)\right], \Gamma_h^2 B_T (B_r + B_{\tilde{\eta}})^2\right\}
\end{aligned}
$$

where in (i) we used (3), and in (ii) we used the Cauchy Schwarz inequality. We see that the iterate $\mathrm{E}\left[|\tilde{h}_{m+1}(x)|^2\right]$ is bounded by the previous iterate $\mathrm{E}\left[|\tilde{h}_m(x)|^2\right]$ or by a constant, thus, $\mathrm{E}\left[|\tilde{h}_m(x)|^2\right]$ is bounded. We denote this bound by $B_{\tilde{h}}$. Therefore, we can conclude easily that $\mathrm{E}\left[V_m^2(\theta_m)\right]$ is bounded since $\mathrm{E}[\tilde{h}_m^2(x)]$ is bounded for all $x \in \mathcal{X}$. Formally,

$$
\begin{aligned}
\mathrm{E}\left[V_m^2(\theta_m)\right] &= \mathrm{E}\left[\left(\sum_{n=t_m}^{t_{m+1}-1} \tilde{d}(x_n, x_{n+1}) \psi(x_n, u_n|\theta_m)\right)^2\right]. \\
&\leq B_T B_\psi^2 (2B_{\tilde{h}} + B_{\tilde{\eta}} + B_r)^2,
\end{aligned}
$$

which concludes the proof. □

The following lemma proves the continuity of several functions and an operator which will be used in proving the continuity of $v(y)$.

Lemma B.6. *Under Assumptions 2.1(ii) and 2.2 we have*

1. *$\psi(x, u|\theta)$ is continuous with respect to θ.*
2. *$P(y|x, \theta)$ is continuous with respect to θ.*

3. *$\pi(x|\theta)$ is continuous with respect to θ.*

4. *$\eta(\theta)$ is continuous with respect to θ.*

5. *For a stopping time $T(x) = \min\{k > 0 | x_k = x\}$, and for a function $g(x, u, \theta)$ continuous for all $x \in \mathcal{X}$, $u \in \mathcal{U}$, and $\theta \in \mathbb{R}^K$, and absolutely bounded with a constant B_g, we have*

$$\mathrm{E}_\theta \left[\sum_{k=0}^{T(x)} g(x_k, u_k, \theta) \Big| x_0 \right] \tag{B.9}$$

is continuous with respect to θ.

6. *$h(x|\theta)$, $T(\theta)$, $C(\theta)$, and $D^{(x)}(\theta)$ are continuous with respect to θ.*

Proof. For brevity, all continuities in this proof are with respect to θ.

1. According to Assumption 2.1(ii) for all $x \in \mathcal{X}$ and $u \in \mathcal{U}$, $\mu(u|x, \theta)$ is twice differentiable, thus $\nabla\mu(u|x, \theta)$ is continuous and so is $\psi(u|x, \theta)$.

2. $P(y|x, \theta)$ is a compound function of an integral and continuous function, thus it is continuous.

3. According to Lemma 2.1, for each $\theta \in \mathbb{R}^K$ we have a unique solution for the following non-homogenous linear equation system in $\{\pi(i|\theta)\}_{i=1}^{|\mathcal{X}|}$

$$\begin{cases} \sum_{i=1}^{|\mathcal{X}|} \pi(i|\theta) P_{ij}(\theta) = \pi(j|\theta), \quad j = 1, \ldots, |\mathcal{X}| - 1, \\ \sum_{i=1}^{|\mathcal{X}|} \pi(i|\theta) = 1, \end{cases}$$

   or in matrix form $M(\theta)\pi(\theta) = b$. Thus, using Cramer's rule we have $\pi(i|\theta) = Q(i, \theta)/\det[M(\theta)]$, where $Q(i, \theta)$ and $\det[M(\theta)]$ are polynomials function of entries in $M(\theta)$, thus $\pi(\theta)$ is continuous. We note that $\det[M(\theta)]$ is not zero since by Assumption 2.1(i) the system has a unique solution for all $\theta \in \mathbb{R}^K$.

4. The variable $\eta(\theta)$ is a linear combination of continuous functions, thus continuous.

5. For a fixed $N = 0, 1, \ldots$ we have

$$\begin{aligned} \mathrm{E}_\theta \left[\sum_{k=0}^{N} g(x_k, u_k, \theta) \Big| x_0 \right] = \sum_{\nu_0 \in \mathcal{U}, \ldots, \zeta_N \in \mathcal{X}, \nu_N \in \mathcal{U}} & \Pr\left(u_0 = \nu_0, \ldots, x_k = \zeta_k, u_N = \nu_N | \, x_0, \theta\right) \\ & \times \left(g(x_0, \nu_o, \theta) + \ldots + g(\zeta_N, \nu_N, \theta)\right) \end{aligned} \tag{B.10}$$

   Both $\pi(\theta)$ and $P(\theta)$ are continuous, therefore $\Pr\left(x_0 = \zeta_0, \ldots, x_k = \zeta_k | \, \theta\right)$ is continuous. We have a finite sum of continuous functions thus $\mathrm{E}_\theta \left[\sum_{k=0}^{N} g(x_k, u_k, \theta) | x_0\right]$ is continuous. Also, looking at (B.10) we see that $\mathrm{E}_\theta \left[\sum_{k=0}^{N} g(x_k, u_k, \theta) | x_0\right]$ is bounded by $B_g(N + 1)$. Define $T_k(x) \triangleq \{x_1 \neq y, \ldots, x_{k-1} \neq y, x_k = y | x_0 = x\}$ for a fixed $k = 0, 1, \ldots$. As in Lemma B.2, there exists a constant $b_T(\theta)$, $0 < b_T(\theta) < 1$, such that for a fixed k we have $\Pr(T_k(x)) \leq b_T^k(\theta)$. Since $\mathcal{P}$ is a closed set, we can find a constant $\tilde{b}_T$, $0 < \tilde{b}_T < 1$, such that $T_k(x) \leq \tilde{b}_T^k$. Define

$$G_N(\theta) \triangleq \sum_{k=0}^{N} \Pr(T_k(x)) \mathrm{E}_\theta \left[\sum_{l=0}^{k} g(x_l, u_l, \theta) \Big| x_0 \right]. \tag{B.11}$$

We claim that the function series $\{G_N(\theta)\}_{N=0}^{\infty}$ convergence uniformly[3] using the Cauchy Criterion for uniform convergence[4], for all $\epsilon > 0$ and N satisfying $2B_g\tilde{b}_T^N/(1-\tilde{b}_T)^2 < \epsilon$. Mathematically,

$$\begin{aligned}\|G_{N+p}(\theta) - G_N(\theta)\| &= \left\| \sum_{k=N+1}^{N+p} \Pr(T_k(x)) \mathrm{E}_\theta \left[\sum_{l=0}^{k} g(x_l, u_l, \theta) \Big| x_0 \right] \right\| \\ &\leq \sum_{k=N+1}^{\infty} \tilde{b}_T^k B_g(k+1) \\ &\leq \frac{2B_g\tilde{b}_T^N}{(1-\tilde{b}_T)^2} \\ &\leq \epsilon.\end{aligned}$$

Thus, we can express $\mathrm{E}_\theta \left[\sum_{k=0}^{T(x)} g(x_k, u_k, \theta) | x_0 \right]$ as

$$\begin{aligned}\mathrm{E}_\theta \left[\sum_{k=0}^{T(x)} g(x_k, u_k, \theta) \Big| x_0 \right] &= \mathrm{E}_\theta \left[\mathrm{E}_\theta \left[\sum_{k=0}^{T(x)} g(x_k, u_k, \theta) \Big| x_0 \right] \Big| T(x) \right] \\ &= \lim_{n\to\infty} \sum_{k=0}^{N} \Pr(T_k(x)) \mathrm{E}_\theta \left[\sum_{l=0}^{k} g(x_l, u_l, \theta) \Big| x_0 \right],\end{aligned}$$

which is continuous since it is a limit of a continuous function series which converges uniformly[5].

6. The continuity of $h(x|\theta)$, $T(\theta)$, $C(\theta)$, and $D^{(x)}(\theta)$ follows immediately from 5.

□

Based on the assumptions of Algorithm 1 and Lemmas B.5, and B.6 we can conclude that the system converges to the desired system of ODE.

C Proof of Theorem 2.6

In this section we find conditions under which Algorithm 1 converges to a neighborhood of a local maximum, and more precisely, conditions that $\limsup_{t\to\infty} \|\nabla\eta(\theta(t))\| \leq \epsilon$, for arbitrary positive ϵ. We begin by establishing a bound on a time dependent ODE of the first order. Then, we prove some properties of the ODE system (5), and conclude by proving the section's main result.

The following lemma will be used later for the linear equations (5), i.e., on the ODEs for $\tilde{\eta}$ and $\tilde{h}(x)$.

Lemma C.1. *Assume the following time dependent ODE*

$$\begin{cases} \dfrac{d}{dt}x(t) = \dfrac{1}{\tau}\left(f(t) - x(t)\right), \\ x(0) = x_0, \end{cases} \tag{C.1}$$

where $df(t)/dt \leq B_f$. *Then,* $\lim_{t\to 0} |x(t) - f(t)| \leq B_f\tau^2$.

[3]We say that a function series $\{G_N(\theta)\}$ *converges uniformly* to $G(\theta)$ on a set E if $\forall\epsilon > 0$, $\exists M(\epsilon)$ such that $\forall N > M(\epsilon)$ and $\forall\theta \in E$ we have $\|G_N(\theta) - G(\theta)\| < \epsilon$.

[4]Cauchy Criteria: A function series $\{G_N(\theta)\}$ converges uniformly in a set E if and only if $\forall\epsilon > 0$, $\exists M(\epsilon)$ such that $\forall N > M(\epsilon)$, $\forall p \geq 1$, and $\forall\theta \in E$ we have $\|G_{N+p} - G_N\| < \epsilon$.

[5]If a continuous function series $\{G_N(\theta)\}$ converges uniformly to $G(\theta)$ on a set E then $G(\theta)$ is continuous on E.

Proof. We write (C.1) in the following way

$$\tau \frac{d}{dt}(x(t) - f(t)) = -(x(t) - f(t)) - \tau \frac{df(t)}{dt}.$$

We define $z(t) \triangleq x(t) - f(t)$ and $g(t) \triangleq -\tau df(t)/dt$, where $|g(t)| \leq \tau B_f$ and $z(0) = z_0 \triangleq x_0 - f(0)$. Thus, we have

$$\begin{cases} \tau \frac{d}{dt} z(t) = -z(t) + g(t), \\ z(0) = z_0. \end{cases}$$

The solution of this ODE is

$$z(t) = z_0 e^{-t/\tau} + e^{-t/\tau} \int_0^t g(s) e^{s/\tau} ds.$$

Thus,

$$\begin{aligned} |z(t)| &\leq |z_0| \, e^{-t/\tau} + e^{-t/\tau} \int_0^t |g(s)| \, e^{s/\tau} ds \\ &\leq |z_0| \, e^{-t/\tau} + B_f \tau^2 \left(1 - e^{-t/\tau}\right), \end{aligned}$$

and taking the limit $t \to \infty$ completes the proof. □

In the following lemma, we prove several properties of the algorithm, which will be used later.

Lemma C.2.

1. *$\nabla \pi(\theta)$ is uniformly bounded.*
2. *$\nabla \eta(\theta)$ is uniformly bounded.*
3. *$\nabla h(x|\theta)$ is uniformly bounded.*
4. *$\dot{\theta}$ is uniformly bounded.*
5. *$\dot{\eta}(\theta)$ is uniformly bounded.*
6. *$\dot{h}(x|\theta)$ is uniformly bounded, for all $x \in \mathcal{X}$.*
7. *$\|C(\theta)\|$ is uniformly bounded.*
8. *$\left\|D^{(x)}(\theta)\right\|$ is uniformly bounded, for all $x \in \mathcal{X}$.*

Proof. 1. From the proof of Lemma B.6, we can write

$$\pi(i|\theta) = Q(i, \theta) / \det[M(\theta)].$$

We note that $\det[M(\theta)] \neq 0$ for all $P(\theta) \in \bar{\mathcal{P}}$, thus there exists a constant, m_0, which satisfies $|M(\theta)| \geq m_0 > 0$, for all $\theta \in \mathbb{R}^K$. In addition, $Q(i, \theta)$ and $\det[M(\theta)]$ are differentiable functions with respect to θ, $\forall x, y \in \mathcal{X}$, and $P(y|x, \theta)$ is a differentiable function with respect to θ with bounded derivative. Therefore, $\|\nabla Q(i, \theta)\|$ and $\|\nabla \det[M(\theta)]\|$ are bounded functions. Thus, we can conclude that

$$\nabla \pi(i|\theta) = \frac{Q(i, \theta) \nabla \det[M(\theta)] - \nabla Q(i, \theta) \det[M(\theta)]}{\det[M(\theta)^2}, \quad i \in \mathcal{X},$$

is bounded. We denote this bound by $B_{\nabla \pi}$.

2. We have $\nabla \eta(\theta) = \sum_{x \in \mathcal{X}} r(x) \nabla \pi(x|\theta) \leq |\mathcal{X}| B_r B_{\nabla \pi}$.

3. We recall the Poisson equation (2). We can write the following system of linear equations in $\{h(x|\theta)\}_{x\in\mathcal{X}}$, namely,

$$\begin{cases} h(x|\theta) = r(x) - \eta(\theta) + \sum_{y\in\mathcal{X}} P(y|x,\theta)h(y|\theta), \quad \forall x \in \mathcal{X}, x \neq x^*, \\ h(x^*|\theta) = 0. \end{cases} \tag{C.2}$$

or in matrix form $N(\theta)h(\theta) = c$. Adding the equation $h(x^*|\theta) = 0$ yields a unique solution for the system (see, for example, Prop. 7.4.1 in [1], Vol. 1). Thus, using Cramer's rule we have $h(x|\theta) = R(x,\theta)/\det[N(\theta)]$, where $R(x,\theta)$ and $\det[N(\theta)]$ are polynomial functions of entries in $N(\theta)$, which are uniformly bounded, and have uniformly bounded derivatives. Continuing in the same steps of the proof of the previous items, we conclude that $\nabla h(x|\theta)$ is uniformly bounded for all $x \in \mathcal{X}$.

4. Looking at the equation for $\dot{\theta}$ in (5), we see that the r.h.s of the equation is composed of bounded terms. We denote this bound by $B_{\dot{\theta}}$.

5. It immediately follows that

$$|\dot{\eta}(\theta)| = |\nabla\eta(\theta)\cdot\dot{\theta}| \leq B_{\nabla\eta}B_{\dot{\theta}}, \quad \forall\theta \in \mathbb{R}^K. \tag{C.3}$$

We denote this bound by $B_{\dot{\eta}}$.

6. It immediately follows that

$$|\dot{h}(x|\theta)| = |\nabla h(x|\theta)\cdot\dot{\theta}| \leq B_{\nabla h}B_{\dot{\theta}}, \quad \forall x \in \mathcal{X}, \quad \theta \in \mathbb{R}^K. \tag{C.4}$$

We denote this bound by $B_{\dot{h}}$.

7. It immediately follows that

$$\|C(\theta)\| = \left\| \mathrm{E}_\theta \left[\sum_{n=0}^{T-1} \psi(x_n, u_n|\theta) \right] \right\| \leq B_T B_\psi.$$

We denote this bound by B_C.

8. It immediately follows that

$$\left\| D^{(x)}(\theta) \right\| = \left\| \mathrm{E}_\theta \left[\sum_{n=0}^{T-1} 1\{x_{n+1} = x\} \psi(x_n, u_n|\theta) \right] \right\| \leq B_T B_\psi, \quad x \in \mathcal{X}.$$

We denote this bound by B_D.

□

The following lemma establishes the main result of this section. It states the conditions under which the ODE system (5) converges to some neighborhood of a stationary point.

Lemma C.3. *If we choose $\Gamma_\eta \geq B_{\dot{\eta}}^2/\epsilon_\eta$ and $\Gamma_h \geq B_{\dot{h}}^2/\epsilon_h$, for some positive ϵ_h and ϵ_η, then*

$$\limsup_{t\to\infty} \|\nabla\eta(\theta(t))\| \leq \epsilon, \tag{C.5}$$

where $\epsilon \triangleq B_C\epsilon_\eta + |\mathcal{X}|B_D\epsilon_h$.

Proof. Using the boundedness of $\dot{\eta}(\theta)$ and $\dot{h}(x\theta)$ from Lemma C.2, and the assumed lower bounds on Γ_h and Γ_η, we apply Lemma C.1 to the variables $\tilde{\eta}$ and $\tilde{h}$. We conclude that there exists a time t_0 such that for all $t \geq t_0$, $|\eta(\theta(t)) - \tilde{\eta}(t)| \leq \epsilon_\eta$ and $|h(x|\theta(t)) - \tilde{h}(x,t)| \leq \epsilon_h$, for all $x \in \mathcal{X}$. Define the

ball $B_\epsilon \triangleq \{\theta : \ \|\nabla\eta(\theta)\| \leq \epsilon\}$, and consider a trajectory starting inside the ball B_ϵ at time t_0. We claim that the trajectory must remain in the ball for all $t \geq t_0$. Assume the trajectory enters the set $S_{\epsilon,\epsilon_2} \triangleq \{\theta : \ \epsilon < \|\nabla\eta(\theta)\| \leq \epsilon + \epsilon_2\}$ at $t_1 > t_0$. Thus, at $t = t_1$ we have

$$\begin{aligned}
\dot{\eta}(\theta) &= \nabla\eta(\theta) \cdot \dot{\theta} \\
&= \nabla\eta(\theta) \cdot \left(T(\theta)\nabla\eta(\theta) + C(\theta)\left(\eta(\theta) - \tilde{\eta}\right) + \sum_{x\in\mathcal{X}} D^{(x)}(\theta)\left(h(x|\theta) - \tilde{h}(x)\right)\right) \\
&\geq \|\nabla\eta(\theta)\|\left(\|\nabla\eta(\theta)\| - G_C\epsilon_\eta - |\mathcal{X}|G_D\epsilon_h\right) \\
&= \|\nabla\eta(\theta)\|\left(\|\nabla\eta(\theta)\| - \epsilon\right) \\
&> 0,
\end{aligned} \tag{C.6}$$

This implies that for $t \geq t_0$ the trajectory does not leave the set S_{ϵ,ϵ_2}. Since this holds for any $\epsilon_2 > 0$, the trajectory never leaves B_ϵ.

Using similar arguments, if at $t = t_0$ we have $\|\nabla\eta(\theta(t))\| > \epsilon$, there exists a time t_1 which $\|\nabla\eta(\theta(t))\| = \epsilon$. Using the claim starting from time $t = t_1$ completes the proof.

□

References

[1] D.P. Bertsekas. *Dynamic Programming and Optimal Control, Vol I.*, 3rd Ed. Athena Scinetific, 2006.

[2] P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer, 1999.

[3] X.R. Cao and H.F. Chen. Pertubation realization, potentials, and sensitivity analysis of markov processes. *IEEE Trans. Automat. Contr*, 42:13821393, 1997.

[4] T. Jaakkola, S.P. Singh, and M.I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *In advances in Neural Information Processing Systems*, volume 7, pages 14681502, San Francisco, CA: Morgan Kaufman, 1995.

[5] H.J. Kushner and G.G. Yin. *Stochastic Approximation Algorithms and Applications.* Springer, 1997.

[6] P. Marbach and J. Tsitsiklis. Simulation-Based Optimization of Markov Reward Processes. *IEEE. Trans. Auto. Cont.*, 46:191–209, 1998.