Supporting Material: Weighted Likelihood Policy Search with Model Selection

Tsuyoshi Ueno *
Japan Science and Technology Agency
ueno@ar.sanken.osaka-u.ac.jp

Kohei Hayashi
University of Tokyo
hayashi.kohei@gmail.com

Takashi Washio
Osaka University
washio@ar.sanken.osaka-u.ac.jp

Yoshinobu Kawahara
Osaka University
kawahara@ar.sanken.osaka-u.ac.jp

Abstract

This supplement provides proofs of Theorem 2 and Lemmas 1-4 in the main text.

1 Proof of Lemma 1

Using the Lagrange multipliers, the function q^* is derived by maximizing the Lagrangian defined by

$$\mathcal{L}(q,\lambda) = \iint q(x_{2:n},u_{1:n}|x_1)\left\{\ln\frac{p_\theta(x_{2:n},u_{1:n}|x_1)R_n}{q(x_{2:n},u_{1:n}|x_1)}\right\}\mathrm{d}x_{2:n}\mathrm{d}u_{1:n} + \lambda\left\{\iint q(x_{2:n},u_{1:n}|x_1)\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}-1\right\}$$

where λ is a Lagrange multiplier. We first derive the function q^* by using the calculus of variations. Let $\delta := \delta(x_{2:n},u_{1:n}|x_1)$ be an arbitrary function of $x_{2:n}$ and $u_{1:n}$ given x_1. We consider how much the functional $\mathcal{L}(q,\lambda)$ changes when we add a small changes $h\delta$ to the function $q(x_{2:n},u_{1:n}|x_1)$. For notational convenience, we define $\mathcal{G}(h;\delta,q,\lambda) := \mathcal{L}(q+\delta,\lambda)$[1]. If the function is twice differentiable with respect to h, then we have

$$\mathcal{G}(h;\delta,q,\lambda) = \mathcal{G}(0;\delta,q,\lambda) + h\cdot\frac{\partial}{\partial h}\mathcal{G}(0;\delta,q,\lambda)\bigg|_{h=0} + O(h^2).$$

Since the function $q^*(x_{2:n},u_{1:n}|x_1)$ must satisfy that the functional is stationary with respect to small variations in the function q, ***i.e.***,

$$\frac{\partial}{\partial h}\mathcal{G}(h;\delta,q,\lambda)\bigg|_{h=0} = 0,$$

for any choice δ. The derivative of $\mathcal{G}(h;\delta,q,\lambda)$ with respect to h can be obtained by

$$\begin{aligned}&\frac{\partial}{\partial h}\mathcal{G}(h;\delta,q,\lambda)\\ &= \iint \delta(x_{2:n},u_{1:n}|x_1)\ln\frac{p_\theta(x_{2:n},u_{1:n}|x_1)\{R_n\}}{q(x_{2:n},u_{1:n}|x_1)}\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}\\ &- \iint q(x_{2:n},u_{1:n}|x_1)\frac{\delta(x_{2:n},u_{1:n}|x_1)}{q(x_{2:n},u_{1:n}|x_1)+h\delta(x_{2:n},u_{1:n}|x_1)}\mathrm{d}x_{2:n} + \lambda\iint\delta(x_{2:n},u_{1:n}|x_1)\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}.\end{aligned}$$

*https://sites.google.com/site/tsuyoshiueno/

[1]We used this notation to emphasize that $G(h;\delta,q,\lambda)$ is a function of h, while δ, q and λ are regarded as auxiliary variables.

Substituting h into 0, we obtain the following equation:

$$\iint \delta(x_{2:n}, u_{1:n}|x_1) \left\{ \ln \frac{p_\theta(x_{2:n}, u_{1:n}|x_1)\{R_n\}}{q(x_{2:n}, u_{1:n}|x_1)} - 1 + \lambda \right\} = 0. \tag{S.1}$$

Since the function $q^*(x_{2:n}, u_{1:n}|x_1)$ satisfies Eq. (S.1) for any variation $\delta(x_{2:n}, u_{1:n}|x_1)$, we can derive

$$q^*(x_{2:n}, u_{1:n}|x_1) = \exp[-1 + \lambda] p_\theta(x_{2:n}, u_{1:n}|x_1)\{R_n\}. \tag{S.2}$$

The Lagrange multipliers can be solved from the constraint:

$$\begin{aligned} 1 &= \int q^*(x_{2:n}, u_{1:n}|x_1) \mathrm{d}x_{2:n} \mathrm{d}u_{1:n} = \exp\left[-1 + \lambda\right] \iint p_\theta(x_{2:n}, u_{1:n}|x_1)\{R_n\} \mathrm{d}x_{2:n} \mathrm{d}u_{1:n} \\ \lambda &= 1 - \ln \iint p_\theta(x_{2:n}, u_{1:n}|x_1)\{R_n\} \mathrm{d}x_{2:n} \mathrm{d}u_{1:n} \end{aligned} \tag{S.3}$$

By plugging Eq. (S.3) into Eq. (S.2), we can conclude

$$q^*(x_{2:n}, u_{1:n}|x_1) = \frac{p_\theta(x_{2:n}, u_{1:n}|x_1)\{R_n\}}{\iint p_\theta(x_{2:n}, u_{1:n}|x_1)\{R_n\} \mathrm{d}x_{2:n} \mathrm{d}u_{1:n}}.$$

2 Proof of Lemma 2

According to Theorem 5.9 in [6], if the following conditions

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i}^{n} \beta^{j-i} \psi_\theta(x_1, u_1) r(x_j, u_j) - \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}} \left[\psi_\theta(x_1, u_1) \sum_{j=1}^{\infty} \beta^{j-1} r(x_j, u_j) \right] \right| \xrightarrow{p} 0 \tag{S.4}$$

$$\inf_{\theta : |\theta - \bar{\theta}| > \epsilon} \left| \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}} \left[\psi_\theta(x_1, u_1) \sum_{j=1}^{\infty} \beta^{i-1} r(x_j, u_j) \right] \right| > \left| \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}} \left[\psi_{\bar{\theta}}(x_1, u_1) \sum_{i=1}^{\infty} \beta^{j-1} r(x_j, u_j) \right] \right| = 0 \tag{S.5}$$

hold for any $\{\theta, \theta'\} \in \Theta \times \Theta$ and $\epsilon > 0$, then any sequence of $\hat{\theta}_n$ such that $G_n^{\theta'}(\hat{\theta}_n) = 0$ converges to the parameter $\bar{\theta}$ in probability. It is obvious that condition (S.5) is satisfied from Assumption 7, thus we discuss whether condition (S.4) is satisfied or not.

Let us consider a stochastic process $\{y_i : i = \{1, 2, \cdots\}\}$ defined by

$$y_i := \psi_\theta(x_i, u_i) \sum_{j=i}^{\infty} \beta^{j-i} r(x_j, u_j).$$

Since the MDP given by Eq. (1) is ergodic, the stochastic process $\{y_i\}$ is also ergodic by the following lemma.

Lemma S.1 *[2, Proposition 6.6]*
Let $\{x_i \in \mathcal{X} : i := \{1, 2, \cdots\}\}$ be a strictly stationary and ergodic stochastic process, and let $\{y_i : i = \{1, 2, \cdots, \}\}$ be a stochastic process defined by

$$y_i := f\left(x_i, x_{i+1}, \cdots\right),$$

where $f : \mathcal{X} \times \mathcal{X} \times \cdots \mapsto \mathbb{R}$ is an arbitrary function. Then, the process $\{y_i\}$ is also ergodic.

From Assumptions 2 and 5, $\left| \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}} \left[\psi_\theta(x_1, u_1) \sum_{j=1}^{\infty} \beta^{j-1} r(x_j, u_j) \right] \right|$ is bounded for any $\{\theta, \theta'\} \in \Theta \times \Theta$. Then, by the pointwise ergodic theorem shown in Theorem 24.1 in [1], we obtain

$$\frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i}^{\infty} \beta^{j-i} \psi_\theta(x_1, u_1) r(x_j, u_j) \xrightarrow{a.s.} \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}} \left[\psi_\theta(x_1, u_1) \sum_{j=1}^{\infty} \beta^{j-1} r(x_j, u_j) \right], \quad \forall \{\theta, \theta'\} \in \Theta \times \Theta,$$

where $\xrightarrow{a.s.}$ denotes the convergence almost surely. The left hand side of the above equation can be expressed as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{\infty}\beta^{j-i}\psi_\theta(x_i,u_i)r(x_j,u_j)\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_\theta(x_i,u_i)r(x_j,u_j)+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=n+1}^{\infty}\psi_\theta(x_i,u_i)\beta^{j-i}r(x_j,u_j).
\end{aligned}$$

About the second term in the right hand side of the above equation, we observe

$$\begin{aligned}
&\frac{1}{n}\left|\sum_{i=1}^{n}\sum_{j=n+1}^{\infty}\beta^{j-i}\psi_\theta(x_i,u_i)r(x_j,u_j)\right|\\
&\leq\frac{1}{n}\sum_{i=1}^{n}\sum_{j=n+1}^{\infty}\beta^{j-i}\left|\psi_\theta(x_i,u_i)r(x_j,u_j)\right|\\
&\leq\frac{BC}{n}\sum_{i=1}^{n}\sum_{j=n+1}^{\infty}\beta^{j-i}=\frac{BC}{n}\sum_{i=1}^{n}\frac{\beta^{n+1-i}}{1-\beta}=\frac{BC}{n}\frac{\beta(1-\beta^n)}{(1-\beta)^2}\xrightarrow{n\to\infty}0,
\end{aligned}$$

where B and C denote such constants that $B:=\sup|\psi_\theta(x,u,\theta)|$ and $C:=\sup|r(x,u)|$ for any $x\in\mathcal{X}$, $u\in\mathcal{U}$ and $\theta\in\Theta$, respectively, which are guaranteed to be bounded according to Assumptions 2 and 5. Therefore, the uniform law of large numbers of $G_n^{\theta'}(\theta)$ is proved, *i.e.*,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta^{j-i}\psi_\theta(x_i,u_i)r(x_j,u_j)\xrightarrow{a.s.}\mathbb{E}_{x_1\sim\mu_{\theta'}}^{\pi_{\theta'}}\left[\psi_\theta(x_1,u_1)\sum_{j=1}^{\infty}\beta^{j-1}r(x_j,u_j)\right],$$

for any $\{\theta,\theta'\}\in\Theta\times\Theta$.

3 Proof of Lemma 3

Applying the Taylor series expansion to estimating equation (7) around the parameter $\bar{\theta}$, we obtain

$$0=\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)+\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\left(\hat{\theta}_n-\bar{\theta}\right)+O_p\left(\left\|\hat{\theta}_n-\bar{\theta}\right\|^2\right),\tag{S.6}$$

Here, high order terms are in total represented as $O_p(\|\hat{\theta}_n-\bar{\theta}\|^2)$ because of the thrice differentiable condition for the function $\pi_\theta(u|x)$ described in Assumption 3. From the assumption in Lemma 3, the estimator $\hat{\theta}_n$ converges to $\bar{\theta}$: $\hat{\theta}_n=\bar{\theta}+o_p(1)$, thus Eq. (S.6) can be rewritten as

$$0=\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)+\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\left(\hat{\theta}_n-\bar{\theta}\right)+o_p\left(1\right).$$

After easy calculation, we derive

$$\sqrt{n}\left(\hat{\theta}_n-\bar{\theta}\right)=-\frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)^{-1}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)+o_p\left(\frac{1}{\sqrt{n}}\right).\tag{S.7}$$

Note that $(1/n)\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)$ satisfies the law of large numbers shown in the following lemma.

Lemma S.2

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\xrightarrow{a.s.}\mathrm{A},\ \ \forall\theta'\in\Theta$$

The proof of Lemma S.2 just follows the proof of Lemma 2 given in Section 2, hence we omit the proof of Lemma S.2. Lemma S.2 implies that $(1/n)\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_i,u_i) = \mathrm{A}+o_p(1)$. From the continues mapping theorem shown in Theorem 2.3 in [6] and Assumption 8, the inverse of $(1/n)\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_i,u_i)$ also converges to A^{-1} almost surely:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\mathrm{K}_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)^{-1} = \mathrm{A}^{-1}+o_p(1). \tag{S.8}$$

Substituting Eq. (S.8) into Eq. (S.7), we have

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n-\bar{\theta}) &= \left(\mathrm{A}^{-1}+o_p(1)\right)\left(-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)+o_p\left(\frac{1}{\sqrt{n}}\right)\\ &= \left(-\frac{1}{\sqrt{n}}\mathrm{A}^{-1}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)\\ &\quad+o_p\left(1\right)\cdot\left(-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)+o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{S.9}$$

We now introduce the following support lemma.

Lemma S.3

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\xrightarrow{d.}N(0,\Sigma),$$

where $\xrightarrow{d}$ denotes the weak convergence (convergence in distribution).

Since the random variable with weak convergence is bounded in probability, Eq. (S.9) can be expressed as

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n-\bar{\theta}) &= \left(-\frac{1}{\sqrt{n}}\mathrm{A}^{-1}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)\\ &\quad+o_p\left(1\right)\cdot\underbrace{\left(-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)\right)}_{=O_p(1)}+o_p\left(\frac{1}{\sqrt{n}}\right)\\ &= -\frac{1}{\sqrt{n}}\mathrm{A}^{-1}\sum_{i=1}^{n}\sum_{j=i}^{n}\beta^{j-i}\psi_{\bar{\theta}}(x_i,u_i)r(x_j,u_j)+o_p(1),\end{aligned}$$

hence we have proved Eq. (9) in Lemma 3. Furthermore, we can derive

$$\sqrt{n}\left(\hat{\theta}_n-\bar{\theta}\right)\sim N\left(0,\mathrm{A}^{-1}\Sigma(\mathrm{A}^{-1})^{\top}\right).$$

4 Proof of Lemma 4

The partial derivative of the lower bound of the expected reward with $q^*_{\theta'}(x_{2:n}, u_{1:n}|x_1)$, *i.e.*, $(\partial/\partial\theta)\mathcal{F}_n(q^*_{\theta'}, \theta)$, can be rewritten by

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathcal{F}_n(q^*_{\theta'},\theta) &= \iint p_{\theta'}(x_{2:n},u_{1:n}|x_1)\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_\theta(x_i,u_i)r(x_j,u_j)\right\}\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}\\ &= \iint p_{\theta'}(x_{2:n},u_{1:n}|x_1)\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\psi_\theta(x_i,u_i)r(x_j,u_j)\right\}\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\iint p_{\theta'}(x_{2:n},u_{1:n}|x_1)\psi_\theta(x_i,u_i)r(x_j,u_j)\mathrm{d}x_{2:n}\mathrm{d}u_{1:n},\end{aligned}$$

where we have used the well-known fact $\iint \pi_\theta(u|x)\psi_\theta(x,u)\mathrm{d}u = 0$ [5].

Let us consider a series S_n defined as

$$S_n = \lim_{n'\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n'}\iint p_{\theta'}(x_{2:n'},u_{1:n'}|x_1)\psi_\theta(x_i,u_i)r(x_j,u_j)\mathrm{d}x_{2:n'}\mathrm{d}u_{1:n'}$$

From Assumption 1, the MDP given by Eq. (1) is ergodic for any θ', then the series S_n converges to

$$S_n \xrightarrow{n\to\infty} \sum_{j=1}^{\infty}\mathbb{E}^{\pi_{\theta'}}_{x_1\sim\mu_{\theta'}}\left[\psi_\theta(x_1,u_1)r(x_j,u_j)\right].$$

Now, we show that the series S_n is corresponding to the partial derivative $(\partial/\partial\theta)\mathcal{F}(q^*_{\theta'},\theta)$ as n goes to ∞. The series S_n can be decomposed as

$$\begin{aligned}S_n = &\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\iint p_{\theta'}(x_{2:n},u_{1:n}|x_1)\psi_\theta(x_i,u_i)r(x_j,u_j)\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}}_{:=(\partial/\partial\theta)\mathcal{F}_n(q^*_{\theta'},\theta)}\\ &+\lim_{n'\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=n+1}^{n'}\iint p_{\theta'}(x_{2:n'},u_{1:n'}|x_1)\psi_\theta(x_i,u_i)r(x_j,u_j)\mathrm{d}x_{2:n'}\mathrm{d}u_{1:n'}. \qquad \text{(S.10)}\end{aligned}$$

In order to calculate the bound of the second term of Eq. (S.10), we introduce the following support lemma for the covariance bound in the stochastic process with uniform mixing.

Lemma S.4 *[4, Theorem 17.2.3.] Suppose that $\{y_i : i = \{\cdots,-1,0,1,\cdots\}\}$ is a strictly stationary process on probabilistic space $(\Omega,\mathcal{F},P)$ with uniform mixing. Let f and g be measurable functions with respect to $\mathcal{F}^k_{-\infty}$ and $\mathcal{F}^\infty_{k+s}$, respectively. If f and g satisfy*

$$\mathbb{E}\left[|f|^p\right] < \infty,\ \ \mathbb{E}\left[|g|^q\right] < \infty,$$

where $p, q > 1$, $p + q = 1$, then

$$\left|\mathbb{E}\left[fg\right] - \mathbb{E}\left[f\right]\mathbb{E}\left[g\right]\right| \le 2\varphi(s)^{1/p}\mathbb{E}\left[|f|^p\right]^{1/p}\mathbb{E}\left[|g|^q\right]^{1/q}.$$

Here, $\mathbb{E}[\cdot]$ denote the expectation over the sample sequence.

Note that from Assumption 4, the MDP satisfies geometrically uniform mixing, *i.e.*, the mixing coefficient decays exponentially fast: $\varphi(s) < D\rho^s$, where $D > 0$ and $\rho \in [0,1)$ are some positive constants. Also note that, from Assumption 2 and 5, there exist some constants B and C such that $\sup|r(x,u)| = B$ and $\sup|\psi_\theta(x,u)| = C$ for any $x \in \mathcal{X}$, $u \in \mathcal{U}$ and $\theta \in \Theta$. From these observations, we have

$$\left|\iint p_{\theta'}(x_{2:n},u_{1:n}|x_1)\psi_\theta(x_i,u_i)r(x_{j+s},u_{j+s})\mathrm{d}x_{2:n}\mathrm{d}u_{1:n}\right| \le 2BCD\rho^s, \qquad \text{(S.11)}$$

where we have also used the fact $\iint \pi_\theta(u|x)\psi_\theta(x,u)\mathrm{d}u = 0$. Thus, using covariance bound (S.11), the second term of Eq. (S.10) can be bounded by

$$\begin{aligned}
&\lim_{n'\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=n+1}^{n'} \iint p_{\theta'}(x_{2:n'}, u_{1:n'}|x_1)\psi_\theta(x_i,u_i)r(x_j,u_j)\mathrm{d}x_{2:n}\mathrm{d}u_{1:n} \right| \\
&\leq \lim_{n'\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=n+1}^{n'} \left| \iint p_{\theta'}(x_{2:n'}, u_{1:n'}|x_1)\psi_\theta(x_i,u_i)r(x_j,u_j)\mathrm{d}x_{2:n}\mathrm{d}u_{1:n} \right| \\
&\leq \lim_{n'\to\infty} \frac{2BCD}{n} \sum_{i=1}^{n} \sum_{j=n+1}^{n'} \rho^{j-i} = \frac{2BCD}{n} \sum_{i=1}^{n} \frac{\rho^{n+1-i}}{1-\rho} = \frac{2BCD}{n} \frac{\rho(1-\rho^n)}{(1-\rho)} \xrightarrow{n\to\infty} 0.
\end{aligned}$$

This result implies that

$$\lim_{n\to\infty} \frac{\partial}{\partial\theta} \mathcal{F}_n(q^*_{\theta'}, \theta) = \sum_{j=1}^{\infty} \mathbb{E}^{\pi_{\theta'}}_{x_1\sim\mu_{\theta'}} \left[\psi_\theta(x_1,u_1)r(x_j,u_j)\right] = \sum_{j=1}^{\infty} a_j \tag{S.12}$$

We introduce Abel's theorem for the power series shown in the following theorem.

Theorem S.5 (Abel's Theorem) *[3, Theorem 18] Let $\{w_i : i \in \{0,1,\cdots\}\}$ be any sequence of real or complex numbers and let*

$$G(\gamma) := \sum_{i=0}^{\infty} w_i \gamma^i$$

be the power series with coefficients w. Suppose that the series $\sum_{i=0}^{\infty} w_i$ converges to S. Then

$$\lim_{\gamma\to 1^-} G(\gamma) = S.$$

From Theorem S.5, if $\sum_{j=1}^{\infty} a_j$ converges, the power series $\sum_{j=1}^{\infty}$ converges to the result of $\sum_{j=1}^{\infty} a_j$ when β approaches 1 from below. Using covariance bound (S.11) again, the convergence of $\sum_{j=1}^{\infty} a_j$ can be easily shown. Thus, we derive

$$\frac{\partial}{\partial\theta}\mathcal{F}(q^*,\theta) = \lim_{\beta\to 1^-} \sum_{j=1}^{\infty} \beta^{i-1} \mathbb{E}^{\pi_{\theta'}}_{x_1\sim\mu_{\theta'}} \left[\psi(x_1,u_1,\theta)r(x_j,u_j)\right].$$

We now consider the series $\{b_n := \psi(x_1,u_1)\sum_{j=1}^{n}\beta^{j-1}r(x_j,u_j) : n = \{1,2,\cdots,\}\}$. Since the sequence b_n converges and $|b_n|$ is dominated for all numbers n from Assumptions 2 and 5, we can exchange the limit of the number n for the expectation by using Lebesgue's dominated convergence theorem. As a consequence it can be shown that

$$\frac{\partial}{\partial\theta}\mathcal{F}(q^*_{\theta'},\theta) = \lim_{\beta\to 1^-} \mathbb{E}^{\pi_{\theta'}}_{x_1\sim\mu_{\theta'}} \left[\psi(x_1,u_1)\sum_{i=1}^{\infty}\beta^{i-1}r(x_i,u_i)\right].$$

5 Proof of Theorem 2

The marginal weighted likelihood $\hat{p}_{\theta'}(x_{2:n}, u_{1:n}|x_1)$ can be expressed as

$$\begin{aligned}
\hat{p}_{\theta'}(x_{2:n}, u_{1:n}|x_1) &= \int \pi_\theta(u_1|x_1)^{Q_1^\beta} \prod_{i=2}^{n} \pi_\theta(u_i|x_i)^{Q_i^\beta} p(x_i|x_{i-1},u_{i-1}) p(\theta|M)\mathrm{d}\theta \\
&= \int \exp\left[L_n^{\theta'}(\theta)\right] p(\theta|M)\mathrm{d}\theta.
\end{aligned} \tag{S.13}$$

Applying the Taylor series expansion to the weighted log-likelihood $L_n^{\theta'}(\theta)$ and the prior $p(\theta|M)$, we obtain

$$L_n^{\theta'}(\theta) = L_n^{\theta'}(\hat{\theta}_n) - \frac{n}{2}(\theta - \hat{\theta}_n)\mathrm{A}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + O_p\left(\left|\theta - \hat{\theta}_n\right|^3\right) \tag{S.14}$$

$$p(\theta|M) = p(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top \left.\frac{\partial}{\partial\theta}p(\theta|M)\right|_{\theta=\hat{\theta}_n} + O_p\left(|\theta - \hat{\theta}_n|^2\right), \tag{S.15}$$

where $\mathrm{A}_n(\hat{\theta}_n) = -(1/n)(\partial^2/\partial\theta\partial\theta^\top)L_n^{\theta'}(\theta)|_{\theta=\hat{\theta}_n} = -(1/n)\sum_{i=1}^n\sum_{j=i}^n \beta^{j-i}\mathrm{K}_{\hat{\theta}_n}(x_i, u_i)r(x_j, u_j)$. Substituting Eq. (S.14) and (S.15) into Eq. (S.13) and simplifying the results lead to the approximation of the marginal weighted likelihood as follows:

$$\begin{aligned}
\hat{p}(x_{2:n}, u_{1:n}|x_1) &= \int \exp\left\{L_n^{\theta'}(\hat{\theta}_n) - \frac{n}{2}(\theta - \hat{\theta}_n)^\top\mathrm{A}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + \cdots\right\} \\
&\quad \times \left\{p(\hat{\theta}_n) + (\theta - \hat{\theta}_n)\frac{\partial}{\partial\theta}p(\theta|M) + \cdots\right\}\mathrm{d}\theta \\
&\approx \exp[L_n^{\theta'}(\hat{\theta}_n)]p(\hat{\theta}_n|M)\int \exp\left[-\frac{n}{2}(\theta - \hat{\theta}_n)^\top\mathrm{A}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)\right]\mathrm{d}\theta \\
&= \exp\left[L_n^{\theta'}(\hat{\theta}_n)\right]p(\hat{\theta}_n|M)(2\pi)^{m/2}n^{-m/2}\left|\mathrm{A}_n(\hat{\theta}_n)\right|^{-1/2}
\end{aligned} \tag{S.16}$$

Here we used the fact that $\hat{\theta}_n$ converges to θ in probability with order $O_p(n^{-1/2})$ and also that the following equations hold:

$$\begin{aligned}
&\int (\theta - \hat{\theta}_n)\exp\left[-\frac{n}{2}(\theta - \hat{\theta}_n)^\top\mathrm{A}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)\right] = 0 \\
&\int \exp\left[-\frac{n}{2}(\theta - \hat{\theta}_n)^\top\mathrm{A}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)\right] = (2\pi)^{m/2}n^{-m/2}\left|\mathrm{A}_n(\hat{\theta}_n)\right|^{-1/2}.
\end{aligned}$$

Taking the logarithm of Eq. (S.16), we obtain

$$\ln\hat{p}(x_{2:n}, u_{1:n}|x_1) \approx L_n^{\theta'}(\hat{\theta}_n) - \frac{1}{2}m\ln n - \frac{1}{2}\ln\left|\mathrm{A}_n(\hat{\theta}_n)\right| + \frac{1}{2}m\ln(2\pi) + \frac{1}{2}\ln p(\hat{\theta}_n|M). \tag{S.17}$$

Note that, from Assumption 1-3 and condition (d), the matrix $\mathrm{A}_n(\theta)$ converges to $\mathrm{A}(\theta) := \mathbb{E}_{x_1\sim\mu_{\theta'}}^{\pi_{\theta'}}\left[\mathrm{K}_\theta(x_1, u_1)\sum_{j=1}^\infty r(x_j, u_j)\right]$ by following the discussion in Section 3. Also, from condition (c) in Theorem 2, using the continues mapping theorem shown in Theorem 2.3 in [6], $-(1/2)\ln|\mathrm{A}_n(\hat{\theta}_n)|$ is bounded by some constant in probability. From condition (b) in Theorem 2, $\ln p(\hat{\theta}_n|M)$ is bounded in probability. Then by ignoring terms with order less than $O_p(1)$ with respect to the sample size n, we can be conclude the result in Theorem 2.

A Proof of Support Lemma S.3

Let $t := [t_1, t_2, \cdots, t_m]^\top \in \mathbb{R}^m$ be a nonzero vector, $t \neq 0$. From if

$$t^\top \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\bar{\theta}}(x_i, u_i) \sum_{j=i}^{n} \beta^{j-i} r(x_j, u_j) \right) \xrightarrow{d} N(0, t^\top \Sigma t) \tag{S.18}$$

holds for any $t \in \mathbb{R}^m \setminus \{0\}$ as $n \to \infty$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\bar{\theta}}(x_i, u_i) \sum_{j=i}^{n} \beta^{j-i} r(x_j, u_j) \xrightarrow{d} N(0, \Sigma)$$

as $n \to \infty$. Thus, we attempt to prove that
$t^\top ((1/\sqrt{n}) \sum_{i=1}^{n} \psi_{\bar{\theta}}(x_i, u_i) \sum_{j=i}^{n} \beta^{j-i} r(x_j, u_j)) t$ converges to the Gaussian $N(0, \Sigma)$.

To prove this, we use the functional central limit theorem under the uniform mixing condition shown as below.

Lemma S.6 *[4, Theorem 18.6.1] Let $\{x_i : i = \{0, 1, \cdots\}\}$ be a strictly stationary and ergodic stochastic process on the probabilistic space $(\Omega, \mathcal{F}, P)$, satisfying the uniform mixing condition, with mixing coefficient $\varphi(i)$, and consider the stochastic process $\{y_i : i = \{0, 1, \cdots\}\}$ defined by*

$$y_i := f(x_i, x_{i+1}, \cdots),$$

where $f : \mathcal{X} \times \mathcal{X} \times \cdots \mapsto \mathbb{R}$ is an arbitrary function. If the following conditions

$$\sum_{i=1}^{\infty} \sqrt{\varphi(i)} < \infty \tag{S.19}$$

$$\left| \sum_{k=1}^{\infty} \mathbb{E}\left[|y_1 - \mathbb{E}[y_1 | x_{0:k}]|^2 \right] \right|^{1/2} < \infty \tag{S.20}$$

hold, then

$$\sigma^2 = \mathbb{E}[y_0^2] + 2 \sum_{k=1}^{\infty} \mathbb{E}[y_0 y_k]$$

converges, and

$$\frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} y_i \xrightarrow{d} N(0, 1).$$

Here, $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot | x_0]$ denote the expectation with respect to the whole sequence of the process $\{x_i\}$, and the conditional expectation with respect to the whole sequence of the process $\{x_i\}$ conditioned on x_0, respectively.

Now consider to assign the following random variable to y_i;

$$y_i := t^\top \left(\psi_{\bar{\theta}}(x_i, u_i) \sum_{j=i}^{\infty} \beta^{j-i} r(x_j, u_j) \right) = \sum_{k=1}^{m} t_k \psi_k(x_i, u_i, \bar{\theta}) \sum_{j=i}^{\infty} r(x_j, u_j), \tag{S.21}$$

where $\psi_k(x, u, \bar{\theta})$ is the k-th entry of the vector $\psi_k(x, u, \theta)$. Then,

$$\begin{aligned}
&\mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}}\left[y_1^2\right] + 2 \sum_{i=2}^{\infty} \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}}\left[y_1 y_i\right] \\
=&t^\top \mathbb{E}^{\pi'_\theta}_{x_1 \sim \mu_{\theta'}} \left[\left(\sum_{j=1}^{\infty} \beta^{j-1} r(x_j, u_j) \right) \left(\sum_{j'=1}^{\infty} \beta^{j'-1} r(x_{j'}, u_{j'}) \right) \psi_{\bar{\theta}}(x_1, u_1) \psi_{\bar{\theta}}(x_1, u_1)^\top \right] t \\
&+ 2 \sum_{i=1}^{\infty} t^\top \mathbb{E}^{\pi_{\theta'}}_{x_1 \sim \mu_{\theta'}} \left[\left(\sum_{j=1}^{\infty} \beta^{j-1} r(x_j, u_j) \right) \left(\sum_{j'=1+i}^{\infty} \beta^{j'-1} r(x_{j'}, u_{j'}) \right) \psi_{\bar{\theta}}(x_1, u_1) \psi_{\bar{\theta}}(x_i, u_i)^\top \right] t \\
=&t^\top \Sigma t,
\end{aligned}$$

If the conditions in Lemma S.6 hold, we can prove Eq.(S.18). Now we see whether the conditions (S.19) and (S.20) in Lemma S.5 are satisfied.

From Assumption 4, the MDP given Eq. (1) satisfies geometrically uniform mixing; there exist some positive constants $D > 0$ and $\rho \in [0, 1)$ such that

$$\sup_{B \in \mathcal{F}_{t+s}^{\infty}, A \in \mathcal{F}_{-\infty}^{t}, P(A) \neq 0} |P(B|A) - P(B)| := \varphi(s) \leq D\rho^{s}.$$

It means

$$\sum_{i=1}^{\infty} \sqrt{\varphi(i)} = \sqrt{D} \sum_{i=1}^{\infty} \rho^{i/2} = \frac{\sqrt{D}\rho^{1/2}}{1 - \rho^{1/2}} < \infty.$$

This proves the condition (S.19).

In order to show the condition (S.20), we rewrite $\sum_{l=1}^{\infty} \mathbb{E}_{x_1 \sim \mu_{\theta'}}^{\pi_{\theta'}} \left[|y_1 - \mathbb{E}^{\pi_{\theta'}} (y_1 | x_{1:l}) |^2 \right]$ as

$$\begin{aligned}
&\sum_{l=1}^{\infty} \mathbb{E}_{x_1 \sim \mu_{\theta'}}^{\pi_{\theta'}} \left[\left| y_1 - \mathbb{E}^{\pi_{\theta'}} [y_1 | x_{1:l}] \right|^2 \right] \\
=& \sum_{l=1}^{\infty} \mathbb{E}_{x_1 \sim \mu_{\theta'}}^{\pi_{\theta'}} \left[\left| \sum_{k=1}^{m} \sum_{j=1}^{\infty} \beta^{j-1} t_k \psi_k(x_1, u_1, \bar{\theta}) \left\{ r(x_j, u_j) - \mathbb{E}^{\pi_{\theta'}} [r(x_j, u_j) | x_{1:l}] \right\} \right|^2 \right] \\
=& \sum_{l=1}^{\infty} \mathbb{E}_{x_1 \sim \mu_{\theta'}}^{\pi_{\theta'}} \left[\left| \sum_{j=1+l}^{\infty} \beta^{j-1} \sum_{k=1}^{m} t_k \psi_k(x_1, u_1, \bar{\theta}) \left\{ r(x_j, u_j) - \mathbb{E}^{\pi_{\theta'}} [r(x_j, u_j) | x_l] \right\} \right|^2 \right],
\end{aligned}$$

where $\mathbb{E}^{\pi_{\theta'}}[\cdot | x_i]$ denotes the conditional expectation over the whole sample sequence conditioned on x_i. Defining a constant $C = \sum_{k=1}^{m} |C_k|$ where
$C_k := \sup |\psi_k(x_1, u_1, \theta) \{ r(x_j, u_j) - \mathbb{E}^{\pi_{\theta'}} [r(x_j, u_j) | x_l] \}| < \infty$, for any $\{x_1, u_1, x_j, u_j\} \in \mathcal{X} \times \mathcal{X} \times \mathcal{U} \times \mathcal{U}$, $k \in \{1, \cdots, m\}$ and $\theta \in \Theta$, we have

$$\sum_{l=1}^{\infty} \mathbb{E}_{x_1 \sim \mu_{\theta'}}^{\pi_{\theta'}} \left[\left| y_1 - \mathbb{E}^{\pi_{\theta'}} [y_1 | x_{1:l}] \right|^2 \right] \leq C^2 \sum_{l=1}^{\infty} \left| \sum_{j=1+l}^{\infty} \beta^{j-1} \right|^2 = C^2 \frac{\beta^2}{(1-\beta)^2 (1-\beta^2)} < \infty,$$

which assures the condition (b). Hence, we can conclude

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i = \frac{1}{\sqrt{n}} \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=i}^{\infty} \beta^{j-i} t_k \psi_k(x_i, u_i, \bar{\theta}) r(x_j, u_j) \xrightarrow{d} \mathcal{N}\left(0, t^{\top} \Sigma t\right).$$

The same argument as in the proof of shows that the tails of can be ignored. This observation implies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=i}^{n} \beta^{j-i} \psi(x_i, u_i, \bar{\theta}) r(x_j, u_j) \xrightarrow{d} N(0, \Sigma).$$

References

[1] P. Billingsley. *Probability and Measure*. John Wiley and Sons, 1995.

[2] L. Breiman. *Probability*. Addison-Wesley, 1968.

[3] G. H. Hardy. *Divergent Series*. Oxford University Press, 1949.

[4] I. A. Ibragimov and I. U. V. Linnik. *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, 1971.

[5] J. Kober and J. Peters. Policy search for motor primitives in robotics. *Machine Learning*, 84(1-2):171–203, 2011.

[6] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.