Supplementary Material to "A Generalized Natural Actor-Critic Algorithm"

Tetsuro Morimura[†], **Eiji Uchibe**[‡], **Junichiro Yoshimoto**[‡], **Kenji Doya**[‡]
†: IBM Research – Tokyo, Kanagawa, Japan
‡: Okinawa Institute of Science and Technology, Okinawa, Japan
tetsuro@jp.ibm.com, {uchibe,jun-y,doya}@oist.jp

Proof of Lemma 1

If the function $\boldsymbol{g}_{\iota,\boldsymbol{\theta}}(s,a;\boldsymbol{\omega})$ defined in Eq. (10) satisfies the three conditions of Eqs. (6), (7), and (8) for estimating function and its solution is equal to the gNG(ι), then Lemma 1 is proven. It is clear that it satisfies the conditions of Eqs. (7) and (8) (under Assumption 1). Thereby, if the following equation holds,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{g}_{\iota,\boldsymbol{\theta}}(s,a;\boldsymbol{\omega}^*)\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{g}'_{\iota,\boldsymbol{\theta}}(s,a;\boldsymbol{\omega}^*) - \nabla_{\boldsymbol{\theta}}\ln\{d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\}\rho(s,s_{+1})\right]\\ &= -\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla_{\boldsymbol{\theta}}\ln\{d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\}\rho(s,s_{+1})\right]\\ &= \mathbf{0}, \end{aligned} \tag{s-1}$$

where $\boldsymbol{\omega}^*$ is the gNG(ι), $\boldsymbol{g}_{\iota,\boldsymbol{\theta}}(s,a;\boldsymbol{\omega})$ satisfies the the condition of Eq. (8) and its unique solution is gNG(ι) and then Lemma 1 is proven. Thus, we prove Eq. (s-1) in the following.

Because of Eq. (13) and

$$\begin{cases}\sum_{a\in\mathcal{A}}\pi(a|s;\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta})\,c = \nabla_{\boldsymbol{\theta}}c = \mathbf{0},\\ \sum_{s\in\mathcal{S}}d_{\boldsymbol{\theta}}(s)\nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s)\,c \quad\;\; = \nabla_{\boldsymbol{\theta}}c = \mathbf{0},\end{cases}$$

we know that

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla_{\boldsymbol{\theta}}\ln\{d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\}\rho(s,s_{+1})\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\{\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s)\}\{b(s)-b(s_{+1})\}\right].$$

Since a time average is equivalent to a state-action space average with the ergodicity of $\mathrm{M}(\boldsymbol{\theta})$ in Assumption 2, the above equation is transformed to

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\{\nabla_{\boldsymbol{\theta}}\ln\pi(a_t|s_t;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_t)\}\{b(s_t)-b(s_{t+1})\}\\ &= \lim_{T\to\infty}\frac{1}{T}\Big[b(s_0)\{\nabla_{\boldsymbol{\theta}}\ln\pi(a_0|s_0;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_0)\} - b(s_T)\{\nabla_{\boldsymbol{\theta}}\ln\pi(a_T|s_T;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_T)\}\\ &\quad + \sum_{t=0}^{T-1}b(s_{t+1})\big\{-\nabla_{\boldsymbol{\theta}}\ln\pi(a_t|s_t;\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_t) + \nabla_{\boldsymbol{\theta}}\ln\pi(a_{t+1}|s_{t+1};\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_{t+1})\big\}\Big]\\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s_{+1}\in\mathcal{S}}\sum_{a_{+1}\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})p(s_{+1}|s,a)\pi(a_{+1}|s_{+1};\boldsymbol{\theta})\\ &\qquad b(s_{+1})\big\{-\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s) + \nabla_{\boldsymbol{\theta}}\ln\pi(a_{+1}|s_{+1};\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_{+1})\big\}\\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s_{+1}\in\mathcal{S}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})p(s_{+1}|s,a)b(s_{+1})\{-\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_{+1})\}\\ &= \sum_{s_{+1}\in\mathcal{S}} d_{\boldsymbol{\theta}}(s_{+1})b(s_{+1})\Big[\nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_{+1}) - \mathbb{E}_{\boldsymbol{\theta}}\{\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s)\mid s_{+1}\}\Big]\\ &= \mathbf{0}.\end{aligned}$$

For the final transformations, we use the result

$$\mathbb{E}_{\boldsymbol{\theta}}\{\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s) \mid s_{+1}\} = \nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s_{+1})$$

in [1]. Therefore, Eq. (s-1) holds. □

References

[1] T. Morimura, E. Uchibe, J. Yoshimoto, J. Peters, and K. Doya. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural Computation*. (in press).