Letter

Supplementary Material for "Bearings-Only Target Motion Analysis via Deep Reinforcement Learning"

Chengyi Zhou, Meiqin Liu, *Senior Member, IEEE*, Senlin Zhang, *Member, IEEE*, Ronghao Zheng, *Member, IEEE*, and Shanling Dong, *Member, IEEE*

Fig. S1 illustrates the network structures of the policy and the soft Q-function. The actor network functions as the parameterized policy, while the parameterized soft Q-function serves as the critic network.

Fig. S2 displays the architecture of the proposed DRL-based estimator.

Fig. S3 shows the simulated target tracking geometry for a constant-velocity target.

Fig. S4 presents the learning curves of the proposed estimator.

Proof of Lemma 1: Introduce a reward function augmented with entropy

$$r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) \triangleq r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\alpha\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P}[\mathcal{H}(\pi(\cdot|\boldsymbol{s}_{i+1}))] \tag{S.1}$$

then, we have

$$\begin{aligned} & r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P}[V(\boldsymbol{s}_{i+1})] \\ &= r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P}[\mathbb{E}_{\hat{\boldsymbol{\xi}}_{i+1}\sim\pi}[Q(\boldsymbol{s}_{i+1},\hat{\boldsymbol{\xi}}_{i+1})-\alpha\log\pi(\hat{\boldsymbol{\xi}}_{i+1}|\boldsymbol{s}_{i+1})]] \\ &= r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P,\hat{\boldsymbol{\xi}}_{i+1}\sim\pi}[Q(\boldsymbol{s}_{i+1},\hat{\boldsymbol{\xi}}_{i+1})]+\alpha\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P}[\mathcal{H}(\pi(\cdot|\boldsymbol{s}_{i+1}))] \\ &= r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P,\hat{\boldsymbol{\xi}}_{i+1}\sim\pi}[Q(\boldsymbol{s}_{i+1},\hat{\boldsymbol{\xi}}_{i+1})]. \end{aligned} \tag{S.2}$$

Therefore, the soft Bellman backup operation can be reformulated as

$$\mathcal{T}^\pi Q(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)=r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P,\hat{\boldsymbol{\xi}}_i\sim\pi}[Q(\boldsymbol{s}_{i+1},\hat{\boldsymbol{\xi}}_i)]. \tag{S.3}$$

Given $|\mathcal{A}|=-\dim(\hat{\boldsymbol{\xi}}_i)$ and $\alpha<\infty$, the second term in (S.1) is bounded. Note that noise is ubiquitous in measurements, which implies that the reward $r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)$ is inherently bounded. Thus, there exist $r_{\min}$ and $r_{\max}$ such that $\bar{r}\in[r_{\min},r_{\max}]$, and further, $|r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)|\le\bar{r}$ with $\bar{r}=\max\{|r_{\min}|,|r_{\max}|\}$. Since

$$Q_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)=r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)+\mathbb{E}[\sum_{j=1}^{\infty}\gamma^j r_\pi(\boldsymbol{s}_{i+j},\hat{\boldsymbol{\xi}}_{i+j})|\boldsymbol{S}_j=\boldsymbol{s}_i,\boldsymbol{A}_j=\hat{\boldsymbol{\xi}}_i] \tag{S.4}$$

we have

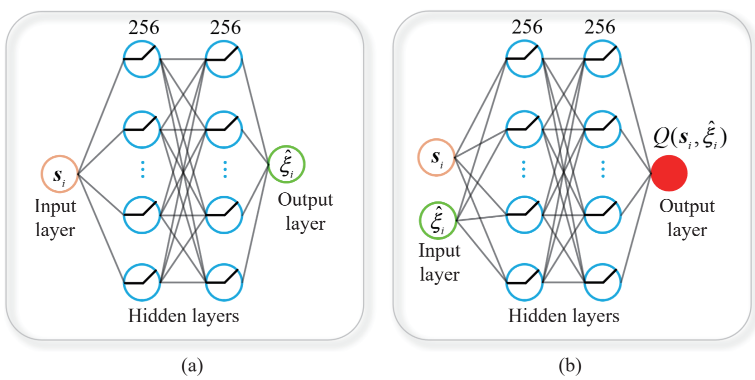

Fig. S1. Actor and critic network structures. (a) Actor network; (b) Critic network.

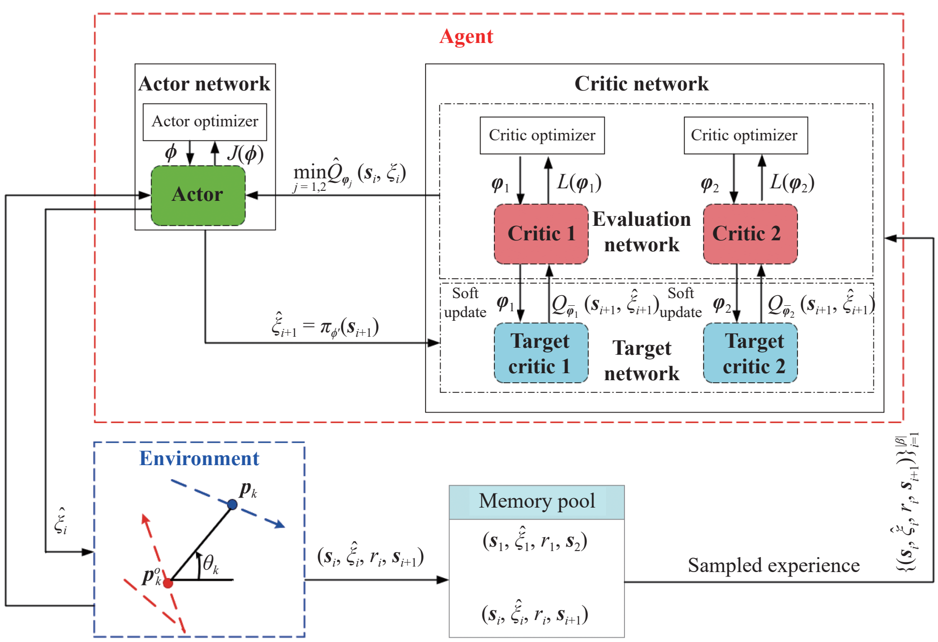

Fig. S2. The architecture of the proposed DRL-based estimator.

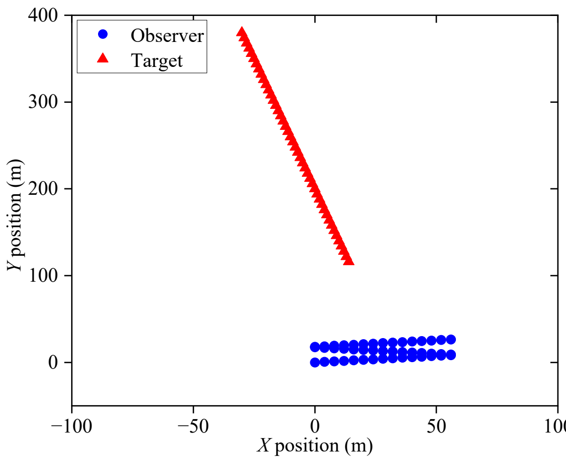

Fig. S3. The simulated tracking scenario with a constant-velocity target.

$$\|Q_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)\|_\infty \le \frac{\bar{r}}{1-\gamma} \tag{S.5}$$

where $\|Q_\pi(\boldsymbol{s},\hat{\boldsymbol{\xi}})\|_\infty = \max_{\boldsymbol{s},\hat{\boldsymbol{\xi}}}|Q_\pi(\boldsymbol{s},\hat{\boldsymbol{\xi}})|$. Hence, the Q-value is bounded in ∞-norm. For any two vectors $Q, Q' \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we have the following contraction proof for $\mathcal{T}^\pi$:

$$\begin{aligned}\|\mathcal{T}^\pi Q - \mathcal{T}^\pi Q'\|_\infty &= \|r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) + \gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P(\boldsymbol{s}_{i+1}|\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)}[Q(\boldsymbol{s}_{i+1},\cdot)] \\ &\quad - r_\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) - \gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P(\boldsymbol{s}_{i+1}|\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)}[Q'(\boldsymbol{s}_{i+1},\cdot)]\|_\infty \\ &= \|\gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim P(\boldsymbol{s}_{i+1}|\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)}[Q(\boldsymbol{s}_{i+1},\cdot) - Q'(\boldsymbol{s}_{i+1},\cdot)]\|_\infty \\ &= \|\sum_{\boldsymbol{s}_{i+1}}\gamma(Q(\boldsymbol{s}_{i+1},\cdot) - Q'(\boldsymbol{s}_{i+1},\cdot))P(\boldsymbol{s}_{i+1}|\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)\|_\infty \\ &\le \sum_{\boldsymbol{s}_{i+1}}\gamma\|Q(\boldsymbol{s}_{i+1},\cdot) - Q'(\boldsymbol{s}_{i+1},\cdot)\|_\infty P(\boldsymbol{s}_{i+1}|\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) \\ &= \gamma\|Q - Q'\|_\infty.\end{aligned} \tag{S.6}$$

By Banach fixed-point theorem, the soft Bellman backup operator is a γ-contraction. Therefore, iterative soft policy evaluation will converge on the unique fixed point of $\mathcal{T}^\pi$. Since $\mathcal{T}^\pi Q_\pi = Q_\pi$ is a fixed point, so that iterative policy evaluation converges on Q_π, i.e., the sequence Q^k will converge to Q_π as $k \to \infty$. In light of this, Lemma 1 is confirmed. ■

Proof of Lemma 2: Let $Q^{\pi_{\text{old}}}$ and $V^{\pi_{\text{old}}}$ be the corresponding soft state-action value and soft state value with $\pi_{\text{old}} \in \Pi$. Define π_{new} as

$$\begin{aligned}\pi_{\text{new}}(\cdot|\boldsymbol{s}_i) &= \underset{\min \pi'\in\Pi}{\arg}\ \mathcal{D}_{KL}(\pi'(\cdot|\boldsymbol{s}_i)\|\exp(Q^{\pi_{\text{old}}}(\boldsymbol{s}_i,\cdot)) - \log Z^{\pi_{\text{old}}}(\boldsymbol{s}_i))) \\ &= \underset{\min \pi'\in\Pi}{\arg}\ J_{\pi_{\text{old}}}(\pi'(\cdot|\boldsymbol{s}_i))\end{aligned} \tag{S.7}$$

Since we can always choose $\pi_{\text{new}} = \pi_{\text{old}} \in \Pi$, there must exist $J_{\pi_{\text{old}}}(\pi_{\text{new}}(\cdot|\boldsymbol{s}_i)) \le J_{\pi_{\text{old}}}(\pi_{\text{old}}(\cdot|\boldsymbol{s}_i))$. Hence, we have

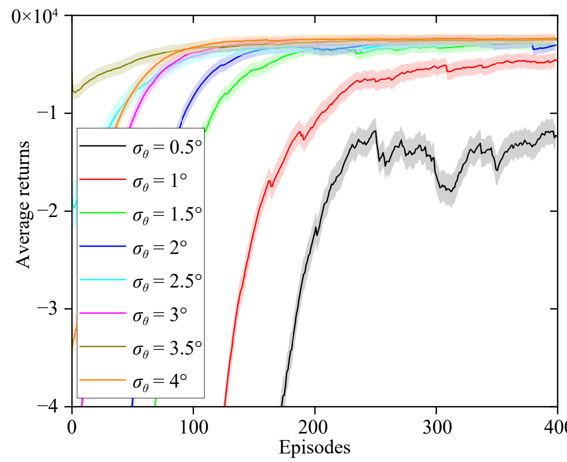

Fig. S4. Learning curves of the proposed estimator at training. The solid lines correspond to the mean and the shaded regions depicts the mean ± the standard deviation over ten different runs.

$$\begin{aligned}&\mathbb{E}_{\hat{\boldsymbol{\xi}}_i\sim\pi_{\text{new}}}[\log\pi_{\text{new}}(\hat{\boldsymbol{\xi}}_i|\boldsymbol{s}_i) - Q^{\pi_{\text{old}}}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) + \log Z^{\pi_{\text{old}}}(\boldsymbol{s}_i)] \\ &\quad \le \mathbb{E}_{\hat{\boldsymbol{\xi}}_i\sim\pi_{\text{old}}}[\log\pi_{\text{old}}(\hat{\boldsymbol{\xi}}_i|\boldsymbol{s}_i) - Q^{\pi_{\text{old}}}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) + \log Z^{\pi_{\text{old}}}(\boldsymbol{s}_i)].\end{aligned} \tag{S.8}$$

It is noted that the partition functon $Z^{\pi_{\text{old}}}$ depends only on the state. Therefore the inequality reduces to

$$\mathbb{E}_{\hat{\boldsymbol{\xi}}_i\sim\pi_{\text{new}}}[Q^{\pi_{\text{old}}}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) - \log\pi_{\text{new}}(\hat{\boldsymbol{\xi}}_i|\boldsymbol{s}_i)] \ge V^{\pi_{\text{old}}}(\boldsymbol{s}_i). \tag{S.9}$$

Then, consider the soft Bellman equation

$$\begin{aligned}Q^{\pi_{\text{old}}}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) &= r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) + \gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim p}[V^{\pi_{\text{old}}}(\boldsymbol{s}_{i+1})] \\ &\le r(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) + \gamma\mathbb{E}_{\boldsymbol{s}_{i+1}\sim p}[\mathbb{E}_{\hat{\boldsymbol{\xi}}_{i+1}\sim\pi_{\text{new}}}[Q^{\pi_{\text{old}}}(\boldsymbol{s}_{i+1},\hat{\boldsymbol{\xi}}_{i+1}) \\ &\quad - \log\pi_{\text{new}}(\hat{\boldsymbol{\xi}}_i)]] \\ &\ \ \vdots \\ &\le Q^{\pi_{\text{new}}}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)\end{aligned} \tag{S.10}$$

where we expand the $Q^{\pi_{\text{old}}}$ repeatedly by applying the soft Bellman equation and the bound in (S.9). The convergence to $Q^{\pi_{\text{new}}}$ can be inferred from Lemma 1. Consequently, Lemma 2 holds. ■

Proof of Theorem 1: Let π_k denote the policy at iteration k. Lemma 2 ensures the monotonic increase of the sequence Q^{π_k}. Given the bounded reward and entropy of Q^π, the sequence π_k converges to some π^*. Upon convergence, $J_{\pi^*}(\pi^*(\cdot|\boldsymbol{s}_i)) < J_{\pi^*}(\pi(\cdot|\boldsymbol{s}_i))$ holds for all $\pi \in \Pi$ and $\pi \ne \pi^*$. Based on the proof of Lemma 2, we can establish the following inequality:

$$\mathbb{E}_{\hat{\boldsymbol{\xi}}_i\sim\pi^*}[Q^{\pi^*}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) - \log\pi^*(\hat{\boldsymbol{\xi}}_i|\boldsymbol{s}_i)] > V^\pi(\boldsymbol{s}_i) \tag{S.11}$$

which implies that $Q^\pi(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) < Q^{\pi^*}(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i)$ for any $(\boldsymbol{s}_i,\hat{\boldsymbol{\xi}}_i) \in \mathcal{S}\times\mathcal{A}$. Therefore, it must be the case that π^* is optimal in Π. In summary, Theorem 1 stands. ■