

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

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

then, we have

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

Therefore, the soft Bellman backup operation can be reformulated as

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

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

$$Q_{\pi}(s_i, \hat{\xi}_i) = r_{\pi}(s_i, \hat{\xi}_i) + \mathbb{E} \left[\sum_{j=1}^{\infty} \gamma^j r_{\pi}(s_{i+j}, \hat{\xi}_{i+j}) | S_j = s_i, A_j = \hat{\xi}_i \right] \quad (\text{S.4})$$

we have

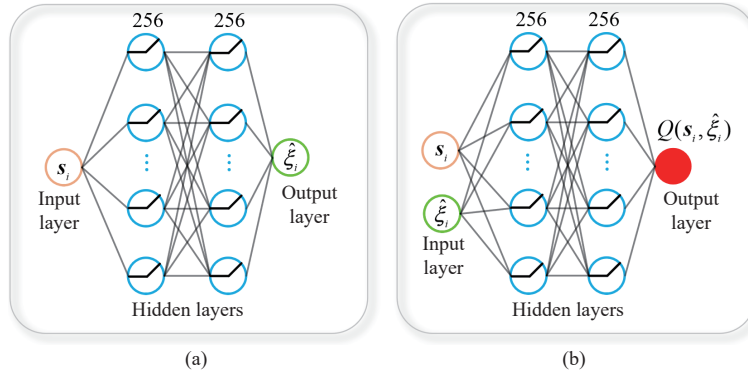


Fig. S1. Actor and critic network structures. (a) Actor network; (b) Critic network.

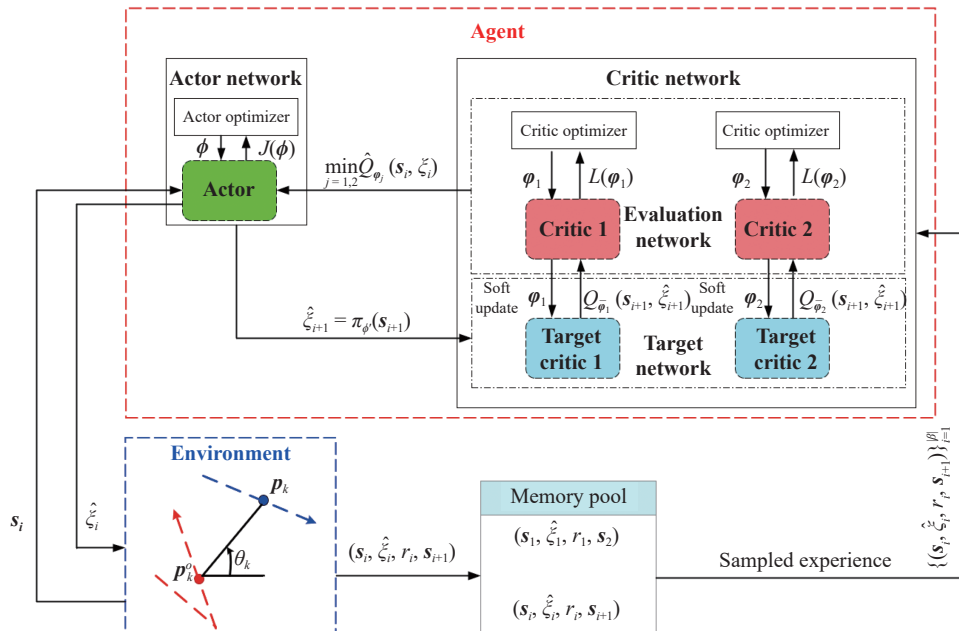


Fig. S2. The architecture of the proposed DRL-based estimator.

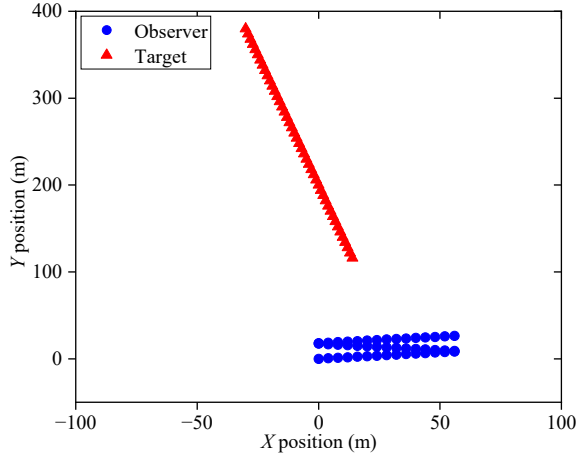


Fig. S3. The simulated tracking scenario with a constant-velocity target.

$$\|Q_\pi(s_i, \hat{\xi}_i)\|_\infty \leq \frac{\bar{r}}{1-\gamma} \quad (\text{S.5})$$

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

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

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}^π . 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 $k \rightarrow \infty$. 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

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

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

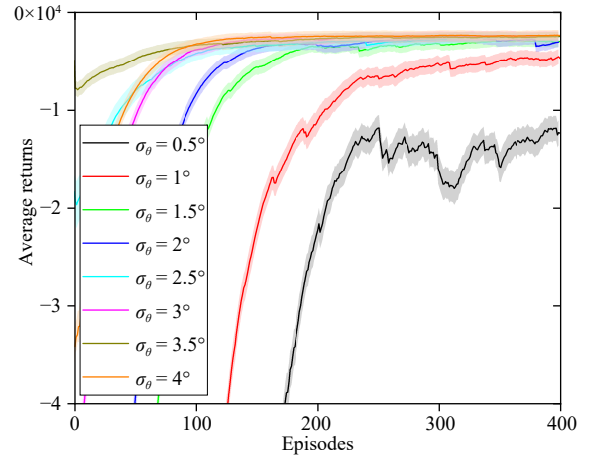


Fig. S4. Learning curves of the proposed estimator at training. The solid lines correspond to the mean and the shaded regions depicts the mean \pm the standard deviation over ten different runs.

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

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

$$\mathbb{E}_{\hat{\xi}_i \sim \pi_{\text{new}}} [\log \pi_{\text{new}}(\hat{\xi}_i|s_i)] \geq V^{\pi_{\text{old}}}(s_i). \quad (\text{S.9})$$

Then, consider the soft Bellman equation

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

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, $J_{\pi^*}(\pi^*(\cdot|s_i)) < J_{\pi^*}(\pi(\cdot|s_i))$ holds for all $\pi \in \Pi$ and $\pi \neq \pi^*$. Based on the proof of Lemma 2, we can establish the following inequality:

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

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