Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

ADVERSARIAL REINFORCEMENT LEARNING: A DUALITY-BASED APPROACH TO SOLVING OPTIMAL CONTROL PROBLEMS

Nan Chen, Mengzhou Liu, Xiaoyan Wang, and Nanyi Zhang

Dept. of Systems Engineering and Engineering Management
and
Centre for Financial Engineering
The Chinese University of Hong Kong, Hong Kong, CHINA

1 PROOF SKETCH OF PROPOSITION 3.1

We present a proof sketch for Proposition 3.1 below. First, we establish one lemma under these assumptions. We omit its proof because it is straightforward.

Lemma 1.1. *Assume that activation function σ in the network (29) is Lipschitz and the state space $\mathbb{S}$ is bounded. Then, the network $F(\boldsymbol{s};\phi)$ is Lipschitz in the hyperparemeter ϕ.*

Now let us turn to the proposition's proof:

Proof Sketch of Proposition 3.1. Given ξ and two sets of hyperparameters ϕ and ϕ', we can show that

$$\left|\max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}\in\mathbb{A}^T} Y(\phi,\boldsymbol{a},\xi)-\max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}\in\mathbb{A}^T} Y(\phi',\boldsymbol{a},\xi)\right|\leq \max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}\in\mathbb{A}^T}|z^{\phi}(\boldsymbol{a},\xi)-z^{\phi'}(\boldsymbol{a},\xi)|, \tag{1}$$

using Assumption A.1. From the construction of the penalty function z (cf. (5)), we know further that the right-hand side of (31) should be bounded from above by

$$\max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}}\left\{\sum_{t=1}^{T}|\rho_t(\boldsymbol{s}_t,\phi_t)-\rho_t(\boldsymbol{s}_t,\phi_t')|+\sum_{t=1}^{T}\mathbb{E}\Big[|\rho_t(\boldsymbol{s}_t,\phi_t)-\rho_t(\boldsymbol{s}_t,\phi_t')|\big|\boldsymbol{s}_{t-1}\Big]\right\}, \tag{2}$$

where $\boldsymbol{s}=\{\boldsymbol{s}_t,t\in\mathbb{T}\}$ is the state trajectory under the randomness ξ and action sequence $\boldsymbol{a}$. The Lipschitz property of deep networks ρ established in Lemma 1.1 ensures that there exists a sufficiently large L such that

$$|\rho_t(\boldsymbol{s}_t,\phi_t)-\rho_t(\boldsymbol{s}_t,\phi_t')|\leq L\|\phi_t-\phi_t'\| \tag{3}$$

for all t. This, together with (1) and (2), implies,

$$\left|\max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}\in\mathbb{A}^T} Y(\phi,\boldsymbol{a},\xi)-\max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}\in\mathbb{A}^T} Y(\phi',\boldsymbol{a},\xi)\right|\leq TL\|\phi-\phi'\|; \tag{4}$$

in other words, $\max_{\boldsymbol{a}} Y(\phi,\boldsymbol{a},\xi)$ is Lipschitz in ϕ.

Furthermore, the infinitesimal perturbation analysis (IPA) in the simulation literature (e.g., (Glasserman 2004; Asmussen and Glynn 2007)) ensures that

$$\nabla_\phi\mathbb{E}[\max_{\boldsymbol{a}} Y(\phi,\xi)]=\mathbb{E}[\nabla_\phi\max_{\boldsymbol{a}} Y(\phi,\xi)]. \tag{5}$$

Consider the gradient inside the expectation on the right-hand side of the above equality. As we change ϕ, the envelope theorem states that, the resulting change in the optimal policy $\boldsymbol{a}^\phi$ has no first-order contribution to

the optimal value Y. (Milgrom and Segal 2002) show the envelope theorem holds under the differentiability of the optimal value function in their Corollary 4. It is easy to check that the conditions of Corollary 4 in (Milgrom and Segal 2002) are satisfied if Assumption A.2 holds. Based on this observation, we have

$$\nabla_\phi Y(\phi,\xi) = -\sum_{t=0}^{T-1} \nabla_\phi z_t^\phi(\boldsymbol{a}^\phi(\xi),\xi) \tag{6}$$

Hence,

$$\nabla_\phi \mathbb{E}\left[\max_{\boldsymbol{a}=\{\boldsymbol{a}_t\}_{t=0}^{T-1}} Y(\phi,\boldsymbol{a},\xi)\middle|\boldsymbol{s}_0=\boldsymbol{s}\right] = \mathbb{E}\left[-\sum_{t=0}^{T-1} \nabla_\phi z_t^\phi(\boldsymbol{a}_t^\phi(\xi),\xi)\right]. \tag{7}$$

□