(Towards a)
Primal-dual view of actor-critic methods

I. Primal-Dual Linear Programming Formulation of MDP

Let S and A denote state and action sets, respectively. Define the transition matrix P, of size $|S||A|\times|S|$, whose entries $P_{(sa,s')}$ specify the conditional probability of going to state s' when starting from state s and taking action a:

$$P_{(sa,s)} = p(s'|s,a) \geq 0, \quad \sum_{s'\in S} p(s'|s,a) = 1 \tag{1}$$

Let $r(s,a,s')$ denote the reward obtained when taking action a in state s and transitioning to state s'. We also define $r(s,a)$ as the reward obtained when taking action a in state s, such that

$$r(s,a) = \sum_{s'} p(s'|s,a) r(s,a,s') \tag{2}$$

Let π denote the policy, such that $\pi(a,s) = p(a|s)$ is the probability of taking action a in state s.

We can consider both average and discounted formulations. For simplicity, we focus on discounted formulation with bounded rewards.

Assumption 1. *There exists some scalar M such that the reward r satisfies*

$$|r(s,a)| \leq M, \quad \forall (s,a) \in S \times A \tag{3}$$

Define the value function as the long-term expected discounted cumulative reward:

$$v^{\pi}(s) = \mathsf{E}\left\{ \sum_{i=0}^{\infty} \gamma^i r(s_i,a_i) \mid a_i \sim \pi(\cdot|s_i),\ s_0 = s \right\} \tag{4}$$

where $0 < \gamma < 1$ is the discount factor. Introduce the optimal value function

$$v^{\star}(s) \triangleq \max_{\pi\in\Pi} v^{\pi}(s), \qquad \forall s \in S. \tag{5}$$

Our goal is to find the policy that maximizes the long-term reward:

$$\pi^\star(s) \triangleq \arg\max_{a\in\mathbb{A}} \left(r(s,a) + \gamma \sum_{s'\in\mathbb{S}} P(s'|s,a) v^\star(s') \right), \quad \forall s \in S \tag{6}$$

Introduce the dynamic programming operator T:

$$(Tv)(s) \triangleq \max_{a\in A} \left(r(s,a) + \gamma \sum_{s'\in\mathbb{S}} P(s'|s,a) v(s') \right) \tag{7}$$

for which the following results are well known.

Lemma 1. *The dynamic programming operator T satisfies the following properties:*

1) (Monotonicity) *For any functions $v : S \to \Re$ and $v' : S \to \Re$, such that*

$$v(s) \le v'(s), \ \forall s \in S \tag{8}$$

we have

$$(T^i v)(s) \le (T^i v')(s), \quad \forall s \in S, \ i = 1, 2, \dots \tag{9}$$

where $(T^i v)(s) = (T(T^{i-1}v))(s) = (T(T \cdots Tv))(s)$.

2) (Convergence) *For any bounded function $v : S \to \Re$, the optimal value function satisfies*

$$v^\star(s) = \lim_{i\to\infty} (T^i v)(s), \quad \forall s \in S \tag{10}$$

3) (Bellman's equation) *The optimal value function satisfies and is the unique solution to*

$$v^\star(s) = (Tv^\star)(s), \quad \forall s \in S \tag{11}$$

Proof: See, e.g., [1]. ■

Let us write $v' \ge v$ to denote that $v'(s) \ge v(s)$, for all s. From Lemma 1.3, we have

$$v \ge Tv \quad \Rightarrow \quad v \ge v^\star = Tv^\star \tag{12}$$

Thus, $v^\star$ is the smallest v that satisfies the constraint $v \ge TJ$. This constraint can be written as a finite set of linear inequalities

$$v(s) \ge r(s,a) + \gamma \sum_{s'\in\mathbb{S}} p(s'|s,a) v(s'), \quad \forall s \in S, \ \forall a \in A \tag{13}$$

which delineates a polyhedron in $\Re^{|S||A|}$. The optimal value is the "shoutheast" corner of this polyhedron.

Write the value in compact vector form:

$$\mathbf{v}^{\pi} \triangleq (v^{\pi}(s))_{s\in S} \tag{14}$$

Hence, the optimal value vector is given by

$$\mathbf{v}^{\star} \triangleq \max_{\pi\in\Pi} \mathbf{v}^{\pi} \tag{15}$$

We can find $\mathbf{v}^{\star}$ by solving the following linear programming problem [1]–[3]:

$$\mathcal{P}_0: \quad \begin{aligned} \underset{\mathbf{v}}{\text{minimize}} \quad & \mathbf{1}^{\top}\mathbf{v} \\ \text{s.t.} \quad & v(s) \geq r(s,a) + \gamma \sum_{s'\in\mathbb{S}} p(s'|s,a)v(s'), \quad \forall s \in S,\ \forall a \in A \end{aligned} \tag{16}$$

Instead of minimizing $\mathbf{1}^{\top}\mathbf{v}$, let us minimize the objective $(1-\gamma)\boldsymbol{\mu}^{\top}\mathbf{v}$, where $0 < \gamma < 1$ and $\boldsymbol{\mu}$ can be seen as the probability distribution over the initial state, given that $\boldsymbol{\mu} \geq 0$ and $\mathbf{1}^{\top}\boldsymbol{\mu} = 1$. Hence, the primal problem becomes

$$\mathcal{P}_1: \quad \begin{aligned} \underset{\mathbf{v}}{\text{minimize}} \quad & (1-\gamma)\boldsymbol{\mu}^{\top}\mathbf{v} \\ \text{s.t.} \quad & v(s) \geq r(s,a) + \gamma \sum_{s'\in\mathbb{S}} p(s'|s,a)v(s'), \quad \forall s \in S,\ \forall a \in A \end{aligned} \tag{17}$$

Although this change has no influence on problem (16) (i.e., it has the same solution as (17)), reference [3] shows that this proportional factor plays an important non-arbitrary role when we formulate its dual problem. In particular, $\boldsymbol{\mu}$ can be seen as a probability distribution over the initial state. Introduce the the reward vector

$$\mathbf{r} \triangleq (r(s,a))_{s\in S, a\in A} \in \Re^{|S||A|} \tag{18}$$

and the marginalization matrix, Φ, of size $|S| \times |S||A|$, such that we can express the inequality constraints in (17) as

$$\Phi^{\top}\mathbf{v} \geq r + \gamma P\mathbf{v} \tag{19}$$

The Lagrangian of (17), with dual variable $\mathbf{d} \geq 0$, is given by

$$L(\mathbf{v}, \mathbf{d}) = (1-\gamma)\boldsymbol{\mu}^{\top}\mathbf{v} + \mathbf{d}^{\top}(\mathbf{r} + \gamma P\mathbf{v} - \Phi^{\top}\mathbf{v}) \tag{20}$$

The dual function is defined as

$$g(\mathbf{d}) = \inf_{\mathbf{v}} L(\mathbf{v}, \mathbf{d}) \tag{21}$$

The first order condition for minimizing the Lagrangian over the primal variable $\mathbf{v}$ is given by:

$$\nabla_{\mathbf{v}} L(\mathbf{v}, \mathbf{d}) = (1-\gamma)\boldsymbol{\mu} + \gamma P^{\top}\mathbf{d} - \Phi\mathbf{d} = \mathbf{0} \tag{22}$$

Hence,

$$g(\mathbf{d}) = \begin{cases} \mathbf{d}^{\top}\mathbf{r} & \text{if } \Phi\mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^{\top}\mathbf{d} \\ -\infty & \text{otherwise} \end{cases} \tag{23}$$

Therefore, the dual problem is given by

$$\mathcal{P}_2 : \begin{aligned} \underset{\mathbf{d}}{\text{maximize}} \quad & \mathbf{d}^{\top}\mathbf{r} \\ \text{s.t.} \quad & \Phi\mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^{\top}\mathbf{d} \\ & \mathbf{d} \geq 0 \end{aligned} \tag{24}$$

Let $\mathbf{d}^{\star}$ denote the solution to (24). By strong duality, we know that the optimal objective value of the dual equals the optimal value function:

$$(1-\gamma)\boldsymbol{\mu}^{\top}\mathbf{v}^{\star} = \mathbf{d}^{\star\top}\mathbf{r} \tag{25}$$

Before unrolling the left side of (25), we introduce the optimal discounted stationary distributions over states conditioned on the initial state:

$$\xi^{\star}(s, x) = \sum_{i=0}^{\infty} \gamma^i p(s_i = s | s_0 = x,\ a_i \sim \pi^{\star}) \tag{26}$$

Now, by using the following relationship due to [4]:

$$v^{\star}(x) = \sum_{s\in S} \xi^{\star}(s, x) \sum_{a\in A} \pi^{\star}(s, a) r(s, a) \tag{27}$$

we can unroll the left side of (25) as follows:

$$\begin{aligned} (1-\gamma)\boldsymbol{\mu}^{\top}\mathbf{v}^{\star} &= (1-\gamma) \sum_{x\in S} \mu(x) \sum_{s\in S} \xi^{\star}(s, x) \sum_{a\in A} \pi^{\star}(s, a) r(s, a) \\ &= \sum_{s\in S} \mu^{\star}(s) \sum_{a\in A} \pi^{\star}(s, a) r(s, a) \end{aligned} \tag{28}$$

where

$$\mu^\star(s) = (1-\gamma)\sum_{x\in S}\mu(x)\xi^\star(s,x), \quad \mu^\star(s) \geq 0,\ \forall s \in S \quad \text{and} \quad \sum_{s\in S}\mu^\star(s) = 1 \tag{29}$$

In order to unroll the right side of (25), we write

$$\begin{aligned}\mathbf{d}^{\star\top}\mathbf{r} &= \sum_{s\in S}\sum_{a\in A} d^\star(s,a) r(s,a) \\ &= \sum_{s\in S} h^\star(s) \sum_{a\in A} \phi^\star(s,a) r(s,a)\end{aligned} \tag{30}$$

where

$$h^\star(s) = \sum_{a\in A} d^\star(s,a), \quad \phi^\star(s,a) = \frac{d^\star(s,a)}{\sum_{a\in A} d^\star(s,a)} \tag{31}$$

We know that $\mathbf{d}^\star \geq 0$ and reference [3, Lemma 1] shows that $\mathbf{1}^\top \mathbf{d}^\star = 1$. Hence, we can see $\mathbf{d}^\star$ as a joint probability distribution over state-action pairs. Moreover, the following properties establish that $h^\star$ and $\phi^\star$ are also probability distributions:

$$h^\star(s) \geq 0, \quad \sum_{s\in S} h^\star(s) = 1, \quad \phi^\star(s,a) \geq 0, \quad \sum_{a\in A} \phi^\star(s) = 1 \tag{32}$$

From (25), (28) and (30), we have:

$$\sum_{s\in S} \mu^\star(s) \sum_{a\in A} \pi^\star(s,a) r(s,a) = \sum_{s\in S} h^\star(s) \sum_{a\in A} \phi^\star(s,a) r(s,a) \tag{33}$$

From (31), (32) and (33), we conclude that we can obtain the optimal policy from the optimal dual variable:

$$\pi^\star(s,a) = \frac{d^\star(s,a)}{\sum_{a\in A} d^\star(s,a)} \tag{34}$$

This conclusion is very interesting: while the primal linear problem searches the optimal value, the dual formulation searches in the policy space. This suggests that we can develop a primal-dual method that finds a saddle point of the Lagrangaian by searching both in the value function and the policy spaces.

II. Saddle-Point formulation of Actor-Critic Methods

If we derive the dual of the dual problem (24), we recover the original primal problem (17). Indeed, since problems (17) and (24) are convex and satisfy Slaters condition [5], strong duality holds and their respective primal and dual optimal values are attained and equal and they form a saddle-point of their

Lagrangian.

Recall from (20) the Lagrangian of the primal problem (17):

$$L_{\mathcal{P}_1}(\mathbf{v}, \mathbf{d}) = (1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} + \mathbf{d}^\top (\mathbf{r} + \gamma P \mathbf{v} - \Phi^\top \mathbf{v}) \tag{35}$$

First order conditions for a saddle point of (35) are:

$$\nabla_{\mathbf{v}} L_{\mathcal{P}_1}(\mathbf{v}, \mathbf{d}) = (1-\gamma)\boldsymbol{\mu} + \Big(\gamma P^\top - \Phi\Big)\, \mathbf{d} = \mathbf{0} \tag{36}$$

$$\nabla_{\mathbf{d}} L_{\mathcal{P}_1}(\mathbf{v}, \mathbf{d}) = \mathbf{r} + \gamma P \mathbf{v} - \Phi^\top \mathbf{v} = \mathbf{0} \tag{37}$$

The Lagrangian of the dual problem (24) is given by:

$$L_{\mathcal{P}_2}(\mathbf{d}, \mathbf{g}, \mathbf{v}) = \mathbf{d}^\top \mathbf{r} + \mathbf{g}^\top \mathbf{d} + \mathbf{v}^\top \Big((1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d} - \Phi \mathbf{d}\Big) \tag{38}$$

with first order conditions for a saddle point given by

$$\nabla_{\mathbf{d}} L_{\mathcal{P}_2}(\mathbf{d}, \mathbf{g}, \mathbf{v}) = \mathbf{r} + \mathbf{g} + \gamma P \mathbf{v} - \Phi^\top \mathbf{v} = \mathbf{0} \tag{39}$$

$$\nabla_{\mathbf{g}} L_{\mathcal{P}_2}(\mathbf{d}, \mathbf{g}, \mathbf{v}) = \mathbf{d} = \mathbf{0} \tag{40}$$

$$\nabla_{\mathbf{v}} L_{\mathcal{P}_2}(\mathbf{d}, \mathbf{g}, \mathbf{v}) = (1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d} - \Phi \mathbf{d} = \mathbf{0} \tag{41}$$

To do: Now, the idea could be to derive standard and derive novel actor-critic methods from these first order conditions.

III. Some Remarks to Explore

Remark 1. *I guess that* (30) *and* (31) *are related to the deterministic policy gradient theorem [6].*

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 4th ed. Athena Scientific, 2012, vol. 2.

[2] D. P. de Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations Research*, vol. 51, no. 6, pp. 850–865, 2003.

[3] T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans, "Dual representations for dynamic programming," *Journal of Machine Learning Research*, pp. 1–29, 2008.

[4] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour *et al.*, "Policy gradient methods for reinforcement learning with function approximation." in Proc. *Advances in Neural Information Processing Systems*, vol. 99, 1999, pp. 1057–1063.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[6] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in Proc. *Int. Conf. on Machine Learning*, Beijing, China, 2014.