A NOTE ON OPTIMIZATION FORMULATIONS OF MARKOV DECISION PROCESSES*

LEXING YING† AND YUHUA ZHU‡

Abstract. This note summarizes the optimization formulations used in the study of Markov decision processes. We consider both the discounted and undiscounted processes under the standard and the entropy-regularized settings. For each setting, we first summarize the primal, dual, and primal-dual problems of the linear programming formulation. We then detail the connections between these problems and other formulations for Markov decision processes such as the Bellman equation and the policy gradient formulation.

Keywords. Markov decision processes; Reinforcement learning; Optimization.

AMS subject classifications. 60J10; 60J22; 90C05.

1. Introduction

Most of the algorithms of Markov decision processes (MDPs) are derived from the fixed-point iteration of the Bellman equation [3]. Examples include value iteration [4,5, 23], policy iteration [3,11], temporal difference (TD) learning [25], Q-learning [33], etc. The analyses of these algorithms in the tabular case and linear function approximation case often leverage the contraction property of the Bellman operator. In the past decade or so, nonlinear approximations such as neural networks have become more popular. However, for nonlinear function approximations, this contraction property no longer holds, often resulting in instability. Many variants and modifications have been proposed to stabilize the training, e.g., DQN [17], A3C [16]. However, theoretical guarantees for these algorithms are still missing.

A second perspective of studying MDPs is based on optimization. For nonlinear approximations, optimization formulations are often more convenient both for algorithmic design and mathematical analysis as they guarantee convergence to at least local minimums. Therefore in recent years, more attention has been given to the optimization framework. One major direction is based on linear programming (LP) [22] and some recent developments include [1,7,29,30]. Another direction is the Bellman residual minimization (BRM) [2], which includes algorithms based on the primal-dual form of the BRM [6,8,27], stochastic compositional gradient (SCGD) methods based on two-scale separation [31,32], algorithms based on the smoothness of the underlying transition dynamics [13,35,36]. The convergence properties of these algorithms have been studied in [14,15,18,28,34].

The contribution of this note is two-fold. First, we summarize the LP problems used in the study of MDPs. Many results in this note are well-known, but we were not able to find a place where these results are summarized in a uniform framework. Second, we point out the connections between the LP problems and other MDP formulations, including the equivalence between the dual problem and the policy gradient formulation and the equivalence between the primal problem and the Bellman equation.

†Department of Mathematics, Stanford University, Stanford, CA 94305, USA (lexing@stanford.edu).
‡Department of Mathematics, Stanford University, Stanford, CA 94305, USA (yuhuazhu@stanford.edu).

1.1. Notation.   A Markov decision process $\mathcal{M}$ with discrete state and action spaces is characterized by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here $\mathcal{S}$ is the discrete state space, with each state usually denoted by s. $\mathcal{A}$ is the discrete action space, with each action usually denoted by $a \in \mathcal{A}$. Throughout the note, $|\mathcal{S}|$ and $|\mathcal{A}|$ are used to denote the size of $\mathcal{S}$ and $\mathcal{A}$, respectively. P is a third-order tensor where, for each action $a \in \mathcal{A}$, $P^a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the transition matrix between the states, i.e., P^a_{st} is the probability of arriving at state t if action a is taken at state s. r is a second-order tensor where, for each action $a \in \mathcal{A}$, r^a_s is the reward at state s if action a is taken. Finally, $\gamma \in [0,1]$ is the discount factor.

Let Δ be the probability simplex over the space of actions, i.e.,

$$\Delta = \left\{ \eta = (\eta^a)_{a \in \mathcal{A}} : \sum_{a \in \mathcal{A}} \eta^a = 1 \text{ and } \eta^a \geq 0 \text{ for } \forall a \in \mathcal{A} \right\}.$$

The set of all valid policies is defined to be

$$\Delta^{|\mathcal{S}|} = \left\{ \pi = (\pi_s)_{s \in \mathcal{S}} : \pi_s \in \Delta \text{ for } \forall s \in \mathcal{S} \right\}.$$

For a policy $\pi \in \Delta^{|\mathcal{S}|}$, the transition matrix $P^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ under the policy π is defined as

$$P^\pi_{st} = \sum_{a \in \mathcal{A}} P^a_{st} \pi^a_s, \tag{1.1}$$

i.e., P^π_{st} is the probability of arriving at state t from state s if policy π is taken. The reward $r^\pi \in \mathbb{R}^{|\mathcal{S}|}$ under the policy π is

$$r^\pi_s = \sum_{a \in \mathcal{A}} r^a_s \pi^a_s, \tag{1.2}$$

i.e., r^π_s represents the expected reward at state s if policy π is taken.

Each policy π induces a discrete Markov process, where at each round m, an action a_m is chosen at state s_m according to a particular policy π, and then the agent arrives at state s_{m+1} according to the distribution of the transition matrix P^{a_m} and receives a reward $r^{a_m}_{s_m}$. The goal of an MDP problem is to maximize the cumulative reward among all possible policies. Depending on whether γ is strictly less than one or not, an MDP can either be **discounted** ($\gamma < 1$) or **undiscounted** ($\gamma = 1$).

When solving the MDPs, entropy regularizer has been proved quite useful in terms of exploration and convergence [9,16,21,24]. In this note, we adopt the following *negative conditional entropy* defined for non-negative $\rho \in \mathbb{R}^{|\mathcal{A}|}$:

$$h(\rho) := \sum_{a \in \mathcal{A}} \rho^a \log \frac{\rho^a}{\sum_{b \in \mathcal{A}} \rho^b}. \tag{1.3}$$

This entropy $h(\rho)$ is both convex and homogeneous of degree one in ρ (see for example Appendix A.1 of [18]). This regularizer has been widely used in the literature [8–10, 16, 21, 24]. Depending on whether this regularizer is used, we call an MDP either **standard** or **regularized**.

1.2. Outline.   The rest of the note is organized as follows. In Section 2, we first derive the primal, dual, and primal-dual problems for the discounted standard MDP. We then show the equivalence between the policy gradient formulation and the

	discounted unregularized	discounted regularized	undiscounted unregularized	undiscounted regularized
primal:	(2.2) /(2.6)	(3.2)/(3.3)/ (3.7)/(3.8)	(4.2).(4.6)	(5.2)/(5.3)/ (5.6)/(5.7)
primal-dual:	(2.3)	(3.1)	(4.3)	(5.1)
dual:	(2.4) /(2.5)	(3.4) /(3.6)	(4.4)/(4.5)	(5.4) /(5.5)
theorems	Theorems 2.1 & 2.2	Theorems 3.1 & 3.2	Theorems 4.1 & 4.2	Theorems 5.1 & 5.2

TABLE 1.1. *Summary of the optimization formulations for different kinds of MDPs.*

dual problem as well as the equivalence between the Bellman equation and the primal problem. Sections 3, 4, and 5 address the discounted regularized MDP, the undiscounted standard MDP, and the undiscounted regularized MDP, respectively, by following the same outline. We summarize all the main results in Table 1.1.

2. Discounted standard MDP

The discounted standard MDP is probably the most studied case in literature [23, 26]. For $\gamma \in (0,1)$, the value function under policy π is a vector $v^\pi \in \mathbb{R}^{|\mathcal{S}|}$, where v_s^π represents the expected discounted cumulative reward starting from state s under the policy π, i.e.,

$$v_s^\pi = \mathbb{E}\left[\sum_{m=0}^\infty \gamma^m r_{s_m}^{a_m} | s_0 = s \right], \tag{2.1}$$

where the expectation is taken over $a_m \sim \pi_{s_m}, s_{m+1} \sim P_{s_m,\cdot}^{a_m}$, for all $m \geq 0$. The value function naturally satisfies the Bellman equation for any $s \in \mathcal{S}$:

$$v_s^\pi = r_s^\pi + \gamma \mathbb{E}^\pi [v_{s_1}^\pi | s_0 = s] = r_s^\pi + \gamma \sum_{t \in \mathcal{S}} P_{st}^\pi v_t^\pi.$$

The goal of an MDP problem is to find the maximum value function among all possible policies.

2.1. LP problems.
Primal problem. The primal problem of finding the maximum value function reads

$$\min_v \ \sum_{s \in \mathcal{S}} e_s v_s, \ \text{s.t.} \ \forall a, \forall s, \ r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s \leq 0, \tag{2.2}$$

where $e \in \mathbb{R}^{|\mathcal{S}|}$ is an arbitrary vector with positive entries. An example with 2 states and 2 actions is illustrated in Figure 2.1, where the pink region represents the constraints and the pink arrow points to the minimization direction. The optimal solution of this minimization problem is the red point. As shall see later in Section 2.2, the optimal solution of the dual problem is the same red point, coming from the opposite direction.

Primal-dual problem. By introducing the Lagrangian multiplier μ_s^a for the inequality constraints, we arrive at the primal-dual

$$\min_{v_s} \max_{\mu_s^a \geq 0} \ \sum_{s \in \mathcal{S}} e_s v_s + \sum_{s,a} (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s)\mu_s^a, \tag{2.3}$$
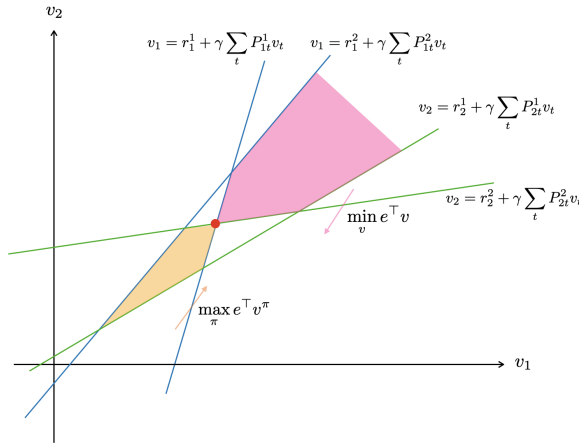
FIG. 2.1. *This plot interprets the primal and dual problem for an MDP with 2 states and 2 actions. The pink region represents the constraints for the primal problem, while the yellow region represents the constraints for the dual problem. The pink and yellow arrows indicate the minimization and maximization directions, respectively. Both formulations end up at the same red point, but from the opposite directions.*

or equivalently in the matrix-vector notation

$$\min_{v}\max_{\mu^a \geq 0}\ e^{\top}v+\sum_{a\in\mathcal{A}}(\mu^a)^{\top}(r^a+\gamma P^a v-v),$$

where $(\cdot)^{\top}$ stands for transpose. This primal-dual problem is, for example, used in [30].

Dual problem. Since the minimum of v_s is taken over a convex function and the maximum of μ_s^a is over a concave function, one can exchange the order of minimum and maximum because of the minimax theorem [19]. The primal-dual problem can also be written as

$$\max_{\mu^a \geq 0}\min_{v}\ e^{\top}v+\sum_{a\in\mathcal{A}}(\mu^a)^{\top}(r^a+\gamma P^a v-v).$$

Taking derivative with respect to v and setting it to be zero gives rise to

$$e=-\sum_{a\in\mathcal{A}}(\gamma(P^a)^{\top}-I)\mu^a,\quad i.e.,\quad \sum_{a\in\mathcal{A}}(I-\gamma(P^a)^{\top})\mu^a=e.$$

Hence the dual problem is

$$\max_{\mu^a \geq 0}\sum_{a\in\mathcal{A}}(r^a)^{\top}\mu^a,\ \text{s.t.}\ \sum_{a\in\mathcal{A}}(I-\gamma(P^a)^{\top})\mu^a=e. \tag{2.4}$$

This dual problem is mentioned, for example, in [34].

The above formulations are summarized as the following theorem.

THEOREM 2.1.   *For the discounted standard MDP, the following three formulations*

are equivalent,

$$\textbf{Primal}: \quad \min_{v \in \mathbb{R}^{|\mathcal{S}|}} \sum_{s \in \mathcal{S}} e^\top v, \ s.t. \ \forall a, \forall s, \ r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s \leq 0,$$

$$\textbf{Primal-dual}: \quad \min_{\substack{v \in \mathbb{R}^{|\mathcal{S}|}}} \max_{\substack{\mu^a \in \mathbb{R}^{|\mathcal{S}|} \\ \mu^a \geq 0}} e^\top v + \sum_{a \in \mathcal{A}} (\mu^a)^\top (r^a + \gamma P^a v - v),$$

$$\textbf{Dual}: \quad \max_{\substack{\mu^a \in \mathbb{R}^{|\mathcal{S}|} \\ \mu^a \geq 0}} \sum_{a \in \mathcal{A}} (r^a)^\top \mu^a, \ s.t. \ \sum_{a \in \mathcal{A}} (I - \gamma (P^a)^\top) \mu^a = e,$$

where $e \in \mathbb{R}^{|\mathcal{S}|}$ is an arbitrary vector with positive entries.

2.2. Equivalences.

Dual problem and policy gradient. The dual problem (2.4) is equivalent to the policy gradient formulation. To see this, let us parameterize $\mu_s^a = w_s \pi_s^a$ with $w_s = \sum_{a \in \mathcal{A}} \mu_s^a$. This ensures that $\pi \in \Delta^{|\mathcal{S}|}$ because $\sum_{a \in \mathcal{A}} \pi_s^a = \sum_{a \in \mathcal{A}} \frac{\mu_s^a}{w_s} = 1$ and $\pi_s^a \geq 0$. By this new parameterization, the constraints in dual become

$$(I - \gamma(P^\pi)^\top) w = e, \quad \text{or} \quad w = (I - \gamma(P^\pi)^\top)^{-1} e,$$

where P_{st}^π is defined in (1.1) as the transition matrix under policy π. By denoting this w as w^π to indicate its π dependence, we can write $\sum_{a \in \mathcal{A}} (r^a)^\top \mu^a = (r^\pi)^\top w^\pi$. As a result, the dual problem (2.4) can be written as

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} r^\pi \cdot (I - \gamma(P^\pi)^\top)^{-1} e, \quad \text{or} \quad \max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top (I - \gamma P^\pi)^{-1} r^\pi,$$

which is equivalent to

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top v^\pi, \ \text{s.t.} \ v^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

Since the constraint can be written as $(I - \gamma P^\pi) v^\pi = r^\pi$, it implies that $v^\pi = r^\pi + \gamma P^\pi v^\pi$. Hence, the dual problem is equivalent to the optimization formulation of the policy gradient method,

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top v^\pi, \ \text{s.t.} \ v^\pi = r^\pi + \gamma P^\pi v^\pi, \tag{2.5}$$

where we recall that $e \in \mathbb{R}^{|\mathcal{S}|}$ is any vector with positive entries. Therefore, **the optimization formulation that the policy gradient method is based on can be viewed as a nonlinear reparameterization of the dual LP problem**. This understanding is also illustrated in Figure 2.1, where the yellow region represents the constraints and the yellow arrow points to the maximum direction. Notice that both the primal and dual problems end up at the same red point from opposite directions.

Primal problem and Bellman equation. Next, we show that the primal problem is equivalent to solving the Bellman equation

$$v_s = \max_{a \in \mathcal{A}} \left(r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t \right). \tag{2.6}$$

Bellman equation to primal problem. The derivation from (2.6) to (2.2) can be found, for example, in [23]. We provide a short derivation here for completeness. Let v^* be the solution to (2.6), then for each s, there exists $a_s^* \in \mathcal{A}$ s.t.,

$$v^* = r^{a^*} + \gamma P^{a^*} v^*,$$

where $r_s^{a^*} \equiv r_s^{a_s^*}, P_{st}^{a^*} \equiv P_{st}^{a_s^*}$. For any v that satisfies the constraints in the primal problem (2.2), the following inequality holds

$$v \geq r^{a^*} + \gamma P^{a^*} v.$$

Subtracting these two equations gives

$$v - v^* \geq \gamma P^{a^*}(v - v^*).$$

Since P^{a^*} is a probability transition matrix, by maximum principle, one has $v - v^* \geq 0$, and thus

$$e^\top v \geq e^\top v^*$$

for all v satisfying the constraints in (2.2). This proves that v^* is the minimizer of the primal problem (2.2).

Primal problem to Bellman equation. Let v^* be the minimizer of the primal problem (2.2). The KKT conditions for (2.2) read

$$
\begin{cases}
r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t^* \leq v_s^*, & \text{for } \forall s, a; \\
\sum_{a \in \mathcal{A}} \left(\mu_s^a - \gamma \sum_{t \in \mathcal{S}} P_{ts}^a \mu_t^a \right) = e_s, & \text{for } \forall s; \\
\mu_s^a \left(r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* \right) = 0 & \text{for } \forall s, a.
\end{cases}
$$

First, we claim it is impossible that there exists s such that for $\forall a$, $r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* > 0$. If it were true, then from the last equation, one would have $\mu_s^a = 0$ for all a. Let $\mu_s^a = w_s \pi_s^a$, then $w_s = 0$ for this s. Inserting it into the second equation, one has

$$w_s - \gamma \sum_{t \in \mathcal{S}} P_{ts}^\pi w_t = -\gamma \sum_{t \in \mathcal{S}} P_{ts}^\pi w_t = e_s.$$

Since both $P_{ts}^\pi, w_t \geq 0$ for all t, then LHS ≤ 0. However, as the RHS $e_s > 0$, we reach a contradiction. Therefore, the claim is true, i.e., there does not exist any s such that $\forall a$, $r_s^a + g \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* > 0$.

Therefore for any fixed s, there exists a_s^*, s.t.,

$$r_s^{a_s^*} + \gamma \sum_{t \in \mathcal{S}} P_{st}^{a_s^*} v_t^* = v_s^*.$$

For all $a \neq a_s^*$, by the first equation, one has

$$r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t^* \leq v_s^*.$$

Combining the above two equations leads to $v_s^* = \max_a (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t^*)$. Therefore, the minimizer v^* also satisfies the Bellman Equation (2.6) for all s.

The above two equivalences are stated in the following theorem.

THEOREM 2.2. *For the discounted standard MDP, the primal formulation* (2.2) *is equivalent to solving the Bellman equation*

$$v_s = \max_{a \in \mathcal{A}} \left(r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t \right), \quad \forall s \in \mathcal{S},$$

and the dual formulation (2.4) *is equivalent to the optimization formulation of the policy gradient method*

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top v^\pi, \ s.t. \ v^\pi = r^\pi + \gamma P^\pi v^\pi.$$

3. Discounted regularized MDP

For $0 < \mu \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, the discounted regularized MDP includes the negative conditional entropy

$$h(\mu_s) = \sum_{a \in \mathcal{A}} \mu_s^a \log \frac{\mu_s^a}{\sum_{b \in \mathcal{A}} \mu_s^b}$$

in the objective function, where for each $s \in \mathcal{S}$, $\mu_s \in \mathbb{R}^{|\mathcal{A}|}$ and for each $a \in \mathcal{A}$, $\mu^a \in \mathbb{R}^{|\mathcal{S}|}$.

3.1. LP problems.

Primal-dual problem. Let us introduce first in the primal-dual problem:

$$\min_{v_s} \sup_{\mu_s^a > 0} \sum_{s \in \mathcal{S}} e_s v_s + \sum_{s,a} \left(r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s \right) \mu_s^a - \sum_{s \in \mathcal{S}} h(\mu_s). \tag{3.1}$$

As $h(\mu_s)$ is convex in μ_s, this objective function is concave in μ_s^a. In connection with the primal-dual problem (2.3) of the standard case, including the extra entropic term allows for replacing the condition $\mu_s^a \geq 0$ with $\mu_s^a > 0$. In the literature, it is common for the entropy term to have a prefactor $\eta > 0$. Here we simply assume $\eta = 1$ as one can always reduce to this case by rescaling the rewards r_s^a.

Primal problem. By introducing $\mu_s^a = w_s \pi_s^a$ with $w_s = \sum_{a \in \mathcal{A}} \mu_s^a$ and $\pi \in \Delta^{|\mathcal{S}|}$, (3.1) becomes

$$\min_{v_s} \sup_{\pi \in \Delta^{|\mathcal{S}|}, \pi_s^a, w_s > 0} \sum_{s \in \mathcal{S}} e_s v_s + \sum_{s,a} (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s) \pi_s^a w_s - \sum_{s \in \mathcal{S}} w_s h(\pi_s),$$

where we use the fact that $h(\cdot)$ is homogeneous of degree one. This is equivalent to

$$\min_{v_s} \left(\sum_{s \in \mathcal{S}} e_s v_s + \sup_{w_s > 0} \sum_{s \in \mathcal{S}} w_s \cdot \max_{\pi_s \in \Delta} \left(\sum_{a \in \mathcal{A}} (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s) \pi_s^a - h(\pi_s) \right) \right).$$

Since the inner optimal over π_s cannot lie on the boundary, it is the same to write the optimal as $\max_{\pi_s \in \Delta}$ and $\sup_{\pi_s \in \Delta, \pi_s > 0}$. The primal problem of the above minimax problem is then given by

$$\min_{v_s} e^\top v, \ s.t. \ \forall s, \ \max_{\pi_s \in \Delta} \sum_{a \in \mathcal{A}} (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s) \pi_s^a - h(\pi_s) \leq 0,$$

or equivalently,

$$\min_v e^\top v, \ s.t. \ \max_\pi (r^\pi + \gamma P^\pi v - h^\pi) \leq v. \tag{3.2}$$

Note that the maximization in the constraint

$$\max_{\pi_s \in \Delta} \left(\sum_{a \in \mathcal{A}} (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s) \pi_s^a - h(\pi_s) \right)$$

is in the form of the Gibbs variational principle. Therefore, the optimizer for this maximization is

$$\pi_s^a = \frac{\exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t - v_s)}{Z_s}$$

where $Z_s = \sum_{a\in\mathcal{A}} \exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t - v_s)$ is the normalization factor and the maximal value is $\log Z_s$. Hence, the constraint is equivalent to $\log Z_s \leq 0$, i.e.,

$$\sum_{a\in\mathcal{A}} \exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t - v_s) \leq 1$$

i.e.

$$e^{-v_s} \cdot \sum_{a\in\mathcal{A}} \exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t) \leq 1, \quad \text{or} \quad \sum_{a\in\mathcal{A}} \exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t) \leq e^{v_s}.$$

Taking log on both sides leads to

$$v_s \geq \log \left(\sum_{a\in\mathcal{A}} \exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t) \right).$$

This implies that the primal problem (3.2) is equivalent to

$$\min_{v_s} \; e^\top v, \text{ s.t. } \forall s, v_s \geq \log \left(\sum_{a\in\mathcal{A}} \exp(r_s^a + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t) \right). \tag{3.3}$$

Dual problem. The supremum and minimum in the primal-dual problem (3.1) can be exchanged because the objective function is convex in v and concave in μ. Now one has,

$$\sup_{\mu_s^a > 0} \min_{v_s} \sum_{s\in\mathcal{S}} e_s v_s + \sum_{s,a} (r_s + \gamma \sum_{t\in\mathcal{S}} P_{st}^a v_t - v_s) \mu_s^a - \sum_{s\in\mathcal{S}} h(\mu_s).$$

Taking derivative in v gives $\sum_{a\in\mathcal{A}} (I - \gamma(P^a)^\top) \mu^a = e$. Therefore, the dual problem takes the form

$$\sup_{\mu^a > 0} \sum_{a\in\mathcal{A}} (r^a)^\top \mu^a - \sum_{s\in\mathcal{S}} h(\mu_s), \text{ s.t. } \sum_{a\in\mathcal{A}} (I - \gamma(P^a)^\top) \mu^a = e. \tag{3.4}$$

The above three formulations are summarized as the following theorem.

THEOREM 3.1. *For the discounted regularized MDP, the following three formulations are equivalent,*

Primal: $\quad \min_{v\in\mathbb{R}^{|\mathcal{S}|}} \; e^\top v, \; s.t. \; v \geq \max_{\pi} (r^\pi + \gamma P^\pi v - h^\pi),$

or equivalently, $\quad \min_{v\in\mathbb{R}^{|\mathcal{S}|}} \; e^\top v, \; s.t. \; v \geq \log \left(\sum_{a\in\mathcal{A}} \exp(r^a + \gamma P^a v) \right),$

Primal − dual: $\quad \min_{v\in\mathbb{R}^{|\mathcal{S}|}} \sup_{\substack{\mu\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \\ \mu>0}} e^\top v + \sum_a (\mu^a)^\top (r^a + \gamma P^a v - v) - \sum_{s\in\mathcal{S}} h(\mu_s),$

Dual: $\quad \sup_{\substack{\mu\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \\ \mu>0}} \sum_{a\in\mathcal{A}} (r^a)^\top \mu^a - \sum_{s\in\mathcal{S}} h(\mu_s), \; s.t. \; \sum_{a\in\mathcal{A}} (I - \gamma(P^a)^\top) \mu^a = e,$

where $e\in\mathbb{R}^{|\mathcal{S}|}$ is an arbitrary vector with positive entries.

3.2. Equivalences.

Dual problem and policy gradient. We claim that the dual problem (3.4) is again equivalent to the policy gradient formulation. As before, let us parameterize $\mu_s^a = w_s \pi_s^a$ with $w_s = \sum_{a \in \mathcal{A}} \mu_s^a$ and $\pi \in \Delta^{|\mathcal{S}|}$. Then the constraints in (3.4) become

$$\forall s, \quad \sum_{a \in \mathcal{A}} \pi_s^a w_s - \gamma \sum_{a,t} P_{ts}^a \pi_t^a w_t = e_s, \quad \text{or} \quad w = (I - \gamma (P^\pi)^\top)^{-1} e.$$

By denoting the solution w by w^π to show its π dependence, one can rewrite (3.4) as

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} r^\pi \cdot w^\pi - \sum_{s \in \mathcal{S}} w_s^\pi \left(\sum_{a \in \mathcal{A}} \pi_s^a \log \pi_s^a \right)$$

By further introducing $h^\pi \in \mathbb{R}^{|\mathcal{S}|}$ as the vector with entry $h_s^\pi = h(\pi_s) = \sum_{a \in \mathcal{A}} \pi_s^a \log \pi_s^a$, we transform (3.4) to

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} (r^\pi - h^\pi) \cdot w^\pi, \quad \text{or} \quad \max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top (I - \gamma P^\pi)^{-1} (r^\pi - h^\pi). \tag{3.5}$$

We can also view $r^\pi - h^\pi$ as a regularized reward by subtracting the entropy function h^π. The value function $v^\pi = \mathbb{E}[\sum_{m \geq 0} \gamma^m (r_{s_m}^{a_m} - h(\pi_{s_m}))]$ under the new reward satisfies the regularized Bellman equation $v^\pi = r^\pi - h^\pi + \gamma P^\pi v^\pi$. Hence, the optimization formulation of the policy gradient method for this regularized discounted MDP is

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top v^\pi, \quad s.t., \quad v^\pi = r^\pi - h^\pi + \gamma P^\pi v^\pi, \tag{3.6}$$

which is clearly equivalent to (3.5) by letting $v^\pi = (I - \gamma P^\pi)^{-1} (r^\pi - h^\pi)$.

Primal problem and Bellman equation. The regularized Bellman equation is

$$v = \max_{\pi \in \Delta^{|\mathcal{S}|}} r^\pi + \gamma P^\pi v - h^\pi. \tag{3.7}$$

In each component,

$$v_s = \max_{\pi_s \in \Delta} \sum_{a \in \mathcal{A}} (r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t) \pi_s^a - h(\pi_s).$$

By the Gibbs variational principle, the RHS is equal to $\log(\exp(\sum_{a \in \mathcal{A}} (r_s + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t)))$. Therefore, the regularized Bellman equation could also be written as the following log-sum-exp form

$$v_s = \log \left(\sum_{a \in \mathcal{A}} \exp \left(r_s^a + \gamma \sum_{t \in \mathcal{S}} P_{st}^a v_t \right) \right). \tag{3.8}$$

Below we show that the primal problem (3.2) is equivalent to solving (3.7).

Bellman equation to primal problem. Let v^* be the solution of (3.7). Then there exists π^* s.t.,

$$v^* = r^{\pi^*} + \gamma P^{\pi^*} v^* - h^{\pi^*}.$$

For any v that satisfies the constraints of the primal problem, the following inequality holds

$$v \geq r^{\pi^*} + \gamma P^{\pi^*} v - h^{\pi^*}.$$

Subtracting these two equations gives rise to

$$v - v^* \geq \gamma P^{\pi^*} (v - v^*).$$

Again by maximum principle, one has $v - v^* \geq 0$ and hence

$$e^\top v \geq e^\top v^*$$

for all v that satisfy the constraints in (3.2). This proves that v^* is the minimizer of the primal problem (3.2).

Primal problem to Bellman equation. Let v^* be the minimizer of the primal problem (3.2). We now prove v^* is also the solution to the Bellman Equation (3.7) by contradiction. Assume that v^* does not satisfy (3.7). Then there must exist $\bar{s}$, s.t. for $\forall \pi$

$$v_{\bar{s}}^* \geq (r^\pi + \gamma P^\pi v^* - h^\pi)_{\bar{s}} + \delta$$

with some constant $\delta > 0$. Let us define $\bar{v}$ s.t., $\bar{v}_{\bar{s}} = v_{\bar{s}}^* - \delta$ and $\bar{v}_s = v_s^*$ for $s \neq \bar{s}$. We claim that for $\forall \pi$

$$\bar{v}_s \geq (r^\pi + \gamma P^\pi \bar{v} - h^\pi)_s, \quad \forall s.$$

First, for $s \neq \bar{s}$ the above inequality holds because $\bar{v}_s = v_s^*$ and v_s^* satisfies the constraints in the primal problem (3.2). For $s = \bar{s}$, one has $\bar{v}_{\bar{s}} = v_{\bar{s}}^* - \delta \geq (r^\pi + \gamma P^\pi v^* - h^\pi)_{\bar{s}}$. Since

$$(P^\pi v^*)_{\bar{s}} = \sum_{t \in \mathcal{S}} P_{\bar{s}t}^\pi v_t^* = \sum_{t \neq \bar{s}} P_{\bar{s}t}^\pi v_t^* + P_{\bar{s}\bar{s}}^\pi (v_{\bar{s}}^* - \delta) + \delta = \sum_{t \in \mathcal{S}} P_{\bar{s}t}^\pi \bar{v}_t + \delta,$$

$\bar{v}_{\bar{s}} \geq (r^\pi + \gamma P^\pi \bar{v} - h^\pi)_{\bar{s}} + \delta$. This completes the proof of the claim. This means that $\bar{v}$ also satisfies the constraints of the primal problem, but $e^\top \bar{v} < e^\top v^*$ by construction, which contradicts with v^* being the minimizer of the primal problem. Therefore, the assumption is wrong and v^* satisfies (3.7).

The above two equivalences are stated in the following theorem.

THEOREM 3.2. *For the discounted regularized MDP, the primal formulations* (3.2) *and* (3.3) *are equivalent to solving the following equations*

$$v = \max_{\pi \in \Delta^{|\mathcal{S}|}} r^\pi + \gamma P^\pi v - h^\pi, \quad \text{or equivalently,} \quad v = \log \left[\sum_{a \in \mathcal{A}} \exp \left(r^a + \gamma \sum_{t \in \mathcal{S}} P^a v \right) \right],$$

and the dual formulation (3.4) *is equivalent to the optimization formulation of the policy gradient method*

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} e^\top v^\pi, \quad s.t., \quad v^\pi = r^\pi - h^\pi + \gamma P^\pi v^\pi.$$

4. Undiscounted standard MDP

In this section, we consider the MDP without discounts, i.e., $\gamma = 1$. Besides, we assume the MDP is unichain, i.e., for each policy π, the MDP induced by policy π is ergodic [20]. Let $\rho^\pi \in \mathbb{R}$ be the average reward under policy π,

$$\rho^\pi = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{m=1}^{T} r_{s_m}^{a_m}\right],$$

where the expectation is taken over $a_m \sim \pi_{s_m}, s_{m+1} \sim P_{s_m,\cdot}^{a_m}$. Notice for an ergodic Markov process, the average-reward is the same for any initial state because the stationary distribution is unique invariant, strictly positive and independent of initial states (Chapter 6 of [12]). Let $w^\pi \in \mathbb{R}^{|\mathcal{S}|} > 0$ be the stationary distribution induced by policy π satisfying $w_s^\pi = \sum_{t \in \mathcal{S}} P_{ts}^\pi w_t^\pi$. Then w^π is the only solution to $(I - (P^\pi)^\top)w^\pi = 0$ with $\sum_{s \in \mathcal{S}} w_s^\pi = 1$. The average reward under policy π could also be written as

$$\rho^\pi = (r^\pi)^\top w^\pi.$$

Define the value function for the non-discount MDP as $v_s^\pi = \mathbb{E}_\pi[\sum_{m=1}^{\infty}(r_{s_m}^{a_m} - \rho^\pi)|s_0 = s]$, then the value function satisfies the average-reward Bellman equation:

$$v_s^\pi = r_s^\pi - \rho^\pi + \sum_{t \in \mathcal{S}} P_{st}^\pi v_t^\pi \tag{4.1}$$

with $\sum_{s \in \mathcal{S}} v_s^\pi w_s^\pi = 1$. Note that without this constraint, there are infinitely many v^π, e.g., $v^\pi + C$ for any constant C still satisfies the above equation. The goal here is to find the maximum average reward among all possible policies.

4.1. LP problems.
Primal problem. The primal problem is

$$\min_{v_s, \rho} \rho \quad \text{s.t.} \quad \forall a, \forall s, \ r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s - \rho \leq 0. \tag{4.2}$$

This can be found, for example, in [18].

Primal-dual problem. By including the Lagrangian multiplier μ_s^a for the inequality constraints, one obtains the primal-dual problem

$$\min_{v_s, \rho} \max_{\mu_s^a \geq 0} \rho + \sum_{s,a}\left(r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s - \rho\right)\mu_s^a, \tag{4.3}$$

or equivalently, in the matrix-vector notation

$$\min_{v_s, \rho} \max_{\mu_s^a \geq 0} \ \rho + \sum_{a \in \mathcal{A}}(\mu^a)^\top(r^a + P^a v - v - \rho\mathbf{1}),$$

where $\mathbf{1}$ is the $|S|$-dimensional vector with all elements equal to 1.

Dual problem. To get the dual problem, we take the derivative with respect to ρ to get

$$1 - \sum_{s,a} \mu_s^a = 0.$$

Taking the derivative with respect to v leads to

$$\sum_{a \in \mathcal{A}} (I - (P^a)^\top)\mu^a = 0.$$

Hence the dual problem is

$$\max_{\mu_s^a \geq 0} \sum_{a \in \mathcal{A}} (r^a)^\top \mu^a, \text{ s.t. } \sum_{a \in \mathcal{A}} (I - (P^a)^\top)\mu^a = 0, \quad 1 - \sum_{s,a} \mu_s^a = 0. \tag{4.4}$$

The above three formulations are summarized as the following theorem.

THEOREM 4.1. *For the undiscounted standard MDP, the following three formulations are equivalent,*

$$\textbf{Primal}: \min_{\substack{v \in \mathbb{R}^{|\mathcal{S}|} \\ \rho \in \mathbb{R}}} \rho, \quad s.t. \ \forall a, \ r^a + P^a v - v - \rho\mathbf{1} \leq 0,$$

$$\textbf{Primal-dual}: \min_{\substack{v \in \mathbb{R}^{|\mathcal{S}|} \\ \rho \in \mathbb{R}}} \max_{\substack{\mu^a \in \mathbb{R}^{|\mathcal{S}|} \\ \mu^a \geq 0}} \rho + \sum_{a \in \mathcal{A}} (\mu^a)^\top (r^a + P^a v - v - \rho\mathbf{1}),$$

$$\textbf{Dual}: \max_{\substack{\mu^a \in \mathbb{R}^{|\mathcal{S}|} \\ \mu^a \geq 0}} \sum_{a \in \mathcal{A}} (r^a)^\top \mu^a, \quad s.t. \ \sum_{a \in \mathcal{A}} (I - (P^a)^\top)\mu^a = 0, \quad \sum_{s,a} \mu_s^a = 1.$$

4.2. Equivalences.

Dual problem and policy gradient. The dual problem (4.4) is equivalent to the policy gradient formulation. Let us again parameterize $\mu_s^a = w_s \pi_s^a$ with $w_s = \sum_{a \in \mathcal{A}} \mu_s^a$ and $\pi \in \Delta^{|\mathcal{S}|}$. By the new parameterization, the constraints become

$$\sum_{a \in \mathcal{A}} (I - (P^a)^\top)\mu^a = 0 \Rightarrow (I - (P^\pi)^\top)w = 0.$$

$1 - \sum_{s,a} \mu_s^a = 0$ also implies that $1 - \sum_{s \in \mathcal{S}} w_s = 0$. Together we conclude that w is the stationary distribution induced by π. By denoting this w by w^π, we can write the dual problem as

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} r^\pi \cdot w^\pi, \quad s.t. \ (I - (P^\pi)^\top)w^\pi = 0, \quad \sum_{s \in \mathcal{S}} w_s^\pi = 1, \tag{4.5}$$

which is exactly the optimization formulation of the policy gradient method.

Primal problem and Bellman equation. Next, we show that the primal problem (4.2) is equivalent to the average reward Bellman equation for v, ρ

$$v_s = \max_a \left(r_s^a - \rho + \sum_{t \in \mathcal{S}} P_{st}^a v_t \right), \quad s \in \mathcal{S}. \tag{4.6}$$

Bellman equation to primal problem. The derivation from (4.6) to (4.2) can be found, for example, in [23]. Here we provide a short proof for completeness. Let v^*, ρ^* be the solution to the Bellman equation (4.6), then for all s, there exists a_s^* s.t.,

$$v^* = r^{a^*} - \rho^* \mathbf{1} + P^{a^*} v^*,$$

where $r_s^{a^*} \equiv r_s^{a_s^*}, P_{st}^{a^*} \equiv P_{st}^{a_s^*}$. For any v, ρ that satisfy the constraints in the primal problem (4.2), the following inequality holds

$$v \geq r^{a^*} - \rho \mathbf{1} + P^{a^*} v.$$

Subtracting these two equations gives

$$v - v^* \geq P^{a^*}(v - v^*) - (\rho - \rho^*)\mathbf{1}$$

Let w^* be the stationary distribution induced by the policy $\pi_s^a = \begin{cases} 1, a = a_s^* \\ 0, a \neq a_s^* \end{cases}$, then $(w^*)^\top = (w^*)^\top P^{a^*}$. Multiplying $(w^*)^\top$ to the last equation yields,

$$(w^*)^\top (\rho - \rho^*)\mathbf{1} \geq 0$$

Since we assume the MDP is unichain, the stationary distribution w^* for any policy is strictly positive. This implies that

$$\rho \geq \rho^*$$

holds for all v, ρ satisfying the constraints in (4.2). This proves that v^*, ρ^* is the minimizer of the primal problem (4.2).

Primal problem to Bellman equation. Let (v^*, ρ^*) be the minimizer of the primal problem (4.2). We now show that (v^*, ρ^*) also satisfies the average reward Bellman Equation (4.6). The KKT conditions of (4.2) are

$$\begin{cases} r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* \leq \rho^*, & \text{for } \forall s, a; \\ 1 - \sum_{s,a} \mu_s^a = 0; \\ \sum_{a \in \mathcal{A}} (\mu_s^a - \sum_{t \in \mathcal{S}} P_{ts}^a \mu_t^a) = 0, & \text{for } \forall s; \\ \mu_s^a (r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* - \rho^*) = 0 & \text{for } \forall s, a. \end{cases}$$

First we claim that it is impossible that there exists s s.t. $\forall a, r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* - \rho^* < 0$. Let $\mu_s^a = w_s \pi_s^a$ with $w_s = \sum_{a \in \mathcal{A}} \mu_s^a$ and $\pi_s \in \Delta$ for all s. Plugging it into the second and third equation gives,

$$1 - \sum_{s \in \mathcal{S}} w_s = 0, \quad w - (P^\pi)^\top w = 0.$$

Therefore w is the unique stationary distribution induced by the policy π. Since we assume the MDP is unichain, the stationary distribution is strictly positive. If there exists s, s.t. for $\forall a, r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* - \rho^* < 0$, then by the last equation of the KKT condition, $\mu_s^a = w_s \pi_s^a = 0$ implies $w_s = 0$ for this s, which contradicts with the unichain assumption. Therefore, the claim is true, i.e., there does not exist any s s.t. $\forall a$, $r_s^a + g \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* > 0$.

Therefore, for $\forall s$, there always exists a_s^*, s.t.,

$$r_s^{a_s^*} + \sum_{t \in \mathcal{S}} P_{st}^{a_s^*} v_t^* - v_s^* = \rho^*.$$

By the first equation, for all $a \neq a^*$,

$$r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* \leq \rho^*.$$

Combining the above two equations gives $r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^* \leq r_s + \sum_{t \in \mathcal{S}} P_{st}^{a_s^*} v_t^* - v_s^* = \rho^*$. This is equivalent to $\rho^* = \max_a r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t^* - v_s^*$, the average reward Bellman Equation (4.6).

The above two equivalences are stated in the following theorem.

THEOREM 4.2. *For the undiscounted standard MDP, the primal formulation* (4.2) *is equivalent to solving the Bellman equation*

$$v_s = \max_a \left(r_s^a - \rho + \sum_{t \in \mathcal{S}} P_{st}^a v_t \right), \quad \forall s \in \mathcal{S},$$

and the dual formulation (4.4) *is equivalent to the optimization formulation of the policy gradient method*

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} (r^\pi)^\top w^\pi, \ s.t. \ (I - (P^\pi)^\top) w^\pi = 0, \ \sum_{s \in \mathcal{S}} w_s^\pi = 1.$$

5. Undiscounted regularized MDP

We again use the the negative conditional entropy

$$h(\mu_s) = \sum_{a \in \mathcal{A}} \mu_s^a \log \frac{\mu_s^a}{\sum_b \mu_s^b}$$

as the regularizer.

5.1. LP problems.

Primal-dual problem. By adding the extra entropic term to the primal-dual problem (4.3) of the undiscounted unregularized case, the primal-dual problem of the undiscounted regularized MDP is

$$\min_{v_s, \rho} \sup_{\mu_s^a > 0} \rho + \sum_{s,a} (r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s - \rho) \mu_s^a - \sum_{s \in \mathcal{S}} h(\mu_s). \tag{5.1}$$

Primal problem. By introducing $\mu_s^a = w_s \pi_s^a$, one obtains

$$\min_{v_s, \rho} \sup_{\pi \in \Delta^{|\mathcal{S}|}, pi_s^a, w_s > 0} \rho + \sum_{s,a} (r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s - \rho) \pi_s^a w_s - \sum_{s \in \mathcal{S}} w_s h(\pi_s),$$

which is equivalent to

$$\min_{v_s, \rho} \left(\rho + \sup_{w_s > 0} \sum_{s \in \mathcal{S}} w_s \cdot \max_{\pi_s \in \Delta} \left(\sum_{a \in \mathcal{A}} (r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s - \rho) \pi_s^a - h(\pi_s) \right) \right).$$

The primal problem of the above minimax problem is

$$\min_{v_s, \rho} \rho \ \ s.t. \ \forall s, \ \max_{\pi_s \in \Delta} \left(\sum_{a \in \mathcal{A}} (r_s^a + \sum_{t \in \mathcal{S}} P_{st}^a v_t - v_s - \rho) \pi_s^a - h(\pi_s) \right) \leq 0,$$

or equivalently,

$$\min_{v,\rho} \rho \quad \text{s.t.} \quad \max_{\pi}(r^\pi - \rho\mathbf{1} + P^\pi v - h^\pi) \le v. \tag{5.2}$$

Note that the constraint

$$\max_{\pi_s \in \Delta}\left(\sum_{a \in \mathcal{A}}(r_s^a + \gamma\sum_{t \in \mathcal{S}}P_{st}^a v_t - v_s - \rho)\pi_s^a - h(\pi_s)\right)$$

is in the form of the Gibbs variational principle. The optimizer for this maximization is

$$\pi_s^a = \frac{\exp(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - v_s - \rho)}{Z_s},$$

where Z_s is the normalization factor, the optimal value is $\log Z_s$. Therefore, the constraint is equivalent to $\log Z_s \le 0$, i.e.,

$$\sum_{a \in \mathcal{A}}\exp(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - v_s - \rho) \le 1$$

i.e.

$$e^{-v_s} \cdot \sum_{a \in \mathcal{A}}\exp(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - \rho) \le 1, \quad \text{or} \quad \sum_{a \in \mathcal{A}}\exp(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - \rho) \le e^{v_s}.$$

Taking log gives

$$v_s \ge \log\left(\sum_{a \in \mathcal{A}}\exp(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - \rho)\right).$$

Hence the primal problem can also be written as

$$\min_{v_s,\rho} \rho \quad \text{s.t.} \quad v_s \ge \log\left(\sum_{a \in \mathcal{A}}\exp(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - \rho)\right), \tag{5.3}$$

which is an alternative formulation of the primal problem. Notice that a similar problem with equality constraints was instead derived in [18]. In practice, the inequality constraints as in (5.3) are often preferred since the feasibility set is then convex.

Dual problem. The supremum and minimum in the primal-dual problem (3.1) can be exchanged because the objective function is convex in v, ρ and concave in μ. Then one has,

$$\sup_{\mu_s^a > 0} \min_{v_s,\rho} \rho + \sum_{s,a}(r_s^a + \sum_{t \in \mathcal{S}}P_{st}^a v_t - v_s - \rho)\mu_s^a - \sum_{s \in \mathcal{S}}h(\mu_s).$$

Taking derivatives in ρ and v leads to

$$1 - \sum_{s,a}\mu_s = 0, \quad \sum_{a \in \mathcal{A}}(I - (P^a)^\top)\mu^a = 0.$$

Hence the dual problem in terms of μ_s^a is

$$\sup_{\mu^a > 0}\sum_{a \in \mathcal{A}}(r^a)^\top\mu^a - \sum_{s \in \mathcal{S}}h(\mu_s) \quad \text{s.t.} \quad \sum_{a \in \mathcal{A}}(I - (P^a)^\top)\mu^a = 0, \quad 1 - \sum_{s,a}\mu_s^a = 0. \tag{5.4}$$

The above three formulations are summarized as the following theorem.

THEOREM 5.1.   *For the undiscounted regularized MDP, the following three formulations are equivalent,*

Primal: $\quad \min\limits_{\substack{v\in\mathbb{R}^{|\mathcal{S}|} \\ \rho\in\mathbb{R}}} \rho, \quad s.t. \ v \geq \max\limits_{\pi}(r^{\pi} - \rho\mathbf{1} + P^{\pi}v - h^{\pi}),$

or equivalently, $\quad \min\limits_{\substack{v\in\mathbb{R}^{|\mathcal{S}|} \\ \rho\in\mathbb{R}}} \rho, \quad s.t. \ v \geq \log\left(\sum\limits_{a\in\mathcal{A}} \exp(r^a + \gamma P^a v - \rho\mathbf{1})\right),$

Primal $-$ dual: $\quad \min\limits_{\substack{v\in\mathbb{R}^{|\mathcal{S}|} \\ \rho\in\mathbb{R}}} \sup\limits_{\substack{\mu\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \\ \mu>0}} \rho + \sum\limits_{a}(\mu^a)^{\top}(r^a + P^a v - v - \rho\mathbf{1}) - \sum\limits_{s\in\mathcal{S}} h(\mu_s),$

Dual: $\quad \sup\limits_{\substack{\mu\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \\ \mu>0}} \sum\limits_{a\in\mathcal{A}}(r^a)^{\top}\mu^a - \sum\limits_{s\in\mathcal{S}} h(\mu_s), \ s.t. \ \sum\limits_{a\in\mathcal{A}}(I - \gamma(P^a)^{\top})\mu^a = 0, \sum\limits_{s,a}\mu_s^a = 1.$

5.2. Equivalences.

Dual problem and policy gradient. The dual problem (5.4) is equivalent to the optimization formulation of the policy gradient method. Let us parameterize $\mu_s^a = w_s\pi_s^a$ with $w_s = \sum_{a\in\mathcal{A}}\mu_s^a$ and $\pi\in\Delta^{|\mathcal{S}|}$. Then the constraint in (5.4) becomes

$$1 - \sum\limits_{s\in\mathcal{S}} w_s = 0, \quad (I - (P^{\pi})^{\top})w = 0,$$

which indicates that w is the stationary distribution of P^{π}. After denoting this solution by w^{π} and plugging it into the objective function in (5.4), we transform the dual problem to

$$\max\limits_{\pi\in\Delta^{|\mathcal{S}|}} (r^{\pi} - h^{\pi})^{\top} w^{\pi}, \ s.t. \ \sum\limits_{s\in\mathcal{S}} w_s^{\pi} = 1, \quad (I - (P^{\pi})^{\top})w^{\pi} = 0. \tag{5.5}$$

We can also view $r^{\pi} - h^{\pi}$ as a regularized reward by subtracting the entropy function h^{π}. The average-reward under the new reward becomes

$$\rho^{\pi} = \lim\limits_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum\limits_{m=1}^{T}(r_{s_m}^{a_m} - h(\pi_{s_m}))\right].$$

In this way, (5.5) is exactly the optimization formulation of the policy gradient method for this undiscounted regularized MDP.

Primal problem and Bellman equation. The regularized average-reward Bellman equation for v, ρ is

$$v = \max\limits_{\pi\in\Delta^{|\mathcal{S}|}} r^{\pi} - \rho\mathbf{1} + P^{\pi}v - h^{\pi}. \tag{5.6}$$

In each component,

$$v_s = \max\limits_{\pi_s\in\Delta}\sum\limits_{a\in\mathcal{A}}(r_s^a - \rho + \sum\limits_{t\in\mathcal{S}} P_{st}^a v_t))\pi_s^a - h(\pi_s).$$

By the Gibbs variational principle, the RHS is equal to $\log(\exp(\sum_{a\in\mathcal{A}}(r_s^a-\rho+$ $\sum_{t\in\mathcal{S}}P_{st}^a v_t)))$. Therefore, the regularized average-reward Bellman equation can be written as the log-sum-exp form for v,ρ

$$v_s=\log\left(\sum_{a\in\mathcal{A}}\exp\left(r_s^a-\rho+\sum_{t\in\mathcal{S}}P_{st}^a v_t\right)\right). \tag{5.7}$$

Next, we show that the primal problem (5.2) is equivalent to solving (5.6).

Bellman equation to primal problem. Let v^*,ρ^* be the solution to the Bellman equation. For v^*,ρ^*, there exists π^* s.t.,

$$v^*=r^{\pi^*}-\rho^*\mathbf{1}+P^{\pi^*}v^*-h^{\pi^*}.$$

Then, for any v satisfying the above constraints, the following inequality holds,

$$v\geq r^{\pi^*}-\rho\mathbf{1}+P^{\pi^*}v-h^{\pi^*}.$$

Subtracting these two equations gives

$$v-v^*\geq P^{\pi^*}(v-v^*)-(\rho-\rho^*)\mathbf{1}$$

Again let w^* be the stationary distribution induced by the policy π^*, then $(w^*)^\top=(w^*)^\top P^{\pi^*}$. Then multiplying $(w^*)^\top$ to the last equation yields

$$(w^*)^\top(\rho-\rho^*)\mathbf{1}\geq 0.$$

Since we assume the MDP is unichain, the stationary distribution for any policy is strictly positive. This implies that

$$\rho\geq\rho^*$$

holds for all v satisfying the constraints in (5.2). This proves that v^* is the minimizer of the primal problem (3.2).

Primal problem to Bellman equation. Let $v^*,\rho^*=\operatorname{argmin}_{v,\rho}\rho$ be the minimizer of the primal problem (5.2). Besides, there exists a policy π^* s.t. $\rho^*=(w^{\pi^*})^\top(r^{\pi^*}-h^{\pi^*})$. This can be seen from multiplying $(w^\pi)^\top$ to the constraint $v^*\geq r^\pi-\rho^*\mathbf{1}+P^\pi v^*-h^\pi$. Due to $(w^\pi)^\top\mathbf{1}=1,(w^\pi)^\top P^\pi=(w^\pi)^\top$, one has

$$\rho^*\geq(w^\pi)^\top(r^\pi-h^\pi)$$

for all π. Let $\pi^*=\operatorname{argmin}_\pi(w^\pi)^\top(r^\pi-h^\pi)$, then the minimizer $\rho^*=(w^{\pi^*})^\top(r^{\pi^*}-h^{\pi^*})$.

We now show that (v^*,ρ^*) is also the solution to the Bellman Equation (5.6) by contradiction. Assume (v^*,ρ^*) does not satisfy (5.6), then there must exist $\bar{s}$, s.t., for $\forall\pi$

$$v_{\bar{s}}^*\geq(r^\pi-\rho^*+P^\pi v^*-h^\pi)_{\bar{s}}+\delta \tag{5.8}$$

with some positive constant $\delta>0$. The inequality (5.8) also holds for π^*, so one can write it in vector form,

$$v^*\geq r^{\pi^*}-\rho^*\mathbf{1}+P^{\pi^*}v^*-h^{\pi^*}+\tilde{\delta},$$

where $\tilde{\delta}$ is a vector with δ on its $\bar{s}$-th element and 0 on all other elements. Then multiplying $(w^{\pi^*})^\top$ to the above equation yields,

$$0 \geq \delta w_{\bar{s}}^{\pi^*}.$$

Note that the RHS is always > 0 because the stationary distribution w^{π^*} is strictly positive, which leads to a contradiction. Hence we conclude that v^*, ρ^* is also the solution to the Bellman Equation (5.6).

The two equivalences are stated in the following theorem.

THEOREM 5.2. *For the undiscounted regularized MDP, the primal formulations* (5.2) *and* (5.3) *are equivalent to solving the following equations,*

$$v = \max_{\pi \in \Delta^{|\mathcal{S}|}} r^\pi - \rho\mathbf{1} + P^\pi v - h^\pi, \quad \text{or eq.} \quad v = \log\left[\sum_{a \in \mathcal{A}} \exp\left(r^a - \rho\mathbf{1} + \gamma\sum_{t \in \mathcal{S}} P^a v\right)\right],$$

and the dual formulation (5.4) *is equivalent to the optimization formulation of the policy gradient method*

$$\max_{\pi \in \Delta^{|\mathcal{S}|}} (r^\pi - h^\pi)^\top w^\pi, \ s.t. \ \sum_{s \in \mathcal{S}} w_s^\pi = 1, \quad (I - (P^\pi)^\top) w^\pi = 0.$$

6. Conclusion

We summarize the primal, primal-dual and dual formulations for discounted standard MDP, discounted regularized MDP, undiscounted standard MDP and undiscounted regularized MDP. For all four MDPs, we present the equivalence between the primal formulation and the Bellman equation and the equivalence between the dual formulation and the policy gradient formulation. We hope these connections between different optimization formulations can be useful to the theoretical understanding of MDP algorithms.

REFERENCES

[1] Y. Abbasi-Yadkori, P.L. Bartlett, and A. Malek, *Linear programming for large-scale Markov decision problems*, JMLR Workshop and Conference Proceedings, MIT Press, 32:496–504, 2014. 1

[2] L. Baird, *Residual algorithms: Reinforcement learning with function approximation*, Machine Learning Proceedings, Elsevier, 30–37, 1995. 1

[3] R. Bellman, *Dynamic programming*, Science, 153(3731):34–37, 1966. 1

[4] D.P. Bertsekas, *Abstract Dynamic Programming*, Athena Scientific, 2018. 1

[5] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-dynamic Programming*, Athena Scientific, 1996. 1

[6] S. Bhatnagar, D. Precup, D. Silver, R.S. Sutton, H.R. Maei, and C. Szepesvári, *Convergent temporal-difference learning with arbitrary smooth function approximation*, Adv. Neur. Inf. Process. Syst., 1204–1212, 2009. 1

[7] Y. Chen, L. Li, and M. Wang, *Scalable bilinear π learning using state and action features*, Proc. Int. Conf. Mach. Learn., 80:834–843, 2018. 1

[8] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song, *SBEED: Convergent reinforcement learning with nonlinear function approximation*, Proc. Int. Conf. Mach. Learn., 80:1125–1134, 2018. 1, 1.1

[9] R. Fox, A. Pakman, and N. Tishby, *Taming the noise in reinforcement learning via soft updates*, Proc. Conf. Uncertain. Artif. Intell., 202–211, 2016. 1.1, 1.1

[10] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, *Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor*, Proc. Int. Conf. Mach. Learn., 80:1851–1860, 2018. 1.1

[11] R.A. Howard, *Dynamic Programming and Markov Processes*, MIT Press, 1961. 1

[12] J.G. Kemeny, J.L. Snell, and A.W. Knapp, *Denumerable Markov Chains: With a Chapter of Markov Random Fields by David Griffeath*, Springer Science & Business Media, 40, 2012. 4

[13] D. Lee and N. He, *Stochastic primal-dual q-learning algorithm for discounted MDPs*, American Control Conference (ACC), IEEE, 4897–4902, 2019. 1

[14] B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik, *Finite-sample analysis of proximal gradient TD algorithms*, Proc. Conf. Uncertain. Artif. Intell. 504–513, 2015. 1

[15] S. Mahadevan, B. Liu, P. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, and J. Liu, *Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces*, arXiv preprint, arXiv:1405.6757, 2014. 1

[16] V. Mnih, A.P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, *Asynchronous methods for deep reinforcement learning*, Proc. Int. Conf. Mach. Learn., 1928–1937, 2016. 1, 1.1, 1.1

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, *Playing atari with deep reinforcement learning*, arXiv preprint, arXiv:1312.5602, 2013. 1

[18] G. Neu, A. Jonsson, and V. Gómez, *A unified view of entropy-regularized Markov decision processes*, arXiv preprint, arXiv:1705.07798, 2017. 1, 1.1, 4.1, 5.1

[19] J.V. Neumann, *Zur theorie der gesellschaftsspiele*, Math. Ann., 100(1):295–320, 1928. 2.1

[20] R. Ortner, *A short note on stationary distributions of unichain Markov decision processes*, arXiv preprint, arXiv:math/0604452, 2006. 4

[21] J. Peters, K. Mülling, and Y. Altun, *Relative entropy policy search*, Proc. AAAI Artif. Intell., 10:1607–1612, 2010. 1.1, 1.1

[22] M.L. Puterman, *Markov decision processes*, in D.P. Heyman and M.J. Sobel (eds.) Handbooks in Operations Research and Management Science, 2:331–434, 1990. 1

[23] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, 2005. 1, 2, 2.2, 4.2

[24] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, *Trust region policy optimization*, Proc. Int. Conf. Mach. Learn., 37:1889–1897, 2015. 1.1, 1.1

[25] R.S. Sutton, *Learning to predict by the methods of temporal differences*, Mach. Learn., 3(1):9–44, 1988. 1

[26] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2018. 2

[27] R.S. Sutton, H.R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, *Fast gradient-descent methods for temporal-difference learning with linear function approximation*, Proc. Int. Conf. Mach. Learn., 993–1000, 2009. 1

[28] R.S. Sutton, C. Szepesvári, and H.R. Maei, *A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation*, Adv. Neur. Inf. Process. Syst., 21(21):1609–1616, 2008. 1

[29] Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu, *Doubly robust bias reduction in infinite horizon off-policy estimation*, arXiv preprint, arXiv:1910.07186, 2019. 1

[30] M. Wang, *Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time*, Math. Oper. Res., 45(2):517–546, 2020. 1, 2.1

[31] M. Wang, E.X. Fang, and H. Liu, *Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions*, Math. Progr., 161(1-2):419–449, 2017. 1

[32] M. Wang, J. Liu, and E.X. Fang, *Accelerating stochastic composition optimization*, J. Mach. Learn. Res., 18(1):3721–3743, 2017. 1

[33] C. Watkins, *Learning from delayed rewards*, PhD thesis, King's College, Cambridge, 1989. 1

[34] Y. Ye, *The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate*, Math. Oper. Res., 36(4):593–603, 2011. 1, 2.1

[35] Y. Zhu, Z. Izzo, and L. Ying, *Borrowing from the future: Addressing double sampling in model-free control*, arXiv preprint, arXiv:2006.06173, 2017. 1

[36] Y. Zhu and L. Ying, *Borrowing from the future: An attempt to address double sampling*, Proc. Mach. Learn. Res., 107:246–268, 2020. 1