An Online Primal-Dual Method for Discounted Markov Decision Processes

Mengdi Wang and Yichen Chen

Abstract—We consider the online solution of discounted Markov decision processes (MDP). We focus on the black-box learning model where transition probabilities and state transition cost are unknown. Instead, a simulator is available to generate random state transitions under given actions. We propose a stochastic primal-dual algorithm for solving the linear formulation of the Bellman equation. The algorithm updates the primal and dual iterates by using sample state transitions and sample costs generated by the simulator. We provide a thresholding procedure that recovers the exact optimal policy from the dual iterates with high probability.

I. INTRODUCTION

Markov decision processes (MDP) are one of the most important mathematical models of sequential decision making. Given a controllable Markov chain and the distribution of state-to-state transition cost, the aim is to find the optimal action to perform at each state in order to minimize the expected overall cost. MDP and their variants are widely applied in engineering systems, artificial intelligence, business and finance.

The MDP has been well-studied as a special case of dynamic programming; see [21] for a textbook review. It can also be viewed as a discrete-time analog of stochastic optimal control [4], [5]. Classical theory and algorithms mainly focus on moderate-size problems with explicit state-to-state dynamics and cost functions. Solving MDP exactly requires full knowledge of the underlying stochastic process and cost distributions, which are often not available in modern big-data applications.

With the presence of modeling uncertainties, approximate and/or online solutions of MDP have been important subjects of study. In practice, either the transition probabilities of the Markov chain or the distribution of the state-to-state transition cost could be unknown. In the context of learning, approximate solutions and estimation methods of MDP are often regarded as reinforcement learning (RL). RL is a rich field that has attracted enormous attention from mathematicians and computer scientists [2], [22].

In this paper, we focus on the black-box learning model where both transition probabilities and state transition costs are unknown. Instead, a simulation oracle is available to generate random state transitions once an action is specified. Such models are often referred as generative models in statistics and learning. The algorithmic goal is to find the optimal policy of the unknown MDP by interacting with the simulator. Our approach is motivated by the linear formulation of the Bellman equation. We consider the minimax problem of the Lagrangian, which is essentially a convex-concave saddle point problem. We propose a stochastic primal-dual algorithm for the saddle point problem. The algorithm updates the primal and dual solutions simultaneously using noisy samples of partial derivatives of the Lagrangian. Each update makes one query to the oracle and observes a single state transition. Our algorithm can be viewed as a stochastic analog of the primal-dual iteration for linear programming.

Our main results are summarized as follows:

1) A stochastic primal-dual algorithm is proposed for finding the optimal policy of an MDP. The algorithm works under the black-box model and does not require prior knowledge of the transition probabilities and cost distributions.

2) A thresholding procedure is proposed to truncate small entries of the dual iterate to zeros. We show that the thresholding procedure recovers the optimal deterministic policy exactly with high probability.

To the authors’ best knowledge, this is the first primal-dual online algorithm for estimating the optimal policy of MDP in the black-box learning environment. Our results indicates that, after observing finitely many state transitions, one can recover the optimal policy without solving the dual variable exactly. It suggests that the linear formulation of an MDP bears convenient structures yet to be fully exploited, especially in the context of estimation and learning.

Section 2 reviews the basics of MDP. Section 3 studies the linear duality of MDP and characterizes properties of the primal and dual solutions. Section 4 presents the online primal-dual algorithm, and Section 5 gives the finite-sample performance bounds. Section 6 illustrates a numerical example on a simulated taxi driver problem.
II. BACKGROUND

A. Markov Decision Processes

Consider a Markov chain on states $S = \{1, \ldots, S\}$ with transition probability matrix $P_a \in \mathbb{R}^{S \times S}$ parameterized by actions $a \in \mathcal{A}$. Here $\mathcal{A} = \{1, \ldots, A\}$ is the action space. Upon a state transition from $i$ to $j$ using action $a$, one incurs a random state transitional cost $g_{ij}a \in \mathbb{R}$.

Let $\mu : S \to \mathcal{A}$ be a policy that maps a state $i \in S$ to a deterministic action $\mu(i) \in \mathcal{A}$. Consider the Markov chain under policy $\mu$. We denote its transition probability matrix by $P_\mu$, where $P_{\mu}(i,j) := P_{\mu(i)}(i,j)$ for $i, j \in S$.

The Markov decision problem (MDP) is to find an optimal policy $\mu^* : S \to \mathcal{A}$ such that the infinite-horizon discounted cost is minimized, regardless of the initial state:

$$
\mu^* = \arg\min_{\mu : S \to \mathcal{A}} \mathbb{E} \left[ \sum_{k=1}^{\infty} \alpha^k g_{i_0,i_{k+1}}\mu(i_k) \right],
$$

where $\alpha \in (0, 1)$ is a discount factor, $(i_0, i_1, \ldots)$ are state transitions generated by the Markov chain under policy $\mu$.

Define the optimal cost vector $x^* \in \mathbb{R}^S$ to be

$$
x^*(i) = \min_{\mu : S \to \mathcal{A}} \mathbb{E} \left[ \sum_{k=1}^{\infty} \alpha^k g_{i_0,i_{k+1}}\mu(i_k) \mid i_0 = i \right],
$$

$$
i \in S.
$$

The value $x^*(i)$ is equal to the optimal expected total cost when the initial state is $i$. The optimal cost vector $x^*$ is often regarded as the optimal value function or optimal cost-to-go.

B. Bellman’s Equation

According to the theory of dynamic programming [4], a vector $x^*$ is the optimal cost vector if and only if it satisfies the following $S \times S$ system of equations, known as the Bellman equation, given by

$$
x^*(i) = \min_{a \in \mathcal{A}} \left\{ \alpha \sum_{j \in S} P_a(i,j)x^*(j) + \sum_{j \in S} P_a(i,j)\mathbb{E}[g_{ija} \mid i, j, a] \right\},
$$

$$
i \in S,
$$

where $P_a$ is the matrix of transition probabilities using a fixed action $a$. We denote the expected state transition cost under action $a$ to be

$$
g_a(i) = \sum_{j \in S} P_a(i,j)\mathbb{E}[g_{ija} \mid i, j, a],
$$

$i \in S$.

We denote the expected state transition cost under policy $\mu$ to be

$$
g_\mu(i) = \sum_{j \in S} P_\mu(i,j)\mathbb{E}[g_{ija} \mid i, j, a],
$$

$i \in S$.

The Bellman equation (1) is equivalent to

$$
x^*(i) = \min_{a \in \mathcal{A}} \{\alpha P_{a,i}x^* + g_a(i)\},
$$

where $P_{a,i}$ is the $i$th row of $P_a$. The Bellman equation can be written in the vector form

$$
x^* = \min_{\mu : S \to \mathcal{A}} \{\alpha P_\mu x^* + g_\mu\}.
$$

When $\alpha \in (0, 1)$, the Bellman equation has a unique fixed point solution $x^*$, and it equals to the optimal value function of the MDP. Moreover, a policy $\mu^*$ is an optimal policy if and only if it attains the minimization in the Bellman equation. The core mathematical problem is to solve the Bellman equation and find the optimal policy $\mu^*$. Note that this is a nonlinear system of fixed point equations.

C. Related Works

The linear program formulation of Bellman’s equation was known at around the same time when Bellman’s equation was invented. Approximate linear programming has been studied for approximating large-scale MDP on a low-dimensional subspace, started with [11] and followed by [1], [23]. A seminal paper [25] shows that the policy iteration for the MDP is a form of simplex method, which is strongly polynomial for the equivalent linear program. In a recent paper [9], the exact primal-dual method was considered for MDP with full knowledge. The exact primal-dual iteration is interpreted as a form of value iteration. Online solution approaches have not been considered in this works.

In the RL literatures, there have been some works about dual temporal difference learning and primal-dual methods, see for examples [13], [15], [16], [24]. However, most of these works focus on evaluating the value function for a fixed policy. This is different from our work, which aims to find the optimal policy.

Our algorithm and analysis is closely related to the class of stochastic approximation (SA) methods. For textbook references on stochastic approximation, please see Kushner and Yin [12], by Benveniste et al. [3], by Borkar [7], by Bertsekas and Tsitsiklis [6]). We will also use the averaging idea by Polyak and Juditsky [20]. Stochastic approximation is regarded as an important class of sampling-based methods for simulation optimization, see [18], [19] for an overview. In particular, our stochastic primal-dual algorithm can be viewed as SA applied to the stochastic saddle point problems; see recent works [8], [10], [17] for convergence rate benchmarks.
III. Duality Analysis of Discounted MDP

A. Linear Formulations of Bellman’s Equation

The Bellman equation for the optimal value function $x^*$ can be equivalently cast into the following $S \times (SA)$ linear programming (LP) problem:

\[
\begin{align*}
\text{minimize} & \quad -e^T x \\
\text{subject to} & \quad (I - \alpha P_a)x - g_a \leq 0, \quad a \in \mathcal{A},
\end{align*}
\]

where $e = (1, \ldots, 1)^T$. Note that the equivalence still holds if $e$ is replaced by an arbitrary vector with positive entries [11].

Its dual problem is

\[
\begin{align*}
\text{maximize} & \quad -\sum_{a \in \mathcal{A}} \lambda_a^T g_a \\
\text{subject to} & \quad \sum_{a \in \mathcal{A}} (I - \alpha P_a^T) \lambda_a = e, \quad \lambda_a \geq 0,
\end{align*}
\]

where the dual variable $\lambda = (\lambda_a)_{a \in \mathcal{A}} \in \mathbb{R}^{S \times A}$, and each $\lambda_a \in \mathbb{R}^S$ is the multiplier corresponding to constraint inequality subset $\alpha P_a x + g_a \geq x$. We denote by $\lambda_{i,a} > 0$ the multiplier associated with the $i$th row of $\alpha P_a x + g_a \geq x$. We denote by $\lambda^*$ the optimal solution to the dual problem or the optimal multiplier.

B. Characterizing the Primal and Dual Solutions

We make the following assumptions regarding the MDP. Under these assumptions, the linear programs associated with the MDP are feasible and non-degenerate.

**Assumption 1:** The discounted MDP has the following properties:

1) The discount factor satisfies $\alpha \in (0,1)$.

2) There exists a unique optimal policy $\mu^*$.

3) There exists a constant $\sigma > 0$ such that $\max_{i,a} E \left[ |g_{i,a}|^2 \right] = \sigma^2$.

Under Assumption 1, there exists a unique optimal policy $\mu^*$ to the MDP. In other words, there exists one optimal action for each state.

Let $P_{\mu^*}$ denote the transition probability matrix under the optimal policy $\mu^*$, i.e.,

$$P_{\mu^*}(i,j) = P_{\mu^*}(i)(i,j), \quad \forall i,j \in S.$$  

Then the optimal value function $x^*$ satisfies

$$x^* = (I - \alpha P_{\mu^*})^{-1} g_{\mu^*}.$$  

By using the complementarity condition of Lagrangian duality, we can show that the optimal dual variable $\lambda$ has exactly $S$ non-zero elements, corresponding to $S$ active row constraints of the primal problem (2). In particular, we have the following lemma.

**Lemma 1:** Let Assumption 1 hold. The optimal dual solution $\lambda^* = (\lambda^*_{i,a})_{i \in S, a \in \mathcal{A}}$ satisfies

$$\left( \lambda^*_{i, \mu^*(i)} \right)_{i \in S} = (I - \alpha P_{\mu^*}^T)^{-1} e,$$

and $\lambda^*_{i,a} = 0$ if $a \neq \mu^*(i)$.

Lemma 1 suggests a critical correspondence between the optimal multiplier $\lambda^*$ and the optimal policy $\mu^*$, which was also pointed out in [21]. In particular, the basis of $\lambda^*$ can be used to yield the optimal policy $\mu^*$ as follows:

$$\mu^*(i) = a, \quad \text{if } \lambda^*_{i,a} > 0.$$  

In other words, finding the optimal policy is equivalent to finding the basis of the optimal dual solution.

Next we provide characterization of the primal and dual solutions. These results will be used to regularize the iterates generated by our stochastic algorithms. They will also be used to establish the finite-sample error bound.

**Lemma 2:** Let Assumption 1 hold. Suppose that $(x^*, \lambda^*)$ is a pair primal and dual solutions to the linear programs (2),(3). Then

$$\|x^*\|_2 \leq \sqrt{S \sigma \over 1 - \alpha}, \quad \|\lambda^*\|_2 \leq \sqrt{S \over 1 - \alpha}.$$  

**Lemma 3:** Let Assumption 1 hold. Suppose that $(x^*, \lambda^*)$ is a pair of primal and dual solutions to the linear programs (2), (3). Then

$$\left( \sum_{a \in \mathcal{A}} \lambda^*_{i,a} \right) \geq 1, \quad \forall i \in S.$$  

The proof of Lemmas 1, 2, 3 rely on duality analysis of linear programs, as well as properties of the Bellman equation. They are deferred to the online supplement.

C. Saddle Point Problem

According to Lagrangian duality, the primal and dual LP programs (2)-(3) are equivalent to the following saddle point problem:

$$\min_{x \in \mathbb{R}^S} \max_{\lambda \geq 0} L(x, \lambda),$$  

\[
\min_{x \in \mathbb{R}^S} \max_{\lambda \geq 0} L(x, \lambda),
\]

where the Lagrangian $L(x, \lambda)$ is a bilinear function

$$L(x, \lambda) = -e^T x + \sum_{a \in \mathcal{A}} \lambda^T_a \left( (I - \alpha P_a)x - g_a \right).$$  

The primal variable $x$ is of dimension $S$, and the dual variable $\lambda = (\lambda_a, a \in \mathcal{A})$ is of dimension $S \cdot A$.

For technical reasons, we will consider a restricted saddle point problem given as

$$\min_{x \in X} \max_{\lambda \in \mathcal{A}} L(x, \lambda),$$  

\[
\min_{x \in X} \max_{\lambda \in \mathcal{A}} L(x, \lambda),
\]

where

$$X = \left\{ \|x\|_2 \leq \sqrt{S \sigma \over 1 - \alpha} \right\}.$$  

\[
X = \left\{ \|x\|_2 \leq \sqrt{S \sigma \over 1 - \alpha} \right\}.
\]
and
\[
\Lambda = \left\{ \lambda \geq 0, \|\lambda\|_2 \leq \frac{\sqrt{S}}{1 - \alpha}, \left( \sum_{a,t \in \mathcal{A}} \lambda_{a,t}^* \right) \geq 1, i \in \mathcal{S} \right\}.
\]

According to Lemmas 2 and 3, we know that the optimal primal and dual solutions belong to the constraints \( X \) and \( \Lambda \), respectively. As a result, the optimal solution of the restricted saddle point problem (5) is equal to the restricted saddle point problem (5) is equal to the

\[
\Lambda = \left\{ \lambda \geq 0, \|\lambda\|_2 \leq \frac{\sqrt{S}}{1 - \alpha}, \left( \sum_{a,t \in \mathcal{A}} \lambda_{a,t}^* \right) \geq 1, i \in \mathcal{S} \right\}.
\]

IV. AN ONLINE PRIMAL-DUAL ALGORITHM

We consider the model-free online learning setting. We do not know the transition probabilities and transition costs. Instead, we have a sampling oracle that generates sample transitions of the system and sample costs. We make the following assumption regarding the sampling oracle.

Assumption 2: The Simulation Oracle \( \mathcal{M} \) takes input \((i, a)\) and generates a state transition to \( j \) such that the next state \( j \) is chosen with probabilities \( P_a(i, j) \).

**Algorithm 1** Stochastic primal-dual algorithm

**Input:** Simulation Oracle \( \mathcal{M} \), \( S, A \), \( \alpha \in (0, 1) \).

**Initialize** \( X \). (1) \( \Lambda \) according to (6)(7).

**Initialize** \( x^{(0)} \) and \( \lambda = (\lambda^{(0)}_u : u \in \mathcal{A}) \) arbitrarily.

for \( k = 1, 2, \ldots, T \) do

Sample \( i_k \) uniformly from \( \mathcal{S} \).

Sample \( a_k \) uniformly from \( \mathcal{A} \).

Sample \( j_k \) and \( g_{i_k j_k a_k} \) conditioned on \((i_k, a_k)\) from \( \mathcal{M} \).

Update the iterates by

\[
x^{(k-\frac{1}{2})} = x^{(k-1)} - \gamma_k \left[ - e + A \lambda^{(k-1)}_{a_k} - \alpha S A \left( A x^{(k-1)} - S A g_{i_k j_k a_k} e_{i_k} \right) e_{j_k} \right],
\]

\[
\lambda^{(k-\frac{1}{2})}_{a_k} = \lambda^{(k-1)}_{a_k} + \gamma_k \frac{S A}{A x^{(k-1)} - S A g_{i_k j_k a_k} e_{i_k}} e_{j_k},
\]

\[
\lambda^{(k-\frac{1}{2})}_u = \lambda^{(k-1)}_u, \quad \forall \ u \neq a_k,
\]

Project the iterates orthogonally to the constraints

\[
x^{(k)} = \Pi_X x^{(k-\frac{1}{2})}, \quad \lambda^{(k)} = \Pi_\Lambda \lambda^{(k-\frac{1}{2})}.
\]

end for

**Output:** Averaged dual iterate \( \hat{\lambda} = \frac{1}{T} \sum_{k=1}^{T} \lambda^{(k)} \) and the truncated pure policy \( \hat{\mu}^{Tr}(i) = \arg \max_{a \in \mathcal{A}} \lambda^{(k)}_{a,i} \) for all \( i \in \mathcal{S} \).

The simulation oracle \( \mathcal{M} \) is able to generate random samples of state-transition tuples \((i, j, a)\). Upon drawing a state-transition tuple \((i, j, a)\), we are able to compute noisy gradients of the Lagrangian function, i.e., noisy samples of partial derivatives \( \partial_x L(x, \lambda) \) and \( \partial_\lambda L(x, \lambda) \). Motivated by the saddle point formulation of Bellman’s equation, we develop an iterative primal-dual method in order to estimate the optimal value function (i.e., the primal solution) and the optimal policy (i.e., the dual solution) directly. Our proposed stochastic primal-dual algorithm is given by Algorithm 1.

Algorithm 1 updates the value function estimates and randomized policies by processing one sample state-transition at a time. The algorithm is able to update the primal and dual variables online as new state-transition tuples are observed. The stepsize \( \gamma_k \) is set to be \( 1/\sqrt{k} \). The update formula for \( x^{(k-\frac{1}{2})} \) and \( \lambda^{(k-\frac{1}{2})} \) involve several vector additions. The update formula for \( x^{(k)} \) and \( \lambda^{(k)} \) involve Euclidean projections onto simple geometric sets, thus are easy to implement.

V. FINITE-SAMPLE PERFORMANCE BOUNDS

In this section, we show that the optimal policy can be recovered from the dual iterate with high probability. Our first result gives an error bound on the duality gap of the iterates generated by Algorithm 1.

**Theorem 1 (Bounds of Duality Gap):** Let \( \hat{\lambda} \) be the averaged dual iterates generated by Algorithm 1 using \( T \) queries from the oracle \( \mathcal{M} \). Then the dual iterate satisfies

\[
\mathbb{E} \left[ \sum_{a \in \mathcal{A}} (\hat{\lambda}_a)^T (g_a + \alpha P_a x^* - x^*) \right] \leq \mathcal{O} \left( \frac{S^2 A (1 + \sigma^2)}{(1 - \alpha)^2 T} \right).
\]

Proof Outline. We consider the quantity

\[
\mathcal{E}^k = \| x^k - x^* \|^2 + SA \sum_{a \in \mathcal{A}} \| \lambda^k_a - \lambda^*_a \|^2.
\]

We analyze the error \( \mathcal{E}^k \) by taking expectation and applying a primal-dual analysis. Then we obtain

\[
\mathbb{E} \left[ \sum_{a \in \mathcal{A}} (\lambda^k_a)^T (g_a + \alpha P_a x^* - x^*) \right] \leq \mathbb{E} [\mathcal{E}^k] - \mathbb{E} [\mathcal{E}^{k+1}] + \gamma_k \Psi,
\]

where \( \Psi > 0 \) is a constant that depends on problem dimensions \( S, A \), the discount factor \( \alpha \) and the second moment of reward \( \sigma^2 \). When we sum \( k \) from 1 to \( T \), the left handside becomes exactly the duality gap. The first term on the right handside is bounded by (6), (7) and Lemma 2. Then by substituting \( \gamma_k = 1/\sqrt{k} \) and a suitable bound of \( \Psi \), we get the desired bound of the duality gap. □
bound for entries of $\hat{\lambda}$ corresponding to inactive primal row constraints. Also note that the duality gap can be large when the state space and action space are very large and the number of samples is small. The duality gap bound reflects the level of difficulty for estimating the optimal policy of large MDP using insufficient data.

Next we consider how to recover the optimal policy $\mu^*$ from the dual iterates $\hat{\lambda}$. Note that the policy space has a discrete nature, which makes it possible to distinguish the optimal one from others when the estimated policy is close enough. We define the minimal action discrimination constant as follows.

Definition 1: the minimal efficiency loss of deviating from the optimal policy $\mu^*$ by making a single wrong action. It is given by

$$\bar{d} = \min_{(i,a):\mu^*(i)\neq a} (\alpha P_{a,i}x^* + g_a(i) - x^*(i)).$$

Under Assumption 1, there exists a unique optimal policy $\mu^*$, therefore $\bar{d} > 0$. A large value of $\bar{d}$ means that it is easy to discriminate optimal actions from suboptimal actions. A small value of $\bar{d}$ means that some suboptimal actions perform similarly to optimal actions. In our second theorem, we will see that the minimal action discrimination constant plays a key role in success recovery of the optimal policy.

Theorem 2 (Recovering Optimal Policy By Truncation): Let $\hat{\mu}^T$ be the truncated pure policy such that

$$\hat{\mu}^T(i) = \arg\max_{a \in A} \hat{\lambda}_{a,i}, \quad i \in S.$$

Then

$$\mathbb{P} (\hat{\mu}^T = \mu^*) \geq 1 - O\left(\frac{S^2A^2(1 + \sigma^2)}{\bar{d}(1 - \alpha)^2\sqrt{T}}\right).$$

Proof. We denote for short that $d_{i,a} = (\alpha P_{a,i}x^* + g_a(i) - x^*(i))$. By using the complementarity condition, we have $d(i,a) > 0$ if $a \neq \mu^*(i)$ and $d(i,a) = 0$ if $a = \mu^*(i)$.

Note that $\sum_{a \in A} \hat{\lambda}_{a,i} \geq 1$ for all $i \in S$. We have

$$\mathbb{P}(\hat{\mu}^T \neq \mu^*)$$

$$= \mathbb{P} (\exists (a,i) \text{ s.t. } \mu^*(i) \neq a, \hat{\lambda}_{a,i} > \hat{\lambda}_{a',i}, \forall a' \in A)$$

$$\leq \mathbb{P} (\exists (a,i) \text{ s.t. } \mu^*(i) \neq a, \hat{\lambda}_{a,i} > \frac{1}{A} \sum_{a \in A} \hat{\lambda}_{a,i})$$

$$\leq \mathbb{P} (\exists (a,i) \text{ s.t. } \mu^*(i) \neq a, \hat{\lambda}_{a,i} > \frac{1}{A})$$

$$\leq \mathbb{P} \left( \sum_{(a,i):\mu^*(i)\neq a} \hat{\lambda}_{a,i}d_{i,a} \geq \frac{1}{A} \sum_{(a,i):\mu^*(i)\neq a} d_{i,a} \right)$$

$$= \mathbb{P} \left( \sum_{(a,i)} \hat{\lambda}_{a,i}d_{i,a} \geq \frac{\bar{d}}{A} \right),$$

where the first inequality uses the positivity of $d_{i,a}$ where $\mu^*(i) \neq a$. By the Markov inequality, we obtain

$$\mathbb{P}(\hat{\mu}^T \neq \mu^*) \leq \frac{\mathbb{E}[\sum_{(a,i)} \hat{\lambda}_{a,i}d_{i,a}]}{\bar{d}/A}.$$
the $T$-sample duality gap is proportional to $1/\sqrt{T}$. Also, we can see that by increasing the number of states, the duality gap increases according. This validate the results of Theorem 1. In Figure 2, we plot the probability of exact recovery of the optimal policy, i.e., $P(\hat{\mu}^T = \mu^*)$, against the number of observed state transitions. This validate the results of Theorem 2.

VII. SUMMARY

We have proposed a new online algorithm for learning the optimal policy of MDP, which is motivated by the linear duality embedded in Bellman equations. The algorithm updates by interacting with a simulation oracle that generates state transitions of the Markov chain once given an action. We show that, after observing a finite number of state transition, the exact optimal policy can be obtained by thresholding the dual iterates with high probability. We conjecture that the error bounds can be further improved, which is left for future research.

REFERENCES