Observation-Based Optimization for POMDPs With Continuous State, Observation, and Action Spaces

Xiaofeng Jiang, *Member, IEEE*, Jian Yang, *Senior Member, IEEE*, Xiaobin Tan, *Member, IEEE*, and Hongsheng Xi

***Abstract*—This paper considers the optimization problem for partially observable Markov decision processes (POMDPs) with the continuous state, observation, and action spaces. POMDPs with the discrete spaces have emerged as a promising approach to the decision systems with imperfect state information. However, in recent applications of POMDPs, there are many problems that have continuous states, observations, and actions. For such problems, due to the infinite dimensionality of the belief space, the existing studies usually discretize the continuous spaces with the sufficient or nonsufficient statistics, which may cause the curse of dimensionality and performance degradation. In this paper, based on the sensitivity analysis of the performance criteria, we have developed a simulation-based policy iteration algorithm to find the local optimal observation-based policy for POMDPs with the continuous spaces. The proposed algorithm needs none of the specific assumptions and prior information, and has a low computational complexity. One numerical example of the complicated multiple-input multiple-output beamforming problem shows that the algorithm has a significant performance improvement.**

***Index Terms*—Continuous spaces, none of the prior information, partially observable Markov decision process (POMDP), sensitivity analysis, simulation-based optimization.**

I. Introduction

In many real applications, the data usually cannot be fully collected, and are corrupted by the observation noises. Partially observable Markov decision processes (POMDPs) allow imperfect information about the data, and thus play important roles in optimizing the performance for such a system [1]–[4]. Typically, POMDPs have finite discrete state spaces, and are studied by transforming them into Markov decision processes (MDPs) with observable continuous belief states, whose dimensions equal the number of states [5]–[7]. However, in recent applications of POMDPs, such as a wideband spectrum sensing in cognitive radio [8], [9], beamforming for multiple-input and multiple-output (MIMO) systems [10], motion planning for the intelligent robot [11], [12], and multitarget tracking in the cognitive radar [13], the state, observation, and action spaces of the POMDPs are usually continuous. The traditional methods are not able to straightforwardly handle such a problem, since the discretization of the continuous spaces usually causes the curse of dimensionality.

The existing studies [14]–[18] consider the MDPs with continuous spaces, where the state is completely observable. The state aggregation method [14], [15] is used to discretize the state spaces. In [16]–[18], the smoothing-based and neural network Q-learning methods are used to approximate the continuous value functions. For POMDPs with the continuous state, observation, and action spaces, the existing studies [11], [19] use nonsufficient statistics and parameterized densities to approximate the belief states and then solve the approximate belief MDPs. The work of [12] develops a general pattern search method to find the optimal action in a continuous convex action space. In [20], the Gaussian mixtures are adopted to approximate the continuous finite-horizon value function, but the number grows exponentially in the value iterations. In [21], [22], the state and observation spaces are discretized, and then the sub-optimal policy is found with the value iteration. The study of [23] proposes a quantity called partial-information state to record the useful information, and then develops a performance difference method to seek the optimal partial-information-state-based policy. However, a restrictive assumption (44) in [23] requires that the state conditional probability distribution under the partial-information state is independent of the policy. With this assumption, the proposed algorithm can only be applied to some special systems.

This paper develops a simulation-based policy iteration algorithm to find the local optimal observation-based policy for POMDPs with the continuous state, observation, and action spaces, which does not have any restrictive assumption and does not need any prior information. Therefore, the proposed algorithm can be applied to most decision problems with the continuous spaces. The numerical results show that the algorithm has a significant performance improvement. Compared with the existing studies, our work focuses on the following three new aspects.

1) We develop an observation-based sensitivity analysis framework that can straightforwardly handle the continuous spaces, without the discretization and approximation, which may cause the curse of dimensionality and performance degradation. Meanwhile, Theorems 1 and 2 show that the policy iteration can be explained by analyzing the performance gradients.
2) The performance gradients with respect to the policy can be estimated from a single sample path. The proposed optimization algorithm does not need any prior information about the system, such as the state transition, observation, and reward functions, which are very difficult to count in the applications with imperfect state information.
3) In contrast to the dynamic approach, we develop a simulation-based policy iteration algorithm based on the gradient estimates, where only the gradient of the policy function and the accumulation of the instantaneous rewards are needed at each iteration. The algorithm may have a lower computational complexity.

Manuscript received November 23, 2017; revised March 18, 2018; accepted July 16, 2018. Date of publication August 1, 2018; date of current version April 24, 2019. This work was supported by National Natural Science Foundations of P.R. China. The development of the optimization method under Grant 61503358, Grant and the experimental design 61233003, and Grant 61673360. Recommended by Associate Editor E. Zhou. *(Corresponding author: Xiaofeng Jiang.)*

The authors are with the Department of Automation, University of Science and Technology of China, Hefei 230000, China (e-mail: jxf@ustc.edu.cn; jianyang@ustc.edu.cn; xbtan@ustc.edu.cn; xihs@ustc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2018.2861910

II. Problem Formulation

Consider a discrete time POMDP with the continuous state, observation, and action spaces that has the following form:

$$\{S, A, O, p(ds'|s,a), q(do|s), r(s,a), \mu_\theta\} \tag{1}$$

where S, O, and A are the state, observation, and action spaces, respectively. In this continuous case, we consider $S = \mathbb{R}^N$, $A = \mathbb{R}^M$, and $O = \mathbb{R}^L$. For an arbitrary probability density function $f(x)$, let dx be an infinitesimal change in x, we use the corresponding probability distribution function such as $P\{dx\}$ to denote $f(x)dx$. The notations $s', s \in S, a \in A, o \in O$ without the time indices are used to denote the arbitrary elements in the corresponding spaces. $p(ds'|s,a)$ is the state transition probability that state s' is reached from state s on action a. At each time slot, though the state cannot be completely observed, an observation o can be obtained according to the observation probability $q(do|s)$. $r(s,a)$ is the instantaneous reward in one time step. μ_θ is a stationary parametrization policy, where θ is the policy parameter.

At each time slot, the action applied to the system is determined by the policy μ_θ, which can be deterministic or randomized. A deterministic policy is a mapping from O to A: $a = \mu_\theta(o)$. A randomized policy is a mapping from O to the set of probability measures on A: $\mu_\theta(da|o)$, which satisfies $\int_{a\in A} \mu_\theta(da|o) = 1$. We assume that the time slot is indexed by $t = 0, 1, \ldots$. Let $s(t), a(t)$, and $o(t)$ be the state, action, and observation at time t, the expected average reward can be defined as

$$V(\theta) = \lim_{T\to\infty} \frac{1}{T} E^\theta \left[\sum_{t=0}^{T-1} r(s(t), a(t))\right]$$

where E^θ is the expectation under the policy μ_θ. The optimization problem of the POMDP with the continuous spaces is of the following form:

$$\max_\theta : \quad V(\theta). \tag{2}$$

The aim is to find an observation-based policy μ_θ to maximize the expected average reward.

III. Sensitivity Analysis

For the optimization problem with continuous state space (2), the sensitivity analysis theory is used to analyze the derivative of the performance criteria. A steady-state expression of the expected average reward $V(\theta)$ can be given as follows:

$$V(\theta) = \int_{s\in S} \pi^\theta(ds) r^\theta(s) \tag{3}$$

where $\pi^\theta(ds)$ is the steady-state probability under policy μ_θ, $r^\theta(s)$ is the expected reward of state s. Equation (3) is an essential base of the sensitivity analysis, since $\pi^\theta(ds), r^\theta(s)$ have different expressions under the deterministic and randomized policies; we show the sensitivity analysis results in the following two sections.

A. Randomized Policy

Considering the randomized policy $\mu_\theta(da|o)$, the expected reward $r^\theta(s)$ and state transition probability $p^\theta(ds'|s)$ can be defined as follows:

$$r^\theta(s) = \int_{o\in O}\int_{a\in A} r(s,a)\mu_\theta(da|o)q(do|s) \tag{4}$$

$$p^\theta(ds'|s) = \int_{o\in O}\int_{a\in A} p(ds'|s,a)\mu_\theta(da|o)q(do|s). \tag{5}$$

We make the following ergodic assumption such as [23]. The ergodicity is an important basis of the infinity horizon performance optimization problems, many excellent works exist, such as the Krylov–Bogolyubov theorem for invariant distribution [24] and the Birkhof–Khinchin theorem for ergodicity [25]. In practice, many systems are stable or can be stabilized, so we can assume that the ergodic assumption holds.

Assumption 1: The internal Markov chain of the POMDP with the continuous state space is ergodic.

Under Assumption 1, the unique steady-state probability $\pi^\theta(ds)$ satisfies

$$\pi^\theta(ds') = \int_{s\in S} p^\theta(ds'|s)\pi^\theta(ds) \tag{6}$$

$$\int_{s\in S} \pi^\theta(ds) = 1 \tag{7}$$

which are the intractable simultaneous integral equations. To obtain a solution of $\pi^\theta(ds)$, we rewrite (6) and (7) as

$$\int_{s\in S} [p^\theta(ds'|s) - \delta(s-s')ds', 1]\pi^\theta(ds) = [0,1] \tag{8}$$

where δ denotes the delta function. (8) is a Fredholm integral function of the first kind that has been widely studied, and the methods of degenerate kernel and successive approximations can be used to find the solution of $\pi^\theta(ds)$. However, in Section IV, we will show that the proposed optimization algorithm does not need to solve this equation. Then, we define the steady-state probability of the observation

$$\pi_o^\theta(do) = \int_{s\in S} \pi^\theta(ds) q(do|s) \tag{9}$$

and the performance potential of the observation

$$g_o^\theta(o) = \lim_{T\to\infty} E^\theta \left[\sum_{t=0}^{T-1} [r(s(t),a(t)) - V(\theta)] \middle| o(0) = o\right], \tag{10}$$

where the subscript "o" denotes that the quantity is with respect to the observation. The performance potential denotes the influence of observation o on the system, and is equivalent to the value function in dynamic programming [4]. Let $g_o^\theta(o,a), g^\theta(s)$, and $g^\theta(s,a)$ be the performance potentials of the corresponding quantities, which have the similar definitions with (10). Let $q_o^\theta(ds|o)$ be the conditional steady-state probability of state s when the observation is o, we have $q_o^\theta(ds|o) = \frac{\pi^\theta(ds)q(do|s)}{\pi_o^\theta(do)}$. Then, we can give the following theorem to show the sensitivity analysis result of the randomized policy.

Theorem 1: Under the randomized policy $\mu_\theta(da|o)$, the performance gradient is

$$\nabla_\theta V(\theta) = \int_{o\in O}\int_{a\in A} \pi_o^\theta(do)[\nabla_\theta \mu_\theta(da|o)] g_o^\theta(o,a).$$

Proof: See Appendix A. ■

B. Deterministic Policy

Under the deterministic policy $a = \mu_\theta(o)$, the definitions of $r^\theta(s)$ and $p^\theta(ds'|s)$ are different from (4) and (5), which can be given by

$$r^\theta(s) = \int_{o\in O} r(s, u_\theta(o))q(do|s) \tag{11}$$

$$p^\theta(ds'|s) = \int_{o\in O} p(ds'|s, u_\theta(o))q(do|s). \tag{12}$$

Then, we can give the following theorem to show the sensitivity analysis result of the deterministic policy.

Theorem 2: Under the deterministic policy $a = \mu_\theta(o)$, the performance gradient is

$$\nabla_\theta V(\theta) = \int_{s\in S}\int_{s'\in S}\int_{o\in O} \pi^\theta(ds)q(do|s) \cdot [\nabla_\theta p(ds'|s, u_\theta(o))]g^\theta(s') + \int_{s\in S}\int_{o\in O} \pi^\theta(ds)q(do|s)\nabla_\theta r(s, u_\theta(o))$$

Proof: By substituting (11) and (12) into (41), we can straightforwardly obtain the result. ■

IV. Sensitivity-Based Optimization

Based on Theorems 1 and 2, we develop a gradient algorithm to find the optimal policy $\mu_\theta(da|o)$ or $a = \mu_\theta(o)$. However, it is difficult to straightforwardly compute $g_o^\theta(o, a)$ in the gradient equations. First, we need to solve the Fredholm integral equation of the second kind (37) to obtain $g^\theta(s)$, and then use (43) and (44) to compute $g_o^\theta(o, a)$, which are intractable. Consequently, we have developed a simulation-based algorithm to estimate $g_o^\theta(o, a)$ for a POMDP with the continuous spaces. In the proposed algorithm, we do not need any intractable computation and do not even need to know any prior information about the model parameters $p(ds'|s, a)$, $q(do|s)$, and $r(s, a)$.

A. Estimation Under Randomized Policies

To estimate the gradient $\nabla_\theta V(\theta)$, we first generate a single observation and action sample path $\{(o(t), a(t)), t = 0, 1, \ldots\}$, which satisfies the following stationary condition:

$$\begin{aligned} &f(o(t) = o_t, o(t+1) = o_{t+1}, \ldots) \\ &= \int_{s_t, s_{t+1}, \ldots \in S} f(o(t) = o_t, o(t+1) \\ &= o_{t+1}, \ldots | s(t) = s_t, s(t+1) = s_{t+1}, \ldots) \\ &\quad \cdot f(s(t) = s_t, s(t+1) = s_{t+1}, \ldots) ds_t ds_{t+1} \ldots \\ &= \int_{s_t, s_{t+1}, \ldots \in S} f(o(t+k) = o_t, o(t+k+1) \\ &= o_{t+1}, \ldots | s(t+k) = s_t, s(t+k+1) = s_{t+1}, \ldots) \\ &\quad \cdot f(s(t+k) = s_t, s(t+k+1) = s_{t+1}, \ldots) ds_t ds_{t+1} \ldots \\ &= f(o(t+k) = o_t, o(t+k+1) = o_{t+1}, \ldots) \end{aligned}$$

for every $k, t \geq 0$ and $o_t, o_{t+1}, \ldots \in O$, where f is the probability density function. The sample path is thus stationary. Consider Theorem 1, we have

$$\begin{aligned} \nabla_\theta V(\theta) &= \int_{o\in O}\int_{a\in A} \pi_o^\theta(do)\mu_\theta(da|o) l^\theta(o, a) g_o^\theta(o, a) \\ &= E^\theta[l^\theta(o(t), a(t)) g_o^\theta(o(t), a(t))] \end{aligned} \quad (13)$$

where $l^\theta(o, a) = \frac{\nabla_\theta \mu_\theta(da|o)}{\mu_\theta(da|o)}$. This gradient can be viewed as the expectation of $l^\theta(o, a) g_o^\theta(o, a)$ with respect to the steady-state probability $\pi_o^\theta(do)\mu_\theta(da|o)$. Since $l^\theta(o(t), a(t)) g_o^\theta(o(t), a(t))$ only depends on the current observation and action, and the sample path is stationary, we have

$$\begin{aligned} \nabla_\theta V(\theta) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} l^\theta(o(t), a(t)) g_o^\theta(o(t), a(t)) \\ &= E^\theta \left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} l^\theta(o(t), a(t)) g_o^\theta(o(t), a(t)) \right]. \end{aligned} \quad (14)$$

We make the following assumption to give an approximation of the gradient.

Assumption 2: For every s, a, o, $||l^\theta(o, a)||$ and $|r(s, a)|$ are uniformly bounded by $B, R < \infty$, respectively.

Here $||\cdot||$ is the Euclidean norm. Under Assumption 2, $g^\theta(s), g^\theta(s, a)$, and $g_o^\theta(o, a)$ are bounded, and thus $\nabla_\theta V(\theta)$ is bounded. We can give an approximation

$$\nabla_\theta V_T(\theta) = E^\theta \left[\frac{1}{T} \sum_{t=0}^{T-1} l^\theta(o(t), a(t)) g_o^\theta(o(t), a(t)) \right] \quad (15)$$

which satisfies $\lim_{T\to\infty} \nabla_\theta V_T(\theta) = \nabla_\theta V(\theta)$, we can find a fixed T_1 such that

$$||\nabla_\theta V(\theta) - \nabla_\theta V_{T_1}(\theta)|| \leq \epsilon_1 \quad (16)$$

for a constant $\epsilon_1 > 0$.

According to the definition of $g_o^\theta(o, a)$ and its boundness, we can also give an approximation

$$g_{o,T}^\theta(o, a) = E^\theta \left[\sum_{t=0}^{T-1} [r(s(t), a(t)) - V(\theta)] \, | o(0) = o, a(0) = a \right] \quad (17)$$

which satisfies $\lim_{T\to\infty} g_{o,T}^\theta(o, a) = g_o^\theta(o, a)$, we can find a fixed T_2 such that

$$|g_o^\theta(o, a) - g_{o,T_2}^\theta(o, a)| \leq \epsilon_2 \quad (18)$$

for a constant $\epsilon_2 > 0$ and every o, a. The definition of $g_{o,T_2}^\theta(o, a)$ (17) suggests that it can be estimated as

$$\tilde{g}_{o,T_2}^\theta\left(o(t), a(t), \tilde{V}\right) = \sum_{k=0}^{T_2-1} \left[r(s(t+k), a(t+k)) - \tilde{V} \right] \quad (19)$$

where $o(t) = o, a(t) = a$ and $\tilde{V}$ is some estimate of $V(\theta)$. By combining above-mentioned approximations (15) and (17) and estimate (19), we can give an estimate of the gradient $\nabla_\theta V(\theta)$ as

$$F_{T_1,T_2}\left(\theta, \tilde{V}\right) = \frac{1}{T_1} \sum_{t=0}^{T_1-1} l^\theta(o(t), a(t)) \tilde{g}_{o,T_2}^\theta\left(o(t), a(t), \tilde{V}\right) \quad (20)$$

which has the following expectation.

Proposition 1: Under Assumptions 1 and 2, we have

$$\begin{aligned} &||\nabla_\theta V(\theta) - E\left[F_{T_1,T_2}(\theta, \tilde{V})\right]|| \\ &\leq \epsilon_1 + B\epsilon_2 + BT_2\left[V(\theta) - \tilde{V}\right]. \end{aligned}$$

Proof: See Appendix B. ■

Proposition 1 shows that $F_{T_1,T_2}(\theta, \tilde{V})$ is able to become an unbiased estimate of the gradient $\nabla_\theta V(\theta)$ with appropriate T_1, T_2, when $\tilde{V}$ approximates $V(\theta)$.

Remark 1: In the above-mentioned estimation of $\nabla_\theta V(\theta)$, the computation of the estimate $F_{T_1,T_2}(\theta, \tilde{V})$ (20) only needs the policy parameter θ and the value of the instantaneous reward $r(s(t), a(t))$ without the information about $s(t)$, it does not need to know any other prior information, such as $p(ds'|s, a), q(do|s)$, and $r(s, a)$.

However, if the value of the instantaneous reward cannot be given, we need $p(ds'|s,a), q(do|s)$, and $r(s,a)$ to compute an expected reward $r_o^\theta(o(t),a(t),o(t+k)$, and $a(t+k))$ to replace $r(s(t+k)$ and $a(t+k))$ in (19)

$$\begin{aligned} &r_o^\theta(o(t),a(t),o(t+k),a(t+k)) \\ &:= E[r(s(t+k),a(t+k))|o(t),a(t),o(t+k),a(t+k)] \\ &= \int_{s_{t+k}\in S} \frac{Pr\{ds_{t+k},da_{t+k},do_t,da_t,do_{t+k}\}}{Pr\{do_t,da_t,do_{t+k},da_{t+k}\}} r(s_{t+k},a_{t+k}) \end{aligned} \quad (21)$$

where $o(t)=o_t\in O, a(t)=a_t\in A, o(t+k)=o_{t+k}\in O, a(t+k)=a_{t+k}\in A$, and

$$\begin{aligned} &Pr\{do_t,da_t,do_{t+k},da_{t+k}\} \\ &\quad=\int_{s_{t+k}\in S} Pr\{ds_{t+k},da_{t+k},do_t,da_t,do_{t+k}\} \end{aligned} \quad (22)$$

$$\begin{aligned} &Pr\{ds_{t+k},da_{t+k},do_t,da_t,do_{t+k}\} \\ &\quad=\int_{s_t,s_{t+k-1}\in S}\int_{o_{t+k-1}\in O}\int_{a_{t+k-1}\in A} \\ &Pr\{ds_{t+k},do_{t+k},da_{t+k}|s_{t+k-1},o_{t+k-1},a_{t+k-1}\} \\ &\quad\cdot Pr\{ds_t,do_t,da_t,ds_{t+k-1},do_{t+k-1},da_{t+k-1}\} \\ &\quad=\int_{s_t,\ldots,s_{t+k-1}\in S}\int_{o_{t+1},\ldots,o_{t+k-1}\in O}\int_{a_{t+1},\ldots,a_{t+k-1}\in A} \\ &\quad\cdot \pi^\theta(ds_t)\mu_\theta(da_t|o_t)q(do_t|s_t)\prod_{m=1}^{k}[\mu_\theta(da_{t+m}|o_{t+m}) \\ &\quad\cdot q(do_{t+m}|s_{t+m})p(ds_{t+m}|s_{t+m-1},a_{t+m-1})]. \end{aligned} \quad (23)$$

B. Estimation under Deterministic Policies

Consider Theorem 2, under the deterministic policy $a=\mu_\theta(o)$, we have

$$\begin{aligned} \nabla_\theta V(\theta) =& \int_{s\in S}\int_{s'\in S}\int_{o\in O}\pi^\theta(ds)q(do|s)p(ds'|s,u_\theta(o)) \\ &\cdot l^\theta(s',s,o)g^\theta(s') \\ &+\int_{s\in S}\int_{o\in O}\pi^\theta(ds)q(do|s)\nabla_\theta r(s,u_\theta(o)) \\ =& E^\theta[l^\theta(s(t+1),s(t),o(t))g^\theta(s(t+1)) \\ &+\nabla_\theta r(s(t),u_\theta(o(t)))] \end{aligned} \quad (24)$$

where $l^\theta(s',s,o)=\frac{\nabla_\theta p(ds'|s,u_\theta(o))}{p(ds'|s,u_\theta(o))}$. Like Assumption 2, we also make the following assumption to give an approximation of the gradient.

Assumption 3: For every s',s,o, $||l^\theta(s',s,o)||, |r(s,u_\theta(o))|$, and $||\nabla_\theta r(s,u_\theta(o))||$ are uniformly bounded by $B,R,C<\infty$, respectively.

Under Assumption 3, $g^\theta(s)$ is bounded, and thus $\nabla_\theta V(\theta)$ is bounded. Like the analysis of the randomized policy, we can give the following estimate of $\nabla_\theta V(\theta)$:

$$\begin{aligned} F_{T_1,T_2}(\theta,\tilde{V}) =& \frac{1}{T_1}\sum_{t=0}^{T_1-1}[l^\theta(s(t+1),s(t),o(t)) \\ &\cdot \tilde{g}^\theta_{T_2}(s(t+1),\tilde{V})+[\nabla_\theta\mu_\theta(o(t))] \\ &\cdot[\nabla_a r_o^\theta(o(t),a(t))]_{a(t)=\mu_\theta(o(t))}] \end{aligned} \quad (25)$$

where

$$\tilde{g}^\theta_{T_2}(s(t+1),\tilde{V}) = \sum_{k=1}^{T_2}\left[r(s(t+k),a(t+k))-\tilde{V}\right] \quad (26)$$

$$\begin{aligned} r_o^\theta(o(t),a(t)) &= E^\theta[r(s(t),a(t))|o(t),a(t)] \\ &= \int_{s\in S} q_o^\theta(ds|o(t))r(s,a(t)). \end{aligned} \quad (27)$$

T_1 and T_2 are the time lengths that satisfy

$$||\nabla_\theta V(\theta)-\nabla_\theta V_{T_1}(\theta)||\le\epsilon_1,\ \ |g^\theta(s)-g^\theta_{T_2}(s)|\le\epsilon_2 \quad (28)$$

for constants $\epsilon_1,\epsilon_2>0$, where

$$\nabla_\theta V_{T_1}(\theta)=E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1} l^\theta(s(t+1),s(t),o(t))\cdot g^\theta(s(t+1))\right] \quad (29)$$

$$g^\theta_{T_2}(s)=E^\theta\left[\sum_{t=0}^{T_2-1}[r(s(t),a(t))-V(\theta)]|s(0)=s\right]. \quad (30)$$

$F_{T_1,T_2}(\theta,\tilde{V})$ has the following expectation.

Proposition 2: Under Assumptions 1 and 3, we have

$$\begin{aligned} &\left|\left|\nabla_\theta V(\theta)-E\left[F_{T_1,T_2}\left(\theta,\tilde{V}\right)\right]\right|\right| \\ &\quad\le\epsilon_1+B\epsilon_2+BT_2\left[V(\theta)-\tilde{V}\right]. \end{aligned}$$

Proof: See Appendix C. ■

Proposition 2 also shows that $F_{T_1,T_2}(\theta,\tilde{V})$ is able to become an unbiased estimate.

C. Optimization Algorithm

Based on the estimates of the gradients (20) and (25), we can develop a policy iteration algorithm to find the optimal observation-based policy. Let m denote the iteration time. The policy θ_m and estimate $\tilde{V}_m$ are updated with the same step size. The proposed algorithm is listed in Algorithm 1, which can be summarized as follows.

Algorithm 1: Policy Iteration Optimization Algorithm.

Step 1 Set $m=0, t=0$, choose appropriate values of T_1, T_2, initialize θ_m and $\tilde{V}_m$.

Step 2 Generate the sample path along t with θ_m until $t=T_1+T_2-1$; Calculate $F_{T_1,T_2}(\theta_m,\tilde{V}_m)$ and $Z(\theta_m,\tilde{V}_m)$ according to (20), (25) and (31); Then update θ_m and $\tilde{V}_m$:

$$\theta_{m+1}=\theta_m+\gamma_m F_{T_1,T_2}(\theta_m,\tilde{V}_m)$$

$$\tilde{V}_{m+1}=\tilde{V}_m+\kappa\gamma_m Z(\theta_m,\tilde{V}_m)$$

Set $m=m+1, t=0$.

Step 3 If $|\tilde{V}_m-\tilde{V}_{m-1}|\le\epsilon_3$ and $||\theta_m-\theta_{m-1}||\le\epsilon_4$, set $m^*=m$, return $\theta_{m^*},\tilde{V}_{m^*}$; else return Step 2.

The system is initialized in Step 1. The values of T_1 and T_2 can be chosen empirically based on the principles (16) and (18) and Proposition 1. For the initial $\tilde{V}$, an inappropriate value may decrease the convergence rate of the algorithm, and $\frac{1}{T_1+T_2}\sum_{t=0}^{T_1+T_2-1} r(s(t),a(t))$ is suggested as the initial value of $\tilde{V}$. Based on a single sample path, Step 2 computes the estimate of $\nabla_\theta V(\theta)$, and updates the policy θ_m

and estimate $\tilde{V}_m$. Here, Z is defined as

$$Z\left(\theta_m, \tilde{V}_m\right) = \frac{1}{T_1+T_2}\sum_{t=0}^{T_1+T_2-1}\left(r(s(t),a(t)) - \tilde{V}^m\right). \quad (31)$$

γ_m is a positive step size, and $\kappa > 0$ is a positive constant. Since $V(\theta) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} r(s(t),a(t))$, without loss of generality, we can assume that $T_1 + T_2$ are large enough, such that

$$\left|V(\theta) - E^{\theta}\left[\frac{1}{T_1+T_2}\sum_{t=0}^{T_1+T_2-1} r(s(t),a(t))\right]\right| \le \epsilon_5 \quad (32)$$

for a constant $\epsilon_5 > 0$. Then, we have

$$\left|E\left[Z\left(\theta_m, \tilde{V}_m\right)\right] - \left(V(\theta) - \tilde{V}_m\right)\right| \le \epsilon_5 \quad (33)$$

which moves $\tilde{V}$ closer to $V(\theta)$. Step 3 determines whether θ_m and $\tilde{V}_m$ have converged with ϵ_3, ϵ_4 being the thresholds.

D. Convergence Analysis

To guarantee the convergence of the proposed algorithm, the step size γ_m should satisfy the following condition proposed by the Robbins–Monro stochastic approximation algorithm [26].

$$\sum_{m=1}^{\infty}\gamma_m = \infty, \quad \sum_{m=1}^{\infty}\gamma_m^2 < \infty.$$

The policy θ_m and estimate $\tilde{V}_m$ are updated in Algorithm 1; we first rewrite the update equations as follows:

$$\begin{aligned}\theta_{m+1} &= \theta_m + \gamma_m \nabla_\theta V(\theta_m)\\ &\quad + \gamma_m\left[F_{T_1,T_2}\left(\theta_m, \tilde{V}_m\right) - E\left[F_{T_1,T_2}\left(\theta_m, \tilde{V}_m\right)\right]\right]\\ &\quad + \gamma_m\left[E\left[F_{T_1,T_2}\left(\theta_m, \tilde{V}_m\right)\right] - \nabla_\theta V(\theta_m)\right] \quad (34)\\ \tilde{V}_{m+1} &= \tilde{V}_m + \kappa\gamma_m\left(V(\theta_m) - \tilde{V}_m\right)\\ &\quad + \kappa\gamma_m\left[Z\left(\theta_m, \tilde{V}_m\right) - E\left[Z\left(\theta_m, \tilde{V}_m\right)\right]\right]\\ &\quad + \kappa\gamma_m\left[E\left[Z\left(\theta_m, \tilde{V}_m\right)\right] - \left(V(\theta_m) - \tilde{V}_m\right)\right]. \quad (35)\end{aligned}$$

Consider (33), according to [26, Th. 2] and [27, Th. 1], (35) concludes that $|V(\theta_m) - \tilde{V}_m|$ converges to $[0, \epsilon_5]$ with probability 1. Then, consider this result and Propositions 1 and 2, (34) concludes that $||\nabla_\theta V(\theta_m)||$ converges to $[0, \epsilon_1 + B\epsilon_2 + BT_2\epsilon_5]$ with probability 1 according to [26] and [27]. When $\epsilon_1, \epsilon_2, \epsilon_5$ are small enough, the policy θ_m converges to a local optimum with probability 1.

The proposed optimization algorithm converges with the rate of the traditional stochastic approximation algorithm. According to the analysis results of [28], the convergence rate is $E[\varepsilon] = O(t^{-1/2})$, where ε is the error of the estimated gradients and t is the number of iterations. Comparing to the existing work, the proposed algorithm can estimate the gradients via numerous addition operations instead of computing the gradients straightforwardly, and the complexity of one iteration becomes lower. With the theoretical convergence rate of the stochastic approximation algorithm, the proposed algorithm thus has a lower complexity.

V. Numerical Results

To illustrate the proposed algorithm for the POMDP system with the continuous state, observation, and action spaces, we consider the application in [10], which is the constrained beamforming control problem for downlink multiuser MIMO systems with the imperfect channel state information. In the system, a multiantenna base station (BS) communicates to K single antenna mobile users. Let $h_k \in \mathbb{R}$ denote the channel state from the BS to User k, we assume that h_k follows the normal distribution $\mathcal{N}(0, \sigma_k^2)$, all the channel bandwidths are 180 kHz, and the communication process follows a time slotted structure, which is indexed by $t = 0, 1, \ldots$, where the length of each time slot is 0.5 ms. Let $Q_k(t)$ denote the queue length of the data of User k waiting for transmission at time t, the data arrivals follow the Possion distributions with different rates, and let $\lambda_k(t), \lambda_{\text{total}}$ denote the arrival rate of User k and the total arrival rate, respectively, we have $\sum_{k=1}^{K}\lambda_k(t) = \lambda_{\text{total}}$. Let $A_k(t)$ be the arrival data, and it follows the Poisson distribution with rate $\lambda_k(t)$. Since if the waiting data are too many, new arrival data will not be accepted, we make the following configuration. Let $d_k(t)$ be the data that are transmitted to User k, if the total queue length after the data transmission $\sum_{k=1}^{K}[Q_k(t) - d_k(t)] > 100$, we set $\lambda_{\text{total}} = 20$, else we set $\lambda_{\text{total}} = 200$.

Then, we consider a constraint on the power for communication, and the total power consumption at each time slot should be less than P_{total}. To satisfy this constraint, we introduce a new user $K+1$ with $A_{K+1}(t) = Q_{K+1}(t) = h_{K+1}(t) = 0$ for every t. Let $a_k(t)$ be the action of User k, then we allocate $P_k(t) = \frac{|a_k(t)|P_{\text{total}}}{\sum_{k'=1}^{K+1}|a_{k'}(t)|}$ power to User k at time t. The power $P_{K+1}(t)$ is not consumed, and we have the total power consumption $\sum_{k=1}^{K}P_k(t) \le P_{\text{total}}$. The transmission capacity of User k can be given as follows:

$$c_k(t) = 0.5\cdot 180\cdot\log_2\left(1 + \frac{|h_k(t)P_k(t)|^2}{1+\sum_{k'\neq k,K+1}|h_k(t)P_{k'}(t)|^2}\right).$$

Then, the data transmitted are

$$d_k(t) = [c_k(t) + Q_k(t) - |c_k(t) - Q_k(t)|]/2.$$

The queue length at the next time is

$$Q_k(t+1) = Q_k(t) - d_k(t) + A_k(t).$$

The reward is given as

$$r(t) = \sum_{k=1}^{K}[d_k(t) - \tau P_k(t)]$$

where $\tau = 0.3$ is the price of the power.

Let $[Q_1(t), h_1(t), \ldots, Q_K(t), h_K(t), 0, 0]$ be the state of the system, due to the noise, we can only observe $[Q_1(t), \tilde{h}_1(t), \ldots, Q_K(t), \tilde{h}_K(t), 0, 0]$, where $\tilde{h}_k(t) = h_k(t) + n_k$ and n_k is the additive white Gaussian noise. The following randomized policy with parameter tensor θ is considered:

$$a_k(t) \sim \mathcal{N}\left(Q(t)\theta(:,1,k) + \tilde{h}(t)\theta(:,2,k), \sigma^2\right)$$

where $k = 1, \ldots, K+1$, the initial policy is $\theta = [1]_{(K+1)\times 2\times(K+1)}$, $Q(t) = [Q_1(t), \ldots, Q_K(t), 0]$, $\tilde{h}(t) = [\tilde{h}_1(t), \ldots, \tilde{h}_K(t), 0]$, and $\sigma^2 = 1$. If the variance σ^2 approximates 0, the policy becomes deterministic.

We set $T_1 = 1000$, $T_2 = 350$, $P_{\text{total}} = 100$, and $K \in [1, 10]$. The step size is $\gamma_m = c_\gamma(\frac{1}{m})^{0.55}$, where $c_\gamma > 0$ is used to adjust the convergence rate, and is set as 10^{-4}. Then, the proposed algorithm is compared with two existing discretization-based methods, which are the

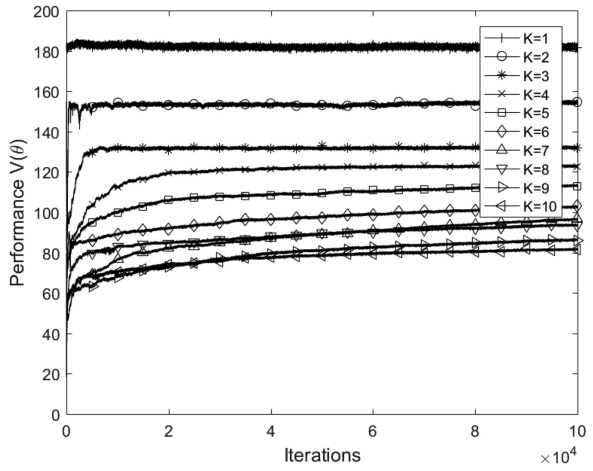

Fig. 1. Performance $V(\theta)$ over iterations for the proposed algorithm with $T_1 = 1000$ and $T_2 = 350$, and different numbers of users K.

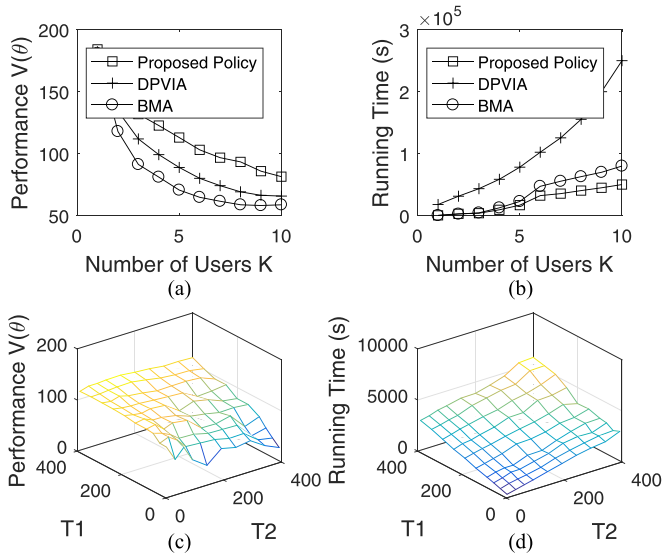

Fig. 2. (a) and (b) Performance $V(\theta)$ and running time over numbers of users K for the proposed algorithm with $T_1 = 1000, T_2 = 350$, DPVIA and BMA. (c) and (d) Performance $V(\theta)$ and running time over T_1, T_2 for the proposed algorithm with $K = 4$.

belief-based myopic algorithm (BMA) [8] and the density projection-based value iteration algorithm (DPVIA) [19]. The BMA is a greedy solution, which first discretizes the continuous spaces, then seeks the optimal action based on the current belief state at each time slot and update the belief with the new observation. The DPVIA uses the particle filtering method to approximate the actual belief state with a density in the exponential family, which is characterized by a few parameters. Then, the value iteration algorithm is performed based on the approximation to find the optimal policy. However, both the BMA and DPVIA need to know $p(ds'|s,a), q(do|s), r(s,a)$ in the model (1), whereas the proposed algorithm does not need these prior information. Due to the complicated expression of the above-mentioned beamforming control problem, it is very difficult to give an analytical expression of p, q, r; we use the discretization-based count from a long enough system log to give the prior information used in the BMA and DPVIA, where both $Q_k(t)$ and $h_k(t)$ are discretized into 100 values, and there are 10^{4K} elements in the discrete state space.

Fig. 1 shows the performance over iterations for the proposed algorithm. Though the dimensionality of the state space increases dramatically as the number of users K increases, the proposed algorithm always converges and is able to find the optimal observation-based policy. Fig. 2(a) and (b) shows the comparisons of the performance and running time for the proposed algorithm, DPVIA and BMA. We can find that the proposed algorithm performs better than the DPVIA and BMA, whereas the running time of the DPVIA and BMA is much larger than that of the proposed algorithm, since the discretization of the continuous spaces causes the curse of dimensionality and performance degradation. Meanwhile, the running time of the proposed algorithm increases linearly as K increases. Fig. 2(c) and (d) shows the performance and running time over T_1, T_2 for the proposed algorithm. We can find that the running time increases as T_1 and T_2 increase, but the optimal performance can be obtained when $T_1 \geq 300$ and $10 \leq T_2 \leq \frac{T_1}{2}$. This result is able to suggest the appropriate values of T_1 and T_2 to seek the optimal policy with the small running time.

VI. Conclusion and Future Work

In this work, POMDPs with the continuous state, observation, and action spaces are studied on the basis of the sensitivity analysis theory without the discretization of the continuous spaces. We have shown that the performance gradients with respect to the randomized policies can be estimated based on a single sample path with none of the prior information, such as the state transition, observation, and reward functions. Based on the gradient estimates, we have developed a simulation-based algorithm to optimize the observation-based policy with the low computational complexity. The proposed algorithm can converge with probability 1. The numerical results show that the algorithm has a significant performance improvement.

However, in our and many existing works, the optimization algorithms for POMDPs rarely concern the global optimization problem, since the reward function is usually not convex or concave, and the computational complexity for seeking a local optimum is usually very high. Since the proposed optimization algorithm has a low computational complexity, in the future work, we will focuses on the global optimization problems for POMDPs with the discrete and continuous spaces on the basis of the sensitivity analysis theory.

Appendix A
Proof of Theorem 1

Computing the derivative of $V(\theta)$ with respect to the policy parameter θ in (3), we have

$$\nabla_\theta V(\theta) = \int_{s\in S} \left[\left(\nabla_\theta \pi^\theta(ds)\right) r^\theta(s) + \pi^\theta(ds)\nabla_\theta r^\theta(s)\right]. \quad (36)$$

Under Assumption 1, we have the following Poisson equation [4], [23]:

$$r^\theta(s) = g^\theta(s) - \int_{s'\in S} p^\theta(ds'|s) g^\theta(s') + V(\theta). \quad (37)$$

Substituting (37) into $(\nabla_\theta \pi^\theta(ds)) r^\theta(s)$ of (36) yields

$$\begin{aligned}\nabla_\theta V(\theta) = &\int_{s\in S} \left(\nabla_\theta \pi^\theta(ds)\right) g^\theta(s) \\ &- \int_{s\in S} \left(\nabla_\theta \pi^\theta(ds)\right) \int_{s'\in S} p^\theta(ds'|s) g^\theta(s') \\ &+ \int_{s\in S} \left(\nabla_\theta \pi^\theta(ds)\right) V(\theta) + \int_{s\in S} \pi^\theta(ds)\nabla_\theta r^\theta(s).\end{aligned} \quad (38)$$

Substituting (6) into the first term of (38) yields

$$\begin{aligned}\int_{s\in S}(\nabla_\theta \pi^\theta(ds))g^\theta(s) &= \int_{s'\in S}\nabla_\theta\left(\int_{s\in S}p^\theta(ds'|s)\pi^\theta(ds)\right)g^\theta(s')\\ &= \int_{s\in S}\int_{s'\in S}(\nabla_\theta\pi^\theta(ds))p^\theta(ds'|s)g^\theta(s')\\ &\quad+\int_{s\in S}\int_{s'\in S}\pi^\theta(ds)(\nabla_\theta p^\theta(ds'|s))g^\theta(s').\end{aligned} \tag{39}$$

Substituting (7) into the third term of (38) yields

$$\int_{s\in S}\left(\nabla_\theta\pi^\theta(ds)\right)V(\theta) = (\nabla_\theta 1)V(\theta) = 0. \tag{40}$$

Substituting (39) and (40) into (38) yields

$$\nabla_\theta V(\theta) = \int_{s\in S}\pi^\theta(ds)\left[\int_{s'\in S}\left(\nabla_\theta p^\theta(ds'|s)\right)g^\theta(s') + \nabla_\theta r^\theta(s)\right]. \tag{41}$$

Substituting (4) and (5) into (41) yields

$$\begin{aligned}\nabla_\theta V(\theta) = &\int_{s\in S}\int_{o\in O}\int_{a\in A}\pi^\theta(ds)q(do|s)[\nabla_\theta\mu_\theta(da|o)]\\ &\cdot\left[\int_{s'\in S}p(ds'|s,a)g^\theta(s')+r(s,a)\right].\end{aligned} \tag{42}$$

Consider the definition of $g^\theta(s,a)$, we have

$$\begin{aligned}g^\theta(s,a) &= r(s,a)-V(\theta)\\ &\quad+\lim_{T\to\infty}E^\theta\left\{E^\theta\sum_{t=1}^{T-1}[r(s(t),a(t))-V(\theta)]\,|s(0)=s,a(0)=a\right\}\\ &= r(s,a)-V(\theta)+\int_{s'\in S}p(ds'|s,a)g^\theta(s').\end{aligned} \tag{43}$$

Consider the definition of $g_o^\theta(o,a)$, we have

$$\begin{aligned}g_o^\theta(o,a) = \lim_{T\to\infty}E^\theta\Bigg\{&E^\theta\left[\sum_{t=0}^{T-1}[r(s(t),a(t))-V(\theta)]\right.\\ &\left.|o(0)=o,a(0)=a,s(0)]|o(0)=o,a(0)=a\right\}\\ &=\int_{s\in S}Pr\{ds|o,a,\theta\}g^\theta(s,a)\\ &=\int_{s\in S}\frac{\pi^\theta(ds)q(do|s)\mu_\theta(da|o)}{\int_{s'\in S}\pi^\theta(ds')q(do|s')\mu_\theta(da|o)}g^\theta(s,a)\\ &=\int_{s\in S}q_o^\theta(ds|o)g^\theta(s,a).\end{aligned} \tag{44}$$

Substituting (43) and (44) into (42) yields

$$\begin{aligned}\nabla_\theta V(\theta) &= \int_{s\in S}\int_{o\in O}\int_{a\in A}\pi^\theta(ds)q(do|s)[\nabla_\theta\mu_\theta(da|o)]\cdot g^\theta(s,a)\\ &= \int_{o\in O}\int_{a\in A}\pi_o^\theta(do)[\nabla_\theta\mu_\theta(da|o)]g_o^\theta(o,a).\end{aligned}$$

Then, the theorem can be proved.

Appendix B
Proof of Proposition 1

The expectation of $F_{T_1,T_2}(\theta,\tilde{V})$ is

$$\begin{aligned}&E[F_{T_1,T_2}(\theta,\tilde{V})]\\ &= E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}l^\theta(o(t),a(t))\cdot\sum_{k=0}^{T_2-1}\left[r(s(t+k),a(t+k))-\tilde{V}\right]\right]\\ &= E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}l^\theta(o(t),a(t))\cdot\sum_{k=0}^{T_2-1}[r(s(t+k),a(t+k))-V(\theta)]\right]\\ &\quad+E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}l^\theta(o(t),a(t))T_2[V(\theta)-\tilde{V}]\right].\end{aligned} \tag{45}$$

Let F^1 denote the first term of (45), we have

$$\begin{aligned}F^1 &= E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}E^\theta[l^\theta(o(t),a(t))\right.\\ &\qquad\left.\cdot\sum_{k=0}^{T_2-1}[r(s(t+k),a(t+k))-V(\theta)]|o(t),a(t)]\right]\\ &= E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}l^\theta(o(t),a(t))g_{o,T_2}^\theta(o(t),a(t))\right].\end{aligned} \tag{46}$$

According to (16) and (18), we have

$$\begin{aligned}||\nabla_\theta V(\theta)-F^1|| &\le ||\nabla_\theta V(\theta)-\nabla_\theta V_{T_1}(\theta)||+||\nabla_\theta V_{T_1}(\theta)-F^1||\\ &\le \epsilon_1+E^\theta\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}||l^\theta(o(t),a(t))||\right.\\ &\qquad\left.\cdot\left|g_o^\theta(o(t),a(t))-g_{o,T_2}^\theta(o(t),a(t))\right|\right]\\ &\le \epsilon_1+B\epsilon_2.\end{aligned} \tag{47}$$

Let F^2 denote the second term of (45), we have

$$||F^2|| \le BT_2[V(\theta)-\tilde{V}]. \tag{48}$$

Consider (47) and (48), we have

$$\begin{aligned}&||\nabla_\theta V(\theta)-E[F_{T_1,T_2}(\theta,\tilde{V})]||\\ &\le ||\nabla_\theta V(\theta)-F^1||+||F^2||\\ &\le \epsilon_1+B\epsilon_2+BT_2[V(\theta)-\tilde{V}].\end{aligned} \tag{49}$$

APPENDIX C
PROOF OF PROPOSITION 2

The expectation of $F_{T_1,T_2}(\theta,\tilde{V})$ is

$$
\begin{aligned}
&E[F_{T_1,T_2}(\theta,\tilde{V})]\\
&= E^{\theta}\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1} l^{\theta}(s(t+1),s(t),o(t))\right.\\
&\qquad\left.\cdot\sum_{k=1}^{T_2}[r(s(t+k),a(t+k))-V(\theta)]\right]\\
&\quad + E^{\theta}\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1}[\nabla_{\theta}\mu_{\theta}(o(t))]\right.\\
&\qquad\left.\cdot E^{\theta}[\nabla_a r(s(t),a(t))|o(t),a(t)]]_{a(t)=\mu_{\theta}(o(t))}\right]\\
&\quad + E^{\theta}\left[\frac{1}{T_1}\sum_{t=0}^{T_1-1} l^{\theta}(s(t+1),s(t),o(t))T_2[V(\theta)-\tilde{V}]\right].
\end{aligned}
$$

The following proof is similar with the proof of Proposition 1.

REFERENCES

[1] A. R. Cassandra, "A survey of pomdp applications," in *Proc. Work. Notes AAAI Fall Symp. Planning Partially Observable Markov Decis. Processes*, 1998, pp. 17–24.

[2] X. R. Cao and J. Y. Zhang, "The nth-order bias optimality for multi-chain Markov decision processes," *IEEE Trans. Autom. Control*, vol. 53, no. 2, pp. 496–508, Mar. 2008.

[3] Y. J. Li, B. Q. Yin, and H. S. Xi, "Finding optimal memoryless policies of POMDPs under the expected average reward criterion," *Eur. J. Oper. Res.*, vol. 211, no. 3, pp. 556–567, 2011.

[4] X. R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. New York, NY, USA: Springer-Verlag, 2007.

[5] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Oper. Res.*, vol. 21, pp. 1071–1088, 1973.

[6] H. Kurniawati, D. Hsu, and W. S. Lee, "SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces," *Robot., Sci. Syst. IV*, pp. 65–72, 2009.

[7] R. He, E. Brunskill, and N. Roy, "Efficient planning under uncertainty with macro-actions," *J. Artif. Intell. Res.*, vol. 40, pp. 523–570, 2011.

[8] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[9] R. G. Stephen, C. R. Murthy, and M. Coupechoux, "A Markov decision theoretic approach to pilot allocation and receive antenna selection," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 3813–3823, Aug. 2013.

[10] V. K. N. Lau, F. Zhang, and Y. Cui, "Low complexity delay-constrained beamforming for multi-user MIMO systems with imperfect CSIT," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4090–4099, Aug. 2013.

[11] A. Brooks, A. Makarenko, S. Williams, and H. Durrant-Whyte , "Parametric POMDPs for planning in continuous state spaces," *Robot. Auton. Syst.*, vol. 54, no. 11, pp. 887–897, 2006.

[12] K. M. Seiler, H. Kurniawati, and S. P. N. Singh, "An online and approximate solver for POMDPs with continuous action space," *IEEE Int. Conf. Robot. Automat.*, pp. 2290–2297, 2015.

[13] V. Krishnamurthy and D. V. Djonin, "Optimal threshold policies for multivariate POMDPs in radar resource management," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3954–3969, Oct. 2009.

[14] Z. Feng, R. Dearden, N. Meuleau, and R. Washington, "Dynamic programming for structured continuous Markov decision problems," in *Proc. 20th Conf. Uncertainty Artif. Intell.*, 2004, pp. 154–161.

[15] Z. Zamani, S. Sanner, and C. Fang, "Symbolic dynamic programming for continuous state and action MDPs," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1839–1845.

[16] S. Carden, "Convergence of a Q-learning variant for continuous states and actions," *J. Artif. Intell. Res.*, vol. 49, no. 1, pp. 705–731, 2014.

[17] H. V. Hasselt and M. A. Wiering, "Reinforcement learning in continuous action spaces," in *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinforcement Learn.*, 2007, pp. 272–279.

[18] C. Gaskett, D. Wettergreen, and A. Zelinsky, "Q-Learning in continuous state and action spaces," in *Proc. Australas. Joint Conf. Artif. Intell.*, pp. 417–428, 1999.

[19] E. Zhou, M. C. Fu, and S. I. Marcus, "Solving continuous-state POMDPs via density projection," *IEEE Trans. Autom. Control*, vol. 55, no. 5, pp. 1101–1116, May 2010.

[20] J. M. Porta, N. Vlassis, M. T. J. Spaan, and P. Poupart, "Point-based value iteration for continuous POMDPs," *J. Mach. Learn. Res.*, vol. 7, pp. 2329–2367, 2006.

[21] S. Brechtel, T. Gindele, and R. Dillmann, "Solving continuous POMDPs: Value iteration with incremental learning of an efficient space representation," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 370–378.

[22] J. Hoey and P. Poupart, "Solving POMDPs with continuous or large discrete observation spaces," in *Proc. 19th Int. Joint Conf. Artif. Intell.*, 2005, pp. 1332–1338.

[23] X. R. Cao, D. X. Wang, and L. Qiu, "Partial-information state-based optimization of partially observable Markov decision processes and the separation principle," *IEEE Trans. Autom. Control*, vol. 59, no. 4, pp. 921–936, Apr. 2014.

[24] G. D. Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[25] P. Walters, *An Introduction to Ergodic Theory*. New York, NY, USA: Springer-Verlag, 1982.

[26] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[27] J. R. Blum, "Approximation methods which converge with probability one," *Ann. Math. Statist.*, vol. 25, no. 2, pp. 382–386, 1954.

[28] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*. New York, NY, USA: Springer-Verlag, 2012, pp. 421–436.