Research on Adaptive Waveform Selection Algorithm in Cognitive Radar

Bin Wang
Northeastern University, Shenyang, China
wangbin_neu@yahoo.com.cn

Jinkuan Wang, Xin Song and Yinghua Han
Northeastern University at Qinhuangdao, Qinhuangdao, China
sxin78916@mail.neuq.edu.cn

***Abstract*—Cognitive radar is a new framework of radar system proposed by Simon Haykin recently. Adaptive waveform selection is an important problem of intelligent transmitter in cognitive radar. In this paper, the problem of adaptive waveform selection is modeled as stochastic dynamic programming model. Then backward dynamic programming, temporal difference learning and Q-learning are used to solve this problem. Optimal waveform selection algorithm and approximate solutions are proposed respectively. The simulation results demonstrate that the two approximate methods approach the optimal waveform selection scheme and have lower uncertainty of state estimation compared to fixed waveform. The performance of temporal difference learning is better than Q-learning, but Q-learning is more suitable to use in radar scene. Finally, the whole paper is summarized.**

***Index Terms*—waveform selection, backward dynamic programming, temporal difference learning, Q-learning**

I. INTRODUCTION

The word radar is an abbreviation for radio detection and ranging which seems to have achieved universal acceptance all over the world. The invention of radar is inspired by the echolocation animals, such as bats and dolphins. The radar's basic function is intimately related to properties and characteristics of electromagnetic waves as they interface with physical objects. With the improvement of modern technology, radar develops rapidly with various needs of people. However, traditional radar is lack of flexibility and not suitable for different environment. The radar environment is usually nonstationary, and adaptive algorithm is the method implemented in modern radar systems for dealing with nonstationarity. In current designs of radar systems, most of the researches have focused on the design of optimal receiver. In order to adapt to different environment, the transmitter of radar should select different waveforms. That means we should consider the design of optimal transmitter.

Cognitive radar, proposed by Simon Haykin in 2006, is a new framework of radar system which is viewed as an intelligent form of radar system. In [1], it is argued that for the radar to be cognitive, adaptivity has to be extended to the transmitter too. Radar-scene analysis, Bayesian target tracking and adaptive radar illumination constitute the basic elements of cognitive radar. In cognitive radar, the radar continuously learns about the environment through experience gained from interactions of the receiver with the environment, the transmitter adjusts its illumination of the environment in an intelligent manner and the whole radar system constitutes a closed-loop dynamic system.

Cognitive radar is different from traditional radar because it can select appropriate waveforms according to different radar environment. So it is an important problem to realize the adaptivity of the transmitter. The design of adaptive transmitter involves adaptive model and adaptive algorithm. The research team of Simon Haykin has done much work on cognitive radar. After proposing the idea of cognitive radar, Simon Haykin suggests that much can be gained by rethinking the design of a radar system as a closed-loop feedback control system. The novel idea has been demonstrated herein in conceptual terms in the context of tracking radar [2]. In [3], a novel optimal radar waveform design problem by combining the signal-to-noise (SNR) and mutual information criteria is formulated. In [4], Arasaratnam develops a square-root extension of the quadrature Kalman filter using matrix triangularizations. And then he have successfully solved the best approximation to the Bayesian filter in the sense of completely preserving second-order information, which is called cubature Kalman filters [5]. N. A. Goodman has also done much work in this field. In [6], he summarizes and demonstrates a framework being developed at the University of Arizona for implementation of closed-loop radar with adaptive waveforms which integrates a Bayesian channel representation, matched illumination techniques, and sequential hypothesis testing. Goodman have proposed and simulated a closed-loop active sensor by updating the probabilities on an ensemble of target hypotheses while

Manuscript received September 9, 2009; revised March 23, 2010; accepted May 12, 2010.

Corresponding author: Bin Wang (wangbin_neu@yahoo.com.cn).

This work was supported by the Fundamental Research Funds for the Central Universities under Grant No. N090604006, and the National Natural Science Foundation of China under Grant No. 60874108 and No. 60904035.

doi:10.4304/jcm.5.6.467-474

adapting customized waveforms in response to prior measurement and compared the performance of two different waveform design techniques [7]. In [8], a new MIMO waveform is proposed which maximizes the mutual information between a Gaussian random target and the received data under AWGN. In [9], the authors present illumination waveforms matched to stochastic targets in the presence of signal-dependent interference. In [10], the authors extend the information-based approach to the signal-dependent interference problem. In [11], an extension to the PDA tracking algorithm to include adaptive waveform selection was developed. In [12], it is shown that tracking errors are highly dependent on the waveforms used and in many situations tracking performance using a good heterogeneous waveform is improved by an order of magnitude when compared with a scheme using a homogeneous pulse with the same energy. In [13], several node teaming algorithms have been considered for cooperative sensing in a mobile scenario. The problem of waveform selection can be thought of as a sensor scheduling problem, as each possible waveform provides a different means of measuring the environment, and related works have been examined in [14], [15]. In [16], Incremental Pruning method is used to solve the problem of adaptive waveform selection for target detection. The problem of optimal adaptive waveform selection for target tracking is also presented in [17]. In [18], genetic algorithm is used to perform waveform selection utilizing the autocorrelation and ambiguity functions in the fitness evaluation. In [19], radar waveform selection algorithm for tracking accelerating targets is considered. In [20], the author uses ADP method to solve the problem of adaptive waveform selection.

In this paper, under the assumption of range-Doppler resolution cell, stochastic dynamic programming model for adaptive transmitter is set up. We use backward dynamic programming method to solve this problem, and optimal waveform selection algorithm is proposed, where two forms of reward function are adopted. Then temporal difference learning waveform selection algorithm is proposed. Considering that the explicit knowledge of state-transition probabilities is unknown, we use Q-learning algorithm to obtain approximate solution. The simulation results demonstrate that these two methods approach the optimal waveform selection scheme and has lower uncertainty of state estimation compared to fixed waveform. The performance of temporal difference learning is better than Q-learning, but Q-learning is more suitable to use in radar scene. Finally, the whole paper is summarized.

Ⅱ. Model for Adaptive Waveform Selection

Generally speaking, for a target, the most important parameters that a radar measures are range, Doppler frequency, and two orthogonal space angles. If we envision a radar resolution cell that contains a certain four-dimensional hypervolume, we may assume different targets fall in different resolution cells. That means if a target measured falls in a resolution cell, then another target fall in another resolution cell and does not interfere with measurements on the first. So as long as each target occupies a resolution cell and the cells are all disjoint, the radar can make measurements on each target free of interference from others.

Through the discussion in [21], we can conclude that angle resolution can be considered independently from range and Doppler resolution in most circumstances. When considering this, the resolution properties of the radar in angle are independent of the resolution properties in range and Doppler frequency.

In our model, we omit angle resolution. We define range-Doppler resolution cell for the waveform selection model.

Radar systems are normally designed to operate between a minimum range $R_{\min}$ and maximum range $R_{\max}$. ΔR is range resolution which is a radar metric that describes its ability to detect targets in close proximity to each other as distinct objects. Targets separated by at least ΔR will be completely resolved in range. The distance between $R_{\min}$ and $R_{\max}$ is divided into N range bins, each of width is ΔR. The relationship between ΔR and N is

$$N = \frac{R_{\max} - R_{\min}}{\Delta R} \qquad (1)$$

Radars use Doppler frequency to extract target radial velocity (range rate), as well as to distinguish moving and stationary targets or objects such as clutter. The Doppler phenomenon describes the shift in the center frequency of an incident waveform.

In heavy clutter environments, we need to consider the problem of adaptive waveform selection and make a trade-off decision between Doppler and range resolution. Actually we can not obtain good Doppler and good range resolution in waveform tailoring simultaneously. So we should define a cost function that describes the cost of observing a target in a particular location for each individual pulse and select the waveform that optimizes this function on a pulse by pulse basis.

We adopt stochastic dynamic model. We make no assumptions about the number of targets that may be present. We divide the area covered by a particular radar beam into a grid in range-Doppler space, with the cells in range indexed by $\tau = 1,...,N$ and those in Doppler indexed by $\upsilon = 1,...,M$. There may be 0 target, 1 target or NM targets. So the number of possible scenes or hypotheses about the radar scene is 2^{NM}.

We roughly describe other variables. The definitions are as follows:

Radar scene: χ;

Model state: X_t;

Measurement variable: Y_t;

Waveform: u_t.

$a_{x'x}$ and $b_{x'x}$ are state transition probability and measurement probability respectively.

$$a_{x'x} = P(x_{t+1} = x' \mid x_t = x) \quad (2)$$

$$b_{x'x}(u_t) = P(Y_{t+1} = x' \mid X_t = x, u_t) \quad (3)$$

$\mathbf{A} = (a_{x'x})_{x',x\in\chi}$ is the state transition matrix. $\mathbf{B}(u_t) = (b_{x'x}(u_t))_{x',x\in\chi}$ is the measurement probability matrix. That means if state of our model is $X_t = x$ and we use the waveform u_t, the probability of measurement $Y_{t+1} = x'$ is $b_{x'x}(u_t)$.

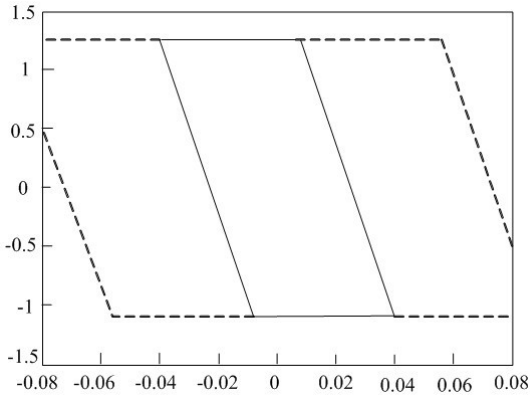

Figure 1. Resolution cell and corresponding parallelogram.

Figure 1 is resolution cell and corresponding parallelogram. We define as a practical resolution cell the parallelogram that contains the resolution cell primitive so that we can handle the problem that when a combined waveform is used.

We assume the matched filter is adpoted in the receiver. $s(t)$ represents baseband signal and $r(t)$ represents the received baseband signal. The matched filter is the one with an impulse response $h(t) = s^*(-t)$, so an output process of our matched filter is

$$x(t) = \int s^*(\lambda - t) r(\lambda) d\lambda \quad (4)$$

The output is given by

$$x(t) = \int s^*(\lambda - t) e^{-j2\pi\nu_0(\lambda - t)} r(\lambda) d\lambda \quad (5)$$

where ν_0 is an expected frequency shift.

The baseband received signal will be modeled as a return from a Swerling target.

Then we consider two situations: there is no target and target is present.

When there is no target

$$x(\tau_0) = \int_0^{\tau_0} n(\lambda) s^*(\lambda - \tau_0) d\lambda \quad (6)$$

The random variable $x(\tau_0)$ is complex Gaussian, with zero mean and variance given by

$$\sigma_0^2 = E\left\{x(\tau_0)x^*(\tau_0)\right\} = 2N_0\xi \quad (7)$$

where ξ is the energy of the transmitted pulse.

When target is present

$$x(\tau_0) = \int_0^{\tau_0} \left[As(\lambda - \tau) e^{j2\pi\nu_d\lambda} + n(\lambda) \right] s^*(\lambda - \tau_0) d\lambda \quad (8)$$

This random variable is still zero mean, with variance given by

$$\sigma_1^2 = E\left\{x(\tau_0)x^*(\tau_0)\right\} = \sigma_0^2 (1 + \frac{2\sigma_A^2 \xi^2}{\sigma_0^2} A(\tau_0 - \tau, \nu_0 - \nu)) \quad (9)$$

where $A(\tau, \nu)$ is ambiguity function, given by

$$A(\tau, \nu) = \frac{1}{\left(\int |s(\lambda)|^2 d\lambda\right)^2} \left| \int s(\lambda) s^*(\lambda - \tau) e^{j2\pi\nu\lambda} d\lambda \right|^2 \quad (10)$$

Finally we can obtain the probability of detection P_d in the case when a target is present in cell (τ, υ).

$$P_d = \frac{1}{|A|} \int_{(\tau_a, \upsilon_a \in A)} e^{-\frac{D}{2\sigma_0^2 (1 + \frac{2\sigma_A^2 \xi^2}{\sigma_0^2} A(\tau_0 - \tau, \nu_0 - \nu))}} d\tau_a d\upsilon_a \quad (11)$$

where ν_0 is an expected frequency shift, ξ is the energy of the transmitted pulse, σ_0^2 is the variance of transmitted signal, σ_A^2 is the variance of magnitude, A is the resolution cell centered on (τ, υ) with volume $|A|$.

Define $\pi = \{u_0, u_1, ..., u_T\}$ where $T+1$ is the maximum number of dwells that can be used to detect and confirm targets for a given beam. So π is a sequence of waveforms that could be used for that decision process. We can obtain different π according to different environment in cognitive radar. That means we should transmit different waveforms according to different radar working conditions. We define $R_t(X_t, u_t)$ is the reward earned when the scene X_t is observed using waveform u_t and γ is discount factor. $R_t(X_t, u_t)$ is called reward function and in it we can consider the importance of detecting a scatter in a cell. Maybe some targets are away from the radar and some targets are approaching the radar. So the importance of their corresponding cells is different. It leads to the different forms of $R_t(X_t, u_t)$.

Different $R_t(X_t, u_t)$ according to different π^* with different $V_t(X_t)$ can be accumulated to form discounted reward. We denote it as $V_t(X_t)$. That is

$$V_t(X_t) = E[\sum_{t=0}^{T} \gamma^t R_t(X_t, u_t)] \quad (12)$$

Then the aim of our problem is to find the sequence that maximize

$$V^*(X_t) = \max_{\pi} E[\sum_{t=0}^{T} \gamma^t R_t(X_t, u_t)] \quad (13)$$

However, the knowledge of the actual state is not available in radar scene. Using the method of [22], we can use another variable to substitute X_t which should

be a sufficient statistic of X_t. Then we can obtain that the optimal control policy π^* that is the solution of (13) is also the solution of

$$V^*(\mathbf{p}(0)) = \max_{\pi} E[\sum_{t=0}^{T} \gamma^t R_t(\mathbf{p}_t, u_t)] \quad (14)$$

where $\mathbf{p}_t$ is the conditional density of the state given the measurements and the controls and $\mathbf{p}_0$ is a priori probability density of the scene. $\mathbf{p}$ is a sufficient statistic for the true state X_t.

So our problem converts to

$$\max_{\pi} E[\sum_{t=0}^{T} \gamma^t R_t(\mathbf{p}_t, u_t)] \quad (15)$$

The refreshment formula of $\mathbf{p}_t$ is given by

$$\mathbf{p}_{t+1} = \frac{\mathbf{B}\mathbf{A}\mathbf{p}_t}{\mathbf{1}'\mathbf{B}\mathbf{A}\mathbf{p}_t} \quad (16)$$

where $\mathbf{B}$ is the diagonal matrix with the vector $(b_{x'x}(u_t))$ the non-zero elements and $\mathbf{1}$ is a column vector of ones. $\mathbf{A}$ is state transition matrix.

This is the waveform selection model that can be used in cognitive radar. If we use some adaptive algorithm, then we can realize the adaptivity of waveform selection.

III. THE OPTIMAL WAVEFORM SELECTION ALGORITHM

If we want to solve this problem using classical dynamic programming, we could have to find the value function $V_t(\mathbf{p}_t)$ using

$$V_t(\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma E\{V_{t+1}(\mathbf{p}_{t+1}) \mid \mathbf{p}_t\}) \quad (17)$$

It can also be written in probability form

$$V_t(\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma \sum_{\mathbf{p}' \in \mathbf{P}} P(\mathbf{p}' \mid \mathbf{p}_t, u_t) V_{t+1}(\mathbf{p}')) \quad (18)$$

With backward dynamic programming, we step forward in time.

Our problem is a finite horizon problem. Solving a finite horizon problem is straightforward. We simply have to compute the value function for each possible state $\mathbf{p}_t \in \mathbf{P}$ which start at the last time period and then step back another time period. So at time period t we have already computed $V_{t+1}(\mathbf{p}_{t+1})$. The critical element that attracts so much attention is the requirement that we compute the value function $V_t(\mathbf{p}_t)$ for all states $\mathbf{p}_t \in \mathbf{P}$. Theoretically, it is the optimal algorithm for waveform selection.

We describe the optimal algorithm as follows.

First, we should initialize the terminal contribution $V_T(\mathbf{p}_T)$. In most circumstances, we can let $V_T(\mathbf{p}_T) = 0$. Then set $t = T - 1$.

Second, we should calculate $V_t(\mathbf{p}_t)$. The formula of $V_t(\mathbf{p}_t)$ is

$$V_t(\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma \sum_{\mathbf{p}' \in \mathbf{P}} P(\mathbf{p}' \mid \mathbf{p}_t, u_t) V_{t+1}(\mathbf{p}')) \quad \text{for all} \quad \mathbf{p}_t \in \mathbf{P} \quad (19)$$

where

$$\mathbf{p}_{t+1} = \frac{\mathbf{B}\mathbf{A}\mathbf{p}_t}{\mathbf{1}'\mathbf{B}\mathbf{A}\mathbf{p}_t} \quad (20)$$

Third, if $t > 0$, decrement t and return to the first step. Else, stop.

Generally speaking, reward function represents the value that we stand in certain place and take some certain action and it can be different forms according to different problems. In the problem of adaptive waveform selection, linear reward function and entropy reward function are usually used.

Linear reward function is usually used in the circumstance that $R(\mathbf{p}, u)$ is required to be a piecewise linear function. The form of this function is simple and easy to calculate. However, it can not reflect the whole value sometimes. Entropy reward function comes from information theory which is usually used in the circumstance that $R(\mathbf{p}, u)$ is not required to be a piecewise linear function. It can reflect the whole value accurately. But it is more complex than linear reward function.

The form of linear reward function is

$$R_1(\mathbf{p}, u) = \mathbf{p}'\mathbf{p} - 1 \quad (21)$$

The form of entropy reward function is

$$R_2(\mathbf{p}, u) = \sum_{x \in \chi} p_x(k) \log(p_x(k)) \quad (22)$$

We can choose different form of reward function according to different problems.

IV. APPROXIMATE SOLUTION FOR WAVEFORM SELECTION ALGORITHM

Backward dynamic programming can be viewed as an optimal adaptive algorithm for waveform selection. When state space and action space are large, it is hard to use this method. That means we hardly find optimal solution for waveform selection. So we need to research on approximate solutions.

We use temporal difference method to find approximate solutions first.

Assume v is an unbiased sample estimate of the value of being in state $\mathbf{p}_t$ and the policy is π. The definition of v is

$$v_t^n = C_t(\mathbf{p}_t^n, u_t^\pi) + C_{t+1}(\mathbf{p}_{t+1}^n, u_{t+1}^\pi) + \ldots + C_T(\mathbf{p}_T^n, u_T^\pi) \quad (23)$$

where C_t is the contribution when in state $\mathbf{p}_t$ and using waveform u_t.

We use standard stochastic gradient algorithm to estimate the value of being in state X_t

$$V_t^n(\mathbf{p}_t) = V_t^{n-1}(\mathbf{p}_t) - \alpha_n[V_t^{n-1}(\mathbf{p}_t) - v_t^n] \quad (24)$$

where α is discount factor.

The temporal differences is

$$D_\tau = C_\tau(\mathbf{p}_\tau, u_\tau) + V_{\tau+1}^{n-1}(\mathbf{p}_{\tau+1}) - V_\tau^{n-1}(\mathbf{p}_\tau) \quad (25)$$

So

$$v_t^n = V_t^{n-1}(\mathbf{p}_t) + \sum_{\tau=t}^{T} D_\tau \quad (26)$$

Substituting (26) into (24), we can obtain

$$V_t^n(\mathbf{p}_t) = V_t^{n-1}(\mathbf{p}_t) - \alpha_{n-1}\sum_{\tau=t}^{T} D_\tau \quad (27)$$

The temporal differences are the errors in our estimates of the value of being in state $\mathbf{p}_\tau$. These errors are stochastic gradients for the problem of minimizing estimation error. Assume λ is artificial discount. The discount form is

$$V_t^n(\mathbf{p}_t) = V_t^{n-1}(\mathbf{p}_t) - \alpha_{n-1}\sum_{\tau=t}^{T} (\gamma\lambda)^{\tau-t} D_\tau \quad (28)$$

Through this formula, we can use this method to update the value of V.

Actually, temporal difference learning is a general class of methods of approximate dynamic programming. Q-learning is a special case of temporal difference learning. In radar scene, explicit knowledge of target state-transition probabilities is unknown. So directly using Bellman's dynamic programming is very hard. The Q-leaning algorithm is a direct approximation of Bellman's dynamic programming, and it can solve the problem that we do not know explicit knowledge of state-transition probabilities. For this reason, Q-learning is very suitable to be used in the problem of adaptive waveform selection in cognitive radar.

We define a Q-factor in our problem. For a state-action pair $(\mathbf{p}_t, u_t)$

$$Q(\mathbf{p}_t, u_t) = \sum_{\mathbf{p}'\in\mathbf{P}} P(\mathbf{p}'|\mathbf{p}_t, u_t)[R_t(\mathbf{p}'|\mathbf{p}_t, u_t) + \gamma V_{t+1}] \quad (29)$$

According to (18), (25) we can derive

$$V_t^* = \max_{u_t} Q(\mathbf{p}_t, u_t) \quad (30)$$

The above establishes the relationship between the value function of a state and the Q-factors associated with a state. Then it should be clear that, if the Q-factors are known, one can obtain the value function of a given state from above formula.

So Q form of Bellman equation is

$$Q(\mathbf{p}_t, u_t) = \sum_{\mathbf{p}'\in\mathbf{P}} P(\mathbf{p}'|\mathbf{p}_t, u_t)[R_t(\mathbf{p}'\Big|\mathbf{p}_t, u_t) + \gamma \max_{u_{t+1}} Q(\mathbf{p}_{t+1}, u_{t+1})] \quad (31)$$

Let us denote the ith independent sample of a random variable X by s^i and the expected value by $E(X)$. X^n represents the estimate of X in the nth iteration. $E(X)$ can be viewed a limitation form of average sample

$$E(X) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} s^i}{n} \quad (32)$$

$$X^n = \frac{\sum_{i=1}^{n} s^i}{n} \quad (33)$$

According (32) and (33), we can derive

$$X^{n+1} = (1-\alpha^{n+1})X^n + \alpha^{n+1}s^{n+1} \quad (34)$$

where

$$\alpha^{n+1} = \frac{1}{n+1} \quad (35)$$

So we can use the following formula to obtain Q

$$Q(\mathbf{p}_t, u_t) = E[R_t(\mathbf{p}'\Big|\mathbf{p}_t, u_t) + \gamma \max_{u_{t+1}} Q(\mathbf{p}_{t+1}, u_{t+1})] \quad (36)$$

where E is the expectation operator. We could use this scheme in a simulator to estimate the same Q-factor. Using this algorithm, Equation (31) becomes

$$Q^{n+1}(\mathbf{p}_t, u_t) \leftarrow (1-\alpha^{n+1})Q^n(\mathbf{p}_t, u_t) + \alpha^{n+1}[R_t(\mathbf{p}'\Big|\mathbf{p}_t, u_t) + \gamma \max_{u_{t+1}} Q^n(\mathbf{p}_{t+1}, u_{t+1})] \quad (37)$$

Obviously, we do not have the transition probabilities in it.

Our Q-learning algorithm is as follows:

Step 1. Initialize the Q-factors to 0 and set $n = 1$.

Then for $t = 0, 1, ..., T$, do step 2-step 5.

Step 2. Simulation action u_t. Let the current state be $\mathbf{p}_t$, and the next state be $\mathbf{p}_{t+1}$.

Step 3. Find the decision using the current Q-factors

$$u_t = \arg\max_{u_t} Q_t^{n-1}(\mathbf{p}_t^n, u_t) \quad (38)$$

Step 4. Update $Q(\mathbf{p}_t, u_t)$ using the following formula:

$$Q^{n+1}(\mathbf{p}_t, u_t) \leftarrow (1-\alpha^{n+1})Q^n(\mathbf{p}_t, u_t) + \alpha^{n+1}[R_t(\mathbf{p}'\Big|\mathbf{p}_t, u_t) + \gamma \max_{u_{t+1}} Q^n(\mathbf{p}_{t+1}, u_{t+1})] \quad (39)$$

Step 5. Update the state using the following equation:

$$\mathbf{p}_{t+1} = \frac{\mathbf{B}\mathbf{A}\mathbf{p}_t}{\mathbf{1}'\mathbf{B}\mathbf{A}\mathbf{p}_t} \quad (40)$$

Step 6. Increment n till $n \geq N$. Otherwise, go to step 2 and do step 2-step 5 again.

Step 7. For each $\mathbf{p}_t \in \mathbf{P}$, select

$$d(\mathbf{p}_t) \in \arg\max_{u_t} Q(\mathbf{p}_{t+1}, u_{t+1}) \quad (41)$$

The policy generated by the algorithm is $\hat{d}$. Stop.

V. SIMULATION

We adopt linear frequency modulation (LFM) signal in the transmitter. The formula of LFM is

$$s(t) = rect(\frac{t}{T})e^{j2\pi(f_c t + \frac{K}{2}t^2)} \quad (42)$$

where f_c is carrier frequency and $rect$ is rectangular signal. The specific form of rectangular signal is as follows, and then we will simulate the linear frequency modulation signal.

$$rect(\frac{t}{T}) = \begin{cases} 1 \quad , & \left|\frac{t}{T}\right| \le 1 \\ 0 \quad , & elsewise \end{cases} \tag{43}$$

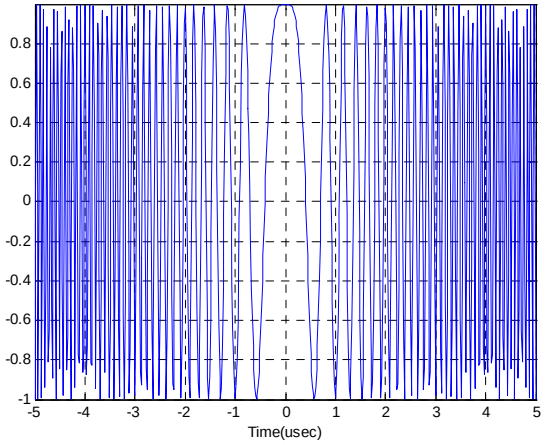

Figure 2. Real part of chirp signal.

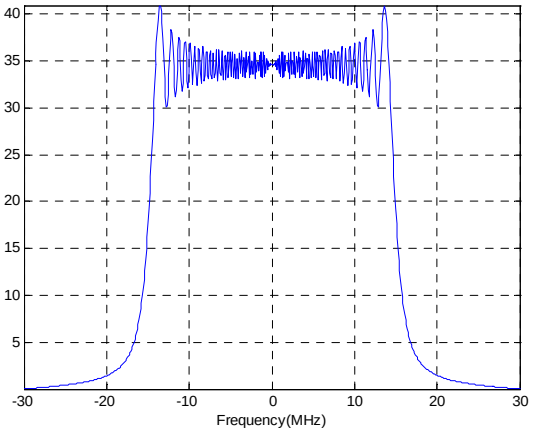

Figure 3. Magnitude spectrum of chirp signal.

Figure 2 is real part of chirp signal and figure 3 is magnitude spectrum of chirp signal.

Actually, the optimal waveforms don't depend on the form of reward function. So we can use the two reward functions as what we need. After a comparison of curve of uncertainty of state estimation using formula (21) and (22), we will use linear reward function as the basis for our reward function in the following experiment. The formula $E(1-\mathbf{p}'\mathbf{p})$ can be considered as the uncertainty in the state estimation. In other words, it can represent tracking errors. This parameter can reflect the performance of our algorithm.

We consider a simple scenario. The state space is 4×4. We consider 5 different waveforms where for each waveform u.

Each hypotheses for the target x, the distribution of x' is given in table 1.

The discount factor $\gamma = 0.9$. The matrix $\mathbf{A}$ is given by equation (44).

TABLE I.
MEASUREMENT PROBABILITIES FOR THE EXAMPLE SCENARIO

	x=1 x'=1,2 3,4	x=2 x'=1,2 3,4	x=3 x'=1,2 3,4	x=4 x'=1,2 3,4
U=1	0.97,0.01 0.01,0.01	0.96,0.01 0.01,0.02	0.01,0.01 0.96,0.02	0.01,0.95 0.02,0.02
U=2	0.96,0.01 0.02,0.01	0.02,0.95 0.01,0.02	0.01,0.01 0.01,0.97	0.02,0.96 0.01,0.01
U=3	0.02,0.95 0.02,0.01	0.02,0.02 0.01,0.95	0.02,0.96 0.01,0.01	0.01,0.02 0.02,0.95
U=4	0.96,0.01 0.01,0.02	0.01,0.96 0.02,0.01	0.97,0.01 0.01,0.01	0.03,0.95 0.01,0.01
U=5	0.01,0.02 0.04,0.03	0.01,0.97 0.01,0.01	0.02,0.01 0.96,0.01	0.04,0.94 0.01,0.01

$$\mathbf{A} = \begin{bmatrix} 0.96 & 0.02 & 0.01 & 0.01 \\ 0.01 & 0.93 & 0.03 & 0.02 \\ 0.02 & 0.03 & 0.95 & 0.02 \\ 0.01 & 0.02 & 0.01 & 0.95 \end{bmatrix} \tag{44}$$

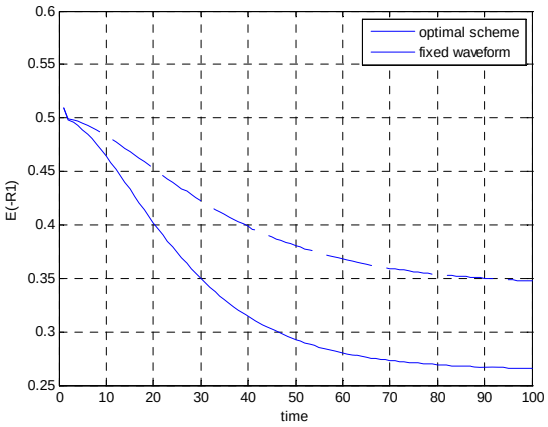

Figure 4. Curve of uncertainty of state estimation using formula (21).

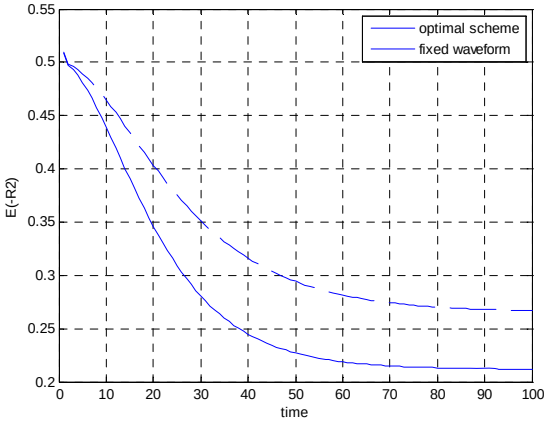

Figure 5. Curve of uncertainty of state estimation using formula (22).

Figure 4 is curve of uncertainty of state estimation using formula (21) and figure 5 is curve of uncertainty of state estimation using formula (22). In fact, the optimal adaptive waveform selection can be viewed as minimizing the uncertainty in the state estimation or target tracking errors. We can see the tracking errors are becoming lower with the increase of time. The tracking errors using BDP method are lower than fixed waveform. Moreover, the advantages of BDP method do not depend on the form of reward function we use.

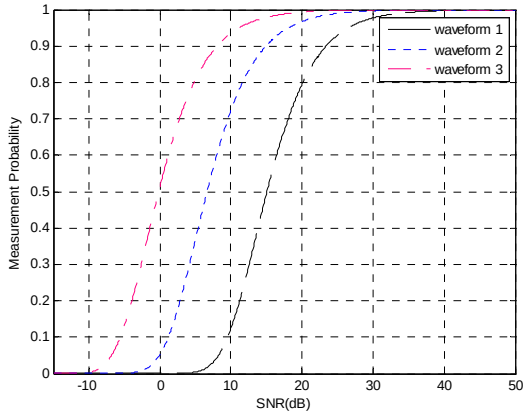

Figure 6. Measure probability versus SNR with three different waveforms.

Figure 6 is curve of measurement probability versus SNR with three different waveforms. From this figure we can see measurement probability is becoming large with the increase of SNR. Under the same SNR, using different waveform corresponds to different measurement probability. So measurement probability can be improved through appropriately scheduling waveform.

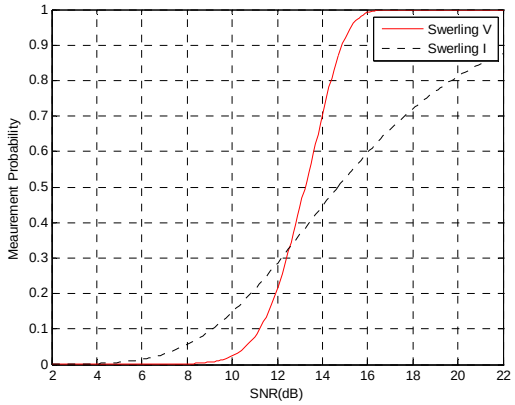

Figure 7. Measurement probability versus SNR with different targets.

Figure 7 is curve of measurement probability versus SNR with different targets. From this figure we can see that under the same SNR, measurement probability is different to different targets. So according to different targets we should select different waveforms. In actuality, path of target is so complex. We should change waveform according to different environment.

Figure 8 is curve of uncertainty of state estimation. From this curve we can see that for all the cases, the uncertainty of state estimation is decreasing with time, no matter how the state is changing with time. Compared to a fixed waveform, temporal difference learning and Q-learning algorithm we proposed has lower uncertainty of state estimation. That means our algorithm will reduce uncertainty in locating targets. Meanwhile our algorithm approaches the optimal waveform selection scheme even though explicit knowledge of state-transition probabilities are unknown. The performance of temporal difference learning is better than Q-learning, but Q-learning is more suitable to use in radar scene.

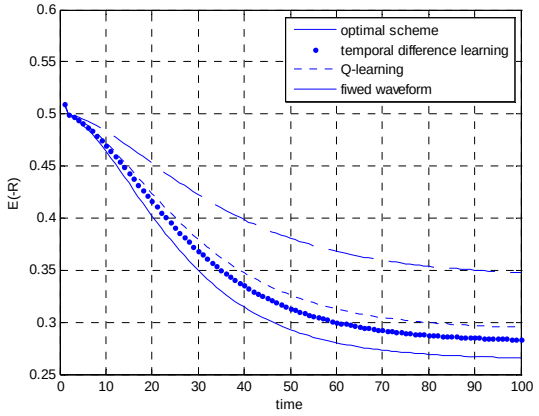

Figure 8. Curve of uncertainty of state estimation.

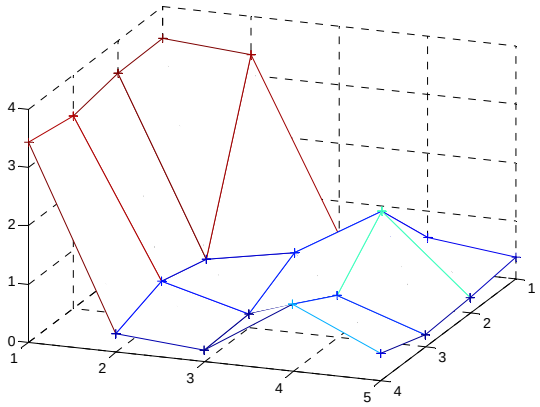

Figure 9. Q value space versus state and waveform.

Figure 9 is the figure of Q value space versus state and waveform. Q value of different state-waveform pair can be obtained in this figure. We can see that the proposed algorithm has lower computational cost.

VI. CONCLUSIONS

Adaptive waveform selection is an important problem in cognitive radar and the problem of adaptive waveform selection can be viewed as a stochastic dynamic programming problem. In this paper, under the assumption of range-Doppler resolution cell, stochastic dynamic programming model for adaptive transmitter is set up. Then backward dynamic programming, temporal difference learning and Q-learning are used to solve this problem. Backward dynamic programming can be viewed as optimal algorithm for waveform selection. Temporal

difference learning and Q-learning are approximate solutions. The advantages of temporal difference learning and Q-learning over fixed waveform have been shown with simulations. The two approximate algorithms can minimize the uncertainty of state estimation compared to fixed waveform and approaches the optimal waveform selection scheme. The performance of temporal difference learning is better than Q-learning, but Q-learning is more suitable to use in radar scene. Q-learning can solve the problems in which explicit knowledge of state-transition probabilities are unknown. Research on algorithms which approach the optimal waveform selection scheme and has lower computational cost is an important problem.

REFERENCES

[1] S. Haykin, "Cognitive radar: a way of the future", *IEEE Signal Processing Magazine*, 2006, 23(1), pp. 30-40.

[2] S. Haykin, "Cognition is the key to the next generation of radar systems", *13th IEEE Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, 2009, pp. 463-467.

[3] S. Haykin, Y. B. Xue and T. Davidson, "Optimal waveform design for cognitive radar", *42nd Asilomar Conference on Signals, Systems and Computers*, 2008, Pacific Grove, CA, pp.3-7.

[4] I. Arasaratnam and S. Haykin, "Square-Root Quadrature Kalman Filtering", *IEEE Tran. Signal Processing*, 2008, 56(6), pp. 2589-2593.

[5] I. Arasaratnam and S. Haykin, "Cubature Kalman filters", *IEEE Tran. Automatic Control*, 2009, 54(6), pp. 463-467.

[6] N. A. Goodman, "Closed-Loop Radar with Adaptively Matched Waveforms", *International Conference on Electromagnetics in Advanced Applications*, 2007, Torino, pp. 468-471.

[7] N. A. Goodman, Phaneendra R. Venkata and Mark A. Neifeld, "Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors", *IEEE Journal of Selected Topics in Signal Processing*, 2007, 1(1), pp. 105-113.

[8] T. B. Butler and N. A. Goodman, "Multistatic target classification with adaptive waveforms", *IEEE radar conference*, 2008, Rome, pp.1-6.

[9] R. A. Romera and N. A. Goodman, "Waveform design in signal-dependent interference and application to target recognition with multiple transmissions", *IET Radar, Sonar & Navigation*, 2009, 3(4), pp.328-340.

[10] R. Romero and N. A. Goodman, "Information-theoretic matched waveform in single dependent interference", *IEEE radar conference*, 2008, Rome, pp.1-6.

[11] D. J. Kershaw and R. J. Evans, "Waveform selective probabilistic data association", *IEEE Transactions on Aerospace and Electronic Systems*, 1997, 33(4), pp. 1180-1188.

[12] C. Rago, P. Willett and Y. Bar-Shalom, "Detecting-tracking performance with combined Waveforms", *IEEE Transactions on Aerospace and Electronic Systems*, 1998, 34(2), pp. 612-624.

[13] Lingfeng Wang, A. Doufexi, C. Williams and J. McGeehan, "Cognitive Node Selection and Assignment Algorithms for Weighted Cooperative Sensing in Radar Systems", *IEEE Wireless Communications and Networking Conference*, 2009, Budapest, pp.1-6.

[14] Y. He and E. K. P. Chong, "Sensor scheduling for target tracking in sensor networks", *43rd IEEE Conference on Decision and Control*, Paradise, Island, Bahamas, 2004, pp. 743-748.

[15] V. Krishnamurthy, "Algorithms for optimal scheduling of hidden Markov model sensors", *IEEE Trans. on Signal Processing*, 2002, 50(6), pp.1382-1397.

[16] B. F. La Scala, W. Moran and R. J. Evans, "Optimal adaptive waveform selection for target detection", *The International Conference on Radar*, Adelaide, SA, Australia, 2003, pp. 492-496.

[17] B. La Scala, M. Rezaeian and B. Moran, "Optimal adaptive waveform selection for target tracking", *International Conference on Information Fusion*, 2005, pp. 552-557.

[18] C. T. Capraro, I. Bradaric, G.T. Capraro and Tsu Kong Lue, "Using genetic algorithms for radar selection", *IEEE Radar Conference*, Inc., Utica, NY, 2008, pp. 1-6.

[19] C. O. Savage and B. Moran, "Waveform Selection for Maneuvering Targets Within an IMM Framework", *IEEE Transactions on Aerospace and Electronic Systems*, 2007, 43(3), pp. 1205-1214.

[20] B. Wang, J. K. Wang and J. Li, "ADP-based optimal adaptive waveform selection in cognitive radar", *International Symposium on Intelligent Information Technology Applications Workshops*, Shanghai, China, 2008, pp. 788-790.

[21] P. Z. Peebles, *Radar Principles*, John Wiley & Sons, Inc, 1998.

[22] D. Bertsekas, *Dynamic Programming and Optimal Control*, volume1, Athena Scientific, 2nd edition, 2001.

Bin Wang was born in Hebei, China, in 1982. He received M.S. degree in communication and information system in Northeastern University in China in 2008. Since March 2008, he has been working for his Ph.D. degree in Northeastern University. His research interests are in the area of cognitive radar and adaptive waveform selection.

Jinkuan Wang received his Ph.D. degree from University of Electro-Communications, Japan, in 1993. He is currently a professor in the college of information science and engineering in Northeastern University, China, since 1998. His main interests are in the area of intelligent control and adaptive array.

Xin Song was born in Jilin, China, in 1978. She received her Ph.D. degree in communication and information system in Northeastern University in China in 2008. She is now a teacher in Northeastern University at Qinhuangdao, China. Her research interests are robust adaptive beamforming and wireless communication.

Yinghua Han was born in Jilin, China, in 1979. She received the M.S. and Ph.D. degree in college of information science and engineering from Northeastern University, Shenyang, China, in 2005 and 2008, respectively. Since 2003, she is with engineering optimization and smart antenna institute. Her research interests include array signal processing and mobile wireless communication systems.