CAN REINFORCEMENT LEARNING ALWAYS PROVIDE THE BEST POLICY?

Zhansheng Duan *Huimin Chen*

Department of Electrical Engineering
University of New Orleans
2000 Lakeshore Drive, New Orleans, LA 70148

ABSTRACT

Reinforcement learning deals with how to find the best policy under uncertain environment to maximize some notion of long term reward. In sequential decision making, it is often expected that the best policy can be designed by choosing appropriate reward or penalty for each action. In this paper, we provide a counterexample to show that the best sequential decision rule can not be obtained by the choice of any reward function in the reinforcement learning framework. In fact, the best policy, namely, the randomized sequential probability ratio test, can only be learned via a rather unconventional formulation of the reinforcement learning. The implication to the design of classifier combining method is also discussed.

Index Terms— Reinforcement learning, sequential decision, classifier combining.

1. INTRODUCTION

In most statistical signal processing problems, one acquires a fixed number of observations to make the inference about the prevailing hypothesis. An alternative approach is to fix the desired performance levels and allow the number of samples to vary so that one achieves this performance with minimum number of samples. Sequential tests, on the average, offer substantial savings over tests with fixed sample size in terms of the number of samples required to achieve a given level of performance. In an automatic target recognition (ATR) system, a large volume of sensor data is assessed to identify target type with constraints on computation. High data rates and real time processing requirements for wide area surveillance have given rise to a staged decision strategy as the de facto approach to ATR. Each stage of the ATR system computes discrimination statistics to reduce the false alarms while maintaining high probability of detection. Screening false alarms reduces the data rate faced by subsequent stages. Sequential decision making is commonly adopted in ATR systems even for combining the outputs from different classifiers working in parallel.

The sample size savings of sequential tests come with a cost. The design of such sequential procedures requires exact knowledge of the conditional density function for the observations under each hypothesis. As a result, sequential tests are not robust to the variations of sample statistics. Reinforcement learning, as a powerful machine learning tool, is concerned with how to find the best policy under uncertain environment so as to maximize some notion of the long term reward [1]. It has been applied to many sequential decision making problems [2, 3]. The major advantage of reinforcement learning over supervised learning lies in that correct input/output pairs are never presented, nor sub-optimal decisions explicitly corrected. By choosing appropriate reward or penalty for each action, one expects to find the best decision making policy in terms of a general set of decision performance metrics. Considerable efforts have been made on the design of optimal sequential detector using reinforcement learning algorithms [4, 5]. However, no theoretical support exists to our best knowledge that guarantees the best policy obtained by reinforcement learning converges to the optimal detector with known conditional densities. In this paper, we provide a counterexample where the best sequential decision rule can not be obtained in the reinforcement learning framework no matter what reward or penalty is assigned to each action. In fact, the best sequential decision rule given by the randomized sequential probability ratio test can only be obtained via a rather unconventional formulation of the reinforcement learning.

The rest of the paper is organized as follows. Section 2 presents a simple sequential hypothesis testing problem to be used throughout the paper. Section 3 provides the sufficient statistic for the sequential decision making of this problem. Reinforcement learning formulation is given Section 4. Section 5 shows the negative result that the optimal decision rule can not be learned by choosing an appropriate reward in reinforcement learning. Concluding remarks are in Section 6.

This work was supported in part by ARO grant W911NF-04-1-0274, ASEE Air Force Summer Faculty Fellowship Program, UNO office of research sponsored program under investing in research excellence (IRE), Navy through Planning Systems Contract # N68335-05-C-0382, and NSFC grant 60602026. Z. Duan is also with the College of Electronic and Information Engineering, Xi'an Jiaotong University.

2. SEQUENTIAL DECISION PROBLEM

Suppose that an object can only be one of the two known types. Based on sensor data, we need to decide whether it is of type H or type T. We assume that a classifier can provide the object's type with the confusion matrix given as follows.

		Decision	
		"H"	"T"
Truth	H	$1-p$	p
	T	p	$1-p$

For the sake of convenience, we assume that $p = 0.1$. With constraints on the false alarm and miss probabilities such as

$$P_{FA} \leq \alpha = 0.01,\ P_M \leq \beta = 0.01, \tag{1}$$

we want to declare the object's type as quickly as possible based on the consecutive outputs from the classifier. In sequential decision context, given the classifier's outputs up to now, one can either declare the type of the object or wait for the next output.

Remarks: The above formulation of the sequential decision problem using the output from a single classifier can be readily extended to the case of combining the outputs from multiple classifiers with improved accuracy. The only difference is that the sequential decision making with a single classifier is over time, while the sequential decision making with multiple classifiers is along the classifiers at a single time. However, different classifiers may have different confusion matrices, which will make the subsequent derivations complicated.

3. LIKELIHOOD RATIO TEST STATISTIC

In this section, we derive the likelihood ratio test statistic which will be used by the optimal sequential decision rule. Denote by Z_i the output of the i-th classification where $Z_i \in \{\text{"}H\text{"}, \text{"}T\text{"}\}$. Let Z^i be the observation sequence up to time i, i.e., Z^i=$\{Z_1, Z_2, \cdots, Z_i\}$. Suppose that during the first N observations, the number of Hs being declared is K and the number of Ts being declared is $N-K$, then the likelihood ratio test statistic up to the N-th observation is

$$T\left(Z^N\right) = \prod_{i=1}^{N} \frac{P\left(Z_i|T\right)}{P\left(Z_i|H\right)} = \frac{\binom{N}{K} p^K (1-p)^{N-K}}{\binom{N}{K} (1-p)^K p^{N-K}} = \left(\frac{p}{1-p}\right)^n.$$

The log-likelihood ratio test statistic can be written as

$$L\left(Z^N\right) = \log\left(T\left(Z^N\right)\right) = n \log\left(\frac{p}{1-p}\right)$$

where $n = 2K - N$ is the difference between the number of Hs being declared and the number of Ts being declared. Thus for an arbitrary output sequence of length N, the likelihood ratio test boils down to compare n with a threshold to achieve a certain test power.

4. REINFORCEMENT LEARNING FORMULATION

In this section, we provide the reinforcement learning formulation of the binary sequential decision problem where one maximizes certain notion of the reward in order to achieve the desired detection performance metrics. At time t, an agent observes Z_t and chooses the best decision from a set of possible actions, i.e., $\{\text{"}H\text{"}, \text{"}T\text{"}, \text{wait}\}$. The best decision should maximize the total expected future reward $R = \sum_{i=1}^{t} \gamma^i r_i$ where γ is a discount constant and r_i is the expected reward at time i. Denote by R_+ the reward for making a correct decision on the type of the object and R_- the penalty for making an incorrect decision. For this sequential decision example, the best policies to maximize various reward functions have been obtained in [6]. We are interested in finding the optimal policy for any given $R_+ > 0$ and $R_- < 0$.

Note that given an observation sequence of length N, the posterior probability that the object is of type H can be written as

$$P\left(H|n\right) = \frac{1}{1 + \exp\left(-n \log\left(\frac{1-p}{p}\right)\right)}.$$

Similarly, the posterior probability that the object is of type T can be written as

$$P\left(T|n\right) = \frac{1}{1 + \exp\left(n \log\left(\frac{1-p}{p}\right)\right)}$$

where we assume that an object has equal prior between the two types. Thus calculating the expected reward only requires the knowledge of n.

Objective 1: Maximizing expected reward per observation.

We call n the state which starts at $n = 0$. For any given $\mu > 0$, denote by $R_H(\mu)$ the expected reward for declaring "H" when $n = \mu$, then we have

$$\begin{aligned} R_H(\mu) &= R_+ \cdot P\left(H|\mu\right) + R_- \cdot P\left(T|\mu\right) \\ &= \left(R_+ - R_-\right) \cdot P\left(H|\mu\right) + R_-. \end{aligned}$$

Similarly, $R_T(-\mu)$ is the expected reward for declaring "T" when $n = -\mu$. We have

$$\begin{aligned} R_T(-\mu) &= R_+ \cdot P\left(T|-\mu\right) + R_- \cdot P\left(H|-\mu\right) \\ &= \left(R_+ - R_-\right) \cdot P\left(T|-\mu\right) + R_-. \end{aligned}$$

Due to the symmetry of the problem, the expected rewards should satisfy

$$P\left(T|-\mu\right) = P\left(H|\mu\right).$$

Thus the expected reward $R(\mu)$ for a given decision policy μ satisfies

$$R(\mu) = R_H(\mu) = R_T(-\mu). \tag{2}$$

Let $t_\mu(n)$ be the expected number of observations to reach a decision if the current state is n and the policy is to declare H when $n = \mu$ (and to declare T when $n = -\mu$). Denote by $P_{\text{next}}(T|n)$ the probability that the next output from the classifier is T while the current state is n. We have

$$P_{\text{next}}(T|n) = p \cdot P(H|n) + (1-p) \cdot P(T|n).$$

Similarly, the probability that the next output is H given the current state n can be obtained as follows.

$$P_{\text{next}}(H|n) = (1-p) \cdot P(H|n) + p \cdot P(T|n).$$

Now we can write $t_\mu(n)$ in a recursive form

$$t_\mu(n) = 1 + P_{\text{next}}(H|n) \cdot t_\mu(n+1) + P_{\text{next}}(T|n) \cdot t_\mu(n-1) \tag{3}$$

with boundary condition $t_\mu(\mu) = 0$, $t_\mu(-\mu) = 0$.

Divide $R(\mu)$ by $t_\mu(0)$, one obtains the expected reward per observation. Thus we can choose the optimal policy which yields the largest expected reward per observation according to the first objective.

Objective 2: Maximizing the total expected future reward

To avoid long delay in order to make an accurate decision, one can associate a fixed penalty c for asking a new observation. In addition, the future reward should be discounted as opposed to the standard reinforcement learning formulation where the future reward is more important than the current one.

Let V_n be the total expected future reward given that the current state is n and the decision policy is μ, i.e., one declares H when $n = \mu$ and declares T when $n = -\mu$. Denote by c the penalty (cost) with each request to see another observation. Note that a decision will be made when $n = \pm\mu$. In this case, the sequential decision terminates so that the total expected future reward is equal to the expected reward for the final action. Thus we have

$$V_\mu = V_{-\mu} = (R_+ - R_-) \cdot P(H|\mu) + R_-. \tag{4}$$

When $-\mu < n < \mu$, we need to ask for one more observation. The total expected future reward of these states is the total expected future reward of the next state minus the cost associated with the action, that is

$$V_n = V_{n+1} P_{\text{next}}(H|n) + V_{n-1} P_{\text{next}}(T|n) - c \tag{5}$$

We should choose the decision policy that yields the largest V_0 according to the second objective.

The questions is, based on objective 1 or 2, whether one can assign appropriate R_+, R_-, c, and γ to obtain the best policy in terms of having the quickest decision satisfying the constraints given by α and β. It is often expected that the best policy can be obtained at least for the symmetric case $\alpha = \beta$.

5. NEGATIVE RESULT

5.1. Existence of the Best Policy

It is well known that sequential probability ratio test (SPRT) is optimal in the sense that it minimizes the expected delays for both hypotheses H and T among all sequential schemes satisfying the constraints of false alarm and miss probabilities [7, 8, 9, 10]. Specifically, for the problem of identifying the object's type, the optimal decision rule is as follows.

$$n \begin{cases} > \frac{\log\left(\frac{\beta}{1-\alpha}\right)}{\log\left(\frac{p}{1-p}\right)} & \text{declare } H \\ < \frac{\log\left(\frac{1-\beta}{\alpha}\right)}{\log\left(\frac{p}{1-p}\right)} & \text{declare } T \\ \in \left[\frac{\log\left(\frac{1-\beta}{\alpha}\right)}{\log\left(\frac{p}{1-p}\right)}, \frac{\log\left(\frac{\beta}{1-\alpha}\right)}{\log\left(\frac{p}{1-p}\right)}\right] & \text{ask for one more observation} \end{cases}$$

Since n can only take integer values, randomized policy has to be used to achieve the minimum delay for general constraints on the false alarm and miss probabilities. Thus the optimal policy for our example is extended as below.

$$n \begin{cases} = \mu_0 & \text{declare } H \\ = \mu_1 & \text{declare } T \\ \in (\mu_1, \mu_0) & \text{ask for one more observation} \end{cases} \tag{6}$$

The threshold μ_0 is randomly chosen between μ_0' and $\mu_0' + 1$ with

$$P\{\mu_0 = \mu_0'\} = p_0$$
$$P\{\mu_0 = \mu_0' + 1\} = 1 - p_0$$

where

$$\mu_0' = \left\lfloor \frac{\log\left(\frac{\beta}{1-\alpha}\right)}{\log\left(\frac{p}{1-p}\right)} \right\rfloor.$$

Similarly, the threshold μ_1 is randomly chosen between μ_1' and $\mu_1' - 1$ with

$$P\{\mu_1 = \mu_1'\} = p_1$$
$$P\{\mu_1 = \mu_1' - 1\} = 1 - p_1$$

where

$$\mu_1' = \left\lceil \frac{\log\left(\frac{1-\beta}{\alpha}\right)}{\log\left(\frac{p}{1-p}\right)} \right\rceil.$$

The two probabilities p_0 and p_1 are chosen to achieve the desired α and β.

For example, when $\alpha = \beta = 0.01$, the optimal policy is to randomly choose the decision rule with $\{\mu_0 = 2, \mu_1 = -2\}$ and that with $\{\mu_0 = 3, \mu_1 = -3\}$. When $\mu_1 = -3$ and the true type is type H, we have

$$P_{FA}^{\mu_1=-3} = P_{\mu_1=-3}(\text{“}T\text{”}|H) = 0.0014.$$

Using Wald's formula [9], the expected number of observations before reaching a decision is

$$E_{\mu_0=3,\ \mu_1=-3}[N|H] = 3.7397.$$

Similarly, when the true type is T, we have

$$P_M^{\mu_0=3} = P_{\mu_0=3}(\text{“}H\text{”}|T) = 0.0014,$$

$$E_{\mu_0=3,\ \mu_1=-3}[N|T] = 3.7397.$$

When $\mu_1 = -2$ and the true type is H, we have

$$P_{FA}^{\mu_1=-2} = P_{\mu_1=-2}(\text{“}T\text{”}|H) = 0.0122,$$

$$E_{\mu_0=2,\ \mu_1=-2}[N|H] = 2.4390.$$

Similarly, when the true type is T, we can get

$$P_M^{\mu_0=2} = P_{\mu_0=2}(\text{“}H\text{”}|T) = 0.0122,$$

$$E_{\mu_0=2,\ \mu_1=-2}[N|T] = 2.4390.$$

Thus we can obtain p_0 to achieve $P(\text{“}H\text{”}|T) \le 0.01$ as follows.

$$\begin{aligned} P(\text{“}H\text{”}|T) &= P_{\mu_0=2}(\text{“}H\text{”}|T)\cdot p_0 + P_{\mu_0=3}(\text{“}H\text{”}|T)\cdot(1-p_0) \\ &= 0.0122\cdot p_0 + 0.0014\cdot(1-p_0) \\ &\le 0.01, \end{aligned}$$

which leads to

$$p_0 \le 0.7963.$$

Similarly, we have $p_1 \le 0.7963$.

When $p_0 = p_1 = 0.7963$, the expected number of observations before reaching a decision is

$$\begin{aligned} E[N|H] &= E_{\mu_0=2,\ \mu_1=-2}[N|H]\cdot p_0 + E_{\mu_0=3,\ \mu_1=-3}[N|H] \\ &\quad\cdot(1-p_0) \\ &= 2.704. \end{aligned}$$

Similarly, we have $E[N|T] = 2.704$.

5.2. Unattainable Policy Using Reinforcement Learning

Theorem: The optimal policy to achieve the minimum delay of decision with certain α and β (e.g., $\alpha = \beta = 0.01$) is unattainable by reinforcement learning with any reward R_+, penalty R_-, cost per observation c.

Proof Sketch: From Eq. (3), we can see that once the policy μ is given, the expected time $t_\mu(0)$ is irrelevant to the choice of R_+ and R_-. In addition, from Eq. (4), we can see that the expected reward when $n = \pm\mu$ depends only on R_+ and R_-. That is, once R_+ and R_- are given, the expected reward per sample is a univariate function of μ. Similarly, from equations (4)–(5), the total expected future reward is also a univariate function of μ once c, R_+ and R_- are given, where different combinations of c, R_+ and R_- can only change the peak of V_0. Thus the optimal policy by maximizing either objective 1 or 2 will be a sequential strategy with fixed μ (which is an integer) without the need of any randomization. However, as shown in Section 5.1, the best decision policy provided by sequential probability ratio test requires random selection between $\mu = 2$ and $\mu = 3$.

5.3. Extension of Reinforcement Learning

To obtain the best policy by maximizing the expected reward per sample, consider, for example $1 \le \mu \le 5$, we have

μ	$t_\mu(0)$
1	1.0000
2	2.4390
3	3.7397
4	4.9985
5	6.2498

From Eq. (2), when $R_+ = 1$ and $R_- = -20$, we have

μ	expected reward/sample
1	-1.100
2	0.305
3	0.260
4	0.199
5	0.160

In this case, $\mu = 2$ is the best policy.

Alternatively, when $R_+ = 1$ and $R_- = -40$, we have

μ	expected reward/sample
1	-3.100
2	0.205
3	0.252
4	0.199
5	0.160

In this case, $\mu = 3$ is the best policy.

To obtain the best policy by maximizing the total expected future reward, from Eq. (5), when $R_+ = 1$, $R_- = -20$ and $c = 0.2$, we have

μ	V_0
1	-1.3000
2	0.2561
3	0.2233
4	-0.0029
5	-0.2503

In this case, $\mu = 2$ is the best policy.

Alternatively, when $R_+ = 1$, $R_- = -20$ and $c = 0.1$,

we have

μ	V_0
1	-1.2000
2	0.5000
3	0.5973
4	0.4970
5	0.3747

In this case, $\mu = 3$ is the best policy.

Thus to achieve the minimum detection delay with given α and β constraints, one has to use *randomized* reward, e.g., assigning $R_- = -20$ with probability p_0 and $R_- = -40$ with probability $1 - p_0$, to obtain the best policy by maximizing the expected reward per sample. Alternatively, one can use *randomized* penalty for asking one more observation, e.g., assigning $c = 0.2$ with probability p_0 and $c = 0.1$ with probability $1 - p_0$, to obtain the best policy by maximizing the total expected future reward. This is an unconventional extension of the reinforcement learning formalism to sequential decision problem.

Note that different R_+, R_- and c may lead to the same best policy. However, there is no general design guideline to assign randomized reward and penalty in order to achieve the desired performance level in terms of α and β. The above example can also be viewed as combining multiple classifiers to achieve better decision accuracy. In this case, a decision will be made when the difference between the outputs from different classifiers is either μ or $-\mu$. Using reinforcement learning, the average delay to reach a decision will be longer compared with the optimal sequential decision strategy for many desired performance levels. In the reinforcement learning formulation, we only considered to maximize the expected reward per sample or the total expected future reward. In fact, some other objectives can also be adopted [11], however, the relationship between the design parameters and the desired performance metrics deserves further research.

6. CONCLUDING REMARKS

In this paper, we studied the problem whether one can learn the optimal sequential decision rule with appropriately chosen reward or penalty for each action in a reinforcement learning framework. Using a binary sequential hypothesis testing example, we showed that certain policy can not be learned no matter how to manipulate the reward and penalty unless one extends the reinforcement learning formalism by allowing random reward. Thus when combining multiple classifiers to achieve better decision accuracy, one may have longer delay even when using the optimally designed reinforcement learning framework than that adopts the best combining method.

Acknowledgment

Stimulating discussions with Dr. Erik Blasch from Air Force Research Lab are gratefully acknowledged.

7. REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[2] M. L. Littman, *Algorithms for Sequential Decision Making*, Ph.D. dissertation, Brown University, RI, 1996.

[3] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic Programming*, Athena Scientific, Cambridge, MA, 1996.

[5] C. Guo and A. Kuh, "Temporal difference learning applied to sequential detection," *IEEE Transactions on Neural Networks*, vol. 8, pp. 278–287, 1997.

[6] D. MacKay, I. Murray, and P. Latham, "Solution of a toy problem by reinforcement learning," http://www.inference.phy.cam.ac.uk/mackay/RLcoin.pdf, March 2006.

[7] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, pp. 117–186, 1945.

[8] A. Wald, *Sequential Analysis*, Wiley and Sons, New York, NY, 1947.

[9] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, pp. 326–339, 1948.

[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, NY, 2nd edition, 1994.

[11] R. S. Sutton, "Open theoretical questions in reinforcement learning," in *Proceedings of the Fourth European Conference on Computational Learning Theory*. 1999, pp. 11–17, Springer-Verlag.