The convergence limit of the temporal difference learning

Ryosuke Nomura

the University of Tokyo

September 3, 2013

Outline

- **Reinforcement Learning**
- **Convergence limit**
- **Construction of the feature vector**
- **Numerical experiments**
- **Conclusion**
- **References**

Reinforcement Learning

Reinforcement learning is one of machine learning, which deals with a problem that an agent decides an optimal policy in an environment.

We consider a finite state space, and an agent gets a reward when he moves from a state to next state.

Our purpose is to evaluate an expectation of a cumulative reward and to find an optimal policy which maximizes or minimizes the reward.

Setting

S: finite state space

state sequence $(s_t)_{t=0,1,\ldots}$: Markov chain on S

s_0: follows some probability distribution on S

$P \in \mathcal{M}_{|S|}(\mathbb{R})$: transition probability matrix

has a stationary distribution $d(s)$

reward sequence $(r_t)_{t\in\mathbb{N}}$:

sequence of uniformly bounded random variables

$p(r_{t+1}|s_0, s_1, \ldots, s_{t+1}) = p(r_{t+1}|s_t, s_{t+1})$

$E[r_{t+1}|s_t = s]$ is independent of t

cumulative reward: $R_t = \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}$

where, γ: discount rate $(0 \leq \gamma \leq 1)$

Setting

value function:

$$V^*(s) = E\left[R_t | s_t = s\right] = E\left[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s\right]$$

Here, let

$$R(s) = E[r_{t+1} | s_t = s],$$

we have

$$V^* = R + \gamma P V^*.$$

We solve this problem when observations are given, and R and P are unknown.

Temporal difference learning

$\phi_k \in \mathbb{R}^{|S|}, k = 1, \ldots, K$: feature vectors
$\Phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}^{|S| \times K}$: feature matrix
$w_0 \in \mathbb{R}^K$: initial values of parameters
V_t: estimator of V^* defined as follows

$$V_t = \sum_{k=1}^{K} w_t(k)\phi_k = \Phi w_t$$

$w_t = (w_t(k))_{k=1,\ldots,K}$ is updated as the following rules:

$$\begin{cases} \delta_t = r_t + \gamma V_{t-1}(s_t) - V_{t-1}(s_{t-1}) \\ w_t = w_{t-1} + a_t \delta_t \phi(s_{t-1}) \end{cases}$$

where, $\phi(s_t) = (\phi_k(s_t))_{k=1,\ldots,K}$.

cf. $$0 = R + \gamma P V^* - V^*$$

Notation and Assumption

Notation:

$\boldsymbol{D} \in \boldsymbol{\mathcal{M}}_{|S|}(\mathbb{R})$:
diagonal matrix whose elements are $\boldsymbol{d(s)'s}$
$\tilde{\phi}_k(s) = \phi_k(s) - \gamma \sum_{s'} P(s, s')\phi_k(s')$
$\tilde{\Phi} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_K) = (I_{|S|} - \gamma P)\Phi \in \mathbb{R}^{|S| \times K}$

Assumption:

$\Phi^T D \tilde{\Phi}$: invertible

Convergence limit

Consider next rules:

$$\begin{aligned} w_t &= w_{t-1} + \alpha \left(\Phi^T D R - \Phi^T D \tilde{\Phi} w_{t-1} \right) \\ &= w_{t-1} + \alpha \Phi^T D \tilde{\Phi} (\hat{w} - w_{t-1}) \end{aligned}$$

where, $\hat{w} = (\Phi^T D \tilde{\Phi})^{-1} \Phi^T D R$.

Then, we have

$$w_t - \hat{w} = \left(I_K - \alpha \Phi^T D \tilde{\Phi} \right)^t (w_0 - \hat{w})$$

where, I_K **is a** $K \times K$ **identity matrix.**

Convergence limit

Theorem. Under Assumption, w_t converges to $\hat{w}$ for small enough $\alpha > 0$ as $t \to \infty$.

outline of proof.

Lemma 1. Under Assumption, every eigenvalue of $\Phi^T D\tilde{\Phi}$ has positive real part.

Lemma 2. There exists some positive number α such that the absolute value of every eigenvalue of $I_K - \alpha\Phi^T D\tilde{\Phi}$ is less than 1.

Motivation

The limit V_∞ of an estimator V_t is the form of

$$V_\infty = \Phi\hat{w} = \Phi(\Phi^T D\tilde{\Phi})^{-1}\Phi^T DR,$$

and if the true value V^* is expressed as linear combination of feature vectors, then the limit consists with the true value, but it is not true in general.

Then, I propose the construction of the feature vector related to this limit to converge the true value.

Property of the limit

Fact. The limit V_∞ satisfies the following equation:

$$V_\infty^T D(I - \gamma P)(V^* - V_\infty) = 0$$

proof.

$$\begin{aligned}
&V_\infty^T D(I - \gamma P)V_\infty \\
&= \hat{w}^T \Phi^T D(I - \gamma P)\Phi(\Phi^T D\tilde{\Phi})^{-1}\Phi^T DR \\
&= \hat{w}^T \Phi^T DR \\
&= V_\infty^T D(I - \gamma P)V^*
\end{aligned}$$

Construction of the feature vector

Here, consider the following algorithm:

1. **Let a limit $V_1 \neq 0$ be an initial vector.**
2. **Obtain $D(I - \gamma P)(V^* - V_1)$.**
3. **Obtain the limit V_2 for two feature vectors, V_1 and $D(I - \gamma P)(V^* - V_1)$.**
4. **Repeat 2 and 3.**

Then, the limit V_t converges to the true value V^* as $t \to \infty$.

Numerical experiments

Consider the model as follows:

$$\boldsymbol{R} = \begin{pmatrix} -2 \\ 6 \\ -6 \\ 4 \end{pmatrix}, \ \boldsymbol{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}, \ \gamma = 0.8$$

and we use an initial feature vector

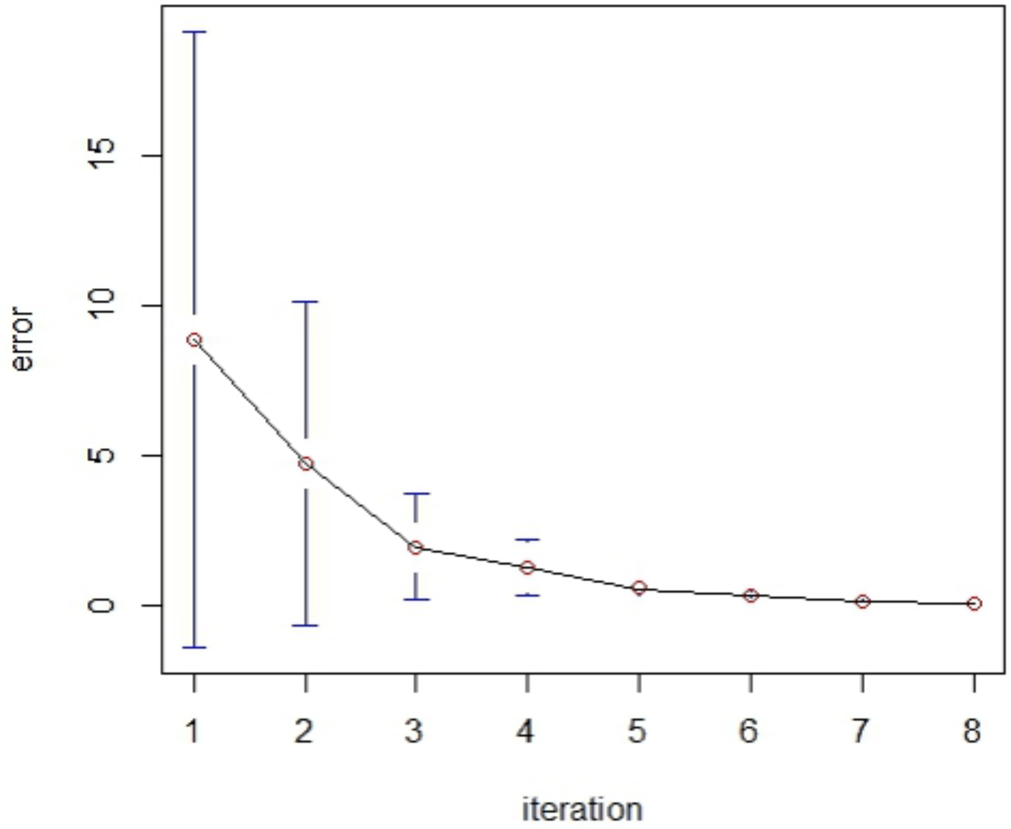
$$\phi^T = (1, 0, 0, 0).$$

Numerical experiments

Black line is a theoretical line and the value of y-axis is $||V_t - V^*||_2$.

Red points are obtained by 100 simulations, and 500 observations are given in each simulation.
The value of y-axis is $||V_t - \hat{V}||_2$, where $\hat{V} = (I - \gamma \hat{P})^{-1} R$ and $\hat{P}$ is an estimator of P.

Numerical experiments



Numerical experiments

Iteration	1	2	3	4
Error	8.86	4.72	1.95	1.27
S.D.	10.26	5.38	1.76	0.922

Iteration	5	6	7	8
Error	0.587	0.358	0.161	0.0965
S.D.	0.229	0.0996	0.0261	0.0106

Conclusion

When the true value is not expressed as linear combination of feature vectors, I constructed the feature vector based on the limit vector, and proposed the algorithm where the estimator converges to the true value.

References

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

[2] **R. S. Sutton, A. G. Barto (1998).** *Reinforcement Learning: An Introduction*. **MIT Press.**

[3] **T. Ueno, S. Maeda, M. Kawanabe, S. Ishii (2011). Generalized TD learning.** *Journal of Machine Learning Research*, **pages 1977–2020.**

[4] **H. Yu and D. P. Bertsekas (2006). Convergence results for some temporal difference methods based on least squares.** *Technical report, LIDS REPORT 2697*.