למידה במערכות דינמיות אביב תש"פ

תרגיל בית 4

מגישים:

יאיר נחום 034462796

דר ערבה 205874951

* 1. Let define the following two state MDP with zero rewards-

Choosing the linear subspace - , for some :

As requested,

* 1. The features in this case-

Thus,

* 1. The Bellman operator in this case-

The transition probabilities-

And-

* 1. The stationary distribution upholds-

We get-

* 1. The projection of into subspace is of form - where -

And upholds -

The Weighted Euclidean Norm for is -

Where -

We get-

After derivation and equaling to zero we get-

The projection operator -

After calculating with MATLAB -

* 1. The Bellman Operator applied on -

Projecting the result to gives us-

* 1. After 1st iteration -

After 2nd iteration -

Proceeding to the nth iteration (n being even)-

For-

diverges as

* 1. As hinted, we will initially analyze the behavior of -

From tutorial 9-

For , the Greedy policy w.r.t satisfies-

In our case - , and we get-

* 1. For : therefore the result from section **a** is-
  2. The TD(0) algorithm-

The TD(0) algorithm converges to the solution of the linear equation -

The online TD(h) algorithm will be described similarly -

And converges to the solution of the linear equation-

* 1. We shall calculate the estimation of based on the data-a sequence of N state pairs . We choose by distribution and receive by the transition probability

The projected h-step bellman equation-

The solution (minimizing the Euclidean norm)-

We recall -

The estimators are-

As seen in the result above, we would need to observe the rewards and state before using for estimation.

* 1. The error bound for policy evaluation in the 1-step case proved in lecture-

The error bound for policy evaluation in this case-

Since we get a tighter error bound for the h-step case.

For , as seen from result above, we get and-

Which is reached when *.*

However, online estimation is not possible for very large h, since observation time (observing the rewards and state before using for estimation) will massively increase the algorithm's running time.