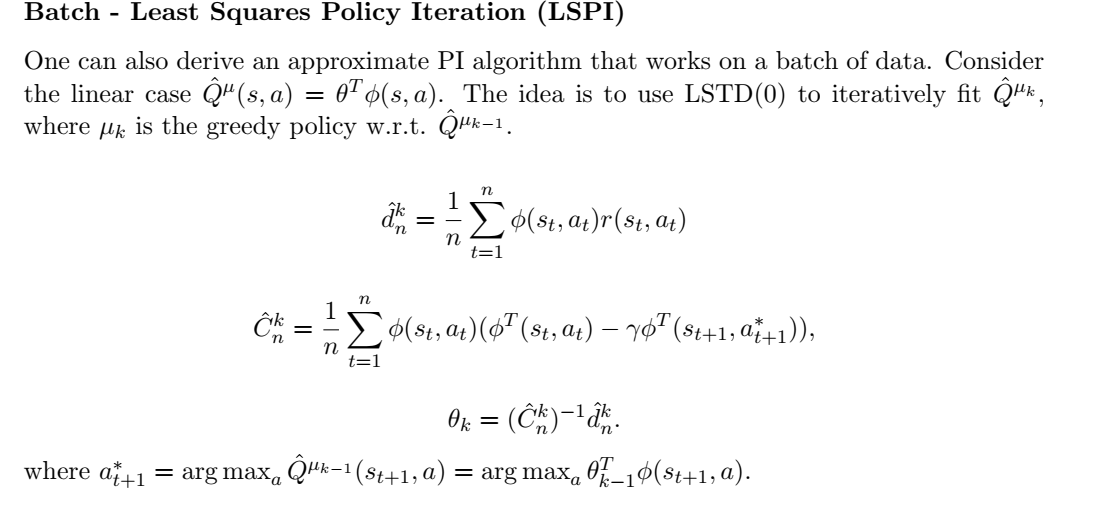
**3 – LSPI**

1. Since the given samples are not truncated at the terminal state and the environment doesn’t reset when reached to the terminal state (not online per sample but rather we collect the data offline and give it to the LSPI) we need to stop accumulating the samples when we reach the terminal state (given by the done state). And start a new data collection and a new LSPI iteration.
2. From program prints on mean and variance:  
   data mean [-3.00931294e-01 7.02421111e-05]  
   data std [0.51958566 0.04022475]  
   when the first component relates to the location, and the second to the speed.  
   It is latter on subtracted from the data to standardize the data (of the different components to be normalized)
3. The size of the weights vector is as the size of the features vector . In our case the size of the features vector (as we estimate Q) is   
   In our case, when we used 12x10 features per state, we get   
   If there is a bias coefficient added to the features vector per state, we have:
4. Will be implemented before and in the LSPI iterations loop to evaluate the success rate and build the average performance per iteration plot.  
   plot:
5. See implementation in code.

In class we saw we need to do the following per iteration:  
**Critic**:

* 1. convert the state-action space to features space.
  2. By using the current and features vector for the next state in each tuple of (state, action, reward, next state) estimate for every action and apply the greedy policy over it to find the next best action (current optimal estimated Q according to current policy).
  3. Accumulate according to the below formulas the vector and matrix (From the lecture notes):  
     
  4. Apply inverse on matrix A and multiple with vector to get the new estimation of coefficients (new estimate)

**Actor**:

* 1. Apply greedy bellman operator on the new estimation and get a new greedy (or maybe close to greedy) policy.