How to win the battle against Glossy Buckthorn using RL

Problem Definition

- Having the population and the seed bank in a 9 cell environment (a 3×3 grid map), we are looking for optimal policy
- No model of the system/environment is available, only data!
- Using methods like LSTD-Q, we can learn the model and approximate the state-action value function
- Using methods like LSPI, we can learn the optimal policy

Background

- LSPI has two steps:
  - Policy evaluation $Q^\pi = R + \gamma P Q^\pi$
  - Policy improvement $\pi(s) = \mathrm{argmax}_{a \in A} Q^\pi(s,a).$
- Based on tabular representation
- Alternative: approximation

$$\hat{Q}(s,a) = \sum_{j=1}^{k} \phi_j(s,a) w_j$$

- LSTDQ is used to calculate w_j
- Policy is not explicit anymore
- Good set of features matters!

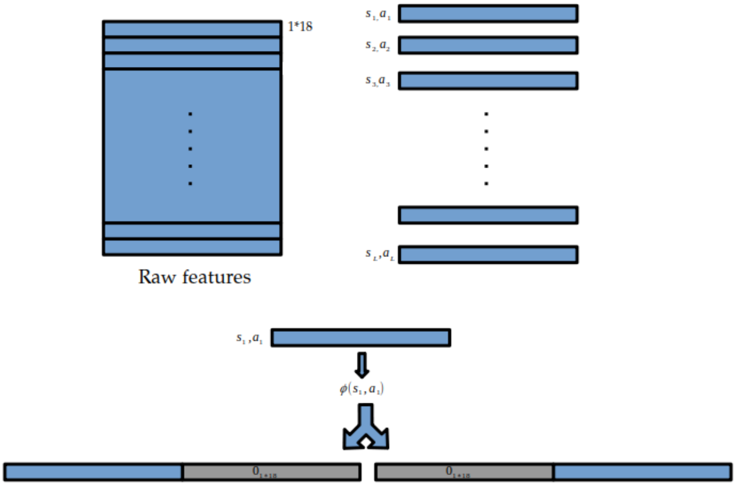
LSTDQ

- Based on TD
- w^π is calculated
- Easy formulation:

$$w^\pi = (\Phi^T(\Phi - \gamma P\Phi))^{-1}\Phi^T R$$

- But, what's Φ?

Creating $\phi(s, a)$



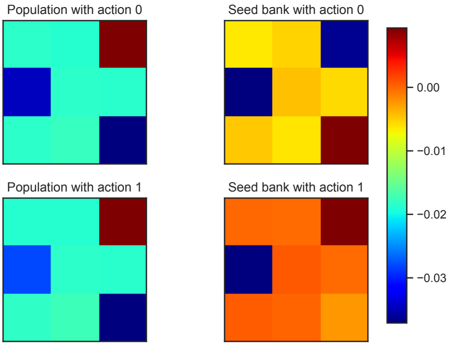
Evaluation

- Two tests:
  - The bigger set to train the system, and small set to test
  - The bigger set used as the input for a k-fold cross validation, choosing k, and test on the small data set
- Bellman Error is used for evaluation

$$\begin{aligned} BE^{\pi} &= (TQ^{\pi}) - Q^{\pi} \\ &= R + \gamma\phi(s', a')w - \phi(s, a)w \end{aligned} \tag{1}$$

Method	BE
LSTDQ	4935580.30
LSTDQ with 4-fold	4905981.01
LSTDQ with 5-fold	4601911.70
LSTDQ with 6-fold	4919249.69
LSTDQ with 10-fold	4886872.99

Results


Population with action 0
Seed bank with action 0
Population with action 1
Seed bank with action 1
0.00
−0.01
−0.02
−0.03