FPT Education
FPT UNIVERSITY

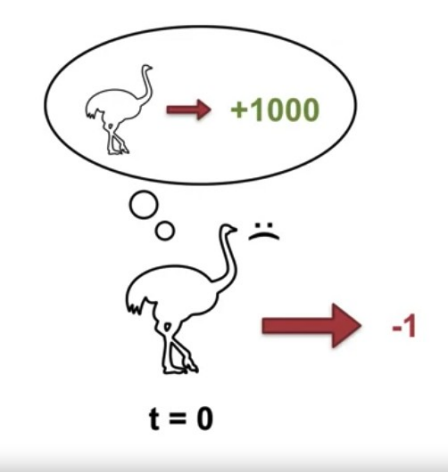
Exploration under Function Approximation

Objectives

- Describe how optimistic initial values can be used with function approximation
- Understand how Epsilon greedy can be used with function approximation

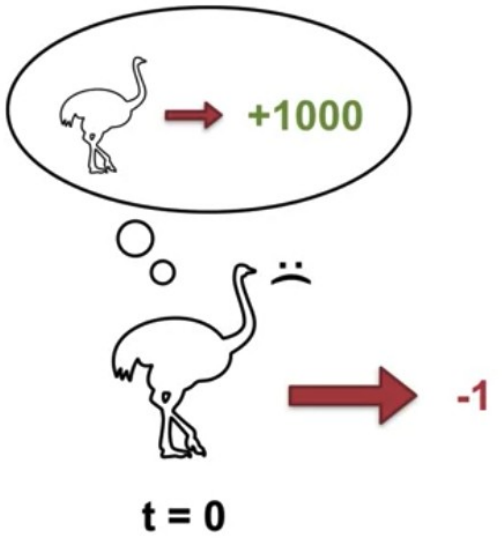
Optimistic Initial Values in the Tabular

- We initialize our values to be greater than the true values.
- This is like the agent imagining that it can get more reward by taking that action than it actually car



Optimistic Initial Values in the Tabular

- This is straightforward to implement in a tabular setting where the update to each state action pair is independent of all the other state action pairs



$$Q(s,a) \leftarrow 1000 \quad \forall s,a$$

FPT Education
FPT UNIVERSITY

Optimistic Initial Values in the Tabular

- How to initialize values optimistically under function approximation:
  - In function approximation, optimistic initial values corresponds to initializing the weights such that the resulting values are optimistic.
  - In some cases this is straightforward, for example, when the features are binary, we simply initialize each weight to be the largest possible return.
  - Then, as long as each state has at least one feature active, the value will be optimistic and likely overly so.

Optimistic Initial Values in the Tabular

- How to initialize values optimistically under function approximation:
  - Non-linear: the relationship between the final values and the features can be quite complicated.

Linear

$$q_\pi(s,a) \approx \hat{q}(s,a,\mathbf{w}) = \mathbf{w}^T\mathbf{x}(s,a)$$

$$\mathbf{w} \leftarrow \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ \vdots \end{bmatrix}$$

Non-linear

$$q_\pi(s,a) \approx \hat{q}(s,a,\mathbf{w}) = \mathbf{NN}(s,a,\mathbf{w})$$

$$\mathbf{w} \leftarrow ???$$

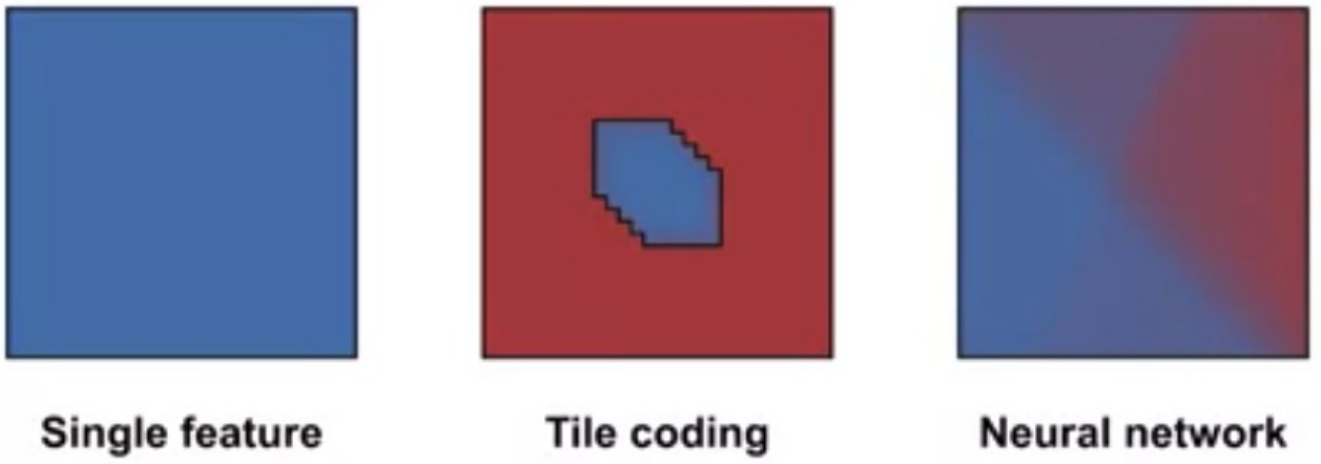
FPT Education
FPT UNIVERSITY

Optimistic Initial Values in the Tabular

- Depending on how our features generalize, optimistic initial values may not result in the same kind of systematic exploration we see in the tabular case.
- Consider an extreme example, where we have only one feature that is always one.
  - We can initialize optimistically but every update will change the value for all states.
  - This means that before some states are even visited, the value will already have decreased such that it is no longer optimistic.

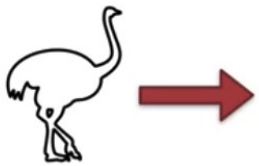
Optimistic Initial Values in the Tabular

- Function approximation with tile coding can produce such localized updates.
- Neural networks and also provide local updates, but neural networks may also generalize ac



Epsilon Greedy

- Epsilon greedy is generally applicable and easy to use even in cases with non-linear function approximation.
- The only thing Epsilon greedy needs are the action value estimates, $1 - \epsilon$ they are initialized or appr



$$A_t = \underset{a}{\text{argmax}}\ \hat{q}(S_t, a, \mathbf{w})$$

Epsilon Greedy

- However, Epsilon greedy is not a directed exploration method.
- It relies on randomness to discover better actions near states followed by the current policy.
- It is therefore not as systematic as exploration methods that rely on optimism.

Epsilon Greedy

ε-greedy

$1 - \epsilon$ ϵ

$A_t = \underset{a}{\text{argmax}}\ \hat{q}(S_t, a, \mathbf{w})$ $A_t = \text{Random action}$

Summary

- Describe how optimistic initial values can be used with function approximation
- Understand how Epsilon greedy can be used with function approximation

FPT Education
FPT UNIVERSITY

Q & A