Assignment 1

Reinforcement Learning, WS22

Team Members		
Last name	First name	Matriculation Number
Amering	Richard	1331945
Michael	Mitterlindner	11824770

1 Task - Proof that iterative Valuefunction evaluation will converge

1.1 "Proof"

Start with the regular update rule that requires iteration over states s and subsequent states s':

$$V_{new}(s) \leftarrow r(s) + \gamma \sum_{s'} p(s'|s) V_{old}(s')$$

This equation can be rewritten in matrix-vector form to get rid of the iteration over states and subsequent states:

$$\mathbf{V(s)} \leftarrow \mathbf{r} + \gamma P_{s,s'} \cdot \mathbf{V(s')}$$

The regular scalar quantities have been replaced with their vector counterparts. **s** is a vector comprising the entire state space (all possible states). $\mathbf{V(s)}$ is the vector valued value-function, simply giving a value to each state. $P_{s,s'}$ is the state transition probability matrix, which is a square matrix collecting probabilities $p(s'|s)$ of all possible state transitions in the state space. Its first row contains $p(s'_1|s_1), p(s'_2|s_1)..p(s'_m|s_1)$, where m is the size of the state space, or the number of possible states. This pattern continues with $p(s'_2|s_1), p(s'_2|s_2)..$ in the consecutive rows. The sum of each row is equal to one, which means in other words that it is guranteed to have a subsequent state to each state (the state space can not be left). The subsequent state to any state can be the original state also, meaning that a state transition leads back to the state.

Since $\mathbf{V(s')}$ is the value function of the entire state space, it is actually the same as $\mathbf{V(s)}$. We can therefore rewrite the equation like in the following:

$$\mathbf{V(s)} \leftarrow \mathbf{r} + \gamma P_{s,s'} \cdot \mathbf{V(s)}$$

We now have to prove only that iterative multiplication of γ times the state transition probability matrix followed by addition of the reward-vector gives a contraction. This can be simply seen by the fact that the value of each entry in the state transition probability matrix is in the interval $[0, 1]$ and also gamma is in the interval $[0, 1]$. Iterative multiplication of this matrix with the value-function will therefore move any value-function-vector towards the origin, since each component of the vector is scaled down by a factor in the probability matrix. This implies that any two subsequent value-function vectors will move closer together upon iterative application of the probability matrix. Therefore, the mapping happening by the state transition matrix is L-Lipschitz, with L < 1:

$$||P_{s,s'}(V(s)_n) - P_{s,s'}(V(s)_{n+1})|| \leq L||V(s)_n) - V(s)_{n+1}||$$

The addition of the reward vector after each matrix vector multiplication does not change this, since it is only a parallel offset that acts on any point in the value-function space in the same way. It does not affect the distance between two points.

From this it can be seen that the iterative application of the state transition probability matrix, followed by addition of the reward-vector gives a contraction.