Ex11

Ex11.1

- Q: S x A → R
- r: S x A → R
- Q-Learning update formula, etc.
- T, r don‘t need to be known for Q-Learning
- It learns that implicitly
- Formal definition != implementation details

- What is S?, A?

Overview

Grp1	OK, r unnecessary	1
Grp3	Q-Learning OK, S/A missing	0.5
Grp4	OK, better formally: S=… A=…	1
Grp5	Informal, no S,A	0
Grp6	OK, better formally: S=… A=…	1
Grp8	OK, better formally: S=… A=…	1
Grp9	OK!	1

Ex11.1

- Don‘t assume you can‘t place at „_“

Initialise $Q(s,a)$ arbitrary for all $s \in S$ and $a \in A$

Repeat

select best action a_t *with the greedy policy:*

$$a_t = \pi(s_t) = \underset{a \in A}{argmax}\, Q(s_t, a)$$

apply a_t *in the world and observe* s_{t+1} and immediate reward r_t*:*

$$s_t \rightarrow s_{t+1}$$

$$r_t$$

adapt the value function for state s_t

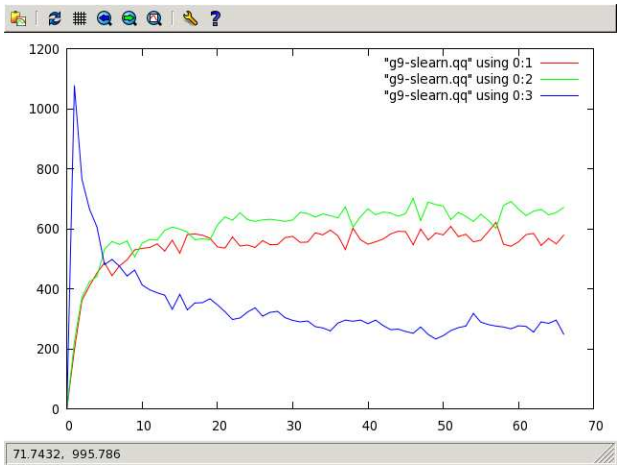
$$Q_{k+1}(s_t, a_t) := (1-\alpha)\, Q_k(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a \in A} Q_k(s_{t+1}, a \quad) \right]$$

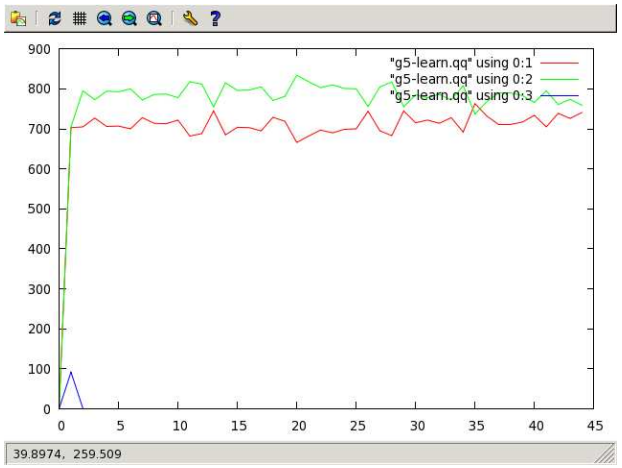
Until (Q_{k+1}-$Q_k < \varepsilon$) or (s is terminal)

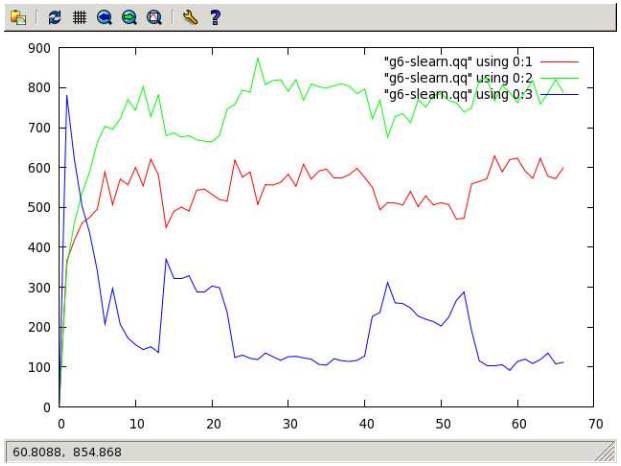
Overview

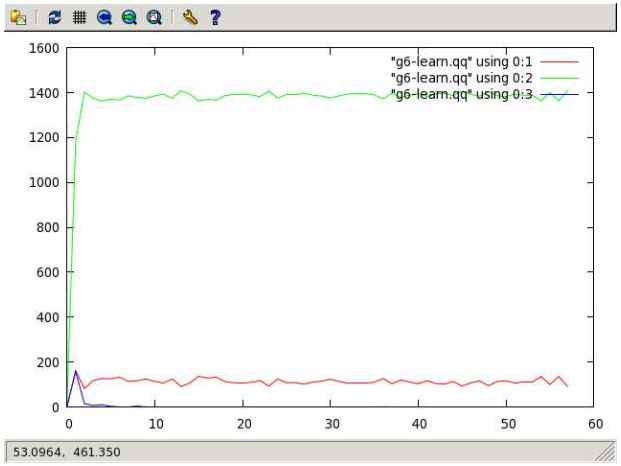
Grp1	Assumes transition model, updates the future G:learns at beginning S:no learning	1
Grp3	Bug in random moves (perhaps works on win) G:OK, S:still good	2
Grp4	e-Greedy + e-traces (we assume) G:good, S: ok, some „bumps“	2
Grp5	Only actions on „_“ G: ok, S:quick	2
Grp6	Only actions on „_“ G: quick, S:some „bumps“, but ok	2
Grp9	e-Greedy G:quick, S:still good	2

Learning Graphs









Next task

- Ant system
- Ex 12.1 = formulate!
- Give formulae for pheromone dropping, etc.
- Formally!