Aalto University
School of Electrical Engineering

ELEC-E8125 Reinforcement Learning
Reinforcement learning in discrete domains

Ville Kyrki

8.10.2019

Today

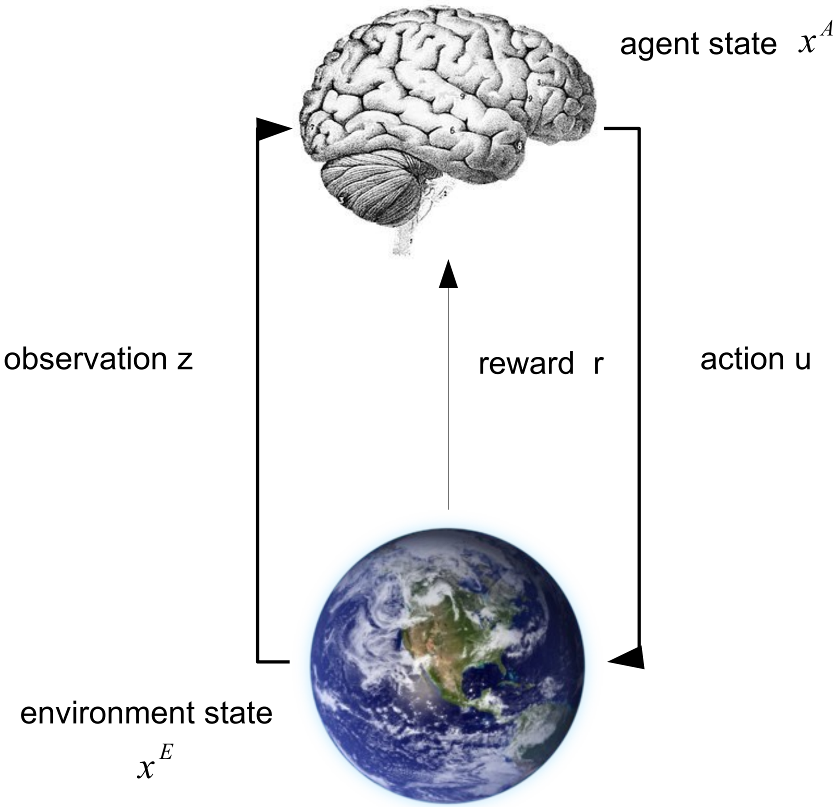
- Reinforcement learning
- Policy evaluation vs control problems
- Monte-Carlo and Temporal difference

Learning goals

- Understand basic concepts of RL.
- Understand Monte-Carlo and temporal difference approaches for policy evaluation and control.

- Be able to implement MC and TD.

Reinforcement learning



RL
MDP with **<u>unknown</u>** Markovian dynamics
$P(x_{t+1}|x_t, u_t)$

Unknown reward function
$r_t = r(x_{t+1}, x_t)$

Solution similar, e.g.
$u^*_{1,\ldots,T} = max_{u_1,\ldots,u_T} \sum_{t=1}^{T} r_t$

Learning must **<u>explore</u>** policies

Aalto University
School of Electrical Engineering

How does this then relate to the control case?

Reinforcement learning

- MDP with unknown dynamics (*T*) and reward function (*r*)

figure out the dynamics

- Model based RL: Estimate MDP, apply MDP methods.
  - Estimate MDP transition and reward functions from data.

not efficient

- Can we do without *T* and *r*?
  - Can we evaluate a policy (construct value function) if we have multiple episodes (in episodic tasks) available?

Monte-Carlo policy evaluation

only for episodic tasks, which has end.

- Complete episodes give us samples of return *R*.
- Learn value of particular policy from episodes under that policy.

$$V_\pi(x) = E_\pi\left[R_t \mid x_t = x\right] \qquad R_t = \sum_{k=0}^{H} \gamma^k r_{t+k+1}$$

- Estimate value as empirical mean return.
  - Each time state *s* visited in an episode,

$$N(x) = N(x) + 1 \quad S(x) = S(x) + R_t \quad V(x) = S(x) / N(x)$$

- When number of episodes approaches infinity, *V(x)* converges $V(x) \rightarrow V_\pi(x)$

Empirical mean approaches true mean.

Every-visit vs first-visit, incremental and running mean

for situation where memory is not enough

- First-visit version
  - Instead of every "visit" of state *s*, only update *N(x)* and *S(x)* on first visit per episode.
  - Both approaches converge to $V_\pi(x)$.

- *S(x)* does not need to be stored

$$V(x) = V(x) + \frac{1}{N(x)}\left(R_t - V(x)\right)$$

Check this

- We can also track a running mean

Why?

$$V(x) = (1-\alpha)V(x) + \alpha R_t = V(x) + \alpha\left(R_t - V(x)\right)$$

more attention to the recent data

Can we do without episodes?

Temporal difference (TD) – learning without episodes

infinite sequence

- For each state transition, update a guess towards a guess:

explicitely make the markovian assumption

$$V(x_t) = V(x_t) + \alpha \left(\underline{r_{t+1} + \gamma V(x_{t+1})} - V(x_t) \right)$$

- Approach called TD(0)

imminidiate reward

Estimated return.

- Compare to MC

$$V(x_t) = V(x_t) + \alpha \left(R_t - V(x_t) \right)$$

True return.

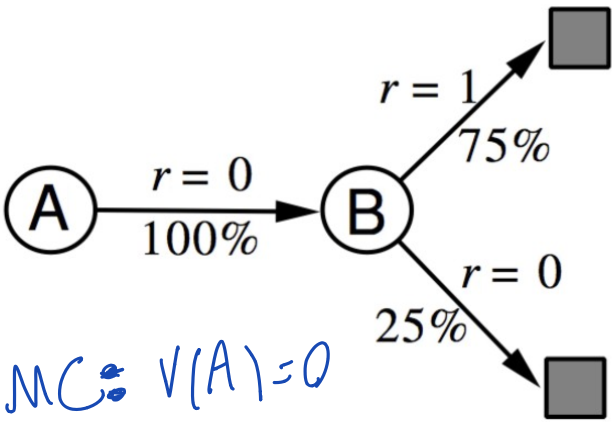
for system which aren't Markovian, MC is better approach.

What if we have limited data?

Batch learning

- For limited number of trials available:
  - Sample episode *k.*
  - Apply MC or TD(0) to episode *k.*

A, 0, B, 0
B, 1
B, 1
B, 1
B, 1
B, 1
B, 1
B, 0



$V(B) = \frac{0+1+1+1+1+1+1+0}{8} = \frac{6}{8} = \frac{3}{4}$

MC: $V(A) = 0$

TD(0): $V(A) = 0 + \alpha\left(0 + \gamma \times \frac{3}{4} - 0\right)$
$= 0.75$

What is *V(A)*?

Beware small sample numbers.

MC vs TD

- MC
  - Needs full episodes. Only works in episodic environments.
  - High variance, zero bias → good but slow convergence.
  - Does not exploit Markov property → often better in non-Markov env.

- TD (esp. TD(0))
  - Can learn from incomplete episodes and on-line after each step.
  - Works in continuing environments.
  - Low variance, some bias → often more efficient than MC, discrete state TD(0) converges, more sensitive to initial value.
  - Exploits Markov property → often more efficient in Markov env.

Can we combine these / find intermediate solution?

More efficient than MC, less sensitive to Markovian property compared to TD(0)

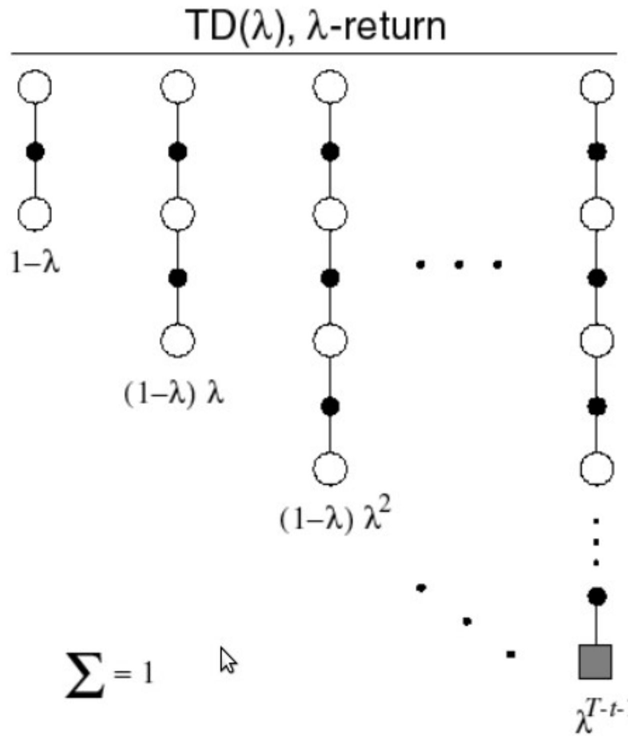
λ-return

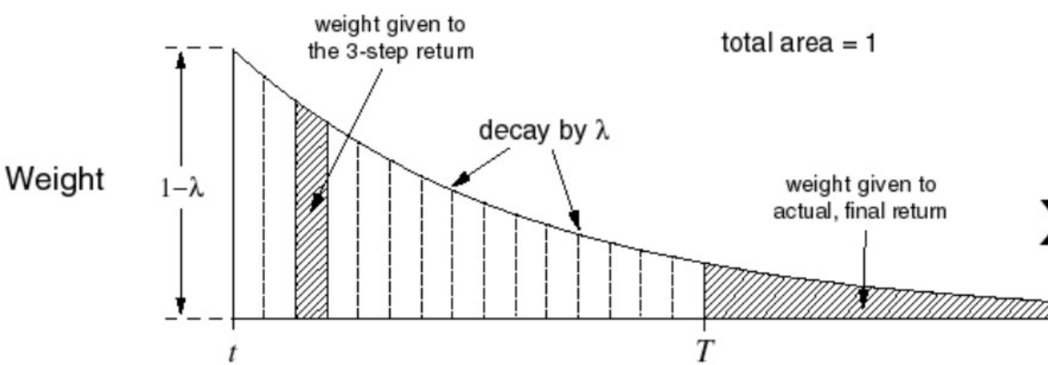
k-step return: $R_t^{(k)} = \sum_{i=1}^{k} \gamma^{i-1} r_{t+i} + \gamma^k V(x_{t+k})$

- Combine returns in different horizons.

$$R_t^{\lambda} = (1-\lambda) \sum_{k=1}^{\infty} \lambda^{k-1} R_t^{(k)}$$

$$V(x_t) = V(x_t) + \alpha \left(R_t^{\lambda} - V(x_t) \right)$$





Requires complete episodes!

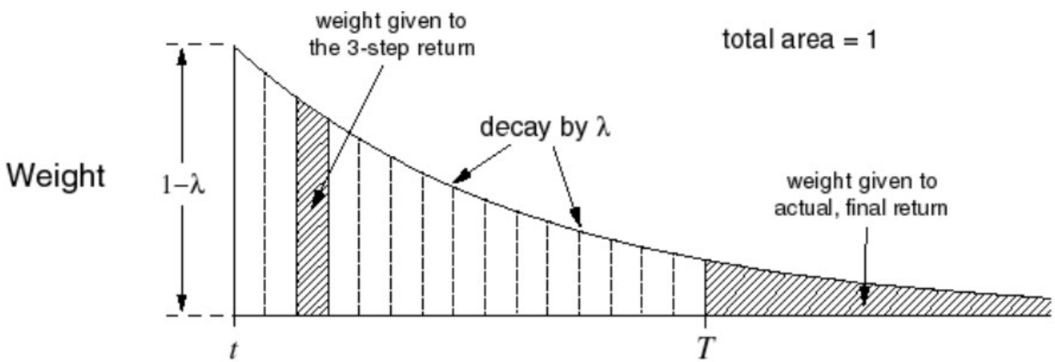
Causes and effects – eligibility traces

- Which state is the “cause” of a reward?
- Frequency heuristic: most frequent states likely.
- Recency heuristic: most recent states likely.
- *Eligibility trace* for a state combines these:

$$E_t(x) = \gamma \lambda E_{t-1}(x) + \mathbf{1}(x_t = x)$$

$\lambda = 0 \rightarrow TD(0)$

$\lambda = 1 \rightarrow MC$



Backward-TD(λ)

- Extend TD time horizon with decay (λ).

- After episode, update

$$V(x) = V(x) + \alpha E_t(x)\left(r_{t+1} + \gamma V(x_{t+1}) - V(x_t)\right)$$

- TD(1) equal to MC.

What if $\lambda = 0$

$$E_t(x) = \gamma \lambda E_{t-1}(x) + \mathbf{1}(x_t = x)$$

- Eligibility traces way to implement *backward* TD(λ), *forward* TD(λ) requires episodes.

Slightly different in on-line case.

Control / decision making?

- So far we only found out how to estimate value functions for a particular policy.

- Can we use this to optimize a policy?

Monte-Carlo Policy iteration

- Can we implement greedy policy improvement as in previous lecture?

$$\pi'(x) = arg\,max_u \sum_{x'} \underline{T(x,u,x')} \left(\underline{r(x,u,x')} + \gamma V(x') \right)$$

- Greedy policy improvement using action-value function *Q(x,u)* does not require model.

$$\pi'(x) = arg\,max_u Q(x,u)$$

- Estimate *Q(x,u)* using MC (empirical mean).

Exploration-exploitation trade-off:
How can we ensure that we try different actions?

Ensuring exploration

- Simple approach: ε-greedy exploration:
  - Explore: Choose action at random with probability ε.
  - Exploit: Be greedy with probability 1-ε.

$$\pi(u|x) = \begin{cases} \epsilon/m + 1 - \epsilon & \text{if } u = \arg\max_{u'} Q(x, u') \\ \epsilon/m & \text{otherwise} \end{cases}$$

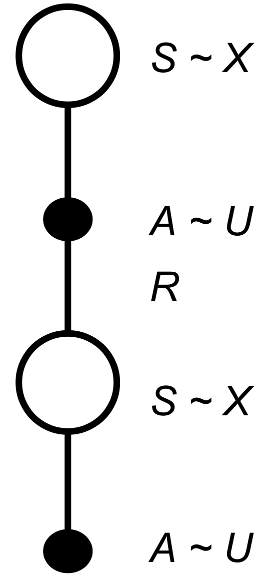
- How to converge to optimal policy?
  - Idea: reduce ε over time.
  - For example, for *k*:th episode $\epsilon = \frac{a}{a+k}$ "Greedy in Limit with Infinite Exploration" (GLIE)

constant

Can we use TD instead of MC for control?

SARSA (XURXU ☺)

- Idea: Apply TD to *Q(X,U)*.
  - With ε-greedy policy improvement.
  - Update each time step.



$$Q(x,u) = Q(x,u) + \alpha\left(r + \gamma Q(x',u') - Q(x,u)\right)$$

Compare with

$$V(x_t) = V(x_t) + \alpha\left(r_{t+1} + \gamma V(x_{t+1}) - V(x_t)\right)$$

- SARSA converges under
  - GLIE policy,
  - $\sum_{t=0}^{\infty} \alpha_t = \infty \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty$

For example $\alpha_t = 1/t$

SARSA(λ)

- Instead of TD(0) update in SARSA, use TD(λ) update.
- Backward SARSA(λ)

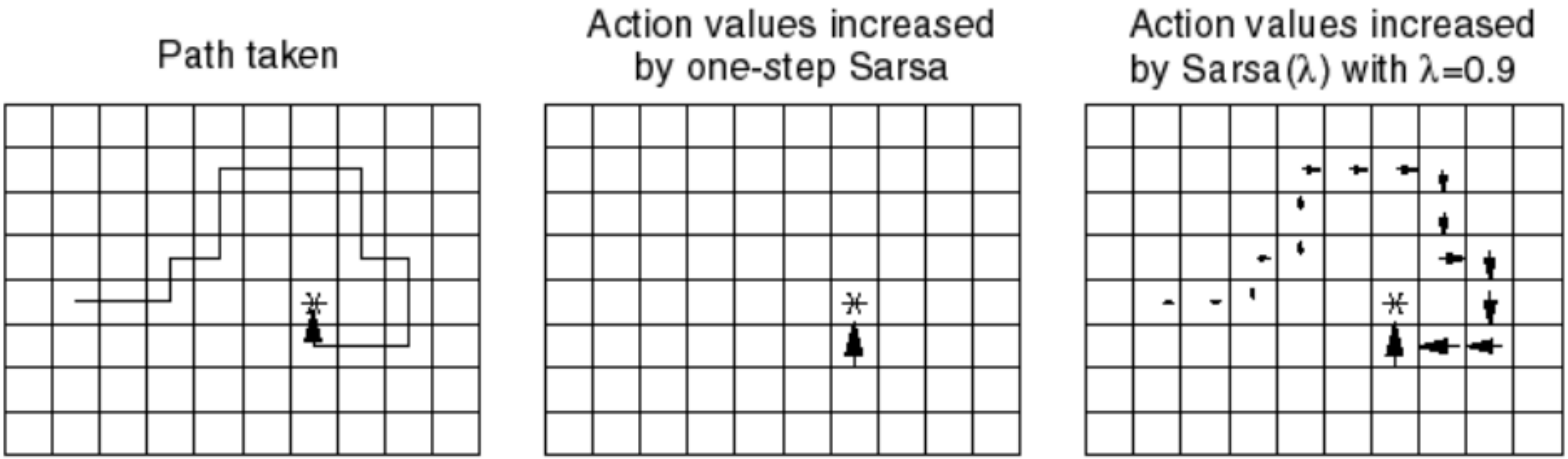
$$E_t(x,u)=\gamma\lambda E_{t-1}(x,u)+\mathbf{1}(x_t=x,u_t=u)$$

$$Q(x,u)=Q(x,u)+\alpha E_t(x,u)\left(r_{t+1}+\gamma Q(x_{t+1},u_{t+1})-Q(x_t,u_t)\right)$$

Compare to

$$E_t(x)=\gamma\lambda E_{t-1}(x)+\mathbf{1}(x_t=x)$$

$$V(x)=V(x)+\alpha E_t(x)\left(r_{t+1}+\gamma V(x_{t+1})-V(x_t)\right)$$


Path taken
Action values increased by one-step Sarsa
Action values increased by Sarsa(λ) with λ=0.9

On-policy vs off-policy learning

- *On-policy learning* (methods so far)
  - Use a policy while learning how to optimize it.
  - “Learn on the job”.

- *Off-policy learning*
  - Use another policy while learning about optimal policy.
  - Can learn from observation of other agents.
  - Can learn about optimal policy when using exploratory policy.

Q-learning

- Use ε-greedy *behavior policy* to choose actions.
- *Target policy* is greedy with respect to *Q*.

$$\pi(x) = arg\,max_u Q(x,u)$$

- Update target policy greedily:

$$Q(x,u) = Q(x,u) + \alpha\left(r + \gamma\, max_{u'} Q(x',u') - Q(x,u)\right)$$

- *Q* converges to *Q**.

Assume we take greedy action at next step.

Summary

- In reinforcement learning, dynamics and reward function of MDP are unknown.

- MC approaches sample returns from full episodes.
- TD approaches sample estimated returns (biased).

Next: Extending state spaces

- What to do if
  - discrete state space is too large?
  - state space is continuous?

- Readings
  - Sutton & Barto, ch. 9-9.3, 10-10.1