Robotics
Exercise 10

Marc Toussaint
Lecturer: Peter Englert
TAs: Matt Bernstein, Danny Driess, Hung Ngo
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

January 24, 2017

1 The value function in Markov Decision Processes

On slide 5 of the RL lecture we defined MDPs as

$$P(s_{0:T+1}, a_{0:T}, r_{0:T}; \pi) = P(s_0) \prod_{t=0}^{T} P(a_t|s_t; \pi)\ P(r_t|s_t, a_t)\ P(s_{t+1}|s_t, a_t)\ , \tag{1}$$

where $P(a_0|s_0; \pi)$ described the agent's policy. We assume a deterministic agent and write $a_t = \pi(s_t)$. The *value* of a state s is defined as the expected discounted sum of future rewards,

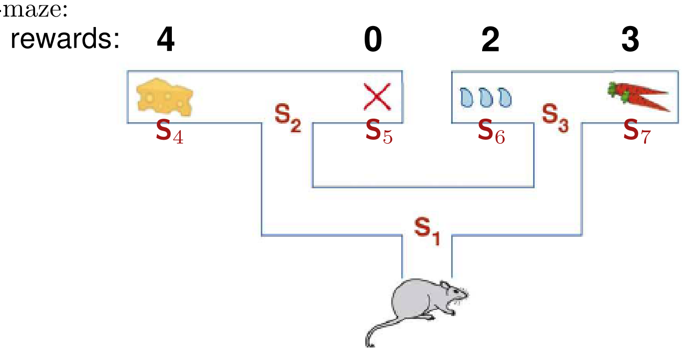
$$V^\pi(s) = \mathrm{E}\{r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots \mid s_0 = s\} \tag{2}$$

given that the agent starts in state s and from there executes the policy π.

a) Prove

$$V^\pi(s) = \mathrm{E}\{r_0 \mid s, \pi(s)\}\ +\ \gamma \textstyle\sum_{s'} P(s' \mid s, \pi(s))\ V^\pi(s')\ . \tag{3}$$

b) Consider the following T-maze:



We distinguish 7 states $s_1, .., s_7$ in the maze. The first 3 states are the T-junctions; the last 4 states receive rewards $(4, 0, 2, 3)$. At each T-junction we have two possible actions: left, right. Everything is deterministic. Assume a discounting $\gamma = 0.5$.

Compute (by hand) the value function V^π over all states when π is the *random policy* (50/50 left/right at each junction).

c) Bellman's principle of optimality says that the optimal policy π^* has a value function $V^{\pi^*}(s) = V^*(s)$,

$$V^*(s) = \max_a \Big[\mathrm{E}\{r_0 \mid s, a\}\ +\ \gamma \textstyle\sum_{s'} P(s' \mid s, a)\ V^\pi(s')\Big]\ . \tag{4}$$

Compute (by hand) the optimal value function V^* over all states for the example above.

d) Now consider continuous state s and action a. Let the policy be stochastic and linear in features $\phi(s) \in \mathbb{R}^k$, that is,

$$\pi(a|s; \beta) = \mathcal{N}(a|\phi(s)^\top \beta, \phi(s)^\top \Sigma \phi(s))\ . \tag{5}$$

The covariance matrix $\phi(s)^\top \Sigma \phi(s)$ describes that each action $a_t = \phi(s_t)^\top(\beta + \epsilon_t)$ was generated by adding a noise term $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ to the parameter β. We always start in the same state $\widehat{s}$ and the value $V^\pi(s_0)$ is

$$V^\pi(\widehat{s}) = \mathrm{E}\{\sum_{t=0} H r_t \mid s_0 = \widehat{s}\} \tag{6}$$

(no discounting, but only a finite horizon H.)

Optimality now requires that $\frac{\partial V^{\pi(\beta)}}{\partial \beta} = 0$. Assume that $a \in \mathcal{R}$ is just 1-dimensional and $\Sigma \in \mathbb{R}$ just a number. Try to prove (see slide 18) that we can derive

$$\beta^* = \beta^{\text{old}} + \Big[\mathrm{E}_{\xi|\beta}\{\sum_{t=0}^H W(s_t) Q^{\pi(\beta)}(s_t, a_t, t)\}\Big]^{-1} \mathrm{E}_{\xi|\beta}\{\sum_{t=0}^H W(s_t) e_t Q^{\pi(\beta)}(s_t, a_t, t)\} \ , \quad W(s) = \phi(s)(\phi(s)^\top \Sigma \phi(s))^{-1} \phi(s)^\top$$

from $\frac{\partial V^{\pi(\beta^*)}}{\partial \beta} = 0$. (In the scalar case, $W(s) = 1$.) As a first step, derive

$$\frac{\partial}{\partial \beta} \log \pi(a|s)$$

Then insert $a_t = \phi(s_t)^\top(\beta^{\text{old}} + \epsilon_t)$ and solve

$$\mathrm{E}_{\xi|\beta}\{\sum_{t=0}^H \frac{\partial}{\partial \beta} \log \pi(a_t|s_t) \ Q(s_t, a_t, t)\} = 0$$

for β. This shows how you can get the optimal policy parameters β^* based on samples generated with β.