AUP - Theory

Fabrice Beaumont

6. Mai 2023

Inhaltsverzeichnis

1 Unnamed

Consider environments as discounted Markov Decision Processes (MDPs) $(S, \mathcal{A}, r, p, \gamma)$:

- S - set of states
  s'_t - baseline state at time t
- $\mathcal{A}$ - set of actions
  $a^{\text{noop}} \in \mathcal{A}$ - special no-operation action
- $r : S \times \mathcal{A} \to \mathbb{R}$ - reward function
- $p : S \times S \times \mathcal{A} \to [0, 1]$ - transition function
- $\gamma \in (0, 1)$ - discount factor (sometimes written as γ_r - reachability discount factor w.r.t. reward function r)
- $d : S \times \{s'_t\} \to \mathbb{R}$ - deviation
- $\beta \in \mathbb{R}$ - **deviation penalty (learned parameter)**

At time step t, the agent receives the current state s_t, outputs the action a_t drawn from its policy $\pi(a_t|s_t)$ and receives reward $r(s_t, a_t)$.

DEF: Intrinsic pseudo-reward

By adding a penalty for impacting the environment to the reward function we can implement an intrinsic pseudo-reward. Therefore, we subtract at time t an impact penalty, which is a scaled deviation penalty from the deviation of the current state from the baseline state s'_t.

$$r_\beta(s_t, a_t) := r(s_t, a_t) - \beta \ \cdot \ d(s_{t+1}, s'_{t+1})$$

DEF: Inaction rollout

An **inaction rollout** from state s_t is a sequence of states obtained by following the inaction policy (a^{noop}) starting from that state. Thus state $s_{t+2}^{(t)}$ denotes the state at time step $t+2$, after arriving at state s_t at time step t and performing the no-operation action for two time steps.
This allows for an easy comparison to environment state after choosing the no-operation action at every time step starting from the baseline state: $s_{t+2}^{\prime(t)}$

DEF: Reachability

Let $\gamma_r \in (0,1]$. We define a **reachability** $R : S \times S \to [0,1]$ to get from state x to state y ($x \neq y$).
TODO: Kapitel 2.2. - erste Gleichung - Ist der Erwartungswert von der Potenz des reachability discount factors gemeint? Mit Ergebnis 1 falls $N_\pi(x,y)$ endlich und null sonst? Oder ist etwas anderes gemeint (value function).
TODO: Vermutung: Es sollte $R(x,y) := \max_\pi \gamma_r^{N_\pi(x,y)} \mathbb{E}[y]$. Begruengung: Siehe rekursive Formel: ist γ_r mal Erwartungswert des naechsten States z und erwarte $N_\pi(x,y)$ rekursive Aufrufe bis zum Ziel.
Its is defined as follows:

$$
\begin{aligned}
R(x,y) &:= \gamma_r \max_a \sum_{z \in S} p(z|\, x,a) R(z,y) \\
&\stackrel{?}{=} \gamma_r^n \max_{a_1} \sum_{z_1 \in S} p(z_1|\, x,a_1) \Big(\max_{a_2} \sum_{z_2 \in S} p(z_2|\, z_1,a_2) \ldots \max_{a_n} \sum_{z_n \in S} (0 + p(y|\, z_n,a_n) * 1) \Big)
\end{aligned}
$$

where $n = N_\pi(x,y)$. And it is $R(y,y) = 1$.

Special case: **Undiscounted reachability** ($\gamma_r = 1$), which computes whether y is reachable in any number of steps. In this cased it is (see paper for proof):

$$R(x,y) = \max_\pi \mathbb{P}\big(N_\pi(x,y) < \infty\big)$$

TODO: Wie passt das zu: $R(x,y) := \max_\pi \mathbb{E}\gamma_r^{N_\pi(x,y)} = \max_\pi \mathbb{E}1$?

The **unreachability** (UR) **deviation measure** $d_{\text{UR}} : S \times S \to [0,1]$ is then defined as:

$$d_{\text{UR}}(x,y) := 1 - R(x,y)$$

$d_{\text{UR}}(x,y)$ close to 1 means low reachability, high unreachability.
Note: The undiscounted unreachability measure only penalizes irreversible transitions[a], while the discounted measure also penalizes reversible transitions.

[a] $d_{\text{UR}}(x,y) = 1$ if unreachable, 0 else.

The unreachability deviation measure is often used to compute the unreachability to the baseline state s'_t from a state s_t: $d_{\text{UR}}(s_t, s'_t)$.

DEF: Relative reachability

The **relative reachability (RR)** measure $d_{\text{RR}} : S \times S \to [0,1]$ is the average reduction in reachability of all states s from the current state s_t compared to the baseline s'_t:

$$d_{\text{RR}}(x,y) := \frac{1}{|S|} \sum_{s \in S} \max\left(R(s'_t, s) - R(s_t, s),\ 0\right)$$

DEF: Attainable utility

The **attainable utility measure** $d_{\text{VD}} : S \times S \to \mathbb{R}$ denotes the average gain in reward by obtaining a state x compared to a state y. To define it, we define the **value** $V_r : S \to \mathbb{R}$ of a state x according to a reward function r. Therefore let x_t^π denote the state obtained from x by following policy π for t steps. It is

$$V_r(x) := \max_\pi \sum_{t=0}^{\infty} \gamma_r^k \, r(x_t^\pi)$$

TODO: Was ist k? x is die reward function r? Die ist aber fuer States UND Aktionen definiert. Ist die Summe aller rewards fuer alle States und Aktionen gemaess π gemeint?] $r(x_t^\pi) = \sum_{t_1=0}^{\infty} r(x_{t_1}, a_{t_1})$ sodass $(x_{t_1}, a_{t_1}) \in \pi$. TODO: Vermutung fuer k: Wie beim inaction rollout time step difference zum aktuellen time step. Frage: Sollte das values eines states nicht abhaengig vom time step des states sein?