Zpětovazební učení

Pasivní učení

Policy je fixní, učíme se jenom Utility stavů

- Direct Utility Estiamtion:
  - Jendoduše necháme agenta, aby provedl několik průchodů prostorem a stavům přiřadíme průměr z utilit
- Adaptive Dynamic Programing
  - To samé jako DUE, ale utility vždy přepočítáme na základě Belmanových rovnic
  - $U^{\pi}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi s)U(s')$
- Teporal-difference Learning
  - Nemusím přepočívat úplně všehcny utility, jenom ty, které se mohli změnit
  - $U^{\pi}(s) = U^{\pi}(s) + \alpha(R(s) + \gamma U^{\pi}(s') - U^{\pi}(s)$

Aktivní učení

Polici upravujem podle toho, jak se nám to zrovna líbí. Musíme udržovat exploraci a exploataci abychom nedostali greedy agenta. Preferujeme nové stavy před starými.

Q-Learning a SARSA

Najednou máme matici Q, takovou, že $Q(s, a)$ říká jakou hodnotu má provedení a v s

$$U(s) = max_a Q(s, a)$$

Update learningu provadíme

$$Q(s, a) = Q(s, a) + \alpha(R(s)\gamma max_{a'} Q(s', a') - Q(s, a)$$

Pro SARSu update vypadá

$$Q(s, a) = Q(s, a) + \alpha(R(s)\gamma Q(s', a') - Q(s, a)$$