

Climbing Game Optimale Strategieën

Max Knobbout

Een vraag tijdens het practicum was of het mogelijk is om met ϵ -greedy te convergeren naar de optimale joint actie (a_1, a_1) (of mogelijk een andere optimum). Dit kunnen we oplossen door simpelweg de Bellman equation uit te schrijven (wat eenvoudig is omdat we stateless zijn) en te kijken naar het Pareto optimum. Gegeven een kleine setting voor α z.d.d. convergentie gegarandeerd is, en $t \mapsto \infty$, laat $Q_m^*(a_n)$ de optimale Q waarde zijn voor agent m en actie a_n . Een optimale strategie voldoet aan de volgende eis:

$$(a_i, a_j) \text{ is optimaal} \iff Q_1^*(a_i) = \max_n(Q_1^*(a_n)) \text{ en } Q_2^*(a_j) = \max_n(Q_2^*(a_n))$$

In andere woorden, een joint actie is een stabiele strategie als voor beide agenten geldt dat hun Q -waarden voor deze acties hoger zijn dan de resterende Q -waarden, en ze dus geen reden hebben om af te wijken. Dit optimum is zowel een *Nash equilibrium* als een *Pareto optimum*. De optimale Q waarden kunnen we, zoals eerder gezegd, uitschrijven met de Bellman equation. Laten we ϵ_m denoteren als de exploratie van agent m . De rewards voor agent 1, gegeven een actie, kunnen we zien als stochastisch met kansparameter ϵ_2 (en vice versa). Dit ziet er als volgt uit:

$$Q_1^*(a_1) = (1 - \epsilon_2)11 + 0.5\epsilon_2(-30 + 0) = \underline{11 - 26\epsilon_2}$$

$$Q_1^*(a_2) = (1 - \epsilon_2)7 + 0.5\epsilon_2(-30 + 0) = \underline{7 - 21\epsilon_2}$$

$$Q_1^*(a_3) = (1 - \epsilon_2)6 + 0.5\epsilon_2(5 + 0) = \underline{6 - 3.5\epsilon_2}$$

$$Q_2^*(a_1) = (1 - \epsilon_1)11 + 0.5\epsilon_1(-30 + 0) = \underline{11 - 26\epsilon_1}$$

$$Q_2^*(a_2) = (1 - \epsilon_1)7 + 0.5\epsilon_1(-30 + 6) = \underline{7 - 19\epsilon_1}$$

$$Q_2^*(a_3) = (1 - \epsilon_1)5 + 0.5\epsilon_1(0 + 0) = \underline{5 - 5\epsilon_1}$$

Om bijvoorbeeld te bepalen of (a_1, a_1) een optimale strategie is, moet gelden volgens de definitie van optimaliteit dat $11 - 26\epsilon_2 \geq 7 - 21\epsilon_2$ (agent 1), $11 - 26\epsilon_2 \geq 6 - 3.5\epsilon_2$ (agent 1), $11 - 26\epsilon_1 \geq 7 - 19\epsilon_1$ (agent 2) en $11 - 26\epsilon_1 \geq 5 - 5\epsilon_1$ (agent 2). Dit is verder elementaire wiskunde, los de vergelijking op en te zien valt dat moet gelden $\epsilon_1 \leq \frac{2}{7}$ en $\epsilon_2 \leq \frac{2}{9}$.

Verdere optimale strategieën kunnen getoond worden door te kleuren in het 2-dimensionale vlak (waarbij x-as ϵ_1 , y-as ϵ_2 en een kleur een joint actie). Te zien is dan waarschijnlijk ook dat er maar 3 of 4 optimale strategieën (=kleuren) zijn. Het diagonaal van dit vlak is dan hoe de meeste van jullie de opdracht hebben geïmplementeerd: $\epsilon_1 = \epsilon_2$.