Zaawansowane Metody Inteligencji Obliczeniowej

Lab 10: Gradient polityki

Michał Kempka Marek Wydmuch

13 maja 2021

"Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)",
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

1 Wprowadzenie

Dotąd wszystkie metody, które omawialiśmy były oparte o estymację funkcji wartości stanów lub stanów i akcji i konstruowały politykę w oparciu o te estymaty. Na tych laboratoriach rozważymy nową rodzinę metod, która uczy się sparametryzowanej polityki i nie bazuje na funkcji wartości.

Oznaczmy wektor parametrów polityki jako $\boldsymbol{\theta} \in \mathbb{R}^{d'}$. Stąd: $\pi(a|s, \boldsymbol{\theta}) = P(A_t = a|S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta})$ będzie oznaczać prawdopodobieństwo podjęcia akcji a w stanie s w kroku t z parametrami polityki θ. Będziemy rozważać metody uczące się parametrów polityki korzystając z gradientów metryk jakości $J(\boldsymbol{\theta})$, którą będziemy starać się maksymalizować:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta}).$$

Metody, które korzystają z tego schematu nazywamy metodami **gradientu polityki (ang. policy gradient)**. Metody które łączą podejście gradientu polityki i estymacji funkcji wartości nazywamy metodami **aktor-krytyk (ang. actor-critic)**, gdzie "aktor" uczy się polityki a "krytyk" uczy się funkcji wartości.

2 Aproksymacja polityki

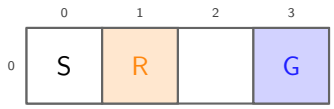
W metodach gradientu polityki polityka $\pi(a|s, \boldsymbol{\theta})$ może być sparametryzowana w dowolny sposób tak długo jak $\pi(a|s, \boldsymbol{\theta})$ jest różniczkowalne (czyli jeśli $\nabla \pi(a|s, \boldsymbol{\theta})$ istnieje względem parametrów $\boldsymbol{\theta}$). By zapewnić eksploracje, wymagamy by polityka nigdy nie stała się deterministyczna dla żadnego s, a i $\boldsymbol{\theta}$.

Naturalnym i częstym sposobem parameteryzacji polityki jest przypisanie prawdopodobieństwa wybrania dla każdej akcji a w stanie s, używając estymatora preferencji akcji $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$ oraz funkcji soft-max (znormalizowanej funkcji wykładniczej):

$$\pi(a|s, \boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_{a' \in \mathcal{A}} e^{h(s,a',\boldsymbol{\theta})}},$$

gdzie e to liczba Eulera. $h(s, a, \boldsymbol{\theta})$ mogą być dowolnym estymatorem, np. modelem liniowym albo siecią neuronową. Taki sposób parametryzacji polityki ma następujące zalety nad dotychczasowo stosowanym przez nas polityką ϵ-greedy: 1) Może zbiegnąć do deterministycznej polityki, gdzie w polityce ϵ-greedy jest zawsze z prawdopodobieństwem ϵ wybierana losowa akcja. 2) Pozwala na naturalne wyznaczenie stochastycznej polityki, która w niektórych problemach może być lepsza niż deterministyczna (np. w grze w pokera, często optymalnie jest blefować z danym prawdopodobieństwem), metody oparte o estymację funkcji wartości nie mają naturalnego sposobu wyznaczania takiej polityki. 3) W niektórych wypadkach polityka może być prostsza do aproksymacji niż funkcja wartości i może uczyć się o wiele szybciej niż metody oparte o estymacji funkcji wartości. 4) Pozwala łatwiej zastosować wiedzę a priori dotyczącą środowiska. 5) Przy parametryzacji poliktyki, wartości prawdopodobieństw dla akcji zmieniają się w sposób płynny w miarę uczenia się paramterów, podczas

gdy prawdopobieństwa w polityce ϵ-greedy mogą zmieniać się drastycznie nawet przy małej zmianie estymacji funkcji wartości w momencie gdy ta zmiana powoduje, że inna akcja ma największą wartość niż poprzednio.



Rysunek 1: Diagram przedstawiający proste środowisko z 4 nierozróżnialnymi stanami. Dopuszczalne akcje to ruch w lewo i w prawo, co krok przyznawana jest kara -1. S to stan początkowy, R to stan w którym akcje są odwrócone (akcja w lewo powoduje ruch w prawo, a akcja w prawo powoduje ruch w lewo, G to stan terminalny.

Przykłady środowiska w których aproksymacja polityki sprawdzi się dużo lepiej zaprezentowane są na Rysunku 1.

Pytanie: Jakie jest optymalne prawdopodobieństwo ruchu w prawo dla środowiska z Rysunku 1? Znając dynamikę tego środowiska skorzystaj z równania Bellmana by wyznaczyć $V(0)$ jako funkcję prawdopodobieństwa ruchu w prawo.

Odpowiedź: Korzystając z równania Bellmana wyznaczamy V(s) jako funkcja V(s'):

$$V(0) = -1 + (1-p)V(0) + pV(1) \tag{1}$$
$$V(0) = -1 + V(0) - pV(1) + pV(1) \tag{2}$$
$$0 = -1 - pV(0) + pV(1) \tag{3}$$
$$V(0) = V(1) - \frac{1}{p} \tag{4}$$
$$V(1) = -1 + (1-p)V(2) + pV(0) \tag{5}$$
$$V(1) = -1 + (1-p)V(2) + p(V(1) - \frac{1}{p}) \tag{6}$$
$$V(1) = V(2) + \frac{2}{p-1} \quad \text{takie samo przekształcenie jak w wypadku V(0)} \tag{7}$$
$$V(2) = -1 + (1-p)V(1) + pV(3) \tag{8}$$
$$V(3) = 0 \tag{9}$$
$$V(2) = -1 + (1-p)(V(2) + \frac{2}{p-1}) \tag{10}$$
$$V(2) = -\frac{3}{p} \tag{11}$$
$$V(0) = V(1) - \frac{p}{1} \tag{12}$$
$$V(0) = V(2) + \frac{2}{p-1} - \frac{p}{1} \tag{13}$$
$$V(0) = -\frac{3}{p} + \frac{2}{p-1} - \frac{1}{p} \tag{14}$$
$$\tag{15}$$

Teraz wystarczy znaleźć maximum tej funkcji, jest ona wklęsła w przedziale od 0 do 1, więc znajdźmy miejsce gdzie pochodna jest równa 0:

$$\frac{d}{dp}\left(-\frac{3}{p} + \frac{2}{p-1} - \frac{1}{p}\right) = \frac{3}{p^2} - \frac{2}{(p-1)^2} + \frac{1}{p^2} = \frac{4}{p^2} - \frac{2}{(p-1)^2}$$

$$\frac{4}{p^2} - \frac{2}{(p-1)^2} = 0 \tag{16}$$
$$p = 2 - \sqrt{2} \approx 0.5858 \tag{17}$$

Pytanie: Ile powinien wynosić ϵ w polityce ϵ-greedy by polityka była optymalna dla środowiska z Rysunku 1?

Odpowiedź:

$$p = 1 - \epsilon + \frac{\epsilon}{2} \tag{18}$$

$$2 - \sqrt{(2)} = 1 - \epsilon + \frac{\epsilon}{2} \tag{19}$$

$$\epsilon = 2(\sqrt{2} - 1) \approx 0.8284 \tag{20}$$

$$\tag{21}$$

Tak jak już powiedzieliśmy potrzebujemy miarę jakości $J(\boldsymbol{\theta})$. Dla epizodycznych środowisk możemy ją zdefiniować jako $J(\boldsymbol{\theta}) = v_{\pi_\theta}(s_0)$, gdzie v_{π_θ} to prawdziwa funkcja wartości stanu dla sparamteryzowanej polityki $\pi_{\boldsymbol{\theta}}$.

Taka miara jakości wydaje się być początkowo trudna do optymalizacji z punktu widzenia parametrów polityki, ponieważ zależy ona zarówno od akcji jak i rozkładu stanów w których te akcje są wybierane, obie te kwestie zależą od parametrów polityki. Mając dany stan, policzenie efektów danej akcji i jej oczekiwanej nagrody może być wyznaczony w relatywnie prosty sposób. Problemem jest jednak efekt polityki na rozkład stanów, który jest nieznaną funkcją środowiska. Czy możemy więc policzyć gradient dla funkcji miary jakości, gdy zależy on od nieznanego efektu zmiany polityki na rozkład stanów? Na szczęście odpowiedź brzmi tak, a dostarcza nam ją twierdzenie o gradiencie polityki:

$$\nabla v_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s,a) \right], \text{for all } s \in S \tag{22}$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \right] \tag{23}$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} P(s',r|s,a)(r + v_\pi(s')) \right] \tag{24}$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} P(s'|s,a) \nabla v_\pi(s') \right] \tag{25}$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} P(s'|s,a) \right. \tag{26}$$

$$\left. \sum_{a'} [\nabla \pi(a'|s') q_\pi(s',a') + \pi(a'|s') \sum_{s''} P(s''|s',a') \nabla v_\pi(s'')] \right] \tag{27}$$

$$= \sum_{s' \in \mathcal{S}} \sum_{k=0}^{\infty} P(s \to s', k, \pi) \sum_a \nabla \pi(a|s') q_\pi(s',a) \,, \tag{28}$$

Gdzie $P(s \to s', k, \pi)$ to prawdopodobieństwo przejścia ze stanu s do stanu s' w k krokach działając zgodnie polityką π.

$$\nabla J(\boldsymbol{\theta}) = \nabla v_\pi(s_0) \tag{29}$$
$$= \sum_s \left(\sum_{k=0}^{\infty} P(s_0 \to s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_\pi(s,a) \tag{30}$$
$$= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) \tag{31}$$
$$= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_s' \eta(s')} \sum_a \nabla \pi(a|s) q_\pi(s,a) \tag{32}$$
$$= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) \tag{33}$$
$$\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) \tag{34}$$
$$\propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s) \tag{35}$$
$$\tag{36}$$

Powyższy wynik pokazuje, że gradient funkcji wartości jest proporcjonalny do następującego równania, które nie zawiera pochodnej dystrybucji stanów. Wielkość tej proporcji nie jest dla nas specjalnie istotny, gdyż i tak kontrolujemy nasze uaktualnianie wielkością kroku α. Zauważ, że:

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s') \sum_a q_\pi(s,a) \nabla \pi(a|s, \boldsymbol{\theta}) \tag{37}$$
$$= \mathbb{E}\left[\sum_a q_\pi(S_t, a) \nabla \pi(a|S_t, \boldsymbol{\theta}) \right]. \tag{38}$$
$$\tag{39}$$

Powyższe równanie zawiera sumę po akcjach, ale środkowe wyrażenie nie jest ważone po $\pi(a|S_t, \boldsymbol{\theta})$, dlatego wprowadzamy takie ważenie bez zmieniania równania:

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_a \pi(a|s, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})} \right] \tag{40}$$
$$= \mathbb{E}\left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \right] \tag{41}$$
$$\tag{42}$$

gdzie $A_t \sim \pi$.

3 Actor-Critic

Do twierdzenie o gradiencie polityki można zawrzeć porównanie z dowolnym baselinem:

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s') \sum_a (q_\pi(s,a) - b(s)) \nabla \pi(a|s, \boldsymbol{\theta}) \tag{43}$$
$$\tag{44}$$

Jednym z naturalnych wyborów na baseline jest aproksymacja funkcja wartości stanu $\hat{v}(S', \boldsymbol{w})$, gdzie $\boldsymbol{w}$ jest wektorem wag.

W ten sposób dochodzimy do reguły aktualizacji metody actor-critic zaprezentowanej w Algorytmie 1.

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha(R_{t+1} + \gamma \hat{v}(S_{t+1}, \boldsymbol{w}) - \hat{v}(S_t, \boldsymbol{w})) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})} \tag{45}$$
$$= \boldsymbol{\theta}_t + \alpha \delta_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})}. \tag{46}$$

Zauważ że:

$$\nabla \ln x = \frac{\nabla x}{x}$$

Algorytm 1: Pseudokod dla algorytmu One-step Actor-Critic

Inicjalizacja:
Różniczkowalna polityka $\pi(a|s, \theta)$
Różniczkowalna funkcja wartości stanu $v(a|s, \boldsymbol{w})$
Zainicjalizowane wagi $\boldsymbol{w} \in \mathbb{R}^d_{\boldsymbol{w}}$ i $\boldsymbol{\theta} \in \mathbb{R}^d_{\theta}$
Zadana szybkość uczenia $\alpha_{\boldsymbol{\theta}}, \alpha_{\boldsymbol{w}} \in \mathcal{R}^+$
repeat
  if S *nieustawiony lub terminalny* **then**
    Rozpocznij nowy epizod
    $S \leftarrow S_0$
  $A \sim \pi(\cdot|S, \theta)$
  Wykonaj akcję A, zaobserwuj nagrodę R i następnik S'
  $\delta \leftarrow R + \gamma v(S', \boldsymbol{w}) - v(S, \boldsymbol{w})$
  $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_{\boldsymbol{w}} \delta \nabla v(S, \boldsymbol{w})$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \delta \nabla \ln \pi(A|S, \boldsymbol{\theta})$
  $S \leftarrow S'$
until *warunek stopu*;

Literatura

[1] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, third edition.

[2] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

"Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)",
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20