Mnogoruki pljačkaš (bandit)

- Konceptualno - *slot* mašina. Ne postoji konkretan budžet koji ulažemo, no imamo pri povlačenju ručice povratnu vrijednost, svaki put (realna vrijednost).
- Opisujemo *bandita* uređenim parom (m, s), gdje je m srednja povratna vrijednost pljačkaša, a s standardna devijacija.
- Poredajmo n takvih pljačkaša.
- **Cilj** - **maksimizovati dobit** ne znajući srednje povratne vrijednosti pljačkaša, odlučiti koja mašina daje najviše dobiti.
- **Faze** u pronalaženju:
  - *Test* faza - veliki broj iteracija (reda 10000) - ***eksploracija***
    - Nasumično biramo mašine
    - Povlačimo poteze
    - Računamo m
  - *Igranje najbolje mašine* - ***eksploatacija***
- Potrebno je napraviti balans između eksploracije i eksploatacije, jer sistem može da bude varijantan u vremenu.
- Pravimo politiku odlučivanja:
  - *Pohlepna* (*greedy*) politika
  - ϵ - *pohlepna* politika ($\epsilon \in (0, 1)$)
    - Biraj mašine nasumično vjerovatnoćom ϵ
    - Biraj najbolju mašinu vjerovatnoćom $1 - \epsilon$
  - *Softmax* politika
- Neka je q trenutna srednja vrijednost mašine. Definišemo $q^+ = p \cdot q + (1 - p) \cdot q$ kao narednu srednju vrijednost mašine. Ako uvedemo smjenu $\alpha = 1 - p$, naredna vrijednost postaje $q^+ = q + \alpha(1 - q)$. Suštinski, ova estimacija trenutne vrijednosti bolja je nego računanje prave srednje vrijednosti jer ne pamtimo broj iteracija.
- Idealno bi bilo pogoditi $q \approx m$, odnosno da naša procijenjena vrijednost bude što bliža sredini.

Domaći 1

- Probati sa manjim ϵ (mijenjamo "intenzitet" *greedy* politike)
- Naučeno q - q nakon iteracija
- Varijabilne karakteristike bandita
- Napraviti konkretne bandite i konkretno okruženje. Prikazati konvergenciju q ka m za svakog bandita