Reinforcement Learning for Business, Economics, and Social Sciences

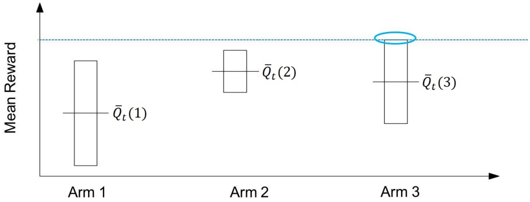
Unit 1-3: Upper Confidence Bound

Davud Rostam-Afschar (Uni Mannheim)

How to learn by being optimistic in the face of uncertainty?

Optimism in the Face of Uncertainty

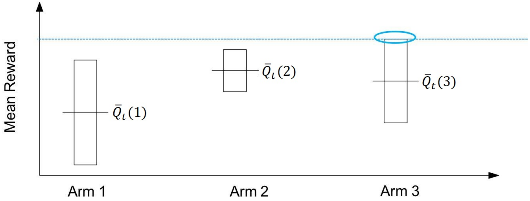
Which action should we pick?



- The more uncertain we are about an action-value...
- The more important it is to explore that action
- It could turn out to be the best action!

Optimism in the Face of Uncertainty

Which action should we pick?

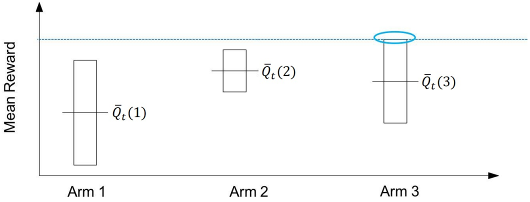


After picking arm 3:

- We become less uncertain about its value
- We are more likely to pick another action
- The idea is to always try the best arm

Optimism in the Face of Uncertainty

Which action should we pick?

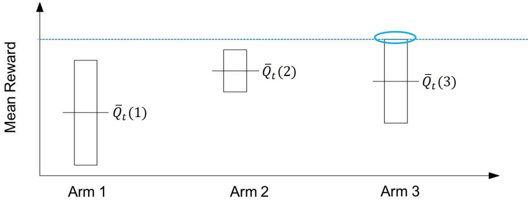


After picking arm 3:

- We become less uncertain about its value
- We are more likely to pick another action
- The idea is to always try the best arm, where "best" includes

Optimism in the Face of Uncertainty

Which action should we pick?

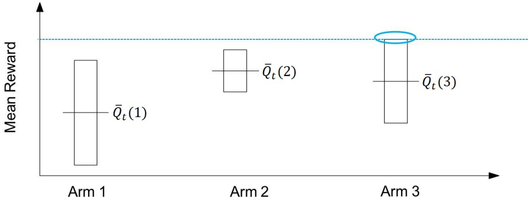


After picking arm 3:

- We become less uncertain about its value
- We are more likely to pick another action
- The idea is to always try the best arm, where "best" includes
    - exploitation (average observed reward) and

Optimism in the Face of Uncertainty

Which action should we pick?

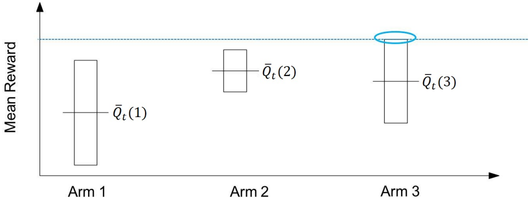


After picking arm 3:

- We become less uncertain about its value
- We are more likely to pick another action
- The idea is to always try the best arm, where "best" includes
  - exploitation (average observed reward) and
  - exploration (uncertainty about observed reward)

Optimism in the Face of Uncertainty

Which action should we pick?



After picking arm 3:

- We become less uncertain about its value
- We are more likely to pick another action
- The idea is to always try the best arm, where "best" includes
  - exploitation (average observed reward) and
  - exploration (uncertainty about observed reward)

(Auer, Cesa-Bianchi, and Fischer, 2002)

- Theorem:
  An optimistic strategy that always selects $\underset{a}{\operatorname{argmax}} U_t(a)$ will converge to a^*.

- Proof by contradiction:
  - Suppose that we converge to suboptimal arm a after infinitely many trials
  - Then $\overline{Q}(a) = U_\infty(a) \geq U_\infty(a') = \overline{Q}(a') \quad \forall\, a'$
  - But $\overline{Q}(a) \geq \overline{Q}(a') \quad \forall\, a'$ contradicts our assumption that a is suboptimal

Convergence

- Theorem:
  An optimistic strategy that always selects $\underset{a}{\mathrm{argmax}}\, U_t(a)$ will converge to a^*.

- Proof by contradiction:
  - Suppose that we converge to suboptimal arm a after infinitely many trials
  - Then $\overline{Q}(a) = U_\infty(a) \geq U_\infty(a') = \overline{Q}(a') \quad \forall\ a'$
  - But $\overline{Q}(a) \geq \overline{Q}(a') \quad \forall\ a'$ contradicts our assumption that a is suboptimal

- Problem: We can't compute an upper bound with certainty since we are sampling

Upper Confidence Bounds

- Estimate an upper confidence $U_t(a)$ for each action value
  $\Rightarrow$ Such that with high probability

$$q(a) \leq \underbrace{\overline{Q}_t(a)}_{\text{estimated mean}} + \underbrace{U_t(a)}_{\substack{\text{estimated} \\ \text{Upper Confidence}}}$$

Upper Confidence Bounds

- Estimate an upper confidence $U_t(a)$ for each action value

  $\Rightarrow$ Such that with high probability

$$q(a) \leq \underbrace{\overline{Q}_t(a)}_{\text{estimated mean}} + \underbrace{U_t(a)}_{\substack{\text{estimated}\\ \text{Upper Confidence}}}$$

- Upper confidence depends on number of times action a has been selected:
  - Small $N_t(a) \Rightarrow$ large $U_t(a)$ (estimated value is uncertain)
  - Large $N_t(a) \Rightarrow$ small $U_t(a)$ (estimated value is accurate)

Upper Confidence Bounds

- Estimate an upper confidence $U_t(a)$ for each action value
  $\Rightarrow$ Such that with high probability

$$q(a) \leq \underbrace{\overline{Q}_t(a)}_{\text{estimated mean}} + \underbrace{U_t(a)}_{\substack{\text{estimated}\\ \text{Upper Confidence}}}$$

- Upper confidence depends on number of times action a has been selected:
  - Small $N_t(a) \Rightarrow$ large $U_t(a)$ (estimated value is uncertain)
  - Large $N_t(a) \Rightarrow$ small $U_t(a)$ (estimated value is accurate)
- Select action maximizing Upper Confidence Bound (UCB):

$$a_t = \arg\max_{a \in \mathcal{A}} \left[\overline{Q}_t(a) + U_t(a) \right]$$

Hoeffding's Inequality

- Let $X_1, \ldots, X_t$ be *i.i.d.* random variables in $[0,1]$, and let

$$\overline{X}_t = \frac{1}{t}\sum_{\tau=1}^{t} X_\tau \quad \text{be the sample mean. Then}$$

$$\mathbb{P}\left[\mathbb{E}[X] > \overline{X}_t + u\right] \leq e^{-2tu^2}$$

- We will apply **Hoeffding's Inequality** to rewards of the bandit conditioned on selecting action a

$$\mathbb{P}\left[Q(a) > \overline{Q}_t(a) + U_t(a)\right] \leq e^{-2N_t(a)U_t(a)^2}$$

Calculating Upper Confidence Bounds

- Pick a probability p that true value exceeds UCB
- Now solve for $U_t(a)$

$$e^{-2N_t(a)U_t(a)^2} = p$$

$$U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$$

Calculating Upper Confidence Bounds

- Pick a probability p that true value exceeds UCB
- Now solve for $U_t(a)$

$$e^{-2N_t(a)U_t(a)^2} = p$$

$$U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$$

- Reduce p as we observe more rewards, e.g. $p = t^{-c},\ c = 4$
  - (note: c is a hyper-parameter that trades-off explore/exploit)
- Ensures we select optimal action as $t \to \infty$

$$U_t(a) = \sqrt{\frac{2 \log t}{N_t(a)}}$$

UCB: Three-Arm Example for $t = 100$

Arm a	Pulls $N_{100}(a)$	Empirical mean $\overline{Q}_t(a)$	Exploration bonus $\sqrt{\frac{2 \log t}{N_t(a)}}$	$\mathrm{UCB}_a =$ $\overline{Q}_t(a) + \text{bonus}_t(a)$
1	30	0.70	$\sqrt{2 \ln(100)/30} = 0.554$	1.254
2	50	0.50	$\sqrt{2 \ln(100)/50} = 0.429$	0.929
3	20	0.60	$\sqrt{2 \ln(100)/20} = 0.679$	1.279

Next selection: Arm 3 (highest UCB of 1.279)

Upper Confidence Bound (UCB)

- Choose a with highest Hoeffding bound

UCB(T)

$Q_t(a) \leftarrow 0, t \leftarrow 0, N_t(a) \leftarrow 0 \quad \forall a$

Repeat until $t = T$

Execute $\underset{a}{\operatorname{argmax}} \overline{Q}_t(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$

Receive $R_t(a)$

$Q_t(a) \leftarrow Q_t(a) + R_t(a)$

$\overline{Q}_t(a) \leftarrow \frac{N_t(a)\overline{Q}_t(a) + R_t(a)}{N_t(a)+1}$

$t \leftarrow t + 1, N_t(a) \leftarrow N_t(a) + 1$

Return $Q_t(a)$

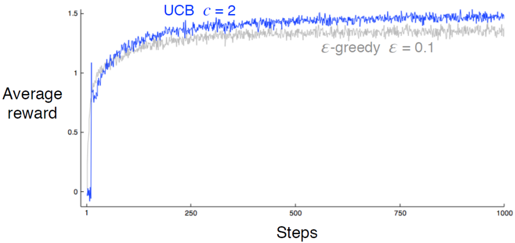
Exploration vs Exploitation

Upper Confidence Bound (UCB)

- A clever way of reducing exploration over time
- Estimate an upper bound on the true action values
- Select the action with the largest (estimated) upper bound:

$$a_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$$

- where $c > 0$ controls the degree of exploration



- **Theorem:** Although Hoeffding's bound is probabilistic, UCB converges.
- **Idea:** As t increases, the term $\sqrt{\frac{2 \log t}{N_t(a)}}$ increases, ensuring that all arms are tried infinitely often.
  The higher $N_t(a)$, the more confident in the estimate for action a.
- Expected cumulative regret: $\text{Loss}_T = \mathcal{O}(\log T)$
  - Logarithmic regret

References I

AUER, P., N. CESA-BIANCHI, AND P. FISCHER (2002): "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, 47, 235–256.

Takeaways

What is Upper Confidence Bound (UCB)?

- Uses a probabilistic upper bound to guide action selection
- At each step, select action with highest empirical mean plus exploration bonus
- Ensures that all actions are tried enough times
- It and converges to the optimal arm
- Achieves logarithmic regret
- UCB often outperforms ε-greedy strategies in practice