Reinforcement Learning Regret Bound Analysis

Pei-Yuan Wu

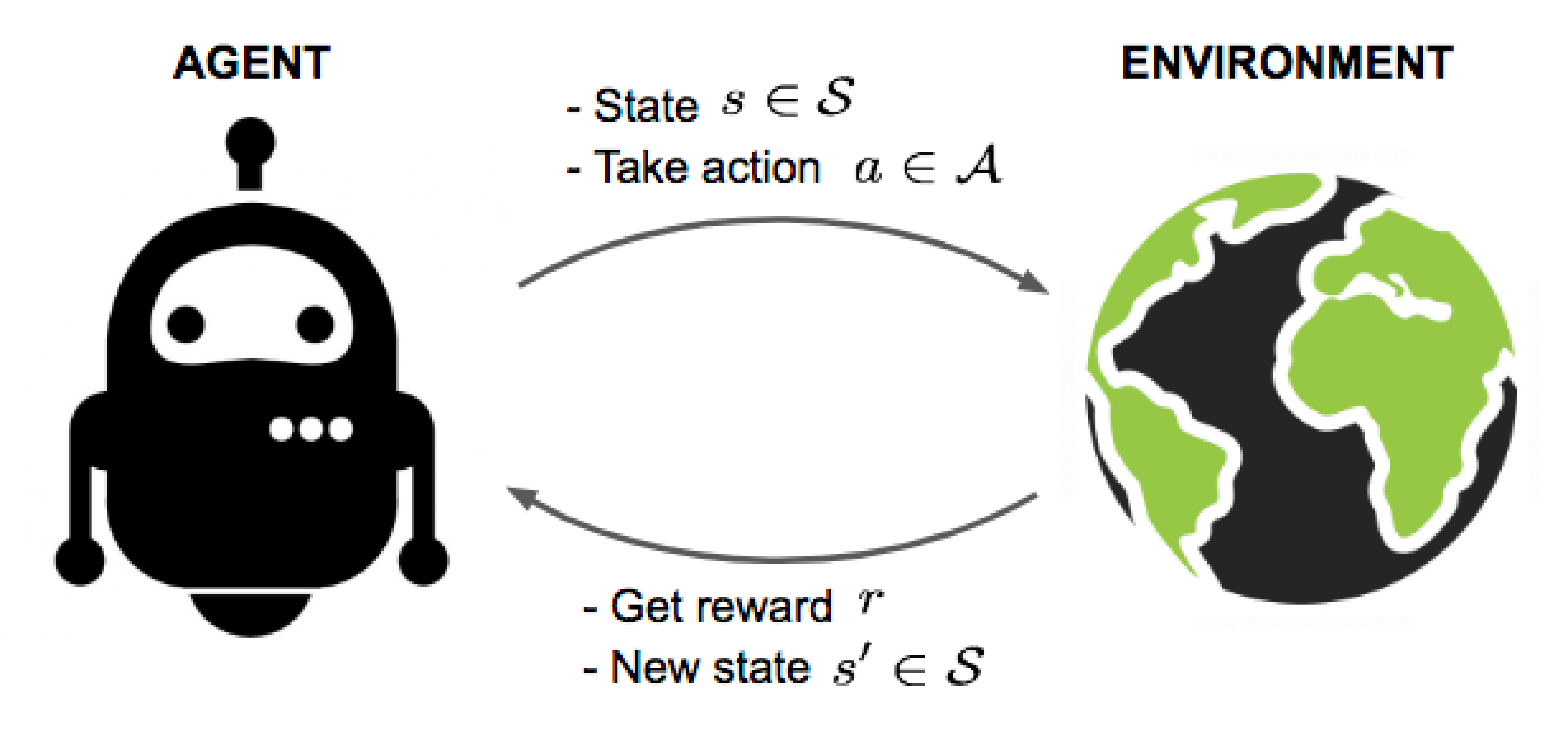
Dept. Electrical Engineering

National Taiwan University

Outline

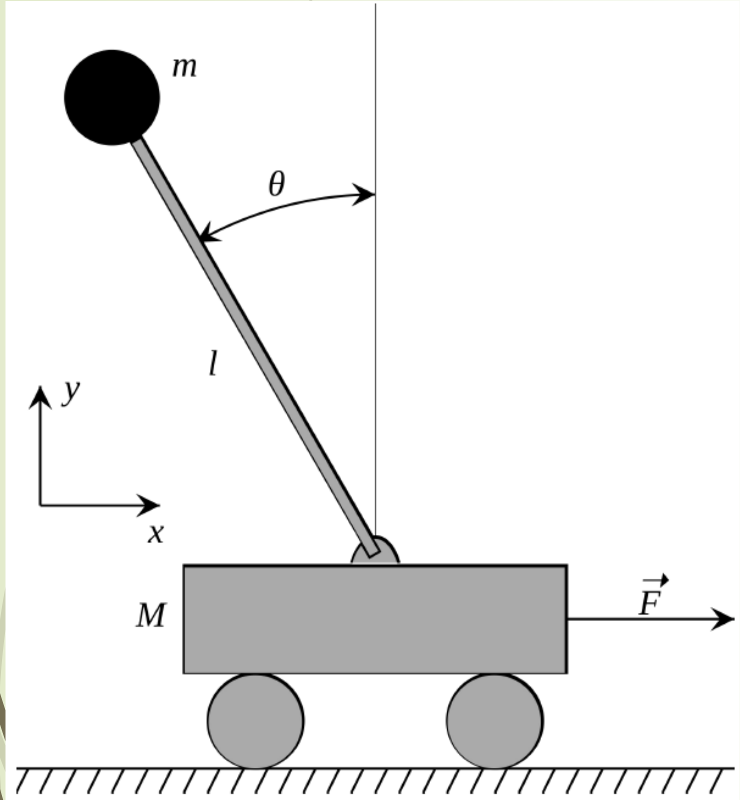
- Reinforcement Learning Basics
  - Markov Decision Process
  - Value function and Q-function
  - Bellman optimality equations
  - Value iteration, policy iteration
- Bandit problem
- Linear UCB
- Pessimistic Value Iteration

Reinforcement Learning



https://lilianweng.github.io/posts/2018-02-19-rl-overview/

Cart-Pole Problem



- **Objective**: Balance a pole on top of a movable cart
- **State:** angle, angular speed, position, horizontal velocity
- **Action:** horizontal force applied on the cart
- **Reward:** 1 at each time step if the pole is upright

Markov Decision Process (MDP)

- Mathematical formulation of the RL problem
- **Markov property**: Current state completely characterizes the state of the world

Defined by $(S, A, r, \mu, \mathbb{P}, \gamma)$

S: set of possible states (for simplicity assume $|S|$ is countable)

A: set of possible actions (for simplicity assume $|A|$ is finite)

$r: S \times A \to [0,1]$: reward function

$\mu \in \Delta(S)$: initial state distribution

$\mathbb{P}: S \times A \to \Delta(S)$: transition probability

$\gamma \in [0,1)$: discount factor

$\Delta(S)$: Collection of probability distributions on S

Markov Decision Process

- At time step $t = 0$, environment samples initial state $s_0 \sim \mu$.
- Then, for $t = 0,1,2, \ldots$ until done:
  - Agent selects action a_t
  - Environment samples reward $r_t = r(s_t, a_t)$
    - Sometimes we assume noisy observation $r_t = r(s_t, a_t) + \epsilon_t$
  - Environment samples next state $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$
  - Agent receives reward r_t and next state s_{t+1}
- Trajectory (up to time t): $\tau_t = (s_0, a_0, r_0, s_1, \ldots, s_t)$
- A policy π specifies what action to take by the agent.
  - Generally speaking, a policy may be stochastic and depends on the past trajectory $a_t \sim \pi(\cdot \mid \tau_t)$
  - Stationary policy only depends on the current state $\pi: S \to \Delta(A), a_t \sim \pi(\cdot \mid s_t)$
  - Deterministic stationary policy $\pi: S \to A, a_t = \pi(s_t)$
- **Objective:** find the optimal policy π^* that maximizes the cumulative discounted rewards $\sum_t \gamma^t r_t$ (formal definition given later)

Value function and Q-function

- With MDP $(S, A, r, \mu, \mathbb{P}, \gamma)$, following a policy π produces sample trajectory (path)

$$\tau_\infty = (s_0, a_0, r_0, s_1, \dots)$$

where $s_0 \sim \mu$, $\mathrm{a_t} \sim \pi(\cdot\,|s_0, a_0, r_0, \dots, s_t)$, $r_t = r(s_t, a_t)$

- The **value function** at state s is the expected cumulative reward from following the policy from state s

$$V^\pi(s) = \mathbb{E}\left[\!\left[\sum_{t\geq 0} \gamma^t r_t \,|\pi, s_0 = s\right]\!\right]$$

- The **Q-value function** at state s and action a is the expected cumulative reward from taking action a in state s and then following the policy:

$$Q^\pi(s,a) = \mathbb{E}\left[\!\left[\sum_{t\geq 0} \gamma^t r_t \,|\pi, s_0 = s, a_0 = a\right]\!\right]$$

Bellman consistency equation

- **Lemma (Bellman consistency equation):**
  Suppose that π is a stationary policy. Then V^π and Q^π satisfy the following Bellman consistency equations: for all $s \in S$ and $a \in A$,
$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} [\![Q^\pi(s,a)]\!]$$
$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} [\![V^\pi(s')]\!]$$

Optimal policy

- The **optimal value function** V^* and **optimal Q-value function** Q^* are defined as

$$Q^*(s,a) = \max_{\pi \in \Pi} Q^\pi(s,a)$$
$$V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$$

Here Π is the collection of all non-stationary and stochastic policies.

- **Theorem:** There exists a stationary and deterministic policy π^* such that for all $s \in S$ and $a \in A$,

$$Q^{\pi^*}(s,a) = Q^*(s,a)$$
$$V^{\pi^*}(s) = V^*(s)$$

We refer to such π^* as an **optimal policy**.

Bellman optimality equation

- The **greedy policy** w.r.t. $Q \in \ell^\infty(S \times A)$ is the stationary and deterministic policy $\pi_Q: S \to A$, defined as
$$\pi_Q(s) = \operatorname*{argmax}_{a \in A} Q(s,a)$$
where ties are broken in some arbitrary (and deterministic) manner.
- **Theorem (Bellman Optimality Equations):** For any $Q \in \ell^\infty(S \times A)$, we have that $Q = Q^*$ iff Q satisfies the **Bellman optimality equation**:
$$Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\!\left[\max_{a' \in A} Q(s',a') \right]\!\right]$$
Furthermore, π_Q is an optimal policy.
- Denote **Bellman optimality operator** $\mathbb{T}: \ell^\infty(S \times A) \to \ell^\infty(S \times A)$ as
$$(\mathbb{T}Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\!\left[\max_{a' \in A} Q(s',a') \right]\!\right]$$
Then the Bellman optimality equation can be written as $Q = \mathbb{T}Q$.

Value Iteration

How to compute Q^*?

- **Lemma (contraction):**
  For any $\mathrm{Q}, \mathrm{Q}' \in \ell^\infty(S \times A)$, holds
  $$\|\mathbb{T}Q - \mathbb{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$
- Since $\mathbb{T}$ is a contraction, and that Q^* is the unique fixed point of $\mathbb{T}$, one can compute Q^* by first setting some $\mathrm{Q}^{(0)} \in \ell^\infty(S \times A)$, and then iteratively applying the fixed point mapping
  $$\mathrm{Q}^{(k+1)} \leftarrow \mathbb{T}\mathrm{Q}^{(k)}$$

  It follows that
  $$\left\|Q^{(k)} - Q^*\right\|_\infty \leq \gamma^k \left\|Q^{(0)} - Q^*\right\|_\infty$$
- **Proposition (Q-Error Amplification):**
  For any $\mathrm{Q} \in \ell^\infty(S \times A)$, holds
  $$V^{\pi_Q}(s) \geq V^*(s) - \frac{2\|Q - Q^*\|_\infty}{1 - \gamma}$$

What’s the problem with this?

Must compute $Q(s, a)$ for every state-action pair, which is computationally infeasible if the state-action space $S \times A$ is large.

Policy Iteration

- Start from an arbitrary policy π_0.
- Then for $k = 0,1,2, \ldots$ until done:
  - *Policy evaluation*: Compute Q^{π_k}.
  - *Policy improvement*: Update the policy:

$$\pi_{k+1} = \pi_{Q^{\pi_k}}$$

Namely, $\pi_{k+1}(s) = \operatorname*{argmax}_{a \in A} Q^{\pi_k}(s, a)$.

- **Proposition:**
  1. $Q^{\pi_{k+1}} \geq \mathbb{T} Q^{\pi_k} \geq Q^{\pi_k}$.
  2. $\|Q^{\pi_{k+1}} - Q^*\|_\infty \leq \gamma \|Q^{\pi_k} - Q^*\|_\infty$.

Episodic MDP

Discounted time independent MDP

Defined by $(S, A, r, \mu, \mathbb{P}, \gamma)$

- S: set of possible states
- A: set of possible actions
- $r: S \times A \to [0,1]$: reward function
- $\mu \in \Delta(S)$: initial state distribution
- $\mathbb{P}: S \times A \to \Delta(S)$: transition probability
- γ: discount factor

Goal: Find policy π that maximizes

$$\mathbb{E}\left[\!\left[\sum_{t\geq 0} \gamma^t r(s_t, a_t) \,|\pi, s_0 = s\right]\!\right]$$

Episodic time dependent MDP

Defined by $\left(S, A, (r_h)_{\mathrm{h=0}}^{\mathrm{H-1}}, \mu, (\mathbb{P}_h)_{\mathrm{h=0}}^{\mathrm{H-1}}, H\right)$

- S: set of possible states
- A: set of possible actions
- $r_{\mathrm{h}}: S \times A \to [0,1]$: reward function at step h
- $\mu \in \Delta(S)$: initial state distribution
- $\mathbb{P}_{\mathrm{h}}: S \times A \to \Delta(S)$: transition probability at step h
- H: horizon

Goal: Find policy π that maximizes

$$\mathbb{E}\left[\!\left[\sum_{h=0}^{H-1} r_h(s_h, a_h) \,|\pi, s_0 = s\right]\!\right]$$

K-armed bandit problem

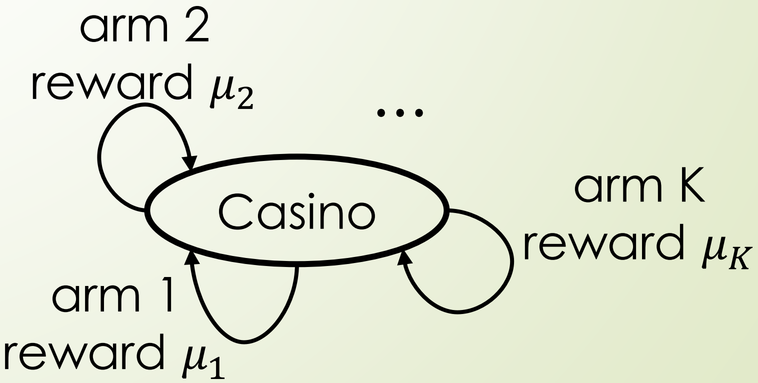
- We have K decisions (the "arms"), where when we play arm $a \in [\![1, K]\!]$ we obtain a random reward with mean $\mu_a \in [0,1]$.
- Every iteration t, the learner will pick an arm $I_t \in [\![1, K]\!]$. Our cumulative regret is

$$R_T = T \max_a \mu_a - \sum_{t=1}^{T} \mu_{I_t}$$

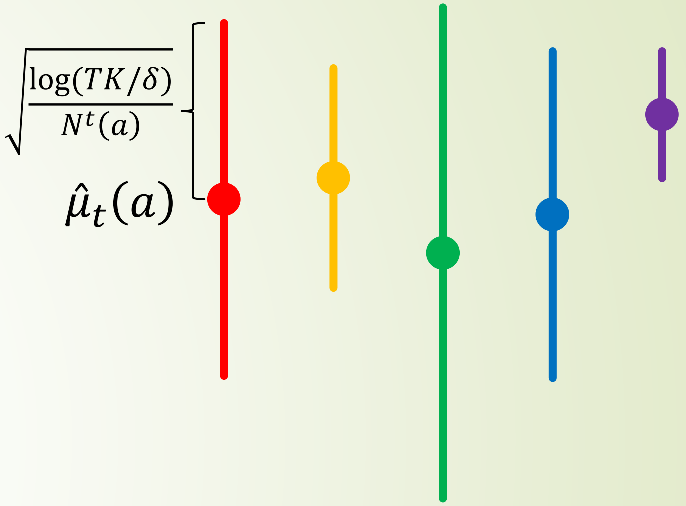
- Can be regarded as a MDP with one state (casino) and K actions (arms)

…


Reward μ_1 μ_2 μ_k



Upper confidence bound (UCB) algorithm



- First play each arm once.
- In iteration t:
  - maintain counts and empirical mean for each arm:

    $$N^t(a) = \sum_{i=1}^{t} 1\{I_i = a\},\ \hat{\mu}_t(a) = \frac{1}{N^t(a)} \sum_{i=1}^{t} 1\{I_i = a\} r_i$$

    From which the UCB is computed as

    $$\hat{\mu}_t(a) + \sqrt{\frac{\log(TK/\delta)}{N^t(a)}}$$

  - Play the arm with the highest UCB.

Algorithm 1 UCB

for $t \leftarrow 1$ to K **do**
  $I_t \leftarrow t$
  Play arm I_t and receive reward r_t
end for
for $t \leftarrow K+1$ to T **do**
  $I_t \leftarrow \arg\max_{a \in [\![1,K]\!]} \left(\hat{\mu}^{t-1}(a) + \sqrt{\frac{\log(TK/\delta)}{N^{t-1}(a)}}\right)$
  Play arm I_t and receive reward r_t
end for

Linear stochastic bandit

- In round t,
  - the learner is given a decision set $D_t \subset \mathbb{R}^d$ from which to choose action $x_t \in D_t$.
  - The learner observes reward $r_t = \langle x_t, \mu_* \rangle + \eta_t$, where $\mu_* \in \mathbb{R}^d$ is fixed but unknown, and η_t is a zero mean random noise.
- **Goal:** Maximize total reward $\sum_{t=1}^{T} \langle x_t, \mu_* \rangle$. In other words, to minimize regret

$$R_T = \sum_{t=1}^{T} \langle x_t^*, \mu_* \rangle - \sum_{t=1}^{T} \langle x_t, \mu_* \rangle = \sum_{t=1}^{T} \langle x_t^* - x_t, \mu_* \rangle$$

where $x_t^* \in \underset{x \in D_t}{\operatorname{argmax}} \langle x, \mu_* \rangle$.

Optimism in the face of uncertainty (OFU) principle

- In round t,
  - Maintains a confidence set $C_{t-1} \subset \mathbb{R}^d$ for the parameter μ_*.
    - ✓ C_{t-1} is calculated from $x_1, x_2, \dots, x_{t-1}$ and $r_1, r_2, \dots, r_{t-1}$ where $\mu_* \in C_{t-1}$ "with high probability".
  - Chooses an optimistic estimate $\tilde{\mu}_t \in \underset{\mu \in C_{t-1}}{\operatorname{argmax}} \max_{x \in D_t} \langle x, \mu \rangle$ and the action $x_t \in \underset{x \in D_t}{\operatorname{argmax}} \langle x, \tilde{\mu}_t \rangle$ which maximizes the reward according to estimate $\tilde{\mu}_t$.

Algorithm 2 OFUL

for $t \leftarrow 1$ to T **do**
  $(x_t, \tilde{\mu}_t) \leftarrow \arg\max_{(x,\mu) \in D_t \times C_{t-1}} \langle x, \mu \rangle$
  Play x_t and observe reward r_t
  Update C_t
end for

Confidence set

How to estimate the confidence set C_t?

- Recall that $r_t = \langle x_t, \mu_* \rangle + \eta_t$.
- Let $\hat{\mu}_t$ be the regularized least-squares estimate of μ_* up to step t, namely

$$\hat{\mu}_t \in \underset{\mu}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T} (\langle x_t, \mu \rangle - r_t)^2 + \rho \|\mu\|_{\Sigma_0}^2 \right\} = \left(\Sigma_0 + \sum_{t=1}^{T} x_t x_t^{\mathrm{T}} \right)^{-1} \left(\sum_{t=1}^{T} r_t x_t \right)$$

where $\Sigma_0 \in \mathbb{R}^{d \times d}$ is positive definite, and $\|v\|_A = \sqrt{v^{\mathrm{T}} A v}$.

- **Theorem:** Assume that each η_t is conditionally R^2-subgaussian, that is,

$$\mathbb{E}\left[e^{\lambda \eta_t} | x_1, \dots, x_t, \eta_1, \dots, \eta_{t-1} \right] \le \exp\left(\frac{\lambda^2 R^2}{2} \right), \forall \lambda \in \mathbb{R}$$

Then for each $\delta < 0$, with probability at least $1 - \delta$, the following holds for all $t \geq 0$:

$$\|\hat{\mu}_t - \mu_*\|_{\Sigma_t} \le \left\| \Sigma_t^{-\frac{1}{2}} \Sigma_0 \right\| \|\mu_*\| + R \sqrt{\log\left(\frac{\det(\Sigma_t)}{\delta^2 \det(\Sigma_0)} \right)}$$

where $\Sigma_t = \Sigma_0 + \sum_{i=1}^{t} x_i x_i^{\mathrm{T}}$.

- Suppose $\|\mu_*\| \le M$, we may take confidence set as

$$C_t = \left\{ \mu \in \mathbb{R}^d \middle| \|\mu - \hat{\mu}_t\|_{\Sigma_t} \le \beta_t \right\}, \beta_t = \left\| \Sigma_t^{-\frac{1}{2}} \Sigma_0 \right\| M + R \sqrt{\log\left(\frac{\det(\Sigma_t)}{\delta^2 \det(\Sigma_0)} \right)}$$

It follows that, with probability at least $1 - \delta$, holds $\mu_* \in C_t$ for all $t \geq 0$.

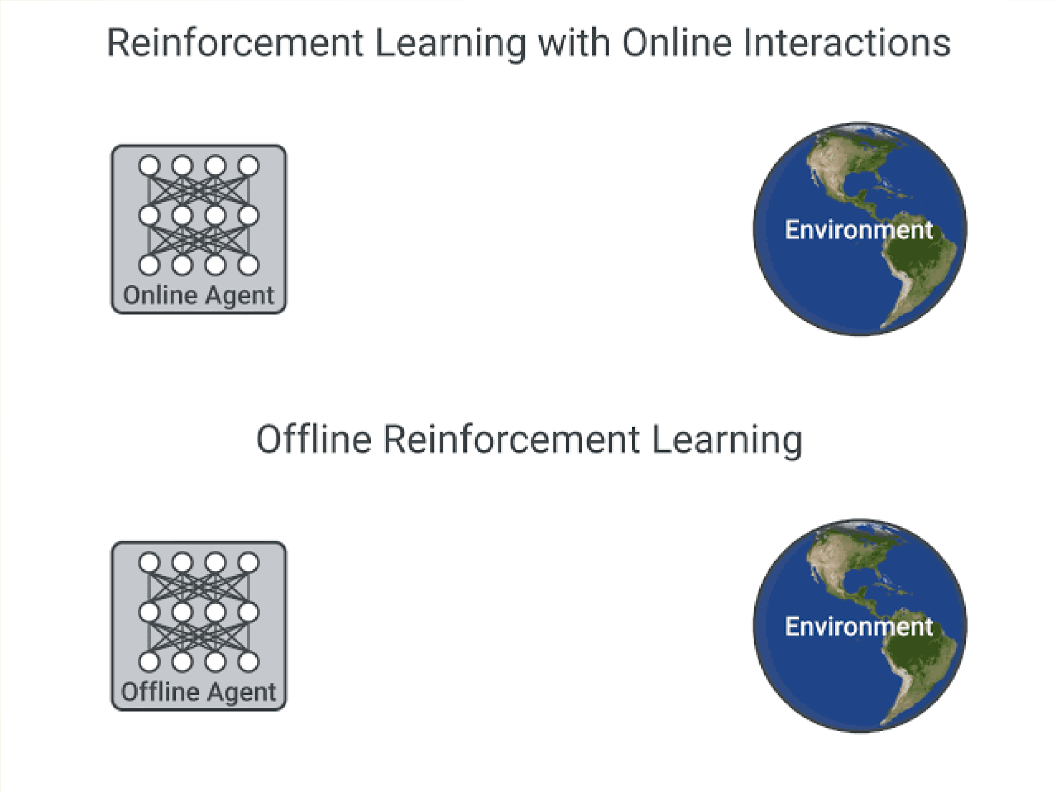
Linear UCB

- $\|\mu - \hat{\mu}_t\|_{\Sigma_t} \leq \beta_t \Leftrightarrow \mu = \hat{\mu}_t + \beta_t \Sigma_t^{-1/2} v$ for some vector v of norm at most 1. Thus
$$\max_{\mu \in C_{t-1}} \langle x, \mu \rangle = \langle x, \hat{\mu}_t \rangle + \beta_t \sqrt{x^{\mathrm{T}} \Sigma_t^{-1} x}$$
- This leads to the linear UCB algorithm.
- **Proposition:** If $\mu_* \in C_{t-1}$, then
$$\langle x_t^*, \mu_* \rangle - \langle x_t, \mu_* \rangle \leq 2\beta_{t-1} \sqrt{x_t^{\mathrm{T}} \Sigma_{t-1}^{-1} x_t}$$

Algorithm 3 LinUCB

$b_0 \leftarrow 0$
for $t \leftarrow 1$ to T **do**
  $\hat{\mu}_{t-1} \leftarrow \Sigma_{t-1}^{-1} b_{t-1}$
  $\beta_{t-1} \leftarrow R\sqrt{\log \frac{\det(\Sigma_{t-1})}{\delta^2 \det(\Sigma_0)}} + \|\Sigma_{t-1}^{-1/2} \Sigma_0\| M$
  $x_t \leftarrow \arg\max_{x \in D_t} \left(\langle x, \hat{\mu}_{t-1} \rangle + \beta_{t-1} \sqrt{x^{\mathsf{T}} \Sigma_{t-1}^{-1} x} \right)$
  Play x_t and observe reward r_t
  $\Sigma_t \leftarrow \Sigma_{t-1} + x_t x_t^{\mathsf{T}}$
  $b_t \leftarrow b_{t-1} + x_t r_t$
end for

Online vs offline RL



https://huggingface.co/learn/deep-rl-course/en/unitbonus3/offline-online