Exploration: UCB-VI

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning
Fall 2022

Today

- Recap
- Why we don’t want to treat MDPs as big contextual bandits
- UCB-VI for tabular MDPs
- UCB-VI for linear MDPs

Recap: UCB

Without context:

For $t = 0,\ldots, T-1$:

Choose the arm with the highest upper confidence bound, i.e.,

$$a_t = \arg \max_{k \in \{1,\ldots,K\}} \hat{\mu}_t^{(k)} + \sqrt{\ln(2TK/\delta)/2N_t^{(k)}}$$

With tabular context ($|\mathscr{X}|$ distinct contexts):

$$\pi_t(x_t) = \arg\max_k \hat{\mu}_t^{(k)}(x_t) + \sqrt{\ln(2TK|\mathscr{X}|/\delta)/2N_t^{(k)}(x_t)}$$

V/Q functions in Finite horizon MDP

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| s_h = s\right]$$

$$Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| (s_h, a_h) = (s,a)\right]$$

Recall: $V_h^\pi(s) \leq H, \qquad Q_h^\pi(s,a) \leq H$

Bellman Consistency Equation:

$$Q_h^\pi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[V_{h+1}^\pi(s')\right]$$

Compute Optimal Policy via VI/DP

VI = DP is a backwards in time approach for computing the optimal policy:

$$\pi^\star = \{\pi_0^\star, \pi_1^\star, \ldots, \pi_{H-1}^\star\}$$

1. Start at $H-1$,

$$Q_{H-1}^\star(s,a) = r(s,a) \qquad \pi_{H-1}^\star(s) = \arg\max_a Q_{H-1}^\star(s,a)$$

$$V_{H-1}^\star = \max_a Q_{H-1}^\star(s,a) = Q_{H-1}^\star(s, \pi_{H-1}^\star(s))$$

2. Assuming we have computed $V_{h+1}^\star$, $h \leq H-2$, i.e., assuming we know how to perform optimally starting at $h+1$, then:

$$Q_h^\star(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V_{h+1}^\star(s')$$

$$\pi_h^\star(s) = \arg\max_a Q_h^\star(s,a), \qquad V_h^\star = \max_a Q_h^\star(s,a)$$

Summary on Finite horizon MDP

$$\mathcal{M} = \{S, A, r, P, H\},$$
$$r : S \times A \mapsto [0,1], \;\; H \in \mathbb{N}, \;\; P : S \times A \mapsto \Delta(S)$$

Comparing to the infinite horizon, discounted MDP:

1. Policy will be time dependent
2. DP takes H steps to compute $\pi^\star$
   - total computation time is $O(H|S|^2|A|)$
   - no need to use contraction argument and no discount factor
3. Extension to non-stationary setting works immediately:
   (i.e. with a non-stationary transition model: $P_0(s'|s,a), P_1(s'|s,a), \ldots P_{H-1}(s'|s,a)$)

Today

- Recap
- Why we don’t want to treat MDPs as big contextual bandits
- UCB-VI for tabular MDPs
- UCB-VI for linear MDPs

Exploration in MDP: make it a bandit and do UCB?

Q: given a discrete MDP, how many unique policies we have?

$$\left(|A|^{|S|}\right)^{H}$$

So treating each policy as an "arm" and running UCB gives us regret $\tilde{O}(\sqrt{|A|^{|S|H}N})$

This seems bad, so are MDPs just super hard or can we do better?

An example of MDP as contextual bandit

$$S = \{a, b\}, \quad A = \{1,2\}, \quad H = 2$$

$|A|^{|S|H} = 2^4 = 16$

All state transitions happen with probability 1/2 for all actions

Reward function: $\begin{aligned} r(a,1) = r(b,1) = 0 \\ r(a,2) = r(b,2) = 1 \end{aligned}$

Suppose we have a lot of data already on a policy $\pi^{(1)}$ that always takes action 1 and a policy $\pi^{(2)}$ that always takes action 2 (note $\pi^{(2)} = \pi^\star$)

What do we know about a policy $\pi^{(3)}$ which always takes action 1 in the first time step, and always takes action 2 at the second time step?

Everything: we have a lot of data on every state-action reward and transition!

If we treat the MDP as a contextual bandit, we treat $\pi^{(3)}$ as a new "arm" about which we know nothing…

Today

- Recap
- Why we don’t want to treat MDPs as big contextual bandits
- UCB-VI for tabular MDPs
- UCB-VI for linear MDPs

UCBVI: Tabular optimism in the face of uncertainty

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n_1, \ldots, \widehat{P}^n_{H-1}$

Design reward bonus $b^n_h(s, a), \forall s, a, h$

Optimistic planning with learned model: $\pi^n = \text{Value-Iter}\left(\{ \widehat{P}^n_h, r_h + b^n_h \}_{h=1}^{H-1} \right)$

Collect a new trajectory by executing π^n in the real world $\{P_h\}_{h=0}^{H-1}$ starting from s_0

Model Estimation

Let us consider the **very beginning** of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h$$

Let's also maintain some statistics using these datasets:

$$N_h^n(s,a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a, h, \quad N_h^n(s,a,s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}, \forall s, a, h$$

Estimate model $\widehat{P}_h^n(s' \mid s,a), \forall s, a, s', h$:

$$\widehat{P}_h^n(s' \mid s,a) = \frac{N_h^n(s,a,s')}{N_h^n(s,a)}$$

Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h, \quad N_h^n(s,a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a, h,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\log\left(SAHN/\delta\right)}{N_h^n(s,a)}}$$

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\{\widehat{P}_h^n\}_h$ **and** $\{r_h + b_h^n\}_h$

$$\widehat{V}_H^n(s) = 0, \forall s \qquad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, \quad H \right\}, \forall s, a$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s \qquad \left\| \widehat{V}_h^n \right\|_\infty \leq H, \forall h, n$$

UCBVI: Put All Together

For $n = 1 \to N$:

1. Set $N_h^n(s,a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s,a,h$

2. Set $N_h^n(s,a,s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}, \forall s,a,a',h$

3. Estimate $\widehat{P}^n$: $\widehat{P}_h^n(s' \mid s,a) = \dfrac{N_h^n(s,a,s')}{N_h^n(s,a)}, \forall s,a,s',h$

4. Plan: $\pi^n = VI\left(\{\widehat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s,a) = cH\sqrt{\dfrac{\log(SAHN/\delta)}{N_h^n(s,a)}}$

5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

High-level Idea: Exploration Exploitation Tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

1. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$ is small?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

2. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$ is large?

Not obvious

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \widehat{V}_{h+1}^n \right]$$ must be large

We collect data at steps where bonus is large or model is wrong, i.e., exploration

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N} \left(V^\star - V^{\pi^n} \right) \right] \leq \widetilde{O}\left(H^2 \sqrt{SAN} \right)$$

Today

- Recap
- Why we don’t want to treat MDPs as big contextual bandits
- UCB-VI for tabular MDPs
- UCB-VI for linear MDPs

Linear MDP Definition

Finite horizon time-dependent episodic MDP $\mathscr{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

S & A could be large or even continuous, hence poly(S,A) is not acceptable

$$P_h(s' \mid s, a) = \mu_h^\star(s') \cdot \phi(s, a), \quad \mu_h^\star \in S \mapsto \mathbb{R}^d, \phi \in S \times A \mapsto \mathbb{R}^d$$

$$r(s, a) = \theta_h^\star \cdot \phi(s, a), \quad \theta_h^\star \in \mathbb{R}^d$$

Feature map ϕ is known to the learner!

(We assume reward is known, i.e., $\theta^\star$ is known)

Planning in Linear MDP: Value Iteration

$$P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a), \quad \mu_h^\star \in \mathbb{R}^{S\times d}, \phi(s, a) \in \mathbb{R}^d$$

$$r_h(s, a) = (\theta_h^\star)^\top \phi(s, a), \quad \theta_h^\star \in \mathbb{R}^d$$

$$V_H^\star(s) = 0, \forall s,$$

$$Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s, a) + \left(\mu_h^\star \phi(s, a)\right)^\top V_{h+1}^\star$$

$$= \phi(s, a)^\top \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right)$$

$$= \phi(s, a)^\top w_h$$

$$V_h^\star(s) = \max_a \phi(s, a)^\top w_h, \quad \pi_h^\star(s) = \arg\max_a \phi(s, a)^\top w_h$$

Indeed we can show that $Q_h^\pi(\cdot, \cdot)$

Is linear with respect to ϕ as well, for any π, h

UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\widehat{P}_h^n\}_{h=0}^{H-1}$ from all previous data $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=0}^{n-1}$

2. Design reward bonus $b_h^n(s, a), \forall s, a$

3. Plan: $\pi^{n+1} =$ Value-Iter $\left(\{\widehat{P}^n\}_h, \{r_h + b_h^n\}\right)$

How to estimate $\{\widehat{P}_h^n\}_{h=0}^{H-1}$?

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

Given s, a, note that $\mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\delta(s')\right] = P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$

Penalized Linear Regression:

$$\min_\mu \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$$

$$A_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I \qquad \widehat{\mu}_h^n = (A_h^n)^{-1} \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top$$

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

How to choose $b_h^n(s,a)$?

Chebyshev-like approach, similar to in linUCB:

$$b_h^n(s,a) = \beta\sqrt{\phi(s,a)^\top (A_h^n)^{-1}\phi(s,a)}, \quad \beta = \widetilde{O}(dH)$$

linUCB-VI: Put All Together

For $n = 1 \to N$:

1. Set $A_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$
2. Set $\widehat{\mu}_h^n = (A_h^n)^{-1} \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top$
3. Estimate $\widehat{P}^n$: $\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$
4. Plan: $\pi^n = VI\left(\{\widehat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s, a) = cdH\sqrt{\phi(s, a)^\top (A_h^n)^{-1} \phi(s, a)}$
5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2 d^{1.5}\sqrt{N}\right)$$

No S, A dependence!

Today

- ✓ Recap
- ✓ Why we don't want to treat MDPs as big contextual bandits
- ✓ UCB-VI for tabular MDPs
- ✓ UCB-VI for linear MDPs

Today's summary:

UCB-VI algorithm for tabular MDPs

- Uses UCB idea, but leverages MDP structure

1-minute feedback form: https://bit.ly/3RHtlxy

Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \le \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a)) \cdot \widehat{V}^n_{h+1} \right]$$

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$$

$$\le r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \left(\widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \right) \cdot \widehat{V}^n_1 + P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \left(\widehat{V}^n_1 - V^{\pi^n}_1 \right)$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a)) \cdot \widehat{V}^n_{h+1} \right]$$