Minimax regret upper bounds of UCBVI for RL

Group Study and Seminar Series (Summer 20)

Yingru Li

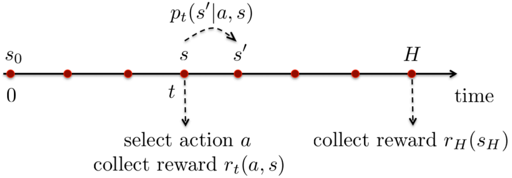
The Chinese University of Hong Kong, Shenzhen, China

July 30, 2020

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. ICML (pp. 263-272).

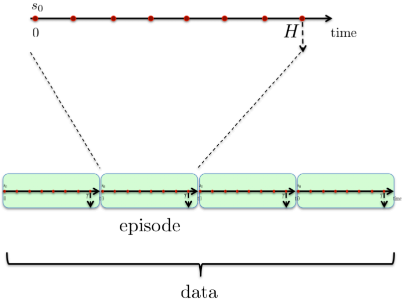
Outline

Finite-horizon episodic RL problems



- Initial state x_1 (could be a r.v.)
- Transition probabilities at time step $h : p\left(y \mid x, a\right)$
- Reward at time step $h : r(x, a)$
- Unknown transition probabilities and reward function
- Objective: quickly learn a policy $\pi^\star$ maximizing over $\pi := \{\pi_1, \pi_2, \cdots, \pi_H\}$

$$V_1^\pi(s) := \mathbb{E}\left[\sum_{h=1}^{H} r\left(s_h, \pi_h(s_h)\right) \mid s_1 = s\right]$$

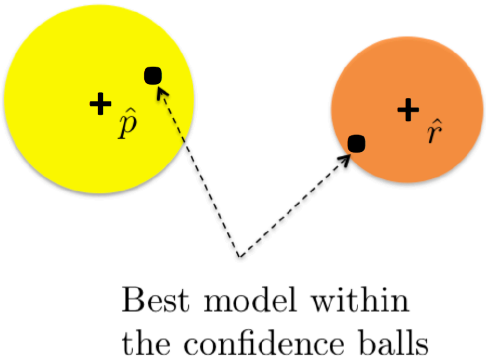


- Data: K episodes of length H (actions, states, rewards)
- Learner: 'the data on previous $K-1$ episodes' $\mapsto \pi_K$
- Performance of the learner: how close π_K is from the optimal policy $\pi^\star$ or **regret up to the K-th episode (time $T = KH$)**:

$$Regret(K) = \sum_{k=1}^{K} \left(V_1^\star\left(x_{k,1}\right) - V_1^{\pi_k}\left(x_{k,1}\right)\right)$$

Algorithm Principle: Optimism face of uncertainty

- *Estimate the unknown system parameters (here $p(\cdot \mid \cdot, \cdot)$ and $r(\cdot, \cdot)$) and build an optimistic reward estimate to trigger exploration.*
- **Estimate**: find confidence balls containing the true model w.h.p.
- **Optimistic reward estimate**: find the model within the confidence balls leading to the highest value.



Outline

UCBVI: Upper Confidence Bound Value Iteration

- UCBVI is an extension of Value Iteration, guaranteeing that **the resulting value function is a (high-probability) upper confidence bound (UCB) on the optimal value function** V^*.
  - At the beginning of episode k, it computes state-action values using empirical transition kernel and reward function.
  - In step h of backward induction (to update $Q_{k,h}(s,a)$ for any (s,a)) , it adds a bonus $b_{k,h}(s,a)$ to the value, and ensures that $Q_{k,h} \leq Q_{k-1,h}$.
- Two variants of UCBVI, depending on the choice of bonus $b_{k,h}$
  - UCBVI-CH using Chernoff-Hoeffding bound
  - UCBVI-BF using Bernstein-Freedman bound
- As more data gathered, the upper confidence bound on the optimal value of initial state get close to the true optimal value.

UCBVI algorithm

Variables to be maintained by the algorithm: for known deterministic reward function

- $\hat{p} = (\hat{p}(s' \mid s, a), s, s' \in \mathcal{S}, a \in \mathcal{A}_s)$: estimated transition probabilities
- $Q = (Q_h(s,a), h \leq H, s \in \mathcal{S}, a \in \mathcal{A}_s)$: estimated Q-function
- $b = (b_h(s,a), h \leq H, s \in \mathcal{S}, a \in \mathcal{A}_s)$: Q-value bonus
- $N = (N(s,a), s \in \mathcal{S}, a \in \mathcal{A}_s)$: number of visits to (s,a) so far
- $N' = (N_h(s,a), h \leq H, s \in \mathcal{S}, a \in \mathcal{A}_s)$: number of visits in the h-step of episodes to (s,a) so far

UCBVI

Algorithm. UCB-VI

Input: Initial state distribution ν_0, precision δ

Initialise the variables $\hat{p}$, N, and N'

For episode $k = 1, 2, \ldots$

1. Optimistic reward:
   a. Compute the bonus: $b \leftarrow \text{bonus}(N, N', \hat{p}, Q, \delta)$
   b. Estimate the Q-function: $Q \leftarrow \text{bellmanOpt}(Q, b, \hat{p})$
2. Initialise the state $s(0) \sim \nu_0$
3. for $h = 1, \ldots, H$, select action
   $a \in \arg\max_{a' \in \mathcal{A}_{s(h-1)}} Q_h(s(h-1), a')$
4. Observe the transition and update $\hat{p}$, N, and N'

UCBVI algorithm: bonus

- UCBVI-CH:

$$b_h(s,a) = \frac{7H}{\sqrt{N(s,a)}} \log(5SAT/\delta)$$

- UCBVI-BF:

$$b_h(s,a) = \sqrt{\frac{8L}{N(s,a)} \operatorname{Var} \widehat{p}(\cdot \mid s,a)\left(V_{h+1}(Y)\right) + \frac{14HL}{3N(s,a)}} + \sqrt{\frac{8}{N(s,a)} \sum_y \widehat{p}(y \mid s,a) \min\left\{\frac{10^4 H^3 S^2 A L^2}{N'_{h+1}(y)}, H^2\right\}}$$

where $L = \log(5SAT/\delta)$.

UCBVI algorithm: Optimistic Bellman operator

$\mathrm{bellmanOpt}(Q, b, \hat{p})$ applies Dynamic Programming with a bonus.

- **Initialization**: $V_{H+1}(s) = 0$ for all (s, a)
- **For step** $h = H, \dots, 1$:
  - for all (s, a) never visited: $Q_h(s, a) = H$
  - for all (s, a) visited at least once so far:
    $Q_h(s,a) \leftarrow \min \Big(Q_h(s,a), H, r(s,a) + \sum_y \hat{p}(y \mid s,a) V_{h+1}(y) + b_h(s,a) \Big)$
  - $V_h(s) = \max_{a \in \mathcal{A}} Q_h(s, a)$
- Q-values $Q_1, Q_2, \cdots, Q_H$

Outline

UCBVI: Regret guarantees

Regret up to time $T = KH$: $\mathrm{Regret}(K) = \sum_{k=1}^{K} \left(V_1^{\star}\left(x_{k,1}\right) - V_1^{\pi_k}\left(x_{k,1}\right)\right)$

Theorem 1.

For any $\delta > 0$, the regret of UCBVI-CH(δ) is bounded w.p. at least $1 - \delta$ by:

$$\mathrm{Regret}^{UCBVI-CH}(K) \leq 20HL\sqrt{SAT} + 250H^2S^2AL^2$$

with $L = \log(5HSAT/\delta)$.

- For $T \geq HS^3A$ and $SA \geq H$, the regret upper bound scales as $\tilde{\mathcal{O}}(H\sqrt{SAT})$

UCBVI: Regret guarantees

Regret up to time $T = KH$: $\mathrm{Regret}(K) = \sum_{k=1}^{K} \left(V_1^{\star}(x_{k,1}) - V_1^{\pi_k}(x_{k,1}) \right)$

Theorem 2.

Consider a parameter $\delta > 0$. Then the regret of UCBVI-BF(δ) is bounded w.p. $1 - \delta$, by

$$\mathrm{Regret}^{UCBVI-BF}(K) \leq 30HL\sqrt{SAK} + 2500H^2S^2AL^2 + 4H^{3/2}\sqrt{KL}$$

where $L = \ln(5HSAT/\delta)$

- For $T \geq H^3S^3A$ and $SA \geq H$, the regret upper bound scales as $\tilde{\mathcal{O}}(\sqrt{HSAT})$
- Achieve regret minimax lower bound

Sketch of proof

Some notations:

- π_k is the policy applied by UCBVI in the k-th episode
- $V_{k,h}$ is the optimistic value function computed by UCBVI in the h -step of the k -th episode
- V_h^π is the value function from step h under π
- $P^\pi = \left(p\left(s' \mid s, \pi(s)\right)\right)_{s,s'}$
- $\hat{P}_k^\pi = \left(\hat{p}_k\left(s' \mid s, \pi(s)\right)\right)_{s,s'}$ where $\hat{p}_k$ is the estimated transitions in episode k

Claim 1: by construction with high probability, $V_{k,h} \geq V_h^\star$. Then:

$$\mathrm{Regret}(K) \leq \widetilde{\mathrm{Regret}}(K) = \sum_{k=1}^{K} \left(V_{k,1}\left(x_{k,1}\right) - V^{\pi_k}\left(x_{k,1}\right)\right)$$

Sketch of proof: Key error decomposition

- Let $\tilde{\Delta}_{k,h} = V_{k,h} - V_h^{\pi_k}$, so that $\widetilde{\mathrm{Regret}}(K) = \sum_{k=1}^{K} \tilde{\Delta}_{k,1}\left(x_{k,1}\right)$
- Backward induction on h to bound $\tilde{\Delta}_{k,1}$: introduce $\tilde{\delta}_{k,h} = \tilde{\Delta}_{k,h}\left(x_{k,h}\right)$ then

$$\tilde{\delta}_{k,h} \leq \left(\hat{P}_k^{\pi_k} - P^{\pi_k}\right) \tilde{\Delta}_{k,h+1}\left(x_{k,h}\right) + \tilde{\delta}_{k,h+1} + \epsilon_{k,h} + b_{k,h} + e_{k,h} \qquad (1)$$

where

$$\begin{cases} \epsilon_{k,h} = P^{\pi_k} \tilde{\Delta}_{k,h+1}\left(x_{k,h}\right) - \tilde{\Delta}_{k,h+1}\left(x_{k,h+1}\right) \\ e_{k,h} = \left(\hat{P}_k^{\pi_k} - P^{\pi_k}\right) V_{h+1}^{\star}\left(x_{k,h}\right) \end{cases}$$

- Concentration + Martingale difference (Azuma-Hoeffding or Bernstein-Freedman) + bounding bonus

Key error decomposition: How and why?

- By algorithm,

$$V_{k,h}(x) = \max_a Q_{k,h}(x,a) \equiv \min\left\{Q_{k-1,h}(x,a), H, r_h(x,a) + [\hat{P}_k V_{k,h+1}](x,a) + b_{k,h}(x,a)\right\},$$

- and we define empirical optimistic bellman operator

$$[\mathcal{T}_{k,h} V_{k,h+1}](x) = \max_a \{r_h(x,a) + [\hat{P}_k V_{k,h+1}](x,a) + b_{k,h}(x,a)\}, \quad \forall x.$$

- Then, we can also write $V_{k,h}(x) = \min\{V_{k-1,h}(x), H, [\mathcal{T}_{k,h} V_{k,h+1}](x)\}$.

Key error decomposition: How and why?

- For simplicity, ignore the subscript k.
- With $\pi(x_h) = a_h$, $b_h = b_h(x_h, \pi(x_h))$, $n_h = N(x_h, \pi(x_h))$ we have the following important decomposition

$$
\begin{aligned}
\tilde{\delta}_h &= V_h(x_h) - V_h^\pi(x_h) = [\mathcal{T}_h V_{h+1}](x_h) - [\mathcal{T}_h^\pi V_{h+1}^\pi](x_h) = [\hat{P}^\pi V_{h+1}](x_h) + b_h - [P^\pi V_{h+1}^\pi](x_h) \\
&= b_h + \underbrace{[(\hat{P}^\pi - P^\pi)V_{h+1}](x_h)}_{\text{Two dependent random variable, could be bound as } \|\hat{P}^\pi - P^\pi\|_1 \|V_{h+1}\|_\infty \text{, bad bound}} \\
&\quad + [P^\pi(V_{h+1} - V_{h+1}^\pi)](x_h) \\
&= b_h + \underbrace{[(\hat{P}^\pi - P^\pi)V_{h+1}^*](x_h)}_{e_h} + \underbrace{[(\hat{P}^\pi - P^\pi)(V_{h+1} - V_{h+1}^*)](x_h)}_{(a)} \\
&\quad + \underbrace{[P^\pi(V_{h+1} - V_{h+1}^\pi)](x_h) - [V_{h+1} - V_{h+1}^\pi](x_{h+1})}_{\text{Martingale difference } \epsilon_h} + \underbrace{[V_{h+1} - V_{h+1}^\pi](x_{h+1})}_{\tilde{\delta}_{h+1}}
\end{aligned}
$$

Key error decomposition: bounding (a)

$$
\begin{aligned}
(a) &= \sum_{y\in\mathcal{S}} \left(\hat{P}^{\pi}(y|x_h) - P^{\pi}(y|x_h)\right)\left(V_{h+1}(y) - V^*_{h+1}(y)\right) \\
&\overset{(I)}{\leq} \sum_{y\in\mathcal{S}} \left[2\sqrt{\frac{p_h(y)(1-p_h(y))L}{n_h}} + \frac{4L}{3n_h}\right]\tilde{\Delta}_{h+1}(y) \\
&\leq 2\sqrt{L}\underbrace{\sum_{y\in\mathcal{S}} \sqrt{\frac{p_h(y)}{n_h}}\tilde{\Delta}_{h+1}(y)}_{(b)} + \frac{4SL}{3n_h},
\end{aligned}
$$

Key error decomposition: bounding (b)

- Typical set:

$$[y]_{k,x,a} := \{y | y \in \mathcal{S}, N_k(x,a)P(y|x,a) \geq \mathcal{O}(1) \cdot LH^2\}$$

-

$$(b) = \underbrace{\sum_{y \in [y]_h} \sqrt{\frac{p_h(y)}{n_h}} \tilde{\Delta}_{h+1}(y)}_{(c)} + \underbrace{\sum_{y \notin [y]_h} \sqrt{\frac{p_h(y)}{n_h}} \tilde{\Delta}_{h+1}(y)}_{(d)}$$

- Now we define another Martingale difference sequence under the typical set,

$$\tilde{\Delta}_{\text{typ},k,h+1}(y) \equiv \sqrt{\frac{\mathbb{I}_{k,h}(y)}{n_{k,h} p_{k,h}(y)}} \tilde{\Delta}_{k,h+1}(y), \forall y \in \mathcal{S},$$

$$\bar{\varepsilon}_{k,h} \equiv [P_h^{\pi_k} \tilde{\Delta}_{\text{typ},k,h+1}](x_{k,h}) - \tilde{\Delta}_{\text{typ},k,h+1}(x_{k,h+1}),$$

Key error decomposition: bounding (c) and (d)

Then the term (c) can be bounded as,

$$
\begin{aligned}
(c) &= \sum_{y \in [y]_h} P^{\pi}(y|x_h) \sqrt{\frac{1}{n_h p_h(y)}} \tilde{\Delta}_{h+1}(y) = \bar{\varepsilon}_h + \sqrt{\frac{\mathbb{I}(x_{h+1} \in [y]_h)}{n_h p_h(x_{h+1})}} \tilde{\delta}_{h+1} \\
&\leq \bar{\varepsilon}_h + \mathcal{O}(1) \cdot \sqrt{\frac{1}{LH^2}} \tilde{\delta}_{h+1},
\end{aligned}
\tag{2}
$$

$$
(d) = \sum_{y \notin [y]_h} \sqrt{\frac{p_h(y) n_h}{n_h^2}} \tilde{\Delta}_{h+1}(y) \leq \mathcal{O}(1) \cdot \frac{S\sqrt{LH^2}}{n_h}
\tag{3}
$$

Then, we deduce,

$$
(b) \leq \mathcal{O}(1) \cdot \frac{S\sqrt{LH^2}}{n_h} + \mathcal{O}(1) \cdot \sqrt{\frac{1}{LH^2}} \tilde{\delta}_{h+1} + \bar{\varepsilon}_h,
\tag{4}
$$

Key error decomposition: Implications

Then we have,

$$(a) \leq \underbrace{\frac{SHL}{n_h} + \frac{SL}{n_h}}_{\equiv c_{4,h}} + \frac{1}{H}\tilde{\delta}_{h+1} + 2\sqrt{L}\bar{\varepsilon}_h$$

Combine with the above,

$$\begin{aligned}\tilde{\delta}_h &\leq \varepsilon_h + 2\sqrt{L}\bar{\varepsilon}_h + b_h + c_{1,h} + c_{4,h} + (1 + \frac{1}{H})\tilde{\delta}_{h+1} \\ &\leq e \sum_{i=h}^{H-1} \left(\varepsilon_i + 2\sqrt{L}\bar{\varepsilon}_i + c_{1,i} + c_{4,i} + b_i\right),\end{aligned}$$

Implications on regret

Corollary 3.

Let $k \in [K]$ and $h \in [H]$. With high probability,

$$\mathrm{Regret}(k) = \sum_{i=1}^{k} \delta_{i,1} \leq \sum_{i=1}^{k} \tilde{\delta}_{i,1} \leq e \sum_{i=1}^{k} \sum_{j=1}^{H-1} \left[\varepsilon_{i,j} + 2\sqrt{L}\bar{\varepsilon}_{i,j} + b_{i,j} + c_{1,i,j} + c_{4,i,j} \right]$$