Exploration in Tabular MDPs

CS 6789: Foundations of Reinforcement Learning

Recap: UCBVI

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 $b_h^n(s,a) = cH\sqrt{\dfrac{\ln(SAHN/\delta)}{N_h^n(s,a)}}$

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

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)\right) \cdot V_{h+1}^\star\right)$

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)\right) \cdot V_{h+1}^\star\right)$

VI with bonus inside the learned model gives optimism, i.e., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V_{h+1}^\star \right)$

VI with bonus inside the learned model gives optimism, i.e., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

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

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)\right) \cdot V_{h+1}^\star\right)$

VI with bonus inside the learned model gives optimism, i.e., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

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

Apply simulation lemma: $\widehat{V}_0^n(s_0) - V^{\pi^n}(s_0)$

1. Model Error using Hoeffing's inequality & Union Bound

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

1. Model Error using Hoeffing's inequality & Union Bound

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

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1 - \delta$:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right)^\top f \right| \leq O(H\sqrt{\ln(\underline{SAHN/\delta})/N_h^n(s, a)}), \forall s, a, h, N$$

V^*

1. Model Error using Hoeffing's inequality & Union Bound

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

Given a fixed function $f : S \mapsto [0, H]$, w/ prob $1 - \delta$:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N_h^n(s, a)}), \forall s, a, h, N$$

Bonus $b_h^n(s, a)$

1. Model Error using Hoeffing's inequality & Union Bound

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

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1 - \delta$:

$$\left| \left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N_h^n(s,a)}), \forall s, a, h, N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

1. Model Error using Hoeffing's inequality & Union Bound

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

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}_h^n(\cdot\,|s,a) - P_h(\cdot\,|s,a) \right)^\top f \right| \leq O\left(H\sqrt{\ln(SAHN/\delta)/N_h^n(s,a)}\right), \forall s, a, h, N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

Intuition:

1. Model Error using Hoeffing's inequality & Union Bound

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

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N_h^n(s,a)}), \forall s, a, h, N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

Intuition:

1. Assume for some i, $s_h^i = s, a_h^i = a$, then $f(s_{h+1}^i)$ is an unbiased estimate of $\mathbb{E}_{s'\sim P_h(\cdot|s,a)} f(s')$

1. Model Error using Hoeffing's inequality & Union Bound

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

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N_h^n(s,a)}), \forall s, a, h, N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

Intuition:

1. Assume for some i, $s_h^i = s, a_h^i = a$, then $f(s_{h+1}^i)$ is an unbiased estimate of $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} f(s')$

2. Note $\widehat{P}_h^n(\cdot|s,a) \cdot f = \frac{1}{N_h^n(s,a)} \sum_{i=1}^{n-1} \mathbf{1}[(s_h^i, a_h^i) = (s,a)] f(s_{h+1}^i)$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

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

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

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

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \quad \forall s$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}^n_h(s) \geq V^\star_h(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

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

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

Inductive hypothesis: $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s), \quad \forall s$

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = \cancel{r_h(s,a)} + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - \cancel{r_h(s,a)} - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}^n_h(s) \geq V^\star_h(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

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

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

Inductive hypothesis: $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s), \quad \forall s$

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = \cancel{r_h(s,a)} + b^n_h(s,a) + \widehat{P}^n_h(\cdot \,|\, s,a) \cdot \widehat{V}^n_{h+1} - \cancel{r_h(s,a)} - P_h(\cdot \,|\, s,a) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) + \widehat{P}^n_h(\cdot \,|\, s,a) \cdot V^\star_{h+1} - P_h(\cdot \,|\, s,a) \cdot V^\star_{h+1}$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}^n_h(s) \geq V^\star_h(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

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

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

Inductive hypothesis: $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s), \quad \forall s$

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - r_h(s,a) - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot V^\star_{h+1} - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$= b^n_h(s,a) + \underline{\left| \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V^\star_{h+1} \right|} \leq b^n_h(s,a)$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}^n_h(s) \geq V^\star_h(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

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

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

$\widehat{V}(s) = \max_a \widehat{Q}(s,a)$

$\geq \widehat{Q}(s, \pi^\star(s))$

$\geq Q^\star(s, \pi^\star(s))$

$= V^\star(s)$

Inductive hypothesis: $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s), \quad \forall s$

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - r_h(s,a) - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot V^\star_{h+1} - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$= b^n_h(s,a) + \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) - b^n_h(s,a) = 0, \quad \forall s, a$$

3. Upper Bounding Regret using Optimism

from VI

$$\text{per-episode regret} := V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$$

This is something
we can control!
And this is related
to our policy π^n

4. Upper bounding Regret via Simulation Lemma

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

π^n inside (P, r)

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

Lemma [Simulation lemma]:

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

Value of π^n inside $(\widehat{P}, r+b)$

4. Upper bounding Regret via 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]:

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

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

4. Upper bounding Regret via 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]:

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

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

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

4. Upper bounding Regret via 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]:

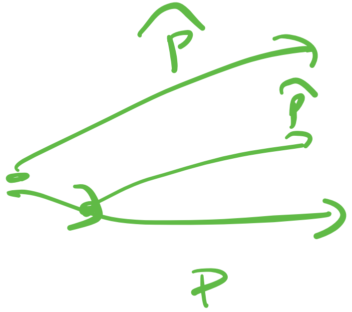
$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \le \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a)\right) \cdot \widehat{V}^n_{h+1} \right]$$

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

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

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

$$+ P \cdot \widehat{V} - P\widehat{V}$$



4. Upper bounding Regret via 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]:

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

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

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

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

$$= \underbrace{b^n_h(s_0, \pi^n(s_0))}_{①} + \underbrace{\left(\widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \right) \cdot \widehat{V}^n_1}_{②} + P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \left(\widehat{V}^n_1 - V^{\pi^n}_1 \right)$$

4. Upper bounding Regret via 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]:

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

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

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

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

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

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

4. Upper bounding Regret via Simulation Lemma

optimism $\widehat{V} \geq V^\star$

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

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

$\mathcal{D}$

depends on $\mathcal{D}$

4. Upper bounding Regret via Simulation Lemma

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

But $\widehat{V}_h^n$ is data-dependent

(this is different from $V_h^\star$) !!!

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

Let's do Holder's inequality

$$\left.\begin{array}{l} \underline{(\hat{P} - P)^T(\hat{V} - V^\star)} \\ + \underline{(\hat{P} - P)^T \cdot V^\star} \end{array}\right\} \Rightarrow \sqrt{S}$$

$$\left(\hat{P}(\cdot|s,a) - P(\cdot|s,a)\right)^T V^\star$$

4. Upper bounding Regret via Simulation Lemma

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

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

But $\widehat{V}_h^n$ is data-dependent
(this is different from $V_h^\star$) !!!

Let's do Holder's inequality

$$\left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)\right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$\leq H$

$a^\top b \leq \|a\|_1 \|b\|_\infty$

4. Upper bounding Regret via Simulation Lemma

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

But $\widehat{V}_h^n$ is data-dependent (this is different from $V_h^\star$) !!!

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

Let's do Holder's inequality

$$\left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)\right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$\leq H\|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \leq H\sqrt{\dfrac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}$, $\forall s,a,h,n$, with prob $1-\delta$

4. Upper bounding Regret via Simulation Lemma

But $\widehat{V}^n_h$ is data-dependent
(this is different from $V^\star_h$) !!!

per-episode regret $:= V^\star_0(s_0) - V^{\pi_n}_0(s_0) \leq \widehat{V}^n_0(s_0) - V^{\pi_n}_0(s_0)$

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

Let's do Holder's inequality

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

$H\sqrt{\frac{\ln(\ldots)}{N_h^n(s,a)}}$

$$\left(\widehat{P}^n_h(\cdot|s,a) - P_h(\cdot|s,a)\right) \cdot \widehat{V}^n_{h+1} \leq \|P_h(\cdot|s,a) - \widehat{P}^n_h(\cdot|s,a)\|_1 \|\widehat{V}^n_{h+1}\|_\infty$$

$\leq H\|P_h(\cdot|s,a) - \widehat{P}^n_h(\cdot|s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n,$ with prob$1-\delta$

4. Upper bounding Regret via Simulation Lemma

But $\widehat{V}_h^n$ is data-dependent
(this is different from $V_h^\star$) !!!

$$\text{per-episode regret} := V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$$

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

Let's do Holder's inequality

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

$$\leq 2\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

$$\left(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a)\right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot\,|\,s,a) - \widehat{P}_h^n(\cdot\,|\,s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$$\leq H\|P_h(\cdot\,|\,s,a) - \widehat{P}_h^n(\cdot\,|\,s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{with prob} 1-\delta$$

4. Upper bounding Regret via Simulation Lemma

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

But $\widehat{V}_h^n$ is data-dependent
(this is different from $V_h^\star$) !!!

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

Let's do Holder's inequality

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

$$\leq 2\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right] = 2H\sqrt{S\ln(SAHN/\delta)} \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[\sqrt{\frac{1}{N_h^n(s,a)}} \right]$$

$$\left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)\right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$$\leq H\|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{with prob} 1-\delta$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\mathbb{E}\left[\text{Regret}_N\right] = \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}\{\text{events don't hold}\}\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\mathbb{E}\left[\text{Regret}_N\right] = \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}\{\text{events don't hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{P}(\text{events don't hold}) \cdot NH$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\mathbb{E}\left[\text{Regret}_N\right] = \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}\{\text{events don't hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{P}(\text{events don't hold}) \cdot NH$$

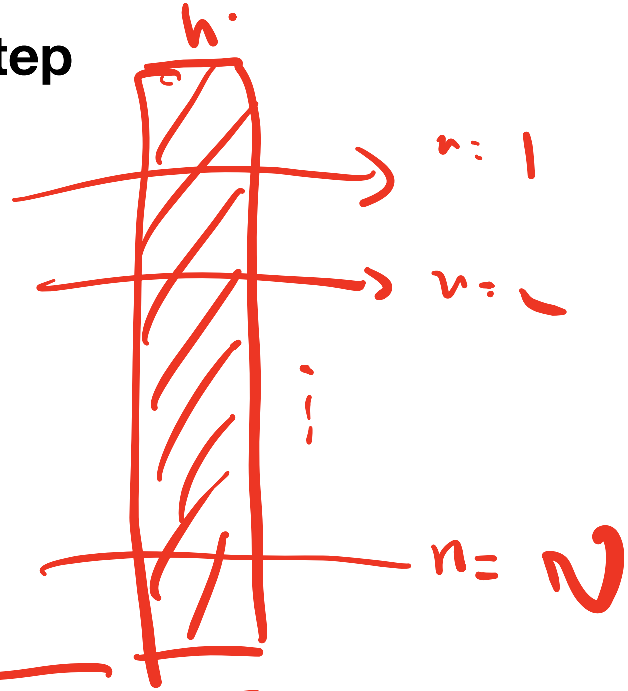
$$\leq H\sqrt{S\ln(SANH/\delta)}\,\mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}}\right] + 2\delta NH$$

5. Final Step

$$\sum_{n=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}}$$

$$= \sum_{h=0}^{H-1} \left[\underbrace{\sum_{n=1}^{N} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}}} \right]$$

$$= \sum_{h} \left[\sum_{s,a} \underbrace{\sum_{n=1}^{N} \mathbb{1}(s_h^n, a_h^n = s,a) \frac{1}{\sqrt{N_h^n(s,a)}}}_{\le \sqrt{N_h^n(s,a)}} \right]$$



5. Final Step

$$\sum_{n=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}} = \sum_{h=0}^{H-1} \sum_{s,a} \sum_{i=1}^{N_h^N(s,a)} \frac{1}{\sqrt{i}} \leq \sqrt{N_h^N(s,a)}$$

5. Final Step

$$\sum_{n=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}} = \sum_{h=0}^{H-1} \sum_{s,a} \sum_{i=1}^{N_h^N(s,a)} \frac{1}{\sqrt{i}} \leq \sum_{h=0}^{H-1} \sum_{s,a} \sqrt{N_h^N(s,a)}$$

5. Final Step

$$\sum_{n=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}} = \sum_{h=0}^{H-1} \sum_{s,a} \sum_{i=1}^{N_h^N(s,a)} \frac{1}{\sqrt{i}} \leq \sum_{h=0}^{H-1} \sum_{s,a} \sqrt{N_h^N(s,a)} \leq \sum_{h=0}^{H-1} \sqrt{SA \sum_{s,a} N_h^N(s,a)}$$

5. Final Step

$$\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}} = \sum_{h=0}^{H-1}\sum_{s,a}\sum_{i=1}^{N_h^N(s,a)}\frac{1}{\sqrt{i}} \leq \sum_{h=0}^{H-1}\sum_{s,a}\sqrt{N_h^N(s,a)} \leq \sum_{h=0}^{H-1}\sqrt{SA\sum_{s,a}N_h^N(s,a)}$$

N

$$\leq \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

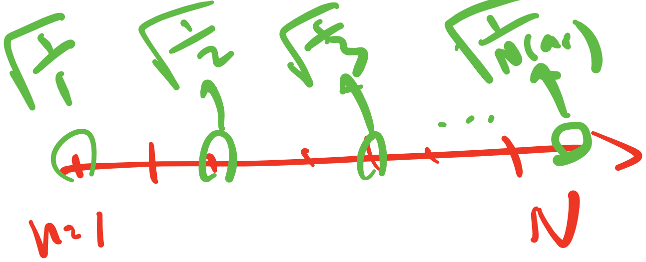
5. Final Step

$$\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}} = \sum_{h=0}^{H-1}\sum_{s,a}\sum_{i=1}^{N_h^N(s,a)}\frac{1}{\sqrt{i}} \leq \sum_{h=0}^{H-1}\sum_{s,a}\sqrt{N_h^N(s,a)} \leq \sum_{h=0}^{H-1}\sqrt{SA\sum_{s,a}N_h^N(s,a)}$$

$$\leq \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

$$\mathbb{E}\left[\text{Regret}_N\right] \leq \underline{2H^2S}\sqrt{\underline{AN\ln(SAHN/\delta)}} + \underline{2\delta NH}$$

failure case



5. Final Step

$$\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n,a_h^n)}} = \sum_{h=0}^{H-1}\sum_{s,a}\sum_{i=1}^{N_h^N(s,a)}\frac{1}{\sqrt{i}} \le \sum_{h=0}^{H-1}\sum_{s,a}\sqrt{N_h^N(s,a)} \le \sum_{h=0}^{H-1}\sqrt{SA\sum_{s,a}N_h^N(s,a)}$$

$$\le \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

≈ 1

$$\mathbb{E}\left[\text{Regret}_N\right] \le 2H^2S\sqrt{AN\ln(SAHN/\delta)} + 2\delta NH$$

Set $\delta = 1/(HN)$

$$\le 2H^2S\sqrt{AN\cdot\ln(SAH^2N^2)} = \widetilde{O}\left(H^2S\sqrt{AN}\right)$$

High-level Idea: Exploration or 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 $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

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

High-level Idea: Exploration or 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 $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

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

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

High-level Idea: Exploration or 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 $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

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

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

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

High-level Idea: Exploration or 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 $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

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

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

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

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

Summary of the Proof of UCB-VI

Bonus $b^n(s, a)$ is related to $\left(\left(\widehat{P}^n_h(\cdot \mid s, a) - P_h(\cdot \mid s, a)\right) \cdot V^\star_{h+1}\right)$

Prove Optimism via Induction: it allows us to focus π^n rather than $\pi^\star$ which is unknown)

Bound per-episode regret via Simulation Lemma (Perf diff of π^n under $\widehat{P}^n$ & P)

Bounding conf-term along traces: $\sum_n \sum_h \sqrt{\frac{1}{N^n_h(s^n_h, a^n_h)}}$