Exploration in Tabular MDPs:

Upper Confidence Bound Value Iteration (UCBVI)

Announcements

1. Scribing Lecture Notes (see Piazza for details)

2. Course Project Website
https://wensun.github.io/CS6789projects.html

Recap:

Generative Model and Statistical Limits

Theorem: (Azar et al. '13) With probability greater than $1-\delta$,

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \gamma \sqrt{\frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N} + \frac{c\gamma}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}},$$

where c is an absolute constant.

Corollary: for $\epsilon < 1$, provided $N \geq \frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{\epsilon^2}$ then

$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \epsilon$ (with prob. greater than $1-\delta$)

$\forall$ s,a

N i.i.d samples

$\hat{P}(s'|s,a)$

plan inside $\hat{P}$.

Recap:

Generative Model and Statistical Limits

Theorem: (Azar et al. '13) With probability greater than $1-\delta$,

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \gamma \sqrt{\frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N} + \frac{c\gamma}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}},$$

where c is an absolute constant.

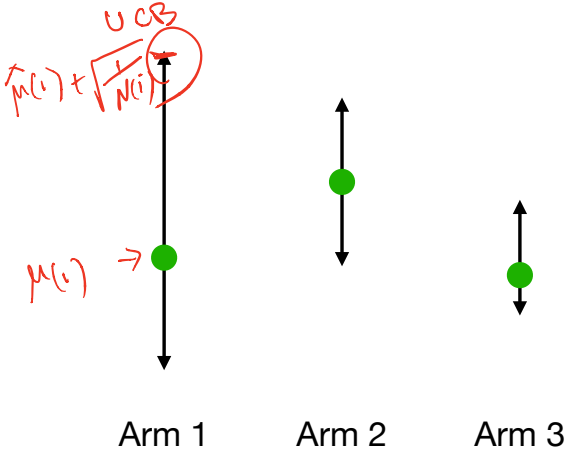
Corollary: for $\epsilon < 1$, provided $N \geq \frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{\epsilon^2}$ then

$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \epsilon$ (with prob. greater than $1-\delta$)

Generative Model: we can reset to anywhere we want (may not be realistic)

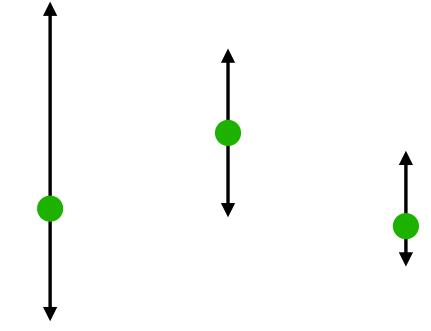
Recap:

Multi-armed Bandits and UCB Algorithm



Recap:

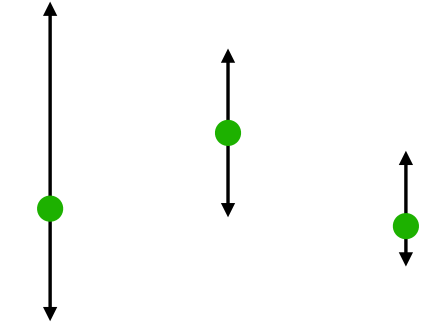
Multi-armed Bandits and UCB Algorithm



$$a^n := \arg\max_i \hat{\mu}^n(i) + \sqrt{\ln(KN/\delta)/N^n(i)}$$

Recap:

Multi-armed Bandits and UCB Algorithm

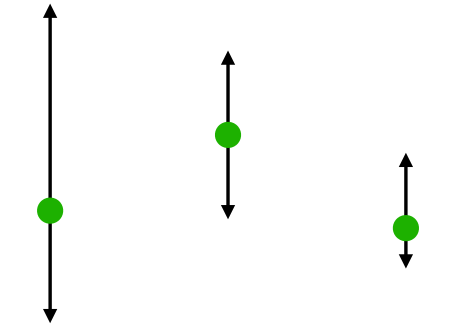


$$a^n := \arg\max_i \hat{\mu}^n(i) + \sqrt{\ln(KN/\delta)/N^n(i)}$$

$$\mathbb{E}\left[N\mu(a^\star) - \sum_{n=1}^{N} \mu(a^n)\right] \leq \widetilde{O}(\sqrt{KN})$$

Recap:

Multi-armed Bandits and UCB Algorithm



$$a^n := \arg\max_i \hat{\mu}^n(i) + \sqrt{\ln(KN/\delta)/N^n(i)}$$

$$\mathbb{E}\left[N\mu(a^\star) - \sum_{n=1}^{N} \mu(a^n)\right] \leq \widetilde{O}(\sqrt{KN})$$

Key step in the proof:

$$\mu(a^\star) - \mu(a^n) \leq \hat{\mu}(a^n) + \sqrt{\frac{\ln(KN/\delta)}{N^n(a_n)}} - \mu(a^n)$$

$UCB(a^n) \geq UCB(a^*) \geq \mu(a^*)$

$UCB(a^n)$

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathscr{M} = \left\{ \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H}, H, \mu, S, A \right\}$

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathscr{M} = \left\{ \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H}, H, \mu, S, A \right\}$

Only reset from μ: we assume it's a delta distribution, all mass at a fixed s_0

$\mu(s_0) = 1$

$\mu(s) = 0$ for $s \neq s_0$

Unknown Transition P (for simplicity assume reward is known)

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathscr{M} = \left\{ \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H}, H, \mu, S, A \right\}$

Only reset from μ: we assume it's a delta distribution, all mass at a fixed s_0

Unknown Transition P (for simplicity assume reward is known)

Different from the Generative Model Setting!

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathscr{M} = \left\{ \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H}, H, \mu, S, A \right\}$

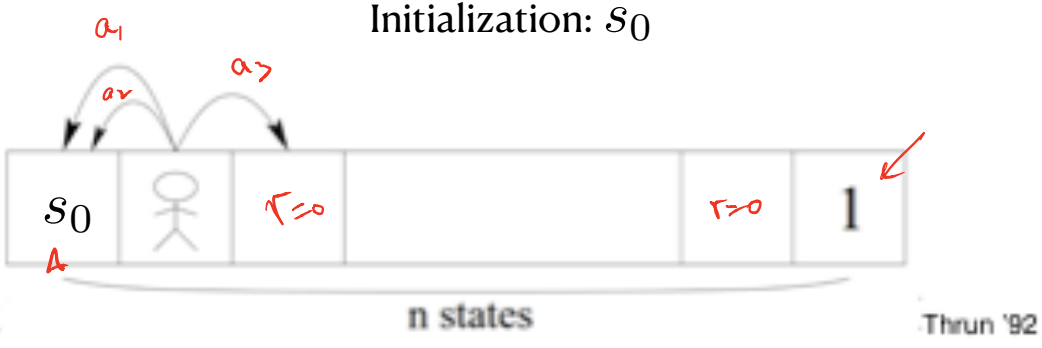
Only reset from μ: we assume it's a delta distribution, all mass at a fixed s_0

Unknown Transition P (for simplicity assume reward is known)

Different from the Generative Model Setting!

EXPLORATION!

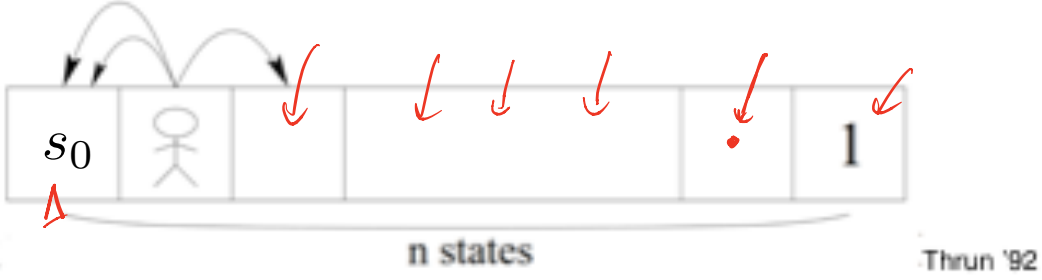
Why we need strategic exploration?



Length of chain is H

Why we need strategic exploration?

Initialization: s_0



Length of chain is H

Probability of random walk hitting reward 1 is $(1/3)^{-H}$

Learning Protocol

Learning Protocol

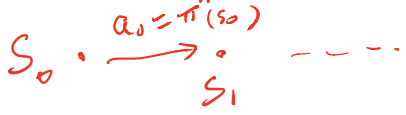
1. Learner initializes a policy π^1

Learning Protocol

1. Learner initializes a policy π^1

2. At episode n, learner executes π^n:

$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$, with $a_h^n = \pi^n(s_h^n)$, $r_h^n = r(s_h^n, a_h^n)$, $s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$



Learning Protocol

1. Learner initializes a policy π^1

2. At episode n, learner executes π^n:
$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$, with $a_h^n = \pi^n(s_h^n)$, $r_h^n = r(s_h^n, a_h^n)$, $s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$

3. Learner updates policy to π^{n+1} using all prior information

$\{s_h^i, a_h^i\}_{h=0, i=1}^{H-1, n}$

Learning Protocol

1. Learner initializes a policy π^1

2. At episode n, learner executes π^n:
$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$, with $a_h^n = \pi^n(s_h^n)$, $r_h^n = r(s_h^n, a_h^n)$, $s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$

3. Learner updates policy to π^{n+1} using all prior information

Performance measure: REGRET

$$\frac{1}{N}\ \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] = \text{poly}(S, A, H)\frac{\sqrt{N}}{N}$$

$N \to 0$, $\frac{\sqrt{N}}{N} \to 0$

Notations for Today

$$= \sum_{s' \in S} P(s' \mid s,a) \cdot f(s')$$

$$\mathbb{E}_{s' \sim P(\cdot \mid s,a)} \left[f(s') \right] := P(\cdot \mid s,a) \cdot f$$

$d_h^\pi(s,a)$: state-action distribution induced by π at time step h (i.e., probability of π visiting (s,a) at time step h starting from s_0)

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

Time-dependent

$\pi_h : S \to A$

$s_0 \quad a_0 = \pi_0(s_0), \quad \ldots \quad a_h = \pi_h(s_h) \quad \ldots$

UCBVI: Optimistic Model-based Learning

Episode

Inside ~~iteration~~ n :

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_0^n, \ldots, \widehat{P}_{H-1}^n$

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

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

> 0

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

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

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

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

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

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

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

$\{s_0^n, a_0^n, \ldots, s_{H-1}^n, a_{H-1}^n\}$

$s_{(n)}^n$

UCBVI—Part 1: Model Estimation

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

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

UCBVI—Part 1: Model Estimation

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

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

Let's also maintain some statistics using these datasets:

UCBVI—Part 1: Model Estimation

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

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

Let's also maintain some statistics using these datasets:

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

UCBVI—Part 1: Model Estimation

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

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

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

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

$$b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N_h^n(s,a)}} \approx c \cdot H \sqrt{\frac{1}{N_h^n(s,a)}}$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

h=0, ---- h=H-1, H

Δ

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

$$\widehat{V}_H^n(s) = 0, \forall s$$

Δ

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

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

$r(s,a) \in [0,1]$

$\|Q^\pi\|_\infty \le H$

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

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

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

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s \qquad h = H, \cdots h = 0$$

$$\pi^n = \{\pi_0^n, \cdots \pi_{H-1}^n\}$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

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

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

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

$$\left\|\widehat{V}_h^n\right\|_\infty \le H, \forall h, n$$

$[0, H]$

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 $\widehat{P}^n$: $\widehat{P}_h^n(s' \mid s,a) = \frac{N_h^n(s,a,s')}{N_h^n(s,a)}, \forall s, a, s', h$ ✓
4. Plan: $\pi^n = VI\left(\{\widehat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N_h^n(s,a)}}$ $\approx H\sqrt{\frac{1}{N(s,a)}}$
5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

$s_0^n = s_0$

Theorem: UCBVI Regret Bound

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

Theorem: UCBVI Regret Bound

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

Remarks:

Note that we consider expected regret here (policy π^n is a random quantity).
High probability version is not hard to get (need to do a martingale argument)

Theorem: UCBVI Regret Bound

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

Remarks:

Note that we consider expected regret here (policy π^n is a random quantity).
High probability version is not hard to get (need to do a martingale argument)

Dependency on H and S are suboptimal; but the **same** algorithm can achieve $H^2\sqrt{SAN}$ in the leading term [Azar et.al 17 ICML]

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., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

$\pi^n, \widehat{P}^n, r_h + b_n$

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

$\pi^n, (\widehat{P}, r)$

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

$\{\widehat{P}^n, r + b_h^n\}$ $\leftarrow \{P, r\}$

High-level Idea: Exploration or Exploitation Tradeoff

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

optimism

$\widehat{V}^n \geq V^*$

small

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

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

High-level Idea: Exploration or Exploitation Tradeoff

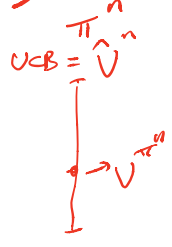
Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

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

$(\widehat{V}_0^n(s_0) \geq V_0^{\pi^n}(s_0))$

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

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

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

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

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

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 $1-\delta$:

$f = V_{h+1}^*$

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

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

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

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 $b_h^n(s, a)$

From now on, assume this event being true

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 $b_h^n(s, a)$

From now on, assume this event being true

Intuition:

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 $b_h^n(s, a)$

$\left| \left(\widehat{P}_{sa} - P_{sa} \right) \cdot V^* \right|$

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 $\mathbb{E}_{s' \sim P_h(\cdot \mid s, a)} f(s')$

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

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

for H-improvement, use Bernstein Rather than Hoeffding'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 $b_h^n(s, a)$

$\mathbb{E}_{s' \sim P_h} V^*(s')$

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 $\mathbb{E}_{s' \sim P_h(\cdot \mid s, a)} f(s')$

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

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

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

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

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

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot V_{h+1}^\star - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

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

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot V_{h+1}^\star - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$= b_h^n(s,a) + \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V_{h+1}^\star \leq H\sqrt{\frac{1}{N_h^n(s,a)}} \simeq b_h^n(s,a)$$

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:

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

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

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

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot\,|\,s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P_h(\cdot\,|\,s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot\,|\,s,a) \cdot V_{h+1}^\star - P_h(\cdot\,|\,s,a) \cdot V_{h+1}^\star$$

$$= b_h^n(s,a) + \left(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a)\right) \cdot V_{h+1}^\star \quad \leq b_h^n(s,a)$$

$$\geq b_h^n(s,a) - b_h^n(s,a) = 0, \quad \forall s, a$$

$\widehat{Q}_h^n \geq Q_h^\star$

$\Rightarrow \widehat{V}_h^n(s) \geq V_h^\star(s), \forall s$

for the case where Reward is unknown

(Here, reward is known)

3. Upper Bounding Regret using Optimism

$\hat{r}_h^n(s,a) = \frac{1}{N_h^n(s,a)} \sum_{i=1}^{n-1} r_h^i \mathbb{1}(s_h^i, a_h^i = s,a)$

$|\hat{r}(s,a) - r(s,a)| \leq \sqrt{\frac{1}{N_h^n(s,a)}}$

π^n, $\hat{P}$, $r + b_h$

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

$V_0^\star(s_0)$: unknown

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

Recall simulation lemma — the lemma measures the difference of a policy under two MDPs

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\le r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

A

$\min\{a, b\} \le a$

Bell-Eqn

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\le r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

A +/− A

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\le r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \left(\widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \right) \cdot \widehat{V}_1^n + P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \left(\widehat{V}_1^n - V_1^{\pi^n} \right)$$

(2) Recursion

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\le r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \left(\widehat{P}_0^n(\cdot\,|\,s_0, \pi^n(s_0)) - P_0(\cdot\,|\,s_0, \pi^n(s_0)) \right) \cdot \widehat{V}_1^n + P_0(\cdot\,|\,s_0, \pi^n(s_0)) \cdot \left(\widehat{V}_1^n - V_1^{\pi^n} \right)$$

(1) (2) Recursion

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

4. Upper bounding Regret via Simulation Lemma

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

$a \cdot b \le \|a\|_1 \|b\|_\infty$

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

Truncate by H

$\le H$

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

$$\|\widehat{P}_{(s,a)} - P(s,a)\|_1 \le \sqrt{\frac{S \ln(\cdots)}{N_h^n(s,a)}}$$

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

$b(s,a) = H\sqrt{\frac{\ln(\cdots/\delta)}{N_h^n(s,a)}}$

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

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

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

$$(\widehat{P}-P)\cdot\widehat{V} = (\widehat{P}-P)\cdot(V^*) + (\widehat{P}-P)\cdot(\widehat{V}-V^*)$$

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

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

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

5. Final Step

$|(\hat{P} - P) \cdot V^*| \leq b_n$ $\quad \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \sqrt{S}\, b_n(s,a)$

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.

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

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

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

5. Final Step

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

5. Final Step

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

5. Final Step

$$\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n} + 1$$

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

5. Final Step

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

CS-inequality

$$\sum_{s,a} 1\cdot\sqrt{N_h^N(s,a)} \le \sqrt{\sum_{s,a}1}\ \sqrt{\sum_{s,a}N_h^N(s,a)}$$

$$\sum_{s,a}1 = SA$$

5. Final Step

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

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

5. Final Step

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

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

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

5. Final Step

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

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

$$\mathbb{E}\left[\text{Regret}_N\right] \le 2H^2S\sqrt{AN\ln(SAHN/\delta)} + 2\delta NH \qquad \text{Set } \delta = 1/(HN)$$

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

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

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

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

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