Risk-Sensitive Reinforcement Learning:

Near-Optimal Risk-Sample Tradeoff in Regret

Presenter: Hao Liang

The Chinese University of Hong Kong, Shenzhen, China

July 2, 2020

Mainly based on:
Fei, Yingjie, et al. "Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regret." arXiv preprint arXiv:2006.13827 (2020).

Background

- Risk-sensitive RL concerns learning policies that take into account risks.
- Effective management of risks in RL is critical to many real-world applications
  - Autonomous driving
  - Real-time strategy games
  - Financial investment
  - Neuroscience: model human behaviors in decision making

Objective

- Maximize a Exponential utility function

$$V = \frac{1}{\beta} \log \left\{ \mathbb{E} e^{\beta R} \right\}, \tag{1}$$

  where R is the return, and $\beta \neq 0$ controls risk preference of the agent.
- (1) admits the Taylor expansion $V = \mathbb{E}[R] + \frac{\beta}{2} \operatorname{Var}(R) + O\left(\beta^2\right)$
  - $\beta > 0$: risk-seeking (favoring high uncertainty in R)
  - $\beta < 0$: risk-averse (favoring low uncertainty in R)
  - $\beta \to 0$: $V = \mathbb{E}[R]$, risk-neutral
- (1) covers the entire spectrum of risk sensitivity by varying β

Challenges

- Non-linearity of the objective function
  - Induces a non-linear Bellman equation
- Designing a risk-aware exploration mechanism
  - How to efficiently explores while adapting to (1) with different β

Contributions

- Propose two provably efficient model-free algorithms that implement risk-sensitive OFU
  - Risk-Sensitive Value Iteration (RSVI): $\tilde{O}\left(\lambda\left(|\beta|H^2\right) \cdot \sqrt{H^3 S^2 AT}\right)$ regret
  - Risk-Sensitive Q-learning (RSQ): $\tilde{O}\left(\lambda\left(|\beta|H^2\right) \cdot \sqrt{H^3 S^2 AT}\right)$ regret
  - $\lambda(u) := \left(e^{3u} - 1\right)/u$
- Establish a regret lower bound showing that the exponential dependence on β and H is unavoidable for any algorithm with an $\tilde{O}\left(\sqrt{T}\right)$ regret

Problem setup

- Episodic MDPs $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$
  - $\mathcal{S}$ and $\mathcal{A}$ are finite discrete spaces, and let $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$
  - $\mathcal{P} = \{P_h\}_{h \in [H]}$ and $\mathcal{R} = \{r_h\}_{h \in [H]}$ are state transition kernels and reward functions
  - Agent does not have access to $\mathcal{P}$ and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic function
- An initial state s_1 is chosen arbitrarily by the environment
- A policy $\pi = \{\pi_h\}_{h \in [H]}$ of an agent is a sequence of functions $\pi_h : \mathcal{S} \to \mathcal{A}$
- For each $h \in [H]$, we define the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π

$$V_h^\pi(s) := \frac{1}{\beta} \log \left\{ \mathbb{E} \left[\exp \left(\beta \sum_{h=1}^{H} r_h\left(s_h, \pi_h\left(s_h\right)\right) \right) \mid s_h = s \right] \right\}. \tag{2}$$

Bellman equations and regret

- Define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$Q_h^\pi(s,a) := \frac{1}{\beta} \log \left\{ \exp\left(\beta \cdot r_h(s,a)\right) \mathbb{E}\left[\exp\left(\beta \sum_{h'=h+1}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \right) \mid s_h = s, a_h = a \right] \right\}$$

- The Bellman equation associated with policy π is given by

$$Q_h^\pi(s,a) = r_h(s,a) + \frac{1}{\beta} \log \left\{ \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)} \left[\exp\left(\beta \cdot V_{h+1}^\pi\left(s'\right)\right) \right] \right\} \tag{3}$$

$$V_h^\pi(s) = Q_h^\pi\left(s, \pi_h(s)\right), \quad V_{H+1}^\pi(s) = 0 \tag{4}$$

- Under some mild regularity conditions, there always exists an optimal policy π^* which gives the optimal value $V_h^*(s) = \sup_\pi V_h^\pi(s)$ for all $(h,s) \in [H] \times \mathcal{S}$

Bellman equations and regret

- The Bellman optimality equation is given by

$$Q_h^*(s,a) = r_h(s,a) + \frac{1}{\beta} \log \left\{ \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[\exp \left(\beta \cdot V_{h+1}^* \left(s' \right) \right) \right] \right\} \tag{5}$$

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a), \quad V_{H+1}^*(s) = 0 \tag{6}$$

- Both Bellman equations are non-linear due to non-linearity of the exponential utility
- s_1^k the initial state, π^k the policy chosen at the beginning of episode k.
- The total regret after K episodes is

$$\mathrm{Regret}(K) := \sum_{k \in [K]} \left[V_1^* \left(s_1^k \right) - V_1^{\pi^k} \left(s_1^k \right) \right]$$

Upper bounds on the value functions and regret

Lemma 1.

For any $(h, s, a) \in \mathcal{S} \times \mathcal{A} \times [H]$, *policy* π *and risk parameter* $\beta \neq 0$, *we have*

$$0 \leq V_h^\pi(s) \leq H \quad \textit{and} \quad 0 \leq Q_h^\pi(s, a) \leq H.$$

Consequently, for each $K \geq 1$, *all policy sequences* $\pi^1, \ldots, \pi^K$ *and any* $\beta \neq 0$, *we have*

$$0 \leq \mathrm{Regret}(K) \leq KH.$$

Proof.

Recall the assumption that the reward functions $\{r_h\}$ are bounded in [0, 1]. The lower bounds are immediate by definition. For the upper bound, we have $V_h^\pi(s) \leq \frac{1}{\beta} \log\{\mathbb{E}[\exp(\beta H)]\} = H$. Upper bounds for Q_h^π and the regret follow similarly. □

Algorithm 1: RSVI

Algorithm 1 RSVI

Input: number of episodes $K \in \mathbb{Z}_{>0}$, confidence level $\delta \in (0,1]$, and risk parameter $\beta \neq 0$

$Q_h(s,a) \leftarrow H - h + 1$ and $N_h(s,a) \leftarrow 0$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$

$Q_{H+1}(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

Initialize datasets $\{\mathcal{D}_h\}$ as empty

for episode $k = 1, \ldots, K$ **do**

  $V_{H+1}(s) \leftarrow 0$ for each $s \in \mathcal{S}$

  for step $h = H, \ldots, 1$ **do** ▷ *value estimation*

    Update w_h via Equation (8)

    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $N_h(s,a) \geq 1$ **do**

      $b_h(s,a) \leftarrow c_\gamma \left| e^{\beta H} - 1 \right| \sqrt{\frac{S \log(2SAT/\delta)}{N_h(s,a)}}$ for some universal constant $c_\gamma > 0$

      $$Q_h(s,a) \leftarrow \begin{cases} \frac{1}{\beta} \log \left[\min\{e^{\beta(H-h+1)}, w_h(s,a) + b_h(s,a)\} \right], & \text{if } \beta > 0; \\ \frac{1}{\beta} \log \left[\max\{e^{\beta(H-h+1)}, w_h(s,a) - b_h(s,a)\} \right], & \text{if } \beta < 0 \end{cases}$$

      $V_h(s) \leftarrow \max_{a' \in \mathcal{A}} Q_h(s,a')$

    end for

  end for

  for step $h = 1, \ldots, H$ **do** ▷ *policy execution*

    Take action $a_h \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h(s_h, a)$ and observe $r_h(s_h, a_h)$ and s_{h+1}

    $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$

    Insert (s_h, a_h, s_{h+1}) into $\mathcal{D}_h$

  end for

end for

Mechanism of RSVI

- Algorithm 1 is inspired by LSVI-UCB. It follows OFU by applying the UCB by incorporating a bonus term to value estimates of state-action pairs.
- Including the value estimation step (Line 6–13) and the policy execution step (Line 14–18)
- In Line 7, the algorithm computes the intermediate value w_h by a least-squares update

$$w_h \leftarrow \underset{w \in \mathbb{R}^{SA}}{\operatorname{argmin}} \sum_{\tau \in [k-1]} \left[e^{\beta \left[r_h(s_h^\tau, a_h^\tau) + V_{h+1}\left(s_{h+1}^\tau\right) \right]} - w^\top \phi\left(s_h^\tau, a_h^\tau\right) \right]^2, \tag{7}$$

  where $\phi(\cdot, \cdot)$ denotes the canonical basis in $\mathbb{R}^{SA}$ and $\left\{ \left(s_h^\tau, a_{h'}^\tau s_{h+1}^\tau \right) \right\}_{\tau \in [k-1]}$ are accessed from the dataset $\mathcal{D}_h$.
- Can be efficiently implemented by computing sample means of $e^{\beta \left[r_h(s,a) + V_{h+1}\left(s'\right) \right]}$ over visited state-action pairs.

Mechanism of RSVI

- In Line 10, the algorithm uses w_h to compute the estimate Q_h, by adding/subtracting bonus b_h and thresholding the sum/difference at $e^{\beta(H-h+1)}$, depending on the sign of β
- The logarithmic-exponential transformation in Line 10 conforms and adapts to the non-linearity in Bellman equations (3) and (4).
- The thresholding operator ensures that Q_h and V_h stays in the range [0, $H-h$+1].
- Subtracting bonus when $\beta < 0$ implements OFU in a risk-sensitive fashion.
- Belong to batch algorithms.

Algorithm 2: RSQ

Algorithm 2 RSQ

Input: number of episodes $K \in \mathbb{Z}_{>0}$, confidence level $\delta \in (0, 1]$, learning rates $\{\alpha_t\}$ and risk parameter $\beta \neq 0$

$Q_h(s, a), V_h(s, a) \leftarrow H - h + 1$ and $N_h(s, a) \leftarrow 0$ for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$

$Q_{H+1}(s, a), V_{H+1}(s, a) \leftarrow 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$

for episode $k = 1, \ldots, K$ **do**

  Receive the initial state s_1

  for step $h = 1, \ldots, H$ **do**

    Take action $a_h \leftarrow \mathrm{argmax}_{a' \in \mathcal{A}} Q_h(s_h, a')$, and observe $r_h(s_h, a_h)$ and s_{h+1}

    $t = N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$

    $b_t \leftarrow c \left| e^{\beta H} - 1 \right| \sqrt{\frac{H \log(SAT/\delta)}{t}}$ for some sufficiently large universal constant $c > 0$

    $w_h(s_h, a_h) \leftarrow (1 - \alpha_t) e^{\beta \cdot Q_h(s_h, a_h)} + \alpha_t e^{\beta [r_h(s_h, a_h) + V_{h+1}(s_{h+1})]}$

    $$Q_h(s_h, a_h) \leftarrow \begin{cases} \frac{1}{\beta} \log \left[\min\{e^{\beta(H-h+1)}, w_h(s_h, a_h) + \alpha_t b_t\} \right], & \text{if } \beta > 0; \\ \frac{1}{\beta} \log \left[\max\{e^{\beta(H-h+1)}, w_h(s_h, a_h) - \alpha_t b_t\} \right], & \text{if } \beta < 0 \end{cases}$$

    $V_h(s_h) \leftarrow \max_{a' \in \mathcal{A}} Q_h(s_h, a')$

  end for

end for

Mechanism of RSQ

- Algorithm 1 requires storage of historical data $\{\mathcal{D}_h\}$ and computation over them (Line 7).
- Q-learning update Q values in an online fashion as each state-action pair is encountered.
- Based on Q-learning with UCB in the work of [38] and use the same learning rates therein

$$\alpha_t := \frac{H+1}{H+t}.$$

- Line 9 updates w_h in an online fashion, in contrast with the batch update of Algorithm 1.

Comparisons

- The bonuses in both algorithms depend on β through a common factor $\left|e^{\beta H} - 1\right|$.
- A careful analysis on the bonuses and the value estimation steps reveals that the effective bonuses is proportional to $\frac{e^{|\beta| H} - 1}{|\beta|}$
- The more risk-sensitive an agent is, the larger bonus it needs to compensate for the uncertainty
- Both algorithms have polynomial time and space complexities in S, A, K and H.
- Algorithm 2 is more efficient than Algorithms 1 in both time and space complexities, since it does not require storing historical data nor computing statistics.

Regret upper bounds for RSVI

Theorem 2.

For any $\delta \in (0,1]$, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by

$$\mathrm{Regret}(K) \lesssim \lambda\left(|\beta| H^2\right) \cdot \sqrt{H^3 S^2 A T \log^2(2SAT/\delta)}$$

Corollary 3.

Under the setting of Theorem 1 and when $\beta \to 0$, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by

$$\mathrm{Regret}(K) \lesssim \sqrt{H^3 S^2 A T \log^2(2SAT/\delta)}$$

Regret upper bounds for RSVI

- Theorem 2 adapts to both risk-seeking and risk-averse settings through a common factor of $\lambda\left(|\beta| H^2\right)$.
- Corollary 3 recovers the regret bound of [4, Theorem 2] under the standard RL setting and is nearly optimal.
- Corollary 3 also reveals that Theorem 2 interpolates between the risk-sensitive and risk-neutral settings.

Proof of Theorem 2: preliminaries

- Let $s_h^k, a_h^k, w_h^k, Q_h^k$ and V_h^k and V_h^k denote the values of s_h, a_h, w_h, Q_h and V_h in episode k
- Let N_h^k and D_h^k denote the value of N_h and D_h at the end of episode $k-1$.

Fact 4.

Consider $x, y, b \in \mathbb{R}$ such that $x \geq y$.
(a) if $y \geq$ g for some $g > 0,$ then $\log(x) - \log(y) \leq \frac{1}{g}(x-y)$
(b) Assume further that $y \geq 0$. If $b \geq 0$ and $x \leq u$ for some $u > 0,$ then $e^{bx} - e^{by} \leq be^{bu}(x-y)$; if $b < 0,$ then $e^{by} - e^{bx} \leq (-b)(x-y)$

Fact 5.

Define $\lambda_0 := \frac{e^{|\beta|H}-1}{|\beta|}$ and $\lambda_2 := e^{|\beta|\left(H^2+H\right)}$. Then we have $\lambda_0\lambda_2 H \leq \frac{e^{3|\beta|H^2}-1}{|\beta|}$.

Proof warmup

- Define $d := SA, l := \log(2dT/\delta)$ for a given $\delta \in (0,1]$.
- Define $\phi(s,a)$ as canonical basis of $\mathbb{R}^{SA}$ and let Λ_h^k be a diagonal matrix in $\mathbb{R}^{d\times d}$ with each (s,a)-th diagonal entry equal to $\max\left\{N_h^{k-1}(s,a), 1\right\}$.
- Fix a tuple $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$ such that $N_h^{k-1}(s,a) \geq 1$ and fix a policy π
- Set $w_h^\pi = e^{\beta\cdot Q_h^\pi(\cdot,\cdot)}$,

$$Q_h^\pi(s,a) = \frac{1}{\beta}\log\left(e^{\beta\cdot Q_h^\pi(s,a)}\right) \quad = \frac{1}{\beta}\log\left(\left\langle \phi(s,a), e^{\beta\cdot Q_h^\pi(\cdot,\cdot)}\right\rangle\right) = \frac{1}{\beta}\log\left(\langle \phi(s,a), w_h^\pi\rangle\right)$$

$$w_h^\pi(s,a) = e^{\beta\cdot Q_h^\pi(s,a)} = \left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta\cdot Q_h^\pi(s_h^\tau, a_h^\tau)}\right]\right\rangle$$

Proof warmup

- Define $d := SA, l := \log(2dT/\delta)$ for a given $\delta \in (0,1]$.
- Define $\phi(s,a)$ as canonical basis of $\mathbb{R}^{SA}$ and let Λ_h^k be a diagonal matrix in $\mathbb{R}^{d\times d}$ with each (s,a)-th diagonal entry equal to $\max\left\{N_h^{k-1}(s,a), 1\right\}$.
- Fix a tuple $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$ such that $N_h^{k-1}(s,a) \geq 1$ and fix a policy π
- Set $w_h^\pi = e^{\beta\cdot Q_h^\pi(\cdot,\cdot)}$,

$$Q_h^\pi(s,a) = \frac{1}{\beta}\log\left(e^{\beta\cdot Q_h^\pi(s,a)}\right) \quad = \frac{1}{\beta}\log\left(\left\langle \phi(s,a), e^{\beta\cdot Q_h^\pi(\cdot,\cdot)}\right\rangle\right) = \frac{1}{\beta}\log\left(\langle \phi(s,a), w_h^\pi\rangle\right)$$

$$w_h^\pi(s,a) = e^{\beta\cdot Q_h^\pi(s,a)} = \left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta\cdot Q_h^\pi(s_h^\tau, a_h^\tau)}\right]\right\rangle$$

Proof warmup

- Define
$$q_1^+ := \begin{cases} \langle \phi(s,a), w_h^k \rangle + b_h^k(s,a), & \text{if } \beta > 0 \\ \langle \phi(s,a), w_h^k \rangle - b_h^k(s,a), & \text{if } \beta < 0, \end{cases}$$
$$q_1 := \begin{cases} \min\left\{e^{\beta(H-h+1)}, q_1^+\right\}, & \text{if } \beta > 0 \\ \max\left\{e^{\beta(H-h+1)}, q_1^+\right\}, & \text{if } \beta < 0 \end{cases}$$
- By the definition of Λ_h^k and $\phi_{h'}^k$, observe that

$$w_h^k(s,a) = \langle \phi(s,a), w_h^k \rangle = \left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau \in [k-1]} \phi_h^\tau \left[e^{\beta\left[r_h^T + V_{h+1}^k\left(s_{h+1}^\tau\right)\right]} \right] \right\rangle.$$

- Define $G_0 := \left(Q_h^k - Q_h^\pi\right)(s,a) = \frac{1}{\beta} \log\{q_1\} - \frac{1}{\beta} \log\{\langle \phi(s,a), w_h^\pi \rangle\}$
- Need to derive upper and lower bounds for G_0.

Proof warmup

- $$G_0 = \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\left\{\left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[e^{\beta\cdot Q_h^\pi(s_h^\tau, a_h^\tau)}\right]\right\rangle\right\}$$
  $$= \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\left\{\left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[\mathbb{E}_{s'\sim P_h(\cdot|s_h^\tau, a_h^\tau)} e^{\beta\left[r_h^\tau + V_{h+1}^\pi(s')\right]}\right]\right\rangle\right\}$$
  $$=: \frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_3\}$$
- In order to control G_0, we define an intermediate quantity

$$q_2 := \left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau\in[k-1]} \phi_h^\tau \left[\mathbb{E}_{s'\sim P_h(\cdot|s_h^\tau, a_h^\tau)} e^{\beta\left[r_h^\tau + V_{h+1}^k(s')\right]}\right]\right\rangle$$

with q_2 replaces the quantity V_{h+1}^π in q_3 by V_{h+1}^k

Proof warmup

- Decompose the error

$$\left(Q_h^k - Q_h^\pi\right)(s,a) = G_0 = (\frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_2\}) + (\frac{1}{\beta}\log\{q_2\} - \frac{1}{\beta}\log\{q_3\}) \quad (8)$$

$$= G_1 + G_2 \quad (9)$$

- $\mathrm{G}_0, \mathrm{G}_1$ and G_2 are all well-defined, according to the following result.

Lemma 6.

We have $q_i \in \left[\min\left\{1, e^{\beta(H-h+1)}\right\}, \max\left\{1, e^{\beta(H-h+1)}\right\}\right]$ *for* $i \in [3]$

Proof warmup

- Control G_1 and G_2

$$\left(Q_h^k - Q_h^\pi\right)(s,a) = G_0 = \left(\frac{1}{\beta}\log\{q_1\} - \frac{1}{\beta}\log\{q_2\}\right) + \left(\frac{1}{\beta}\log\{q_2\} - \frac{1}{\beta}\log\{q_3\}\right) \quad (10)$$

$$= G_1 + G_2 \quad (11)$$

- $\mathrm{G}_0, \mathrm{G}_1$ and G_2 are all well-defined, according to the following result.

Lemma 7.
We have $q_i \in \left[\min\left\{1, e^{\beta(H-h+1)}\right\}, \max\left\{1, e^{\beta(H-h+1)}\right\}\right]$ *for* $i \in [3]$.

Proof warmup

- Control G_1 and G_2

Lemma 8.

For all $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$ that satisifies $N_h^{k-1}(s, a) \geq 1$, there exist universal constants $c_1, c_\gamma > 0$ (where c_γ is used in Line 9 of Algorithm 1) such that

$$0 \leq G_1 \leq c_1 \cdot \frac{e^{|\beta| H} - 1}{|\beta|} \cdot d\sqrt{\iota}\sqrt{\phi(s,a)^\top \left(\Lambda_h^k\right)^{-1} \phi(s,a)}$$

with probability at least $1 - \delta/2$. Furthermore, if $V_{h+1}^k(s') \geq V_{h+1}^\pi(s')$ for all $s' \in \mathcal{S}$, then we have

$$0 \leq G_2 \leq e^{|\beta| H} \cdot \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[V_{h+1}^k(s') - V_{h+1}^\pi(s')\right].$$

Proof of Lemma 8

- Start with case $\beta > 0$. The case $\beta < 0$ follows the same idea.

$$
\begin{aligned}
&\left|q_1^+ - q_2 - b_h^k(s,a)\right| \\
=& \left| \left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau \in [k-1]} \phi_h^\tau \left[e^{\beta\left[r_h^\tau + V_{h+1}^k\left(s_{h+1}^\tau\right)\right]} - \mathbb{E}_{s' \sim P_h\left(\cdot | s_h^\tau, a_h^\tau\right)} e^{\beta\left[r_h^\tau + V_{h+1}^k\left(s'\right)\right]} \right] \right\rangle \right| \\
=& \left| \frac{1}{N_h^{k-1}(s,a)} \sum_{(s,a,s^+) \in \mathcal{D}_h^{k-1}} e^{\beta\left[r_h(s,a) + V_{h+1}^k\left(s^+\right)\right]} - \mathbb{E}_{s' \sim P_h(\cdot|s,a)} e^{\beta\left[r_h(s,a) + V_{h+1}^k\left(s'\right)\right]} \right| \\
\le& c \left|e^{\beta H} - 1\right| \sqrt{\frac{Sl}{N_h^{k-1}(s,a)}} \\
=& c \left|e^{\beta H} - 1\right| \sqrt{S_l} \cdot \sqrt{\phi(s,a)^\top \left(\Lambda_h^k\right)^{-1} \phi(s,a)}
\end{aligned}
$$

Proof of Lemma 8

- The first inequality holds by Lemma 16. Choose $c_\gamma = c$ in the definition of $b_h^k(s,a)$,

$$0 \leq q_1^+ - q_2 \leq 2c \cdot \left| e^{\beta H} - 1 \right| \sqrt{S_l} \cdot \sqrt{\phi(s,a)^\top \left(\Lambda_h^k\right)^{-1} \phi(s,a)}.$$

- Therefore, we have $q_1 \geq q_2$, and thus $G_1 \geq 0$.
- By Lemma 7 and Fact 4(a) (with $g = 1$, $x = q_1$, and $y = q_2$)

$$G_1 \leq \frac{1}{\beta}\left(q_1 - q_2\right) \leq \frac{1}{\beta}\left(q_1^+ - q_2\right).$$

- Control G_2. $V_{h+1}^k\left(s'\right) \geq V_{h+1}^\pi\left(s'\right)$ for all $s' \in \mathcal{S}$ implies that $q_2 \geq q_3$ and therefore $G_2 \geq 0$.

Proof of Lemma 8

- By Fact 4(a) $(\text{with} g = 1, x = q_2,$ and $y = q_3)$ and the fact that $q_2 \geq q_3 \geq 1$

$$
\begin{aligned}
G_2 &\leq \frac{1}{\beta}\left(q_2 - q_3\right) \\
&\leq e^{\beta H}\left\langle \phi(s,a), \left(\Lambda_h^k\right)^{-1} \sum_{\tau \in [k-1]} \phi_h^\tau \left[\mathbb{E}_{s' \sim P_h\left(\cdot \mid s_h^\tau, a_h^\tau\right)}\left[V_{h+1}^k\left(s'\right) - V_{h+1}^\pi\left(s'\right)\right]\right]\right\rangle \\
&= e^{|\beta| H}\mathbb{E}_{s' \sim P_h(\cdot \mid s,a)}\left[V_{h+1}^k\left(s'\right) - V_{h+1}^\pi\left(s'\right)\right]
\end{aligned}
$$

- The second step holds by Fact 4(b) (with $b = \beta, x = r_h^\tau + V_{h+1}^k(s)$, and $y = r_h^\tau + V_{h+1}^\pi(s)\big)$ and $H \geq r_h^\tau + V_{h+1}^k(s) \geq r_h^\tau + V_{h+1}^\pi(s) \geq 0$.
- Case $\beta < 0$ is similar to the previous one. The proof is hence completed.

Proof of Lemma 8

- The following lemmas establishes the dominance of Q_h^k over Q_h^* and V_h^k over V_h^*.

Lemma 9.
On the event of Lemma 8, we have $Q_h^k(s,a) \geq Q_h^\pi(s,a)$ *for all* $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

Lemma 10.
For any $\delta \in (0,1]$, *with probability at least* $1-\delta/2$, *we have* $V_h^k(s) \geq V_h^\pi(s)$ *for all* $(k,h,s) \in [K] \times [H] \times \mathcal{S}$.

Proof of Theorem 2

- Define $\delta_h^k := V_h^k\left(s_h^k\right) - V_h^{\pi_k}\left(s_h^k\right)$ $\zeta_{h+1}^k := \mathbb{E}_{s' \sim P_h\left(\cdot \mid s_h^k, a_h^k\right)}\left[V_{h+1}^k\left(s'\right) - V_{h+1}^{\pi_k}\left(s'\right)\right] - \delta_{h+1}^k$
- For any $(k, h) \in [K] \times [H]$, we have

$$
\begin{aligned}
\delta_h^k = & \left(Q_h^k - Q_h^{\pi_k}\right)\left(s_h^k, a_h^k\right) \\
\leq & c_1 \cdot \frac{e^{|\beta| H} - 1}{|\beta|} \cdot \sqrt{S_l} \sqrt{\phi\left(s_{h'}^k a_h^k\right)^{\top}\left(\Lambda_h^k\right)^{-1} \phi\left(s_{h'}^k a_h^k\right)} \\
& + e^{|\beta| H} \cdot \mathbb{E}_{s' \sim P_h\left(\cdot \mid s_h^k, a_h^k\right)}\left[V_{h+1}^k\left(s'\right) - V_{h+1}^{\pi_k}\left(s'\right)\right] \\
= & c_1 \cdot \frac{e^{|\beta| H} - 1}{|\beta|} \cdot \sqrt{S_l} \sqrt{\phi\left(s_h^k, a_h^k\right)^{\top}\left(\Lambda_h^k\right)^{-1} \phi\left(s_h^k, a_h^k\right)} \\
& + e^{|\beta| H}\left(\delta_{h+1}^k + \zeta_{h+1}^k\right)
\end{aligned}
$$

Proof of Theorem 2

- Noting that $V_{H+1}^k(s) = V_{H+1}^{\pi_k}(s) = 0$ and $\delta_{h+1}^k + \zeta_{h+1}^k \geq 0$, expand the recursion

$$\delta_1^k \leq \sum_{h\in[H]} e^{(|\beta|H)h} \zeta_{h+1}^k + c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot \sum_{h\in[H]} e^{(|\beta|H)(h-1)} \sqrt{S\iota} \sqrt{\phi\left(s_{h'}^k, a_h^k\right)^\top \left(\Lambda_h^k\right)^{-1} \phi\left(s_{h'}^k, a_h^k\right)}$$

- Apply Lemma 10 with π set to π^*

$$\begin{aligned}
\mathrm{Regret}(K) = \sum_{k\in[K]} \left[\left(V_1^* - V_1^{\pi_k}\right)\left(s_1^k\right)\right] &\leq \sum_{k\in[K]} \delta_1^k \\
&\leq e^{|\beta|H^2} \sum_{k\in[K]h\in[H]} \zeta_{h+1}^k \\
&\quad + c_1 \cdot \frac{e^{|\beta|H} - 1}{|\beta|} \cdot e^{|\beta|H^2} \cdot \sqrt{S_l} \sum_{k\in[K]h\in[H]} \sqrt{\phi\left(s_h^k, a_h^k\right)^\top \left(\Lambda_h^k\right)^{-1} \phi\left(s_h^k, a_h^k\right)}
\end{aligned}$$

Proof of Theorem 2

- Proceed to control the two terms.
- Since V_H^K is independent of the new observation, $\left\{\zeta_{h+1}^k\right\}$ is a martingale difference sequence satisfying $\left|\zeta_h^k\right| \leq 2H$ for all $(k,h) \in [K] \times [H]$.
- By the Azuma-Hoeffding inequality, we have for any $t > 0$,

$$\mathbb{P}\left(\sum_{k\in[K]}\sum_{h\in[H]} \zeta_{h+1}^k \geq t\right) \leq \exp\left(-\frac{t^2}{2T \cdot H^2}\right).$$

- With probability $1 - \delta/2$, there holds

$$\sum_{k\in[K]}\sum_{h\in[H]} \zeta_{h+1}^k \leq \sqrt{2TH^2 \cdot \log(2/\delta)} \leq 2H\sqrt{T\iota}.$$

Proof of Theorem 2

- For the second term, apply Lemma 18 and the Cauchy-Schwartz inequality to obtain
$\sum_{k\in[K]h\in[H]}\sqrt{\phi\left(s_h^k,a_h^k\right)^\top\left(\Lambda_h^k\right)^{-1}\phi\left(s_h^k,a_h^k\right)}$
$\leq\sum_{h\in[H]}\sqrt{K}\sqrt{\sum_{k\in[H]}\phi\left(s_{h'}^k a_h^k\right)^\top\left(\Lambda_h^k\right)^{-1}\phi\left(s_{h'}^k a_h^k\right)}\leq H\sqrt{2dK_l}$
- Recall Fact 5 and the fact $\frac{e^{|\beta|H}-1}{|\beta|}\geq H$

$$
\begin{aligned}
\operatorname{Regret}(K) &\leq e^{|\beta|H^2}\cdot 2H\sqrt{T\iota}+c_1\cdot\frac{e^{|\beta|H}-1}{|\beta|}\cdot e^{|\beta|H^2}\cdot H\sqrt{2dSK\iota^2}\\
&\leq (c_1+2)\cdot\frac{e^{|\beta|H}-1}{|\beta|}\cdot e^{|\beta|H^2}\cdot\sqrt{2dHST\iota^2}\\
&\lesssim \lambda\left(|\beta|H^2\right)\cdot\sqrt{H^3S^2AT\log^2(2SAT/\delta)}
\end{aligned}
$$

Regret upper bounds for RSQ

Theorem 11.

For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, and when T is sufficiently large, the regret of Algorithm 2 is bounded by

$$\mathrm{Regret}(K) \lesssim \lambda\left(|\beta| H^2\right) \cdot \sqrt{H^4 SAT \log(SAT/\delta)}$$

Corollary 12.

Under the setting of Theorem 11 and when $\beta \to 0$, with probability at least $1 - \delta$, the regret of Algorithm 2 is bounded by

$$\mathrm{Regret}(K) \lesssim \sqrt{H^4 SAT \log(SAT/\delta)}$$

Regret lower bound

Theorem 13.
For sufficiently large K and H, the regret of any algorithm obeys

$$\mathbb{E}[\mathrm{Regret}(K)] \geq \frac{e^{|\beta|H/2} - 1}{|\beta|} \sqrt{T \log T}.$$

- Exponential dependence on the $|\beta|$ and H and a sub-linear dependence on T through the $\tilde{O}(\sqrt{T})$ factor is essentially indispensable.
- Both Theorems are nearly optimal in their dependence on β, H and T.
- Contrast with Lemma 1, an algorithm must incur a regret that is exponential in H in order to achieve a sublinear regret in T.

Proof of Theorem 13

- Construct a bandit instance as a special case of episodic fixed-horizon MDP problem.
- Establish lower bound on the instance in terms of the logarithmic-exponential objective.
- Start with two important lemmas.
- For each $\rho \in [0, 1]$, let $\mathrm{Ber}(\rho)$ denote the Bernoulli distribution with parameter ρ

Lemma 14.
Let $p, p' \in (0, 1)$ *be such that* $p > p'$. *We have* $D_{\mathrm{KL}}\left(\mathrm{Ber}\left(p'\right) \| \mathrm{Ber}(p)\right) \leq \frac{\left(p-p'\right)^2}{p(1-p)}$.

Proof of Theorem 13

Lemma 15.

Let $K_0 := K_0(K, \pi)$ *be the number of times that the sub-optimal arm is pulled in the* K*-round two-arm bandit problem with policy* π*. When* K *is sufficiently large, we have*

$$\mathbb{E}K_0 \gtrsim \frac{\log K}{D}.$$

Proof of Theorem 13

- Case $\beta > 0$.
- Two-arm bandit problem with K rounds, the reward for pulling arm i

$$X_i = \begin{cases} H & \text{w.p. } p_i \\ 0 & \text{w.p. } 1 - p_i \end{cases}$$

- $p_1 > p_2$ are to be specified later. Let $\Delta := p_1 - p_2 > 0$.
- By Lemma 14, $D_{\mathrm{KL}}\left(X_2 \| X_1\right) \leq \frac{\Delta^2}{p_1(1-p_1)}$.
- Lemma implies 15 $\mathbb{E}K_0 \gtrsim \frac{\log K \cdot p_1(1-p_1)}{\Delta^2}$.

Proof of Theorem 13

- Choose $\Delta = C\sqrt{\frac{\log K \cdot p_1(1-p_1)}{K}}$ for an universal constant $C > 0$.
- Set $p_2 = e^{-\beta H}$. Since $p_1\left(1 - p_1\right) \leq \frac{1}{4}$, we have $\Delta \lesssim \sqrt{\frac{\log K}{K}}$
- By choosing K and H large enough, we can ensure $\Delta \leq e^{-\beta H}$ and $p_1 = p_2 + \Delta \leq \frac{3}{4}$.
- Define X_i^k to be the outcome of arm X_i (if pulled) in round k, and Y^k to be the outcome of the arm actually pulled in round k.

Proof of Theorem 13

- Conditional on K_0, we have

$$
\begin{aligned}
\mathrm{Regret}(K) &= \frac{1}{\beta}\log\left[\mathbb{E}\exp\left(\beta\sum_{k\in[K]}X_1^k\right)\right] - \frac{1}{\beta}\log\left[\mathbb{E}\exp\left(\beta\sum_{k\in[K]}Y^k\right)\right] \\
&\overset{(i)}{=} \frac{1}{\beta}\log\left[\prod_{k=1}^{K}\mathbb{E}\exp\left(\beta X_1^k\right)\right] - \frac{1}{\beta}\log\left[\prod_{k=1}^{K}\mathbb{E}\exp\left(\beta Y^k\right)\right] \\
&\geq \frac{1}{\beta}\log\left[\prod_{k=1}^{K}\mathbb{E}\exp\left(\beta X_1^k\right)\right] - \frac{1}{\beta}\log\left[\prod_{k=1}^{K}\mathbb{E}\exp\left(\beta X_2^k\right)\right] \\
&= \frac{K}{\beta}\log\left[\mathbb{E}\exp\left(\beta X_1\right)\right] - \frac{K}{\beta}\log\left[\mathbb{E}\exp\left(\beta X_2\right)\right] \\
&\geq \frac{K_0}{\beta}\log\left[\mathbb{E}\exp\left(\beta X_1\right)\right] - \frac{K_0}{\beta}\log\left[\mathbb{E}\exp\left(\beta X_2\right)\right]
\end{aligned}
$$

Proof of Theorem 13

Taking expectation over K_0 on both sides

$$
\begin{aligned}
\mathbb{E}[\text{Regret}(K)] &\geq \frac{\mathbb{E}K_0}{\beta}\left(\log \mathbb{E}e^{\beta X_1} - \log \mathbb{E}e^{\beta X_2}\right) \\
&= \frac{\mathbb{E}K_0}{\beta}\log\left(\frac{p_1 e^{\beta H} + (1-p_1)}{p_2 e^{\beta H} + (1-p_2)}\right) \\
&= \frac{\mathbb{E}K_0}{\beta}\log\left(1 + \frac{\Delta\left(e^{\beta H} - 1\right)}{p_2 e^{\beta H} + (1-p_2)}\right) \\
&\geq \frac{\mathbb{E}K_0}{\beta}\log\left(1 + \frac{\Delta\left(e^{\beta H} - 1\right)}{1+1}\right) \\
&\geq \frac{\mathbb{E}K_0}{\beta}\cdot\frac{1}{4}\Delta\left(e^{\beta H} - 1\right) \\
&\gtrsim \frac{1}{\beta}\cdot\frac{\log K \cdot p_1(1-p_1)}{\Delta}\cdot\left(e^{\beta H} - 1\right) \\
&\gtrsim \frac{1}{\beta}\cdot\sqrt{K\log K \cdot p_1(1-p_1)}\cdot\left(e^{\beta H} - 1\right) \\
&\gtrsim \frac{1}{\beta}\cdot\sqrt{K\log K}\cdot\left(e^{\beta H/2} - 1\right) \\
&\gtrsim \frac{1}{\beta}\cdot\sqrt{T\log T}\cdot\left(e^{\beta H/2} - 1\right)
\end{aligned}
$$

Proof of Theorem 13

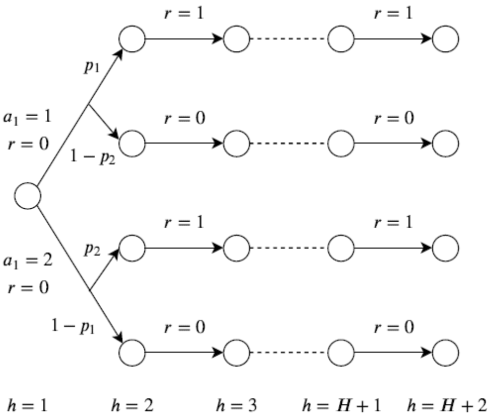

Figure: From bandit model to MDP

Supporting Lemmas of Theorem 2

Lemma 16.

Define

$$\overline{\mathcal{V}}_{h+1} := \left\{\bar{V}_{h+1} : \mathcal{S} \to \mathbb{R} \mid \forall s \in \mathcal{S}, \bar{V}_{h+1}(s) \in \left[\min\left\{e^{\beta(H-h)}, 1\right\}, \max\left\{e^{\beta(H-h)}, 1\right\}\right]\right\}$$

There exists a universal constant $c > 0$ *such that with probability* $1 - \delta$, *we have*

$$\left|e^{\beta\left[r_h\left(s_h^k, a_h^k\right) + V\left(s_{h+1}^k\right)\right]} - \mathbb{E}_{s' \sim P_h\left(\cdot \mid s_h^k, a_h^k\right)} e^{\beta\left[r_h\left(s_h^k, a_h^k\right) + V\left(s'\right)\right]}\right| \leq c\left|e^{\beta H} - 1\right| \sqrt{\frac{S_l}{N_h^k(s,a)}}$$

for all $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$ *and all* $V \in \overline{\mathcal{V}}_{h+1}$

Supporting Lemmas of Theorem 2

Lemma 17.

Let $\{\phi_t\}_{t\geq 0}$ *be a bounded sequence in* $\mathbb{R}^d$ *satisfying* $\sup_{t>0}\|\phi_t\|\leq 1$*. Let* $\Lambda_0\in\mathbb{R}^{d\times d}$ *be a PD matrix with* $\lambda_{\min}(\Lambda_0)\geq 1$*. For any* $t\geq 0$*, we define* $\Lambda_t := \Lambda_0+\sum_{i\in[t]}\phi_i\phi_i^\top$*. Then, we have*

$$\log\left[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right]\leq\sum_{i\in[t]}\phi_i^\top\Lambda_{i-1}^{-1}\phi_i\leq 2\log\left[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right]$$

Lemma 18.

Let $\iota=\log(2dT/\delta)$*. For any* $h\in[H]$*, we have*

$$\sum_{k\in[K]}\left(\phi_h^k\right)^\top\left(\Lambda_h^k\right)^{-1}\phi_h^k\leq 2d\iota$$