STATS 701 – Theory of Reinforcement Learning
Thompson/Posterior Sampling in MDPs

Ambuj Tewari

Associate Professor, Department of Statistics, University of Michigan
tewaria@umich.edu
https://ambujtewari.github.io/stats701-winter2021/

Winter 2021

Outline

Finite Horizon (or episodic) MDP

- A finite horizon MDP M consists of
  - $\mathcal{S}$, state space and $\mathcal{A}$, action space
  - μ_0, the initial state distribution
  - Horizon H: every episode terminates in exactly H steps
  - Transition dynamics $s_{t+1} \sim P_{s_t,a_t}$
  - Reward distributions $r_t \sim R_{s_t,a_t}$
- Need to consider non-stationary policy $\boldsymbol{\pi}$

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_H)$$

- Trajectory

$$\begin{aligned}
&s_1 \sim \mu_0, a_1 \sim \pi_1(s_1), r_1 \sim R_{s_1,a_1},\\
&s_2 \sim P_{s_1,a_1}, a_2 \sim \pi_2(s_2), r_2 \sim R_{s_2,a_2},\\
&\vdots\\
&s_H \sim P_{s_{H-1},a_{H-1}}, a_H \sim \pi_H(s_{H-1}), r_H \sim R_{s_H,a_H}
\end{aligned}$$

Optimal policy

- Value functions are now also indexed by time step within episode:

$$V_M^{\boldsymbol{\pi},h}(s) = \mathbb{E}_M^{\boldsymbol{\pi}}\left[\sum_{t=h}^{H} r_t \middle| s_h = s\right]$$

- Optimal policy $\boldsymbol{\pi}_M^\star$ satisfies, for all $s \in \mathcal{S}, h \in \{1, \ldots, H\}$:

$$V_M^{\boldsymbol{\pi}_M^\star,h}(s) = \max_{\boldsymbol{\pi}} V_M^{\boldsymbol{\pi},h}(s)$$

- Will omit MDP M if it is fixed and clear from context

DP equation for value functions of a policy

- DP equations for finite horizon case

$$V_M^{\boldsymbol{\pi},h} = T_M^{\pi_h} V_M^{\boldsymbol{\pi},h+1}, \quad h = \{1, 2, \ldots, H\}$$

- Base case is $V^{\boldsymbol{\pi},H+1} = 0$
- Here the operator T_M^{π} for a single stationary π is defined as usual:

$$T_M^{\pi} V = R_M^{\pi} + P_M^{\pi} V$$

Regret

- Let's say the agent interacts with a fixed but unknown finite horizon MDP M for T steps
- There are $K = T/H$ episodes each of length H
- Agent chooses policy $\boldsymbol{\pi}^{(k)}$ at the start of episode k (based on available data at that moment)
- Regret in episode k

$$\Delta_k = \sum_{s \in \mathcal{S}} \mu_0(s)(V^{\boldsymbol{\pi}^\star,1}(s) - V^{\boldsymbol{\pi}^{(k)},1}(s))$$

- Overall regret

$$\mathrm{Regret}(T; \mathrm{agent}, M) = \sum_{k=1}^{K} \Delta_k$$

Posterior Sampling: Per Episode Version

- Also called Thompson Sampling because of [Tho33]
- Tends to perform better than optimism based algorithms
- Start with a prior distribution over MDPs
- In every episode:
  - Use collected statistics to create a posterior distribution over MDPs
  - Sample an MDP from this posterior
  - Compute optimal policy for the sampled MDP
  - For time steps within the episode:
    - Choose actions according to the optimal policy for sampled MDP

Posterior Sampling: Per Time Step Version

- Start with a prior distribution over MDPs
- In every episode:
  - For time steps within the episode:
    - Use collected statistics to create a posterior distribution over MDPs
    - Sample an MDP from this posterior
    - Compute optimal policy for the sampled MDP
    - Choose actions according to the optimal policy for sampled MDP

Per Episode vs Per Time Step

- Per time step version does worse, sometimes much worse, than per episode version
- Difference in performance increases as MDP size increases
- Per episode version is also computationally more efficient
- See [RVRK+18], Section 7.5 for details

Bayesian Regret

- Note that worst-case (or frequentist) regret bounds are of the form

$$\sup_{M \in \mathcal{M}} \mathrm{Regret}(T; \mathrm{agent}, M)$$

  for some class $\mathcal{M}$ of MDPs
- It is easier to analyze Bayesian regret of posterior sampling

$$\mathbb{E}_{M \sim f} \left[\mathrm{Regret}(T; \mathrm{agent}, M)\right]$$

- Here f is the prior distribution over MDPs

Posterior Sampling for RL (PSRL)

- **Input:** Prior distribution f
- $t \leftarrow 1$
- **For** episodes $k = 1, 2, \ldots$ **do**
  - sample $\tilde{M}_k \sim f(\cdot|\mathcal{H}_{<k})$
  - compute $\tilde{\boldsymbol{\pi}}^{(k)} = \boldsymbol{\pi}^\star_{\tilde{M}_k}$
  - **For** timesteps $h = 1, \ldots, H$ **do**
    - choose action $a_t = \tilde{\pi}_h^{(k)}(s_t)$
    - observe r_t and s_{t+1}
    - $t \leftarrow t + 1$

For more details see original paper [ORR13]

A Crucial Observation

- (Bayesian) regret analysis of PS rests on a simple but crucial observation
- Let $\mathcal{H}_{<k}$ be the history of all observations available at the start of episode k

$$\mathbb{E}\left[g(\tilde{M}_k)|\mathcal{H}_{<k}\right] = \mathbb{E}\left[g(M)|\mathcal{H}_{<k}\right]$$

  for any $g(\cdot)$ measurable w.r.t. $\mathcal{H}_{<k}$
- The sampled MDP $\tilde{M}_k$ (observed) has the same distribution as the true MDP M (unobserved)!

Regret Equivalence

- Recall per-episode regret

$$\Delta_k = \sum_{s \in \mathcal{S}} \mu_0(s)(V_M^{\pi^\star,1}(s) - V_M^{\pi^{(k)},1}(s))$$

- Consider its proxy

$$\tilde{\Delta}_k = \sum_{s \in \mathcal{S}} \mu_0(s)(V_{\tilde{M}}^{\pi^{(k)},1}(s) - V_M^{\pi^{(k)},1}(s))$$

- Note that by our crucial observation

$$\begin{aligned}\mathbb{E}\left[\Delta_k - \tilde{\Delta}_k \middle| \mathcal{H}_{<k}\right] &= \mathbb{E}\left[\sum_{s \in \mathcal{S}} \mu_0(s)(V_M^{\pi^\star,1}(s) - V_{\tilde{M}}^{\pi^{(k)},1}(s)) \middle| \mathcal{H}_{<k}\right] \\ &= 0\end{aligned}$$

Bounding the Proxy Regret

- So we will focus on bounding

$$\mathbb{E}[\tilde{\Delta}_k] = \mathbb{E}[\sum_{s\in\mathcal{S}} \mu_0(s)(V_{\tilde{M}}^{\boldsymbol{\pi}^{(k)},1}(s) - V_{M}^{\boldsymbol{\pi}^{(k)},1}(s))]$$
$$= \mathbb{E}[V_{\tilde{M}}^{\boldsymbol{\pi}^{(k)},1}(s_{t_k+1}) - V_{M}^{\boldsymbol{\pi}^{(k)},1}(s_{t_k+1})]$$

- Recall DP equations for finite horizon case (with $V^{\boldsymbol{\pi},H+1} = 0$ as base case)

$$V_M^{\boldsymbol{\pi},h} = T_M^{\pi_h} V_M^{\boldsymbol{\pi},h+1}, \quad h = \{1, 2, \ldots, H\}$$

where the operator T_M^{π} for a single stationary π is defined as usual:

$$T_M^{\pi} V = R_M^{\pi} + P_M^{\pi} V$$

Towards the Key Recursion

refer to states within the episodes as $s_1, s_2, \ldots$ instead of $s_{t_k+1}, s_{t_k+2}, \ldots$
denote the non-stationary policy $\boldsymbol{\pi}^{(k)}$ in episode k as $\tilde{\boldsymbol{\pi}}$

$$
\begin{aligned}
V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},1} - V_{M}^{\tilde{\boldsymbol{\pi}},1} &= T_{\tilde{M}}^{\tilde{\pi}_1} V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - T_{M}^{\tilde{\pi}_1} V_{M}^{\tilde{\boldsymbol{\pi}},2} \\
&= T_{\tilde{M}}^{\tilde{\pi}_1} V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - T_{M}^{\tilde{\pi}_1} V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} + T_{M}^{\tilde{\pi}_1} V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - T_{M}^{\tilde{\pi}_1} V_{M}^{\tilde{\boldsymbol{\pi}},2} \\
&= (T_{\tilde{M}}^{\tilde{\pi}_1} - T_{M}^{\tilde{\pi}_1}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} + T_{M}^{\tilde{\pi}_1} (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2}) \\
&= (T_{\tilde{M}}^{\tilde{\pi}_1} - T_{M}^{\tilde{\pi}_1}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} + P_{M}^{\tilde{\pi}_1} (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2})
\end{aligned}
$$

Therefore,

$$
\mathbf{e}_{s_1}^\top (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},1} - V_{M}^{\tilde{\boldsymbol{\pi}},1}) = \mathbf{e}_{s_1}^\top (T_{\tilde{M}}^{\tilde{\pi}_1} - T_{M}^{\tilde{\pi}_1}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} + \mathbf{e}_{s_1}^\top P_{M}^{\tilde{\pi}_1} (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2})
$$

Key Recursion

$$
\begin{aligned}
\mathbf{e}_{s_1}^\top (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},1} - V_{M}^{\tilde{\boldsymbol{\pi}},1}) &= \mathbf{e}_{s_1}^\top (T_{\tilde{M}}^{\tilde{\pi}_1} - T_{M}^{\tilde{\pi}_1}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} + \mathbf{e}_{s_1}^\top P_{M}^{\tilde{\pi}_1} (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2}) \\
&= \mathbf{e}_{s_1}^\top (T_{\tilde{M}}^{\tilde{\pi}_1} - T_{M}^{\tilde{\pi}_1}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} + \mathbf{e}_{s_2}^\top (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2}) \\
&\quad + \underbrace{(\mathbf{e}_{s_1}^\top P_{M}^{\tilde{\pi}_1} - \mathbf{e}_{s_2}^\top)(V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2})}_{\text{mean zero given } M,\, \tilde{M}}
\end{aligned}
$$

We have therefore set up the key recursion

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{e}_{s_1}^\top (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},1} - V_{M}^{\tilde{\boldsymbol{\pi}},1}) \middle| M, \tilde{M}\right] &= \mathbb{E}\left[\mathbf{e}_{s_1}^\top (T_{\tilde{M}}^{\tilde{\pi}_1} - T_{M}^{\tilde{\pi}_1}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} \middle| M, \tilde{M}\right] \\
&\quad + \mathbb{E}\left[\mathbf{e}_{s_2}^\top (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},2} - V_{M}^{\tilde{\boldsymbol{\pi}},2}) \middle| M, \tilde{M}\right]
\end{aligned}
$$

Unrolling the Recursion

Unrolling the key recursion gives

$$\begin{aligned}\mathbb{E}\left[\tilde{\Delta}_k \middle| M, \tilde{M}\right] &= \mathbb{E}\left[\mathbf{e}_{s_1}^\top (V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},1} - V_{M}^{\tilde{\boldsymbol{\pi}},1}) \middle| M, \tilde{M}\right] \\ &= \mathbb{E}\left[\sum_{h=1}^{H} \mathbf{e}_{s_h}^\top (T_{\tilde{M}}^{\tilde{\pi}_h} - T_{M}^{\tilde{\pi}_h}) V_{\tilde{M}}^{\tilde{\boldsymbol{\pi}},h+1} \middle| M, \tilde{M}\right]\end{aligned}$$

Enter Confidence Sets

Similar to UCRL2 analysis (but now confidence sets are only in the analysis, not in the algorithm!), define $\mathcal{M}_k$ as the set of all MDPs M' such that $\forall s, a$,

$$\begin{aligned} \|P_{M'}(\cdot|s,a) - P_M(\cdot|s,a)\|_1 &\le \beta_k(s,a) \\ |R_{M'}(s,a) - R_M(s,a)| &\le \beta_k(s,a) \end{aligned}$$

where

$$\beta_k(s,a) = O\left(\sqrt{\frac{S\log(SAK)}{1 \vee N_{t_k}(s,a)}}\right)$$

Confidence Set Failure Probability

Can easily show that

$$\mathbb{E}[\mathbf{1}_{(M \notin \mathcal{M}_k)}] \leq 1/K$$

Note that $\mathcal{M}_k$ is $\mathcal{H}_{<k}$-measurable which, using the crucial observation again, gives

$$\mathbb{E}[\mathbf{1}_{(\tilde{M}_k \notin \mathcal{M}_k)}] \leq 1/K$$

Sum up Regret over Episodes

Now we sum up regrets over all episodes

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}\tilde{\Delta}_k\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\tilde{\Delta}_k \mathbf{1}_{\left(M,\tilde{M}_k\in\mathcal{M}_k\right)}\right] + \mathbb{E}\left[\sum_{k=1}^{K}\tilde{\Delta}_k \mathbf{1}_{\left(M \text{ or } \tilde{M}_k\notin\mathcal{M}_k\right)}\right] \\
&\le \mathbb{E}\left[\sum_{k=1}^{K}\tilde{\Delta}_k \mathbf{1}_{\left(M,\tilde{M}_k\in\mathcal{M}_k\right)}\right] + H\sum_{k=1}^{K}2\mathbb{E}[\mathbf{1}_{(M\notin\mathcal{M}_k)}] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\mathbb{E}\left[\tilde{\Delta}_k \middle| M,\tilde{M}\right]\mathbf{1}_{\left(M,\tilde{M}_k\in\mathcal{M}_k\right)}\right] + 2H
\end{aligned}$$

Recall that we proved that

$$\mathbb{E}\left[\tilde{\Delta}_k \middle| M,\tilde{M}\right] = \mathbb{E}\left[\sum_{h=1}^{H}\mathbf{e}_{s_{t_k+h}}^{\top}\left(T_{\tilde{M}_k}^{\tilde{\pi}_h^{(k)}} - T_{M}^{\tilde{\pi}_h^{(k)}}\right)V_{\tilde{M}_k}^{\tilde{\boldsymbol{\pi}}^{(k)},h+1} \middle| M,\tilde{M}_k\right]$$

DP Operators Concentrate

On the event $M, \tilde{M}_k \in \mathcal{M}_k$, the two MDPs are close

Therefore $T^{\tilde{\pi}_h^{(k)}}_{\tilde{M}_k}$ and $T^{\tilde{\pi}_h^{(k)}}_{M}$ are also close

Also, value function cannot exceed H (rewards are bounded)

$$\mathbb{E}\left[\sum_k \tilde{\Delta}_k\right] \leq \mathbb{E}\left[\sum_k \sum_{h=1}^{H} |\mathbf{e}^\top_{s_{t_k+h}}(T^{\tilde{\pi}_h^{(k)}}_{\tilde{M}_k} - T^{\tilde{\pi}_h^{(k)}}_{M})V^{\tilde{\boldsymbol{\pi}}^{(k)},h+1}_{\tilde{M}_k}|\mathbf{1}_{(M,\tilde{M}_k \in \mathcal{M}_k)}\right] + 2H$$

$$\leq \underbrace{H \sum_k \sum_{h=1}^{H} \beta_k(s_{t_k+h}, a_{t_k+h})}_{\text{contributes } \tilde{O}(\sqrt{S} \cdot \sqrt{SAT})} + 2H$$

Bayesian Regret Bound for Posterior Sampling

Theorem (from [ORR13])

The Bayesian regret of PSRL in an H horizon problem with bounded rewards is at most $\tilde{O}(HS\sqrt{AT})$.

Regret Analysis of Posterior Sampling: Non-episodic case

- There is a subtlety in the extension of this analysis to the non-episodic case (where we compete against the average reward optimal policy)
- At the start of the episode

$$\mathbb{E}\left[\tilde{\rho}_k | \mathcal{H}_{<k}\right] = \mathbb{E}\left[\rho^\star | \mathcal{H}_{<k}\right]$$

- However, the length of episode k may not be measurable w.r.t. $\mathcal{H}_{<k}$ (see [OVR16] for explanation of this subtlety)
- Redefining the stopping criterion in posterior sampling allows us to prove Bayesian regret bounds [OGNJ17]
- Frequentist aka worst-case regret analysis more difficult and still not fully resolved in the non-episodic setting

Summary

- Posterior sampling replaces optimism with sampling (from the posterior)
- Bayesian regret analysis relies on the equality of the distributions of the true and the sampled MDPs
- Confidence intervals still needed but only in the analysis
- Worst-case/frequentist analysis is technically more challenging
- Works better than optimism in practice (see [OVR17] for more discussion)

References I

- Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain, *Learning unknown markov decision processes: A Thompson sampling approach*, Advances in Neural Information Processing Systems, 2017, pp. 1333–1342.
- Ian Osband, Benjamin Van Roy, and Daniel Russo, *(More) efficient reinforcement learning via posterior sampling*, Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 3003–3011.
- Ian Osband and Benjamin Van Roy, *Posterior sampling for reinforcement learning without episodes*, 2016.
- Ian Osband and Benjamin Van Roy, *Why is posterior sampling better than optimism for reinforcement learning?*, Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 2701–2710.

References II

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen, *A tutorial on thompson sampling*, Foundations and Trends® in Machine Learning **11** (2018), no. 1, 1–96.

William R Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika **25** (1933), no. 3/4, 285–294.