Online Learning in MDPs
Part 1

Ambuj Tewari

Department of Statistics, and
Department of EECS (by courtesy),
University of Michigan, Ann Arbor

RL20 Boot Camp
September 1, 2020

Outline

Outline

Origins of RL

- Minsky first used the term “Reinforcement Learning” [Min61]
- Waltz and Fu independently used the term a few years later [WF65]
- Earliest ML research viewed as directly relevant now Samuel’s checker playing program 1959
- Not much activity in 1970s
- Modern field of RL created in the late 1980s

Beginnings of Regret Analysis

- Progress continued into the 1990s
  - Sutton & Barto 1st edition 1998
  - Kaelbling, Littman, Moore 1996 survey [KLM96]
    "Unfortunately, results concerning the regret of algorithms are quite hard to obtain"
- Sample complexity concerns arose in the early 2000s
  - E^3 [KS02] and R-MAX [BT02]
  - Sham Kakade's thesis 2003 [Kak03]
- UCRL2 paper [JOA10] kicks off regret analysis in MDPs (conference version in NIPS 2008)

Online Learning and Regret

- In online learning, an agent learns from sequential interaction with an environment (often an MDP)
  - Experience arrives bit by bit
  - No separation between learning phase and evaluation phase
- Explore-Exploit trade-off: learning vs earning, estimation vs control
- Regret measures the difference between:
  - some benchmark/competitor/yardstick (typically known only in hindsight), and
  - the agent's actual performance
- This part (Part 1) deals with the fixed MDP case
  - Part 2 will deal with changing MDPs, potentially chosen adversarially

The Hare and the Tortoise

"If the inference/algorithm race is a tortoise-and-hare affair, then modern electronic computation has bred a bionic hare."

– Efron & Hastie, Computer Age Statistical Inference

- Deep RL has taken off in the past 5-6 years
- Google Scholar lists 16,600 papers during 2011-2020 with RL in the title (for 2001-2010 it's 6,400)
- Sutton & Barto 2nd edition 2018 ("twice as large as the first")
- The "theory tortoise" has lots to catch up
- Hope that the RL20 program will breed a faster tortoise!

E^3 (Explicit Explore or Exploit) algorithm

- Makes a distinction between known and unknown based on visitation counts
- In unknown state: take least tried action
- Maintain a partial model: this will be good on the known states
- In a known state: perform two calculations
  - attempted exploitation: is there a high return policy based on the partial model?
  - attempted exploration: is there a policy with non-trivial probability of leaving the known states fast?
- Analysis hinges on two key lemmas
  - Simulation Lemma: Values of a policy in actual MDP restricted to the known states and in partial model are close
  - Explore or Exploit Lemma: At least one of the attempted calculations will succeed

R-MAX

- Retains the distinction between known and unknown states
- But simplifies the algorithm with implicit explore-exploit
- Uses OFU (Optimism in the Face of Uncertainty) principle
- Unknown states are given maximum reward (R-MAX!) with self-loops
- Analysis covers not just MDPs but also (2-player, fixed sum) stochastic games

OFU Principle

- Appears under "Ad-hoc techniques" in [KLM96]
- Sutton & Barto: "a simple trick that can be quite effective on stationary problems"
- Related ideas in adaptive control:
  - cost-biased estimation [CK98]
  - bet-on-the-best principle [BC06]
- The R-MAX paper provided theoretical justification for the OFU principle

E^3, R-MAX and UCRL2

K/U = Known/Unknown state distinction
E/E = Explore/Exploit distinction

	Explicit K/U	Explicit E/E	Explicit OFU
E^3	✓	✓	×
R-MAX	✓	×	✓
UCRL2	×	×	✓

Outline

High Level Description

- Runs in episodes – these are used by the algorithm only
- Actual experience is one long trajectory

$$s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T$$

  generated during interaction with a tabular MDP with S states, A actions, reward function $r(s, a)$ and transition function $p(s'|s, a)$
- In every episode:
  - Use collected statistics to create set of plausible MDPs
  - Pick most optimistic MDP from this set
  - Follow the optimal policy for this MDP until a stopping criterion is satisfied

Set of Plausible MDPs - I

- Let t_k be the start time for episode k
- Visitation count for (s, a) pairs and (s, a, s') triples

$$N_k(s,a) = |\{\tau < t_k \; : \; s_\tau = s, a_\tau = a\}|$$

$$N_k(s,a,s') = \left|\left\{\tau < t_k \; : \; s_\tau = s, a_\tau = a, s_{\tau+1} = s'\right\}\right|$$

- Accumulated reward for (s, a) pairs

$$R_k(s,a) = \sum_{\tau < t_k} r_\tau \mathbf{1}_{(s_\tau = s, a_\tau = a)}$$

- Reward and transition function estimates

$$\hat{r}_k(s,a) = \frac{R_k(s,a)}{1 \vee N_k(s,a)} \quad \hat{p}_k(s'|s,a) = \frac{N_k(s,a,s')}{1 \vee N_k(s,a)}$$

Set of Plausible MDPs - II

- $\mathcal{M}_k$ consists of all MDPs with reward and transition functions close to our estimates

$$\forall s, a, \; |r(s,a) - \hat{r}_k(s,a)| \leq \sqrt{\frac{\log(SAt_k/\delta)}{1 \vee N_k(s,a)}}$$

$$\forall s, a, \; \left\| p(s'|s,a) - \hat{p}_k(s'|s,a) \right\|_1 \leq \sqrt{\frac{S\log(At_k/\delta)}{1 \vee N_k(s,a)}}$$

Optimism and Stopping Criterion

- $\rho^\star(M)$: optimal long term average reward obtainable in MDP M
- Find optimistic MDP $\tilde{M}_k$ such that

$$\tilde{M}_k := \operatorname*{argmax}_{M \in \mathcal{M}_k,\, D(M) \leq D} \rho^\star(M)$$

  and let $\tilde{\pi}_k$ be an optimal policy for $\tilde{M}_k$
- Follow the policy $\tilde{\pi}_k$ until you reach a state s_t such that

$$v_k(s_t, \tilde{\pi}_k(s_t)) \geq 1 \vee N_k(s_t, \tilde{\pi}_k(s_t))$$

- $v_k(s, a)$ is the visitation count within episode k
  (so $N_{k+1} = N_k + v_k$)

Average Reward Criterion

- The long term average reward

$$\rho(M, \pi, s) := \lim \sup_{T \to \infty} \frac{1}{T} \mathbb{E}^{M,\pi} \left[\sum_{t=1}^{T} r_t \middle| s_1 = s \right]$$

- Assume MDP is communicating, i.e., has finite diameter

$$D(M) := \max_{s \neq s'} \min_{\pi} \mathbb{E}^{M,\pi}[T_{s'} | s_1 = s]$$

where $T_{s'}$ = first time you visit s' (under π starting from s)

- Then optimal reward $\rho^\star(M)$ is well defined and independent of start state

$$\forall s, \ \rho^\star(M) = \rho^\star(M, s) := \max_{\pi} \rho(M, \pi, s)$$

Bellman equation

- The optimal policy $\pi^\star$ with (state-independent) gain $\rho^\star$ satisfies

$$\forall s,\ \rho^\star + h^\star(s) = r(s, \pi^\star(s)) + \sum_{s'} p(s'|s, \pi^\star(s)) h^\star(s')$$

- The bias vector $h^\star$ is not unique
  (e.g., can shift it by a constant)
- Relationship with diameter

$$\text{span}(h^\star) \leq D$$

where $\text{span}(h) = \max_s h(s) - \min_s h(s)$

Outline

Regret

- T-step regret of algorithm $\mathcal{A}$ in M starting from s:

$$\Delta(M, \mathcal{A}, s, T) := \underbrace{\rho^\star(M) \cdot T}_{\text{benchmark performance}} - \underbrace{\sum_{t=1}^{T} r_t}_{\mathcal{A}\text{'s performance}}$$

- With probability at least $1 - \delta$, for any s and any $T > 1$,

$$\Delta(M, UCRL2, s, T) \leq 34 \cdot DS\sqrt{AT \log(T/\delta)}$$

in any MDP with S states, A actions, and diameter D.

Reduction to Per Episode Regret

- For simplicity assume deterministic reward $r(s, a)$
- Per episode regret

$$\Delta_k = \sum_{s,a} v_k(s, a)(\rho^\star - r(s, a))$$

- Decompose regret over episodes

$$\Delta = \sum_{k=1}^{m} \Delta_k$$

- Due to the stopping criterion for episodes, can show that $m = O(SA \log T)$

Failure of Confidence Regions

- The set are chosen so that standard concentration arguments give

$$\mathbb{P}\left(M \notin \mathcal{M}(t)\right) \leq \frac{\delta}{15t^6}$$

- This can be used to show that w.h.p.

$$\sum_{k=1}^{m} \Delta_k \mathbf{1}_{(M \notin \mathcal{M}_k)} \leq \sqrt{T}$$

Using Optimism

- Suppose our confidence regions are correct

$$\Delta_k = \sum_{s,a} v_k(s,a)(\rho^\star - r(s,a))$$
$$\leq \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - r(s,a))$$

- Due to optimism, we know that $\tilde{\rho}_k \geq \rho^\star$
- Bellman equation for $\tilde{\pi}_k$

$$\tilde{\rho}_k \mathbf{1} + \tilde{\mathbf{h}}_k = \tilde{\mathbf{r}}_k + \tilde{\mathbf{P}}_k \tilde{\mathbf{h}}_k$$

where

$$\tilde{\mathbf{r}}_k(s) = \tilde{r}_k(s, \tilde{\pi}_k(s)) \qquad \tilde{\mathbf{P}}_k(s,s') = \tilde{p}_k(s'|s, \tilde{\pi}_k(s))$$

Isolating the Dominant Term

$$\Delta_k \leq \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - r(s,a))$$

$$= \underbrace{\sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}_k(s,a))}_{\text{dominant contribution to regret}} + \underbrace{\sum_{s,a} v_k(s,a)(\tilde{r}(s,a) - r(s,a))}_{\text{essentially } \frac{v_k(s,a)}{\sqrt{1 \vee N_k(s,a)}}}$$

Controlling the Dominant Term - I

$$
\begin{aligned}
&\sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}_k(s,a)) \\
&= \sum_{s} v_k(s,\tilde{\pi}_k(s))(\tilde{\rho}_k - \tilde{r}_k(s,\tilde{\pi}_k(s)) \\
&= \mathbf{v}_k^\top(\tilde{\rho}_k \mathbf{1} - \tilde{\mathbf{r}}_k) \\
&= \mathbf{v}_k^\top(\tilde{\mathbf{P}}_k - \mathbf{I})\tilde{\mathbf{h}}_k \qquad \text{recall Bellman equation below}
\end{aligned}
$$

Bellman equation:

$$
\tilde{\rho}_k \mathbf{1} + \tilde{\mathbf{h}}_k = \tilde{\mathbf{r}}_k + \tilde{\mathbf{P}}_k \tilde{\mathbf{h}}_k
$$

Controlling the Dominant Term - II

Transition kernel of $\tilde{\pi}_k$ in the true MDP:

$$\mathbf{P}_k(s,s') = p(s'|s,\tilde{\pi}_k(s))$$

$$\begin{aligned}
&\mathbf{v}_k^\top(\tilde{\mathbf{P}}_k - \mathbf{I})\tilde{\mathbf{h}}_k \\
&= \mathbf{v}_k^\top(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\tilde{\mathbf{h}}_k + \underbrace{\mathbf{v}_k(\mathbf{P}_k - \mathbf{I})}_{\text{would be zero for SD of } \tilde{\pi}_k} \tilde{\mathbf{h}}_k \\
&\leq \underbrace{\mathbf{v}_k^\top(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\tilde{\mathbf{h}}_k}_{\tilde{\mathbf{P}}_k, \mathbf{P}_k \text{ are close}} + \underbrace{\text{martingale diff. seq.} + D}_{\text{overall contribution } \tilde{O}(D\sqrt{T}) + mD}
\end{aligned}$$

Controlling the Dominant Term - III

$$
\begin{aligned}
&\mathbf{v}_k^\top(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\tilde{\mathbf{h}}_k \\
&= \sum_s \sum_{s'} v_k(s,\tilde{\pi}_k(s)) \cdot (\tilde{p}_k(s'|s,\tilde{\pi}_k(s)) - p_k(s'|s,\tilde{\pi}_k(s)) \cdot \tilde{h}_k(s') \\
&= \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{s'} (\tilde{p}_k(s'|s,\tilde{\pi}_k(s)) - p_k(s'|s,\tilde{\pi}_k(s)) \cdot \tilde{h}_k(s') \\
&= \sum_s v_k(s,\tilde{\pi}_k(s)) \cdot \|\tilde{p}_k(\cdot|s,\tilde{\pi}_k(s)) - p_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \cdot \|\tilde{\mathbf{h}}_k\|_\infty \\
&\leq \sum_s v_k(s,\tilde{\pi}_k(s)) \cdot \sqrt{\frac{S\log(At_k/\delta)}{1 \vee N_k(s,\tilde{\pi}_k(s))}} \cdot D \\
&\leq D\sqrt{S\log(AT/\delta)} \underbrace{\sum_{s,a} \frac{v_k(s,a)}{\sqrt{1 \vee N_k(s,a)}}}_{\text{overall contribution } \sqrt{SAT}} = O\left(DS\sqrt{AT\log(T/\delta)}\right)
\end{aligned}
$$

Why $\sqrt{SAT}$?

$$\begin{aligned}
\sum_{k=1}^{m}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{1\vee N_k(s,a)}} &= \sum_{s,a}\sum_{k=1}^{m}\frac{v_k(s,a)}{\sqrt{1\vee N_k(s,a)}} \\
&\leq \sum_{s,a} 3\sqrt{N(s,a)} && \text{fact below \& } v_k \leq N_k \\
&\leq 3\sqrt{SA}\sqrt{\sum_{s,a} N(s,a)} && \text{concavity of square-root} \\
&= 3\sqrt{SAT}
\end{aligned}$$

Fact: For $Z_k = 1 \vee \sum_{i=1}^{k-1} z_k$ and $0 \leq z_k \leq Z_k$, we have

$$\sum_{k=1}^{n}\frac{z_k}{\sqrt{Z_k}} \leq 3\sqrt{Z_{n+1}}$$

Outline

Tightness of the Bound

- The UCRL2 paper [JOA10] also proved a lower bound
- For any algorithm $\mathcal{A}$, any $S, A \geq 10$, $D \geq 20 \log_A S$ and $T \geq DSA$, there is an MDP with S states, A actions, diameter D such that for any s
$$\mathbb{E}\left[\Delta(M, \mathcal{A}, s, T)\right] \geq 0.015 \cdot \sqrt{DSAT}$$
- Gap of roughly $\sqrt{DS}$ between upper and lower bounds
- Recent preprint [TBD19] claims to eliminate the gap by analyzing an improved algorithm called UCRL-V

Posterior Sampling

- 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
  - Follow the optimal policy for this MDP until a stopping criterion is satisfied

Regret Analysis of Posterior Sampling

- It is easier to analyze Bayesian regret of posterior sampling
- 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

Beyond UCRL2 - I

- Other algorithmic ideas: Thompson Sampling, Injecting Random Noise
- Other optimality criteria: discounted infinite horizon, finite horizon
- Model-free vs model-based: do we need to build an (approximate) model of the environment?
- Large/continuous state spaces: Factored MDPs, function approximation (recent work on LQ systems, Linear/Low Rank MDPs)

Beyond UCRL2 - II

- Learning across multiple MDPs: learning to learn, meta-learning, transfer learning, multi-task learning, curriculum learning
- Causality: Can causal knowledge help learn faster? Help with transfer learning?
- Partial Observability: Hard even without learning!
- Multi-agent RL: What is a good goal for learning?

Summary

- How well is an agent learning in an online setup?
- Finite-time regret analysis offers one theoretical approach among many
- UCRL2, like R-MAX, is based on the OFU principle
- Provided a detailed overview of its regret analysis
- Many interesting new research directions!

Thank You!

References I

- Sergio Bittanti and Marco C Campi, *Adaptive control of linear time invariant systems: the "bet on the best" principle*, Communications in Information & Systems **6** (2006), no. 4, 299–320.
- Ronen I Brafman and Moshe Tennenholtz, *R-max-a general polynomial time algorithm for near-optimal reinforcement learning*, Journal of Machine Learning Research **3** (2002), no. Oct, 213–231.
- Marco C Campi and PR Kumar, *Adaptive linear quadratic gaussian control: the cost-biased approach revisited*, SIAM Journal on Control and Optimization **36** (1998), no. 6, 1890–1907.
- Thomas Jaksch, Ronald Ortner, and Peter Auer, *Near-optimal regret bounds for reinforcement learning*, Journal of Machine Learning Research **11** (2010), 1563–1600.
- Sham M Kakade, *On the sample complexity of reinforcement learning*, Ph.D. thesis, University College London, 2003.

References II

- Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore, *Reinforcement learning: A survey*, Journal of artificial intelligence research **4** (1996), 237–285.
- Michael Kearns and Satinder Singh, *Near-optimal reinforcement learning in polynomial time*, Machine learning **49** (2002), no. 2-3, 209–232.
- Marvin Minsky, *Steps toward artificial intelligence*, Proceedings of the IRE **49** (1961), no. 1, 8–30.
- 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 and Benjamin Van Roy, *Posterior sampling for reinforcement learning without episodes*, 2016.

References III

Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis, *Near-optimal optimistic reinforcement learning using empirical bernstein inequalities*, 2019, arXiv preprint 1905.12425v2.

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.

M Waltz and K Fu, *A heuristic approach to reinforcement learning control systems*, IEEE Transactions on Automatic Control **10** (1965), no. 4, 390–398.