Near-optimal Reinforcement Learning in Factored MDPs

Ian Osband Benjamin Van Roy

Mangement Science and Engineering
Stanford University
iosband@stanford.edu

November 7, 2014

Table of contents

Reinforcement Learning

- We imagine an agent taking actions within an environment.
- Actions serve two purposes:
    - Instantaneous loss/reward to the agent.
    - Influence the state of the environement.
- The agent wants to maximize cumulative reward through time.
- How can an agent learn to take “good” actions?

Reinforcement Learning

- We imagine an agent taking actions within an environment.
- Actions serve two purposes:
  - Instantaneous loss/reward to the agent.
  - Influence the state of the environement.
- The agent wants to maximize cumulative reward through time.
- How can an agent learn to take "good" actions?

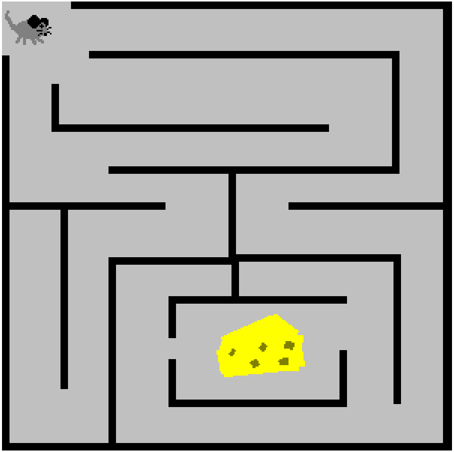
- Multi-armed bandit with added state transitions.
- Statistical estimation + optimal control.

Reinforcement Learning

- We imagine an agent taking actions within an environment.
- Actions serve two purposes:
  - Instantaneous loss/reward to the agent.
  - Influence the state of the environement.
- The agent wants to maximize cumulative reward through time.
- How can an agent learn to take "good" actions?

- Multi-armed bandit with added state transitions.
- Statistical estimation + optimal control.
- ...this could get hard!

Mouse in a maze



What's the best way to the cheese?

Self-driving car

Drive me from A to B

Self-driving car

Drive me from A to B... also don't kill anyone.

Markov decision process (MDP)

- Model decision making in a stochastic environment.
- At time t the agent in state $s_t \in \mathcal{S}$ chooses action $a_t \in \mathcal{A}$.
- Reward $r_t \sim R(s_t, a_t)$ and transition to $s_{t+1} \sim P(\cdot|s_t, a_t)$.
- Given R, P there is some optimal policy $\pi^* : s_t \to a_t^*$.
- Generally this is computed via Dynamic Programming.

Markov decision process (MDP)

- Model decision making in a stochastic environment.
- At time t the agent in state $s_t \in \mathcal{S}$ chooses action $a_t \in \mathcal{A}$.
- Reward $r_t \sim R(s_t, a_t)$ and transition to $s_{t+1} \sim P(\cdot|s_t, a_t)$.
- Given R, P there is some optimal policy $\pi^* : s_t \to a_t^*$.
- Generally this is computed via Dynamic Programming.

- In reinforcement learning, the agent is unsure of P, R.
- The learning algorithm will pick $\pi_t^L : s_t \to a_t^L$.
- We would like to get average reward $\rho(\pi^L)$ close to $\rho(\pi^*)$.

Efficient reinforcement learning

- Will we learn the best policy?

$$\rho(\pi_t^L) \to \rho(\pi^*)$$

Efficient reinforcement learning

- Will we learn the best policy?

$$\rho(\pi_t^L) \to \rho(\pi^*)$$

- How long do we have to wait to do well? (Sample complexity)

$$\forall t > T(\mathrm{MDP}, \pi^L) \quad \rho(\pi_t^L) \geq \rho(\pi^*) - \epsilon$$

Efficient reinforcement learning

- Will we learn the best policy?

$$\rho(\pi_t^L) \to \rho(\pi^*)$$

- How long do we have to wait to do well? (Sample complexity)

$$\forall t > T(\mathrm{MDP}, \pi^L) \quad \rho(\pi_t^L) \geq \rho(\pi^*) - \epsilon$$

- How badly do we do while we're learning? (Regret)

$$\mathrm{Regret}(T, \pi^L) = \sum_{t=1}^{T} r_t^* - r_t \leq f(\mathrm{MDP}, T, \pi^L)$$

Reward and transition functions

We will specialize our analysis two important function classes:

Definition (Reward functions $\in \mathcal{P}^{C,\sigma}_{\mathcal{S}\times\mathcal{A},\mathbb{R}}$)

$\mathcal{P}^{C,\sigma}_{\mathcal{X},\mathbb{R}}$ is the set of functions from $\mathcal{X}$ to σ-sub gaussian measures over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with mean in $[0, C]$.

Definition (Transition functions $\in \mathcal{P}_{\mathcal{S}\times\mathcal{A},\mathcal{S}}$)

$\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ is the set of functions mapping elements of a finite set $\mathcal{X}$ to probability mass functions over a finite set $\mathcal{Y}$.

Regret bounds

- Greedy algorithms may never learn $\rho(\pi_t^L) \nrightarrow \rho(\phi^*)$.
- Naïve exploration leads to regret exponential in $|\mathcal{S}|, |\mathcal{A}|$.
- Efficient algorithms must guide their exploration:

Regret bounds

- Greedy algorithms may never learn $\rho(\pi_t^L) \nrightarrow \rho(\phi^*)$.
- Naïve exploration leads to regret exponential in $|\mathcal{S}|, |\mathcal{A}|$.
- Efficient algorithms must guide their exploration:
  - "Optimism in the face of uncertainty" (OFU).
  - "Posterior sampling" (PS)

Regret bounds

- Greedy algorithms may never learn $\rho(\pi_t^L) \nrightarrow \rho(\phi^*)$.
- Naïve exploration leads to regret exponential in $|\mathcal{S}|, |\mathcal{A}|$.
- Efficient algorithms must guide their exploration:
  - "Optimism in the face of uncertainty" (OFU).
  - "Posterior sampling" (PS)

- Efficent algorithms with Regret $\tilde{O}(|\mathcal{S}|\sqrt{|\mathcal{A}|T})$.
- Lower bound on Regret $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|T})$.
- But in many cases of interest $\mathcal{S}, \mathcal{A}$ are huge...

Regret bounds

- Greedy algorithms may never learn $\rho(\pi_t^L) \nrightarrow \rho(\phi^*)$.
- Naïve exploration leads to regret exponential in $|\mathcal{S}|, |\mathcal{A}|$.
- Efficient algorithms must guide their exploration:
  - "Optimism in the face of uncertainty" (OFU).
  - "Posterior sampling" (PS)
- Efficent algorithms with Regret $\tilde{O}(|\mathcal{S}|\sqrt{|\mathcal{A}|T})$.
- Lower bound on Regret $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|T})$.
- But in many cases of interest $\mathcal{S}, \mathcal{A}$ are huge... ☹

Factored MDPs

Each transition only depends directly on a subset of the MDP.

Factored MDPs

Let $\mathcal{S} \times \mathcal{A} = \mathcal{X} = \mathcal{X}_1 \times .. \times \mathcal{X}_n$, $Z \subseteq [n]$ and $\mathcal{X}[Z] := \bigotimes_{i \in Z} \mathcal{X}_i$

Definition (Factored reward functions $R \in \mathcal{R} \subseteq \mathcal{P}^{C,\sigma}_{\mathcal{X},\mathbb{R}}$)

$\mathcal{R}$ is factored over $\mathcal{X} = \mathcal{X}_1 \times .. \times \mathcal{X}_n$ with scopes $Z_1, ..Z_l \iff$, for all $R \in \mathcal{R}, x \in \mathcal{X}$ there exist,

$$\mathbb{E}[R(x)] = \textstyle\sum_{i=1}^{l} \mathbb{E}\left[R_i(x[Z_i])\right]$$

with each $r_i \sim R_i(x[Z_i])$ and individually observed.

Definition (Factored transition functions $P \in \mathcal{P} \subseteq \mathcal{P}_{\mathcal{X},\mathcal{S}}$)

$\mathcal{P}$ is factored over $\mathcal{X} = \mathcal{X}_1 \times .. \times \mathcal{X}_n$ and $\mathcal{S} = \mathcal{S}_1 \times .. \times \mathcal{S}_m$ with scopes $Z_1, ..Z_m \iff$, for all $P \in \mathcal{P}, x \in \mathcal{X}, s \in \mathcal{S}$ there exist,

$$P(s|x) = \prod_{i=1}^{m} P_i \left(s[i] \,\middle|\, x[Z_i] \right)$$

Factored MDPs

- Agent knows $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^m;\ \{\mathcal{X}_i\}_{i=1}^n;\ \{Z_i^R\}_{i=1}^l;\ \{Z_i^P\}_{i=1}^m\right)$.
- Must learn $\left(\{R_i\}_{i=1}^l;\ \{P_i\}_{i=1}^m\right)$ from experience.
- Algorithms ignoring $\mathcal{G}$ lead to exponential regret bounds.
- $|\mathcal{S}| = \prod_{i=1}^m |\mathcal{S}_i| = |\mathcal{S}_1|^m, \quad |\mathcal{S}||\mathcal{A}| = \prod_{i=1}^n |\mathcal{X}_i| = |\mathcal{X}_1|^n$

Factored MDPs

- Agent knows $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^m;\ \{\mathcal{X}_i\}_{i=1}^n;\ \{Z_i^R\}_{i=1}^l;\ \{Z_i^P\}_{i=1}^m\right)$.
- Must learn $\left(\{R_i\}_{i=1}^l;\ \{P_i\}_{i=1}^m\right)$ from experience.
- Algorithms ignoring $\mathcal{G}$ lead to exponential regret bounds.
- $|\mathcal{S}| = \prod_{i=1}^m |\mathcal{S}_i| = |\mathcal{S}_1|^m, \quad |\mathcal{S}||\mathcal{A}| = \prod_{i=1}^n |\mathcal{X}_i| = |\mathcal{X}_1|^n$

- Efficient complexity bounds exist polynomial in $|\mathcal{X}_i|, |\mathcal{S}_i|$
- No such results for regret...

Factored MDPs

- Agent knows $\mathcal{G} = (\{\mathcal{S}_i\}_{i=1}^m;\ \{\mathcal{X}_i\}_{i=1}^n;\ \{Z_i^R\}_{i=1}^l;\ \{Z_i^P\}_{i=1}^m)$.
- Must learn $(\{R_i\}_{i=1}^l;\ \{P_i\}_{i=1}^m)$ from experience.
- Algorithms ignoring $\mathcal{G}$ lead to exponential regret bounds.
- $|\mathcal{S}| = \prod_{i=1}^m |\mathcal{S}_i| = |\mathcal{S}_1|^m, \quad |\mathcal{S}||\mathcal{A}| = \prod_{i=1}^n |\mathcal{X}_i| = |\mathcal{X}_1|^n$

- Efficient complexity bounds exist polynomial in $|\mathcal{X}_i|, |\mathcal{S}_i|$
- No such results for regret. . . until now.

Posterior Sampling for Reinforcement Learning

Input: Prior ϕ encoding $\mathcal{G}$, $t = 1$

for episodes $k = 1, 2, ..$ **do**

  sample $M_k \sim \phi(\cdot | H_t)$

  compute $\mu_k = \Gamma(M_k, \sqrt{\tau / k})$ $\leftarrow$(ADP planner)

  for timesteps $j = 1, .., \tau$ **do**

    sample and apply $a_t = \mu_k(s_t, j)$

    observe $r_t^1, .., r_t^l$ and $s_{t+1}^1, .., s_{t+1}^m$

    $t = t + 1$

  end for

end for

UCRL-Factored

Input: Graph structure $\mathcal{G}$, confidence δ, $t = 1$

for episodes $k = 1, 2, ..$ **do**

  $d_t^{R_i} = 4\sigma^2 \log\left(4l|\mathcal{X}[Z_i^R]|k/\delta\right)$ for $i = 1, .., l$

  $d_t^{P_j} = 4|\mathcal{S}_j| \log\left(4m|\mathcal{X}[Z_j^P]|k/\delta\right)$ for $j = 1, .., m$

  $\mathcal{M}_k = \{M \,|\mathcal{G}, \overline{R}_i \in \mathcal{R}_t^i(d_t^{R_i}), P_j \in \mathcal{P}_t^j(d_t^{P_j})\ \forall i, j\}$ ←(Confidence sets)

  compute $\mu_k = \tilde{\Gamma}(\mathcal{M}_k, \sqrt{\tau/k}) \leftarrow$ ←(ADP planner)

  for timesteps $u = 1, .., \tau$ **do**

    sample and apply $a_t = \mu_k(s_t, u)$

    observe $r_t^1, .., r_t^l$ and $s_{t+1}^1, .., s_{t+1}^m$

    $t = t + 1$

  end for

end for

Main results

Theorem (Expected regret for PSRL in factored MDPs)

If the prior ϕ is the distribution of M^* and Ψ is the span of the optimal value function:

$$\mathbb{E}\left[\mathrm{Regret}(T, \pi_\tau^{\mathrm{PS}}, M^*)\right] = \tilde{O}\left(\sigma \sum_{i=1}^{d_1} \sqrt{|\mathcal{X}[Z_i^R]|T} + \mathbb{E}[\Psi] \sum_{j=1}^{d_2} \sqrt{|\mathcal{X}[Z_j^P]||\mathcal{S}_j|T}\right) \tag{1}$$

Theorem (High probability regret for UCRL-Factored)

If D is the diameter of M^*, then for any M^* can bound the regret of UCRL-Factored:

$$\mathrm{Regret}(T, \pi_\tau^{\mathrm{UC}}, M^*) = \tilde{O}\left(\sigma \sum_{i=1}^{d_1} \sqrt{|\mathcal{X}[Z_i^R]|T} + CD \sum_{j=1}^{d_2} \sqrt{|\mathcal{X}[Z_j^P]||\mathcal{S}_j|T}\right) \tag{2}$$

with probability at least $1 - \delta$

Clean bounds in the symmetric case

Let $\mathcal{Q}$ be shorthand for the structure $\mathcal{G}$ such that $l + 1 = m$, $C = \sigma = 1$, $|\mathcal{S}_i| = |\mathcal{X}_i| = K$ and $|Z_i^R| = |Z_i^P| = \zeta$ for all suitable i and write $J = K^\zeta$. In this case $\Psi, D \leq \tau$ trivially.

Corollary (Clean bounds for PSRL)

$$\mathbb{E}\left[\mathrm{Regret}(T, \pi_\tau^{\mathrm{PS}}, M^*)\right] \leq 15 m\tau \sqrt{JKT \log(2mJT)} \qquad (3)$$

Corollary (Clean bounds for UCRL-Factored)

$$\mathrm{Regret}(T, \pi_\tau^{\mathrm{UC}}, M^*) \leq 15 m\tau \sqrt{JKT \log(12mJT/\delta)} \qquad (4)$$

with probability at least $1 - \delta$.

Clean bounds in the symmetric case

Let $\mathcal{Q}$ be shorthand for the structure $\mathcal{G}$ such that $l+1=m$, $C=\sigma=1$, $|\mathcal{S}_i|=|\mathcal{X}_i|=K$ and $|Z_i^R|=|Z_i^P|=\zeta$ for all suitable i and write $J=K^\zeta$. In this case $\Psi, D \leq \tau$ trivially.

The key point is that we go from $\mathcal{G}$-agnostic

$$\tilde{O}(|\mathcal{S}|\sqrt{|\mathcal{A}|T}) = \tilde{O}(\sqrt{J^{m/\zeta}K^m T})$$

to the new bounds

$$\tilde{O}(\sum_{j=1}^{m}\sqrt{|\mathcal{X}[Z_j^P]||\mathcal{S}_j|T}) = \tilde{O}(m\sqrt{JKT})$$

which can be exponentially tighter.

Analysis outline

We consider $Regret(T) = \sum_{k=1}^{m} \Delta_k$, regret within each episode.

$$\Delta_k = V^*_{*,1}(s) - V^*_{k,1}(s) = \left(V^k_{k,1}(s) - V^*_{k,1}(s)\right) + \left(V^*_{*,1}(s) - V^k_{k,1}(s)\right) \quad (5)$$

Where $V^M_{\mu,1}(s)$ is the value of empolying policy μ on MDP M for one episode. We write $*, k$ as shorthand for the optimal and algorithmic MDPs at each stage.

$$\left(V^k_{k,1} - V^*_{k,1}\right)(s_{t_k+1}) = \sum_{i=1}^{\tau} \left(\mathcal{T}^k_{k,i} - \mathcal{T}^*_{k,i}\right) V^k_{k,i+1}(s_{t_k+i}) + \sum_{i=1}^{\tau} d_{t_k+1}. \quad (6)$$

Where d_t is a bounded martingale difference and the first term A:

$$A \leq \sum_{i=1}^{\tau} |\overline{R}^k(x_{k,i}) - \overline{R}^*(x_{k,i})| + \frac{1}{2}\Psi_k \|P^k(\cdot|x_{k,i}) - P^*(\cdot|x_{k,i})\|_1 \quad (7)$$

Key lemma

Lemma (Bounding factored deviations)

Let the transition function class $\mathcal{P} \subseteq \mathcal{P}_{\mathcal{X},\mathcal{S}}$ be factored over $\mathcal{X} = \mathcal{X}_1 \times .. \times \mathcal{X}_n$ and $\mathcal{S} = \mathcal{S}_1 \times .. \times \mathcal{S}_m$ with scopes $Z_1, ..Z_m$. Then, for any $P, \tilde{P} \in \mathcal{P}$ we may bound their L1 distance by the sum of the differences of their factorizations:

$$\|P(x) - \tilde{P}(x)\|_1 \leq \sum_{i=1}^{m} \|P_i(x[Z_i]) - \tilde{P}_i(x[Z_i])\|_1$$

Using Azuma-Hoeffding with a union bound we can bound the Bellman error in terms of a concentration which depends on $|\mathcal{X}[Z_i^P]|$ as opposed to $|\mathcal{X}|$. From the previous slide this gives us bounds on regret.

References

Please see arXiv for a full version of the paper.

References

Please see arXiv for a full version of the paper.

Thanks!