Main Contribution

- Extend Information-Directed Sampling (IDS) to solving general reinforcement learning problems
- Propose practical algorithms to efficiently compute the solution in both model-based and model-free manners
- Provide insight into the regret bound and caveat of the methods

Introduction

Information-directed sampling (IDS) was proposed in [2] to address some shortcomings of Thompson sampling (TS) and UCB algorithm in multi-armed bandit problems, including indirect, cumulating, or irrelevant information. It balances current expected reward and the reduction in uncertainty about the optimal action. In particular, the randomized action $s_t^i$ at time $t$ is chosen so that

$$
\pi_{IDS} = \arg\min_{\pi \in \Pi} \mathbb{E}_{\pi}[\Delta_{t_0}(\pi)]
$$

where $\Delta_{t_0}(\pi)$ is the space of all distributions over action space $A$, $\Delta_{t_0}(\pi)$ is the expected instantaneous regret and $\pi^*(s)$ is the information gain by taking action $s$.

IDS Properties

- admits-optimal regret bound under full information
- drastic improvement over TS and UCB in some specific problems
- $\Delta_t$ and $\pi_t$ pose great challenge for computation - analytic formula only exist for a very restricted class of problems

Challenges in Reinforcement Learning

More than 1 state, i.e. under MDP $M = (S, A, \mathbb{P}, \mathbb{R}, \rho)$, $|S| > 1$ and transition probabilities $P$ and reward distribution $\mathbb{R}$ are unknown.

Methods and Algorithms

We have the following conceptual correspondence between bandit and reinforcement learning problems.

<table>
<thead>
<tr>
<th>Bandit Learning</th>
<th>Reinforcement Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action $a_t$</td>
<td>episode $t$</td>
</tr>
<tr>
<td>reward $r_{a_t}$</td>
<td>policy $\mu$</td>
</tr>
<tr>
<td>total reward $J_{a_t}$</td>
<td>optimal policy $\mu^*$</td>
</tr>
</tbody>
</table>

We could use the above relation, treat the reinforcement learning problem as a bandit problem with $|A|^{|S|}$ arms, and apply the IDS on this derived bandit problem. However, neglecting the structure leads to

- intractable computation: $|S|$ is usually large, let alone $|A|^{|S|}$
- loose bound

$$
\mathcal{I}(\text{Regret}(\text{IDSRL}, H)) \leq \mathcal{I}(\mathcal{A})^{|S|} \mathcal{I}(\mathcal{A})^{|H|/2}
$$

Model-Based: IDSRL

- Decompose RL into $|S|$ bandit problems with mean reward $Q^\mu(s, \cdot)$
- Estimate information gain from the entire observable chain via cumulative one-step information gains

To describe the algorithm, we need some notations: $h$: episode, $s$: state, $a$: action, $\pi_s(a)$: cumulative information gain starting from state $s$, time $t$, $\Delta_t(s, a)$: expected immediate regret by taking action $a$ at state $s$, $\pi_{IDS}(s)$: median transition prob. from $s$ to $s'$ by taking action $a$.

We could use the above relation, treat the reinforcement learning problem as a bandit problem with $|A|^{|S|}$ arms, and apply the IDS on this derived bandit problem. However, neglecting the structure leads to

- intractable computation: $|S|$ is usually large, let alone $|A|^{|S|}$
- loose bound

$$
\mathcal{I}(\text{Regret}(\text{IDSRL}, H)) \leq \mathcal{I}(\mathcal{A})^{|S|} \mathcal{I}(\mathcal{A})^{|H|/2}
$$

Model-Free Value Function: IDSVI

- Model the posterior distribution of the parameters $\theta_t$ using $\{\hat{Q}^\mu_t(s, \cdot), \hat{r}^\mu_t(s)\}$
- Compute information ratio based on resulting samples of $Q^\mu$

Example: Deep Sea Exploration

- $N \times N$ grid. The agent starts from the top left cell, and can take action in $A = \{\uparrow, \downarrow\}$ at each step. It will move to either the left or right cell in the next step.
- At each cell, the association of actions with "left" and "right" is unknown.
- Reward is 0 at all but the right bottom cell, in which a treasure is given to you reward $r=1$.

![Figure 1: Deep sea exploration problem](#)

**Expert Agent**: The agent knows about all the assumptions and has the correct prior belief. It is straightforward to know

Table 1: Expected number of episodes to learn an optimal policy

<table>
<thead>
<tr>
<th>Method</th>
<th>Expected Episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal</td>
<td>$o(\sqrt{N})$</td>
</tr>
<tr>
<td>Pure Exploration</td>
<td>$n$</td>
</tr>
<tr>
<td>Dithering</td>
<td>$o(N^2)$</td>
</tr>
<tr>
<td>PSRL</td>
<td>$N^2$</td>
</tr>
<tr>
<td>UCBRL</td>
<td>$N^2$</td>
</tr>
<tr>
<td>IDSRL</td>
<td>$N^2$</td>
</tr>
</tbody>
</table>

**Observations**

- In the simulation, IDSRL consistently performs better than PSRL in terms of early discovery.
- Information-theoretic criteria help balance exploration and speed up searching of the optimal policies in early stages.
- We would like to derive non-trivial regret bound for the algorithm proposed here and at the same time trying to find more effective and efficient algorithms.

**Acknowledgement**

The authors would like to thank Professor Benjamin Van Roy for offering the opportunity of working on this project and Abbas Kazerouni for the helpful feedback that helped improve this work.

**References**