Information Directed Reinforcement Learning

Junyang Qian & Junzi Zhang

Department of Statistics, Institute for Computational & Mathematical Engineering

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 π_t^{IDS} at time t is chosen so that

$$\pi_t^{\text{IDS}} \in \arg\min_{\pi\in\mathcal{D}(\mathcal{A})} \left\{ \Psi_t(\pi) := \frac{(\mathbb{E}_{a\sim\pi}\Delta_t(a))^2}{\mathbb{E}_{a\sim\pi} g_t(a)} \right\},$$

where $\mathcal{D}(\mathcal{A})$ is the space of all distributions over action space $\mathcal{A}$, $\Delta_t(a)$ is the expected instantaneous regret and $g_t(a)$ is the information gain by taking action a.

IDS Properties

- order-optimal regret bound under full information ($\sqrt{(1/2)\log|\mathcal{A}|T}$) and linear bandit feedback ($\sqrt{(1/2)\log(|\mathcal{A}|)dT}$)
- drastic improvement over TS and UCB in some specific problems
- Δ_t and g_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 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho)$, $|\mathcal{S}| > 1$ and transition probabilities $\mathcal{P}$ and reward distribution $\mathcal{R}$ are unknown.

Methods and Algorithms

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

Bandit Learning	Reinforcement Learning
time t	episode h
action a	policy μ
reward $R_{t,a}$	total reward $R_{h,\mu} = \Sigma_{t=1}^{\tau} r_t^{h,\mu}$

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

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

$$\mathbb{E}[\text{Regret(IDSRL}, H)] \leqslant \sqrt{|\mathcal{A}|^{|\mathcal{S}|} H(\mu_0^*) H/2}$$

Model-Based: IDSRL

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

To describe the algorithm, we need some notations (h: episode).

- $\Delta_h(s,a)$: expected immediate regret by taking action a at state s.
- $\tilde{I}_{h,t}(s)$: cumulative information gain starting from state s, time t.
- $Z_{h,s,a}$: next state from s by taking action a.
- $p_h(s,a,s')$: mean transition prob. from s to s' by taking action a.

procedure IDSRL
  for all $s \in \mathcal{S}$ **do** $\tilde{I}_{h,T}(s) \leftarrow 0$
  end for
  for $t \leftarrow T-1$ **to** 0 **do**
    for all $s \in \mathcal{S}$ **do**
      for all $a \in \mathcal{A}$ **do**
        $\tilde{I}_{h,t}(s,a) \leftarrow I(A_{h,s}^*; Z_{h,s,a})$
          $+ \sum_{s'\in\mathcal{S}} p_h(s,a,s')\tilde{I}_{h,t+1}(s)$
      end for
      $\tilde{I}_{h,t}(s) \leftarrow \max_{a\in\mathcal{A}} \tilde{I}_{h,t}(s,a)$
    end for
  end for
  Compute $\Delta_h(s,a) = \mathbb{E}\left[Q_h^*\left(s, A_{h,s}^*\right) - Q_h^*(s,a) | \mathcal{H}_h\right]$
  for all $s \in \mathcal{S}$ **do**
    Solve $\pi_{h,s}^{\text{IDSRL}} \in \arg\min_{\omega(s)\in\mathcal{D}(\mathcal{A})} \frac{(\pi^T \Delta_h(s,\cdot))^2}{\pi^T \tilde{I}_{h,0}(s,\cdot)}$
  end for
end procedure

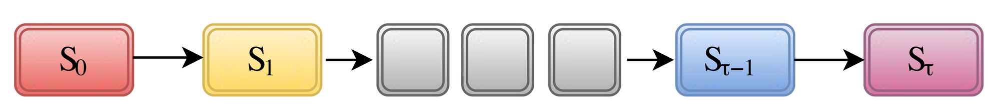
Here, with $|\mathcal{S}|$ optimization problems each with variable dimension $|\mathcal{A}|$, we may be able to obtain poly($|\mathcal{A}|, |\mathcal{S}|$) regret bound. And when the state space can be partitioned, we have the following result.

Theorem 1 (Mutual Information Decomposition). *Suppose that the MDP starts with fixed s_0 and has finite horizon of length T. If the state space is factorized as $\mathcal{S} = \{s_0\} \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_{T-1}$, then we have the following lower bound on mutual information between optimal policy μ^* and the observations obtained by following policy μ.*

$$I(\mu^*; Y_1, \ldots, Y_{T-1}) \geqslant \sum_{t=0}^{T-2} \sum_{s'\in\mathcal{S}_t} p^{(t)}(s_0, \mu, s') I(\mu_{\mathcal{S}_t}^*; Z_{t+1}(s')),$$

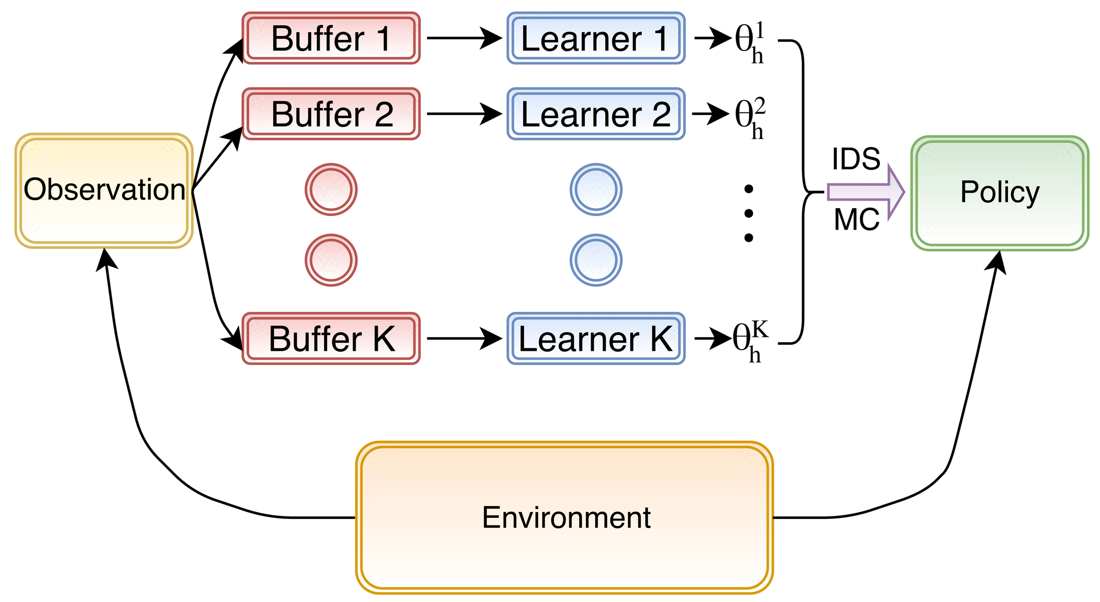
where $p^{(t)}(s_0,\mu,s')$ is the t-step transition probability from s_0 to s' following μ, $\mu_{\mathcal{S}_t}^*$ is optimal policy for states in $\mathcal{S}_t$, and $Z_t(s) \sim \mathbb{P}(Y_t|s)$.

The theorem implies that $\tilde{I}_{h,0}(s_0)$ computed in the algorithm serves as a lower bound for the real mutual information $I(\mu^*; Y_1, \ldots, Y_{T-1})$. Furthermore, in this case, the computation can be simplified in a stepwise manner and is illustrated by the following diagram. We call it *Stepwise IDSRL*.



Model-Free Value Function: IDSVI

- Model the posterior distribution of the parameters θ_h using $\{\theta_h^1, \ldots, \theta_h^K\}$.
- Compute information ratio based on resulting samples of Q_θ.



We can then use empirical mean and variance to approximate the following values.

- $\Delta_h^s(a) := \mathbb{E}_{\tilde{\theta}_h}\left[\tilde{Q}_{\tilde{\theta}_h}^h(s,a^*) - \tilde{Q}_{\tilde{\theta}_h}^h(s,a)\right]$
- $v_h^s(a) := Var_{\tilde{\theta}_h}\left(\tilde{Q}_{\tilde{\theta}_h}^h(s,a)\right)$
- $\pi_{h,s}^{\text{IDSVI}} \in \arg\min_{\omega\in\mathcal{D}(\mathcal{A})} (\pi^T\Delta_h^s)^2/\pi^T v_h^s$

Example: Deep Sea Exploration

- $N \times N$ grid. The agent starts from the top left cell, and can take action in $\mathcal{A} = \{1,2\}$ at each step. It will move to either the left or right cell in the next row.
- At each cell, the association of actions with "left" and "right" is unknown.
- Reward is 0 in all but the right bottom cell, in which a treasure gives you reward 1.

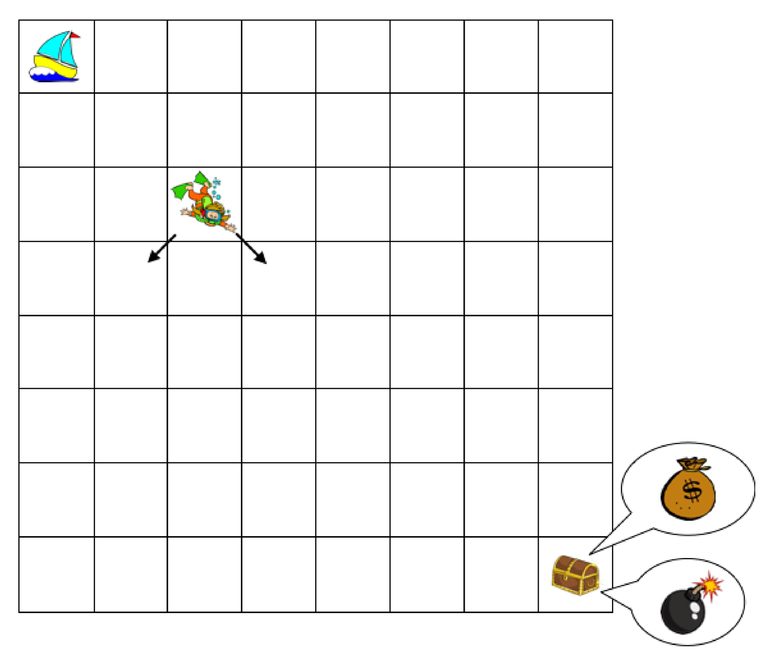


Figure 1: Deep sea exploration problem [1].

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

Method	Expected Episodes
Optimal	$\Theta(N)$
Pure Exploitation	∞
Dithering	$\Theta(2^N)$
PSRL	$\Theta(N)$
UCRL	$\Theta(N)$
IDSRL	$\Theta(N)$

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

Knowledgeable Agent: The agent knows part of the assumptions: each action leads to next row, but is agnostic of which specific cells are more likely (rather than only left or right move in the previous case)

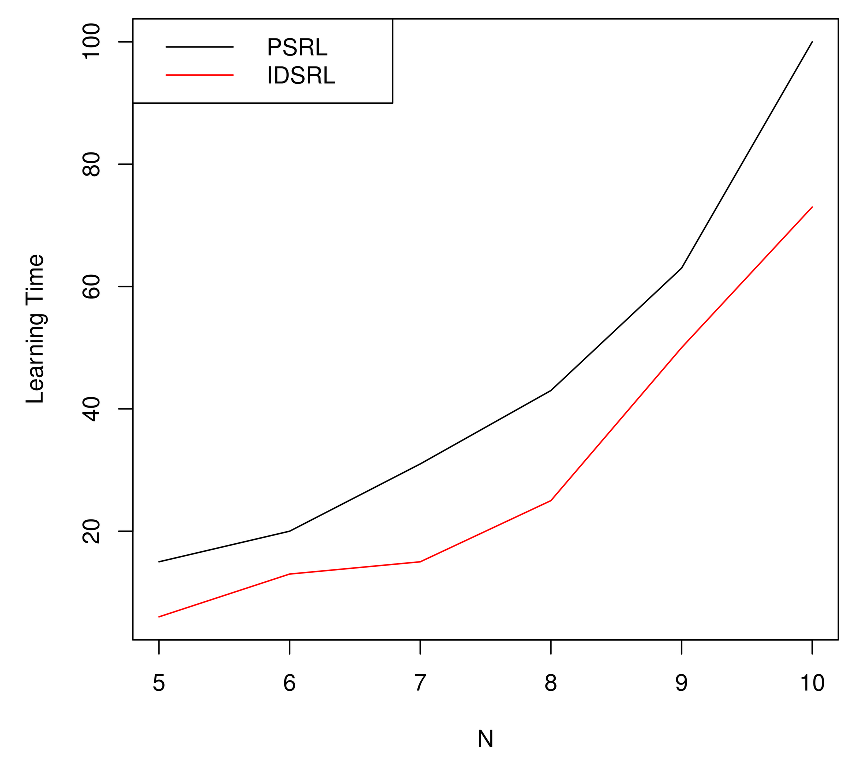


Figure 2: Horizontal axis is the size of deep sea problem, and vertical axis is learning time, i.e. the earliest time that the average reward exceeds 0.5.

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

[1] Ian Osband, Dan Russo, Benjamin Van Roy, and Zheng Wen. Deep exploration via randomized value functions. 2017.

[2] Dan Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, 2014.