

DeepMind

Contextual Information-Directed Sampling

Botao Hao

Joint work with Tor Lattimore and Chao Qin



Data-Efficient RL Agent

Information-directed sampling (IDS) has demonstrated its potential as a data-efficient reinforcement learning algorithm (Lu et al. 2021).

Existing theoretical understanding is limited to the **fixed action set**.

Q: What is the right design of IDS when context or observation is available?



Contextual Bandits

- A finite set of possible contexts \mathcal{S} . The environment samples a sequence of independent contexts $(s_t)_{t=1}^n$ from a distribution ξ over \mathcal{S} .
- Reward:

$$Y_{t,a} = f(s_t, a, \theta^*) + \eta_{t,a},$$

where f is the reward function, θ^* is the unknown parameter and $\eta_{t,a}$ is 1-sub-Gaussian noise.

- The agent receives an observation O_{t,A_t} including an immediate reward Y_{t,A_t} as well as some side information.
- *Bayesian regret* of a policy π is defined as

$$\mathfrak{BR}(n; \pi) = \mathbb{E} \left[\sum_{t=1}^n \max_{a \in \mathcal{A}_t} f(s_t, a, \theta^*) - \sum_{t=1}^n Y_t \right], \quad (1)$$



Conditional IDS or Contextual IDS

- Conditional information ratio:

$$\Gamma_t(\pi(\cdot|s_t)) = \frac{(\Delta_t(s_t)^\top \pi(\cdot|s_t))^2}{\mathbb{I}_t(a_t^*, s_t)^\top \pi(\cdot|s_t)},$$

Conditional IDS finds a **probability distribution**:

$$\pi_t(\cdot|s_t) = \operatorname{argmin}_{\pi(\cdot|s_t) \in \mathcal{P}(\mathcal{A}_t)} \Gamma_t(\pi(\cdot|s_t)).$$

- Marginal information ratio (MIR):

$$\Psi_t(\pi) = \frac{(\mathbb{E}_{s_t} [\Delta_t(s_t)^\top \pi(\cdot|s_t)])^2}{\mathbb{E}_{s_t} [\mathbb{I}_t(\pi^*)^\top \pi(\cdot|s_t)]}.$$

Contextual IDS minimizes MIR to find **a mapping from the context space to the action space**:

$$\pi_t = \operatorname{argmin}_{\pi \in \Pi} \Psi_t(\pi).$$



Conditional IDS or Contextual IDS

Conditional IDS myopically balances exploration and exploitation without taking the context distribution into consideration.

Conditional IDS could either over-explore or under-explore.



Two Popular Bandits Problems

- For contextual bandits with graph feedback, conditional IDS suffers $\Omega(\sqrt{\beta(\mathcal{G})n})$ Bayesian regret lower bound. Contextual IDS can achieve $O(\min\{\sqrt{\beta(\mathcal{G})n}, \delta(\mathcal{G})^{1/3}n^{2/3}\})$ Bayesian regret upper bound for any prior.

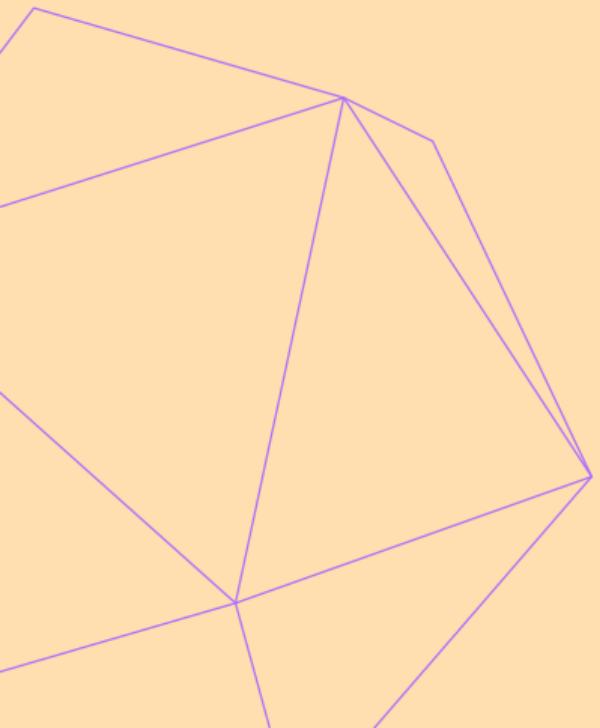
Here, \mathcal{G} is a directed feedback graph over the set of actions, $\beta(\mathcal{G})$ is the independence number and $\delta(\mathcal{G})$ is the domination number of the graph.

In the regime where $\beta(\mathcal{G}) \gtrsim (\delta(\mathcal{G})^2 n)^{1/3}$, contextual IDS achieves better regret bound than conditional IDS.

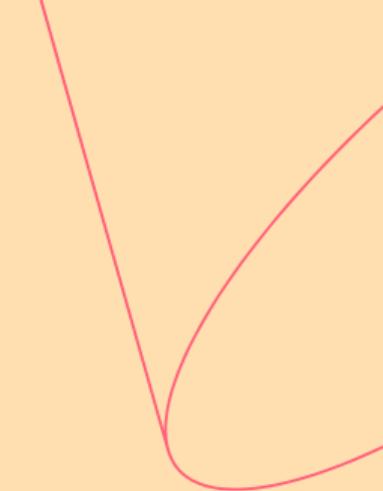
- For sparse linear contextual bandits, conditional IDS suffers $\Omega(\sqrt{nds})$ Bayesian regret lower bound. Contextual IDS can achieve $O(\min\{\sqrt{nds}, sn^{2/3}\})$ Bayesian regret upper bound for any sparse prior. Here, d is the feature dimension and s is the sparsity.

In the data-poor regime where $d \gtrsim sn^{1/3}$, contextual IDS achieves better regret bound than conditional IDS.





DeepMind



Thanks!



Bibliography

