Understanding Information-Directed Sampling: When and How to Use It?

Botao Hao
Deepmind

Stanford RL Forum
Nov. 9, 2022

What is Information-Directed Sampling?

- IDS (Russo and Van Roy, 2014) is a design principle that explicitly balance the trade-off between information and regret.
- IDS minimizes a notion of information ratio:

$$\text{information ratio} = \frac{\Delta^2}{\mathbb{I}}$$

Part I: When can IDS outperform optimism-based algorithms?

Sparse Linear Bandits

- At each round $t \in [n]$, the agent chooses an action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ and receives a reward:

$$Y_t = \langle A_t, \theta^* \rangle + \eta_t.$$

where η_t is 1-sub-Gaussian noise. The notion of sparsity can be defined through the parameter space Θ:

$$\Theta = \left\{ \theta \in \mathbb{R}^d \middle| \sum_{j=1}^{d} \mathbb{1}\{\theta_j \neq 0\} \leq s, \|\theta\|_2 \leq 1 \right\} .$$

Sparse Linear Bandits

- At each round $t \in [n]$, the agent chooses an action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ and receives a reward:

$$Y_t = \langle A_t, \theta^* \rangle + \eta_t.$$

where η_t is 1-sub-Gaussian noise. The notion of sparsity can be defined through the parameter space Θ:

$$\Theta = \left\{ \theta \in \mathbb{R}^d \middle| \sum_{j=1}^{d} \mathbb{1}\{\theta_j \neq 0\} \leq s, \|\theta\|_2 \leq 1 \right\}.$$

- Cumulative regret:

$$\mathfrak{R}_{\theta^*}(n; \pi) = \mathbb{E}\left[\sum_{t=1}^{n} \langle x^*, \theta^* \rangle - \sum_{t=1}^{n} Y_t \right],$$

where x^* is the optimal action.

- Worse-case regret: $\sup_{\theta^*} \mathfrak{R}_{\theta^*}(n; \pi)$; Bayesian regret: $\mathbb{E}_{\theta^*}[\mathfrak{R}_{\theta^*}(n; \pi)]$.

Explorability Constant

Definition. Let $\mathcal{P}(\mathcal{A})$ be the space of probability measures over $\mathcal{A}$. The explorability constant is defined as

$$C_{\min}(\mathcal{A}) = \sup_{\mu \in \mathcal{P}(\mathcal{A})} \sigma_{\min}\Big(\mathbb{E}_{A \sim \mu}\left[AA^\top\right]\Big).$$

Explorability Constant

Definition. Let $\mathcal{P}(\mathcal{A})$ be the space of probability measures over $\mathcal{A}$. The explorability constant is defined as

$$C_{\min}(\mathcal{A}) = \sup_{\mu \in \mathcal{P}(\mathcal{A})} \sigma_{\min}\Big(\mathbb{E}_{A\sim\mu}\left[AA^\top\right]\Big).$$

Remarks.

- When $C_{\min}(\mathcal{A})$ is dimension-free, we say

  "action set $\mathcal{A}$ admits a well-conditioned exploratory policy".

- What is information? Pulling arms according to this exploratory policy, we collect information (well-conditioned data).

Minimax Lower Bound

Theorem.[1] For any policy π, there exists an action set $\mathcal{A}$ with $C_{\min}(\mathcal{A}) > 0$ and s-sparse parameter $\theta^* \in \mathbb{R}^d$ such that

$$\mathfrak{R}_{\theta^*}(n;\pi) \gtrsim \min\left(C_{\min}^{-\frac{1}{3}}(\mathcal{A}) s^{\frac{1}{3}} n^{\frac{2}{3}}, \sqrt{dsn}\right).$$

[1]High-Dimensional Sparse Linear Bandits. (**Hao**, Lattimore, Wang, NeurIPS 2020)

Minimax Lower Bound

Theorem.[2] For any policy π, there exists an action set $\mathcal{A}$ with $C_{\min}(\mathcal{A}) > 0$ and s-sparse parameter $\theta^* \in \mathbb{R}^d$ such that

$$\mathfrak{R}_{\theta^*}(n;\pi) \gtrsim \min\left(C_{\min}^{-\frac{1}{3}}(\mathcal{A}) s^{\frac{1}{3}} n^{\frac{2}{3}}, \sqrt{dsn} \right) .$$

- Data-poor regime: $d^3 \gtrsim n$; data-rich regime: $d^3 \lesssim n$.
- Carefully balancing the trade-off between information and regret is necessary in sparse linear bandits.

[2] High-Dimensional Sparse Linear Bandits. (**Hao**, Lattimore, Wang, NeurIPS 2020)

Minimax Lower Bound

Theorem.[3] For any policy π, there exists an action set $\mathcal{A}$ with $C_{\min}(\mathcal{A}) > 0$ and s-sparse parameter $\theta^* \in \mathbb{R}^d$ such that

$$\mathfrak{R}_{\theta^*}(n;\pi) \gtrsim \min\left(C_{\min}^{-\frac{1}{3}}(\mathcal{A}) s^{\frac{1}{3}} n^{\frac{2}{3}}, \sqrt{dsn}\right).$$

- Data-poor regime: $d^3 \gtrsim n$; data-rich regime: $d^3 \lesssim n$.
- Carefully balancing the trade-off between information and regret is necessary in sparse linear bandits.

Q: Does the optimism optimally balance information and regret?

[3] High-Dimensional Sparse Linear Bandits. (**Hao**, Lattimore, Wang, NeurIPS 2020)

Optimism-Based Algorithms

Optimism-based algorithms π^{opt} choose

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\widetilde{\theta} \in \mathcal{C}_t} \langle a, \widetilde{\theta} \rangle,$$

where $\mathcal{C}_t$ is some sparsity-aware confidence set.

Optimism-Based Algorithms

Optimism-based algorithms π^{opt} choose

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\widetilde{\theta} \in \mathcal{C}_t} \langle a, \widetilde{\theta} \rangle,$$

where $\mathcal{C}_t$ is some sparsity-aware confidence set.

Claim. Let π^{opt} be such an optimism-based algorithm. There exists a sparse linear bandit instance characterized by θ such that for the data-poor regime, we have

$$\mathfrak{R}_\theta(n; \pi^{\mathsf{opt}}) \gtrsim n\,.$$

Information Directed Sampling

IDS takes the action according to

$$\mu_t = \underset{\mu \in \mathcal{P}(\mathcal{A})}{\operatorname{argmin}} \frac{(\Delta_t^\top \mu)^2}{\mathbb{I}_t^\top \mu},$$

where $\mathbb{I}_t \in \mathbb{R}^{|\mathcal{A}|}$ is the *information gain* about the optimal action and $\Delta_t \in \mathbb{R}^{|\mathcal{A}|}$ is *the expected single-round regret*[4].

[4] $\Delta_t(a) := \mathbb{E}_t[\langle x^*, \theta^* \rangle - \langle a, \theta^* \rangle]$

Information Directed Sampling

IDS takes the action according to

$$\mu_t = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{A})} \frac{(\Delta_t^\top \mu)^2}{\mathbb{I}_t^\top \mu} \,,$$

where $\mathbb{I}_t \in \mathbb{R}^{|\mathcal{A}|}$ is the *information gain* about the optimal action and $\Delta_t \in \mathbb{R}^{|\mathcal{A}|}$ is *the expected single-round regret*[5].

Theorem.[6] The following regret bound holds for IDS:

$$\mathfrak{BR}(n; \pi^{\mathrm{IDS}}) \lesssim \min\left\{\sqrt{nds}, \frac{sn^{2/3}}{C_{\min}(\mathcal{A})^{1/3}}\right\} .$$

---

[5] $\Delta_t(a) := \mathbb{E}_t[\langle x^*, \theta^* \rangle - \langle a, \theta^* \rangle]$

[6] Information Directed Sampling for Sparse Linear Bandits. (**Hao**, Lattimore, Deng, NeurIPS 2021)

Information Directed Sampling

IDS takes the action according to

$$\mu_t = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{A})} \frac{(\Delta_t^\top \mu)^2}{\mathbb{I}_t^\top \mu},$$

where $\mathbb{I}_t \in \mathbb{R}^{|\mathcal{A}|}$ is the *information gain* about the optimal action and $\Delta_t \in \mathbb{R}^{|\mathcal{A}|}$ is *the expected single-round regret*[7].

Theorem.[8] The following regret bound holds for IDS:

$$\mathfrak{BR}(n; \pi^{\mathrm{IDS}}) \lesssim \min\left\{\sqrt{nds}, \frac{sn^{2/3}}{C_{\min}(\mathcal{A})^{1/3}}\right\}.$$

Great adaptivity of IDS for sparse linear bandits in the sense that a single policy adapts to different information-regret structures.

[7] $\Delta_t(a) := \mathbb{E}_t[\langle x^*, \theta^* \rangle - \langle a, \theta^* \rangle]$

[8] Information Directed Sampling for Sparse Linear Bandits. (**Hao**, Lattimore, Deng, NeurIPS 2021)

Part II: What is the right form of information ratio to optimize for reinforcement learning?

Contextual Bandits

Suppose $(s_t)_{t=1}^n$ are i.i.d contexts from a distribution ξ.

- Conditional IDS finds a probability distribution:

$$\pi_t(\cdot|s_t) = \underset{\pi(\cdot|s_t)\in\mathcal{P}(\mathcal{A}_t)}{\operatorname{argmin}} \Gamma_t(\pi(\cdot|s_t)) := \underbrace{\frac{\left(\Delta_t(s_t)^\top \pi(\cdot|s_t)\right)^2}{\mathbb{I}_t(a_t^*, s_t)^\top \pi(\cdot|s_t)}}_{\text{conditional information ratio}} .$$

Contextual Bandits

Suppose $(s_t)_{t=1}^n$ are i.i.d contexts from a distribution ξ.

- Conditional IDS finds a probability distribution:

$$\pi_t(\cdot|s_t) = \underset{\pi(\cdot|s_t)\in\mathcal{P}(\mathcal{A}_t)}{\operatorname{argmin}} \Gamma_t(\pi(\cdot|s_t)) := \underbrace{\frac{\left(\Delta_t(s_t)^\top \pi(\cdot|s_t)\right)^2}{\mathbb{I}_t(a_t^*, s_t)^\top \pi(\cdot|s_t)}}_{\text{conditional information ratio}} .$$

- Contextual IDS finds a mapping from the context space to the action space:

$$\pi_t = \underset{\pi\in\Pi}{\operatorname{argmin}} \Psi_t(\pi) = \underbrace{\frac{\left(\mathbb{E}_{s_t\sim\xi}[\Delta_t(s_t)^\top \pi(\cdot|s_t)]\right)^2}{\mathbb{E}_{s_t\sim\xi}[\mathbb{I}_t(\pi^*)^\top \pi(\cdot|s_t)]}}_{\text{marginal information ratio}} .$$

Contextual Bandits

Suppose $(s_t)_{t=1}^n$ are i.i.d contexts from a distribution ξ.

- Conditional IDS finds a probability distribution:

$$\pi_t(\cdot|s_t) = \underset{\pi(\cdot|s_t)\in\mathcal{P}(\mathcal{A}_t)}{\operatorname{argmin}} \Gamma_t(\pi(\cdot|s_t)) := \underbrace{\frac{\left(\Delta_t(s_t)^\top \pi(\cdot|s_t)\right)^2}{\mathbb{I}_t(a_t^*, s_t)^\top \pi(\cdot|s_t)}}_{\text{conditional information ratio}} .$$

- Contextual IDS finds a mapping from the context space to the action space:

$$\pi_t = \underset{\pi\in\Pi}{\operatorname{argmin}} \Psi_t(\pi) = \underbrace{\frac{\left(\mathbb{E}_{s_t\sim\xi}[\Delta_t(s_t)^\top \pi(\cdot|s_t)]\right)^2}{\mathbb{E}_{s_t\sim\xi}[\mathbb{I}_t(\pi^*)^\top \pi(\cdot|s_t)]}}_{\text{marginal information ratio}} .$$

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

Why Conditional IDS Could be Myopic?

Example 1 [UNDER EXPLORATION] Consider a noiseless case.

- Context set 1: k actions where one is the optimal action and the remaining $k-1$ actions yield regret 1.
- Context set 2: a revealing action with regret 1 and one action with no regret. The revealing action provides an observation of the rewards for all the k actions in context set 1.

Why Conditional IDS Could be Myopic?

Example 1 [UNDER EXPLORATION] Consider a noiseless case.

- Context set 1: k actions where one is the optimal action and the remaining $k-1$ actions yield regret 1.
- Context set 2: a revealing action with regret 1 and one action with no regret. The revealing action provides an observation of the rewards for all the k actions in context set 1.
- When context set 2 arrives, conditional IDS will **never play the revealing action** since it incurs high immediate regret with no useful information for the current context set.
- However, this ignores the fact that the revealing action could be informative for the unseen context set 1. Conditional IDS *under-explores* and suffers $O(k)$ regret.

Why Conditional IDS Could be Myopic?

Example 1 [UNDER EXPLORATION] Consider a noiseless case.

- Context set 1: k actions where one is the optimal action and the remaining $k-1$ actions yield regret 1.
- Context set 2: a revealing action with regret 1 and one action with no regret. The revealing action provides an observation of the rewards for all the k actions in context set 1.
- When context set 2 arrives, conditional IDS will **never play the revealing action** since it incurs high immediate regret with no useful information for the current context set.
- However, this ignores the fact that the revealing action could be informative for the unseen context set 1. Conditional IDS *under-explores* and suffers $O(k)$ regret.
- Contextual IDS exploits the context distribution and plays the revealing action in context 2 and only suffers $O(1)$ regret.

Reinforcement Learning

- Finite-horizon time-inhomogeneous MDP: $\mathcal{E} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$.
- Assume r_h is known and deterministic, P_h is unknown and random.
- The expected cumulative regret of an algorithm $\pi = \{\pi^\ell\}_{\ell=1}^L$ with respect to an environment $\mathcal{E}$ is defined as

$$\mathfrak{R}_L(\mathcal{E}, \pi) = \mathbb{E}\left[\sum_{\ell=1}^{L} \left(V_{1,\pi^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi^\ell}^{\mathcal{E}}(s_1^\ell)\right)\right],$$

  where the expectation is taken with respect to the randomness of π^ℓ.
- The Bayesian regret is defined as

$$\mathfrak{BR}_L(\pi) = \mathbb{E}[\mathfrak{R}_L(\mathcal{E}, \pi)],$$

  where the expectation is taken with respect to the prior distribution of $\mathcal{E}$.

Vanilla IDS

- The information ratio for a policy π at episode ℓ is defined as

$$\Gamma_\ell(\pi, \chi) := \frac{(\mathbb{E}_\ell[V_{1,\pi^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi}^{\mathcal{E}}(s_1^\ell)])^2}{\mathbb{I}_\ell^\pi(\chi; \mathcal{H}_{\ell,H})},$$

  where χ is the learning target, $\mathcal{H}_{\ell,H}$ as the history of episode ℓ up to layer H, $\mathbb{I}_\ell^\pi$ is the conditional mutual information.
- At the beginning of each episode ℓ, `vanilla-IDS` computes a stochastic policy (let $\chi = \mathcal{E}$):

$$\pi_{\text{IDS}}^\ell = \operatorname*{argmin}_\pi \Gamma_\ell(\pi, \mathcal{E}).$$

Regret Bound of `Vanilla-IDS`

Theorem.[9] A generic regret bound for `vanilla-IDS` is

$$\mathfrak{BR}_L(\pi_{\text{IDS}}) \leq \sqrt{\mathbb{E}[\Gamma^*]\mathbb{I}\left(\mathcal{E}; \mathcal{D}_{L+1}\right) L}\,.$$

Here, Γ^* is the worst-case information ratio such that $\Gamma_\ell(\pi^\ell_{\text{IDS}}) \leq \Gamma^*$ for any $\ell \in [L]$ a.s. and $\mathcal{D}_{L+1}$ is the entire history.

[9] Regret Bounds for Information-Directed Reinforcement Learning (**Hao**, Lattimore, NeurIPS 2022)

Regret Bound of `Vanilla-IDS`

Theorem.[10] A generic regret bound for `vanilla-IDS` is

$$\mathfrak{BR}_L(\pi_{\text{IDS}}) \leq \sqrt{\mathbb{E}[\Gamma^*]\mathbb{I}\left(\mathcal{E};\mathcal{D}_{L+1}\right)L}\,.$$

Here, Γ^* is the worst-case information ratio such that $\Gamma_\ell(\pi_{\text{IDS}}^\ell) \leq \Gamma^*$ for any $\ell \in [L]$ a.s. and $\mathcal{D}_{L+1}$ is the entire history.

- For tabular MDPs with independent priors[11] across different layers,

$$\mathbb{E}[\Gamma^*] \lesssim SAH^3, \mathbb{I}(\mathcal{E};\mathcal{D}_{L+1}) \lesssim S^2AH\,.$$

[10] Regret Bounds for Information-Directed Reinforcement Learning (**Hao**, Lattimore, NeurIPS 2022)

[11] ρ_h is the prior measure for P_h and $\rho = \rho_1 \otimes \cdots \otimes \rho_H$ as the product prior measure for the whole environment.

Regret Bound of `Vanilla-IDS`

Theorem.[12] A generic regret bound for `vanilla-IDS` is

$$\mathfrak{BR}_L(\pi_{\text{IDS}}) \leq \sqrt{\mathbb{E}[\Gamma^*]\mathbb{I}\left(\mathcal{E}; \mathcal{D}_{L+1}\right) L}\,.$$

Here, Γ^* is the worst-case information ratio such that $\Gamma_\ell(\pi_{\text{IDS}}^\ell) \leq \Gamma^*$ for any $\ell \in [L]$ a.s. and $\mathcal{D}_{L+1}$ is the entire history.

- For tabular MDPs with independent priors[13] across different layers,

$$\mathbb{E}[\Gamma^*] \lesssim SAH^3, \mathbb{I}(\mathcal{E}; \mathcal{D}_{L+1}) \lesssim S^2 AH\,.$$

- The bound for information gain can be sharpen to SAH by the rate-distortion.

[12] Regret Bounds for Information-Directed Reinforcement Learning (**Hao**, Lattimore, NeurIPS 2022)

[13] ρ_h is the prior measure for P_h and $\rho = \rho_1 \otimes \cdots \otimes \rho_H$ as the product prior measure for the whole environment.

Proof Sketch for `Vanilla-IDS`

Step one. Use the mean MDP $\bar{\mathcal{E}}_\ell$ as a bridge:

$$\mathbb{E}_\ell\left[V^{\mathcal{E}}_{1,\pi^*}(s_1^\ell) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathsf{TS}}}(s_1^\ell)\right]$$
$$= \underbrace{\mathbb{E}_\ell\left[V^{\mathcal{E}}_{1,\pi^*}(s_1^\ell) - V^{\bar{\mathcal{E}}_\ell}_{1,\pi^\ell_{\mathsf{TS}}}(s_1^\ell)\right]}_{I_1} + \underbrace{\mathbb{E}_\ell\left[V^{\bar{\mathcal{E}}_\ell}_{1,\pi^\ell_{\mathsf{TS}}}(s_1^\ell) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathsf{TS}}}(s_1^\ell)\right]}_{I_2}.$$

Proof Sketch for `Vanilla-IDS`

Step one. Use the mean MDP $\bar{\mathcal{E}}_\ell$ as a bridge:

$$\mathbb{E}_\ell\left[V^{\mathcal{E}}_{1,\pi^*}(s_1^\ell) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathrm{TS}}}(s_1^\ell)\right]$$
$$= \underbrace{\mathbb{E}_\ell\left[V^{\mathcal{E}}_{1,\pi^*}(s_1^\ell) - V^{\bar{\mathcal{E}}_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s_1^\ell)\right]}_{I_1} + \underbrace{\mathbb{E}_\ell\left[V^{\bar{\mathcal{E}}_\ell}_{1,\pi^\ell_{\mathrm{TS}}}(s_1^\ell) - V^{\mathcal{E}}_{1,\pi^\ell_{\mathrm{TS}}}(s_1^\ell)\right]}_{I_2}.$$

Step two. Denote the value function difference as

$$\Delta_h^{\mathcal{E}}(s,a) = \mathbb{E}_{s'\sim P_h^{\mathcal{E}}(\cdot|s,a)}[V^{\mathcal{E}}_{h+1,\pi^*}(s')] - \mathbb{E}_{s'\sim P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s,a)}[V^{\mathcal{E}}_{h+1,\pi^*}(s')].$$

With the use of state-action occupancy measure, we can derive

$$I_1 = \sum_{h=1}^{H}\mathbb{E}_\ell\left[\sum_{(s,a)}\frac{d^{\bar{\mathcal{E}}_\ell}_{h,\pi^*}(s,a)}{(\mathbb{E}_\ell[d^{\bar{\mathcal{E}}_\ell}_{h,\pi^*}(s,a)])^{1/2}}(\mathbb{E}_\ell[d^{\bar{\mathcal{E}}_\ell}_{h,\pi^*}(s,a)])^{1/2}\Delta_h^{\mathcal{E}}(s,a)\right].$$

Proof Sketch for `Vanilla-IDS`

Step 3. Applying the Cauchy–Schwarz inequality and Pinsker's inequality, we can obtain

$$I_1 \leq \sqrt{SAH^3}\left(\sum_{h=1}^{H}\mathbb{E}_\ell\left[\mathbb{E}_{\pi_{\text{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}\left[\frac{1}{2}D_{\mathrm{KL}}\left(P_h^{\mathcal{E}}(\cdot|s_h^\ell,a_h^\ell)||P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_h^\ell,a_h^\ell)\right)\right]\right]\right)^{1/2},$$

where $\mathbb{E}_{\pi_{\text{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}$ is taken with respect to s_h^ℓ, a_h^ℓ and $\mathbb{E}_\ell$ is taken with respect to π_{TS}^ℓ and $\mathcal{E}$.

Proof Sketch for `Vanilla-IDS`

Step 3. Applying the Cauchy–Schwarz inequality and Pinsker's inequality, we can obtain

$$I_1 \leq \sqrt{SAH^3} \left(\sum_{h=1}^{H} \mathbb{E}_\ell \left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell} \left[\frac{1}{2} D_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s_h^\ell, a_h^\ell) || P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_h^\ell, a_h^\ell) \right) \right] \right] \right)^{1/2},$$

where $\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}$ is taken with respect to s_h^ℓ, a_h^ℓ and $\mathbb{E}_\ell$ is taken with respect to π_{TS}^ℓ and $\mathcal{E}$.

Step 4. It remains to establish the following equivalence of above KL-divergence and the information gain:

$$\sum_{h=1}^{H} \mathbb{E}_\ell \left[\mathbb{E}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell} \left[D_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s_h, a_h) || P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s_h, a_h) \right) \right] \right] = \mathbb{I}_\ell^{\pi_{\mathrm{TS}}^\ell} \left(\mathcal{E}; \mathcal{H}_{\ell,H} \right).$$

A crucial step is to use the linearity of the expectation and the independence of priors over different layers to show

$$\mathbb{P}_{\ell,\pi_{\mathrm{TS}}^\ell}(s_h = s, a_h = a) = \mathbb{P}_{\pi_{\mathrm{TS}}^\ell}^{\bar{\mathcal{E}}_\ell}(s_h = s, a_h = a).$$

How to Compute?

Recall that `Vanilla-IDS` computes

$$\pi^{\ell}_{\text{IDS}} = \operatorname*{argmin}_{\pi:\mathcal{S}\times[H]\to\mathcal{A}} \left[\frac{(\mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi^*}(s^{\ell}_1) - V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)])^2}{\mathbb{I}^{\pi}_{\ell}(\mathcal{E};\mathcal{H}_{\ell,H})} = \frac{\Delta_{\ell}(\pi)}{\mathbb{I}_{\ell}(\pi)} \right].$$

- When $|\mathcal{S}| = 1, H = 1$, it reduces to the bandit case. `Vanilla-IDS` traverses two non-zero components over the *action space*.
- When $|\mathcal{S}| > 1, H > 1$, `Vanilla-IDS` traverses two non-zero components over the *policy space* that the computational time might grow exponentially in $\mathcal{S}$ and H.

How to Compute?

Recall that `Vanilla-IDS` computes

$$\pi_{\mathsf{IDS}}^{\ell} = \underset{\pi:\mathcal{S}\times[H]\to\mathcal{A}}{\operatorname{argmin}} \left[\frac{(\mathbb{E}_{\ell}[V_{1,\pi^*}^{\mathcal{E}}(s_1^{\ell}) - V_{1,\pi}^{\mathcal{E}}(s_1^{\ell})])^2}{\mathbb{I}_{\ell}^{\pi}(\mathcal{E};\mathcal{H}_{\ell,H})} = \frac{\Delta_{\ell}(\pi)}{\mathbb{I}_{\ell}(\pi)} \right].$$

- When $|\mathcal{S}| = 1, H = 1$, it reduces to the bandit case. `Vanilla-IDS` traverses two non-zero components over the *action space*.
- When $|\mathcal{S}| > 1, H > 1$, `Vanilla-IDS` traverses two non-zero components over the *policy space* that the computational time might grow exponentially in $\mathcal{S}$ and H.

Can we have an IDS that can be solved by dynamic programming?

Regularized IDS

- At each episode ℓ, `regularized-IDS` finds the policy:

$$\pi^{\ell}_{\text{r-IDS}} = \operatorname*{argmax}_{\pi} \mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] + \lambda \mathbb{I}_{\ell}\left(\mathcal{E}; \mathcal{H}^{\pi}_{\ell,H}\right) \ ,$$

where $\lambda > 0$ is a tunable parameter.

Regularized IDS

- At each episode ℓ, `regularized-IDS` finds the policy:

$$\pi^{\ell}_{\text{r-IDS}} = \operatorname*{argmax}_{\pi} \mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] + \lambda \mathbb{I}_{\ell}\left(\mathcal{E}; \mathcal{H}^{\pi}_{\ell,H}\right) ,$$

  where $\lambda > 0$ is a tunable parameter.

- Define an *augmented* reward function:

$$r'_h(s,a) = r_h(s,a) + \lambda \int D_{\mathrm{KL}}\left(P^{\mathcal{E}}_h(\cdot|s,a) || P^{\bar{\mathcal{E}}_{\ell}}_h(\cdot|s,a)\right) \mathrm{d}\mathbb{P}_{\ell}(\mathcal{E}) ,$$

  where $\bar{\mathcal{E}}_{\ell}$ is the posterior mean of $\mathcal{E}$.

Regularized IDS

- At each episode ℓ, `regularized-IDS` finds the policy:

$$\pi^{\ell}_{\text{r-IDS}} = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] + \lambda \mathbb{I}_{\ell}\left(\mathcal{E}; \mathcal{H}^{\pi}_{\ell,H}\right) ,$$

  where $\lambda > 0$ is a tunable parameter.

- Define an *augmented* reward function:

$$r'_h(s,a) = r_h(s,a) + \lambda \int D_{\mathrm{KL}}\left(P^{\mathcal{E}}_h(\cdot|s,a) || P^{\bar{\mathcal{E}}_{\ell}}_h(\cdot|s,a)\right) \mathrm{d}\mathbb{P}_{\ell}(\mathcal{E}) ,$$

  where $\bar{\mathcal{E}}_{\ell}$ is the posterior mean of $\mathcal{E}$.

- We prove

$$\mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] + \lambda \mathbb{I}^{\pi}_{\ell}\left(\mathcal{E}; \mathcal{H}_{\ell,H}\right) = \mathbb{E}^{\bar{\mathcal{E}}_{\ell}}_{\pi}\left[\sum_{h=1}^{H} r'_h(s_h,a_h)\right] .$$

Regularized IDS

- At each episode ℓ, `regularized-IDS` finds the policy:

$$\pi^{\ell}_{\text{r-IDS}} = \operatorname*{argmax}_{\pi} \mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] + \lambda \mathbb{I}_{\ell}\left(\mathcal{E}; \mathcal{H}^{\pi}_{\ell,H}\right) ,$$

  where $\lambda > 0$ is a tunable parameter.
- Define an *augmented* reward function:

$$r'_h(s,a) = r_h(s,a) + \lambda \int D_{\mathrm{KL}}\left(P^{\mathcal{E}}_h(\cdot|s,a) || P^{\bar{\mathcal{E}}_\ell}_h(\cdot|s,a)\right) \mathrm{d}\mathbb{P}_{\ell}(\mathcal{E}) ,$$

  where $\bar{\mathcal{E}}_\ell$ is the posterior mean of $\mathcal{E}$.
- We prove

$$\mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] + \lambda \mathbb{I}^{\pi}_{\ell}\left(\mathcal{E}; \mathcal{H}_{\ell,H}\right) = \mathbb{E}^{\bar{\mathcal{E}}_\ell}_{\pi}\left[\sum_{h=1}^{H} r'_h(s_h, a_h)\right] .$$

- Finding $\pi^{\ell}_{\text{r-IDS}}$ = finding the optimal policy based on $\{\bar{\mathcal{E}}_\ell, r'_h\}$.
- Can be solved by any DP solver! And enjoy the same regret bound as `Vanilla-IDS`.

Variance-based Regularized IDS

By Pinsker's inequality,

$$\int D_{\mathrm{KL}}\left(P_h^{\mathcal{E}}(\cdot|s,a)||P_h^{\bar{\mathcal{E}}_\ell}(\cdot|s,a)\right)\mathrm{d}\mathbb{P}_\ell(\mathcal{E}) \geq \sum_{s'} \mathrm{Var}\left(P_h^{\mathcal{E}}(s'|s,a)\right) .$$

Then the *augmented* reward function in terms of variance terms is

$$r_h'(s,a) = r_h(s,a) + \lambda \sum_{s'} \mathrm{Var}\left(P_h^{\mathcal{E}}(s'|s,a)\right) .$$

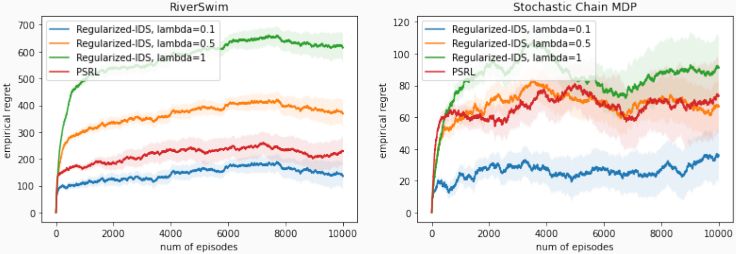

Figure 1: Compare regularized-IDS and PSRL.

Future Directions

- If IDS and vanilla PSRL can achieve $O(\sqrt{SAH^2L})$ rate for tabular MDPs?
- Can we find any interesting RL problem such that IDS can outperform optimism-based principle?
- Extend the information-theoretical analysis to rich classes of RL problems.


thank you!

Why Conditional IDS Could be Myopic?

Example 2 [OVER EXPLORATION]

- Context set 1 contains a single revealing action (hence no regret).
- Context set 2 has k actions. The first is a revealing action and has a (known) regret of $\Theta(\sqrt{k}\Delta)$ with $\Delta = \Theta(1/\sqrt{n})$. Of the remaining actions, one is optimal (zero regret) and the others have regret Δ, with the prior such that the identify of the optimal action is unknown.
- Contextual IDS will **avoid the revealing action in context set 2** because it understands that this information can be obtained more cheaply in context set 1. Its regret is $O(\sqrt{n})$.
- Meanwhile, if the constants are tuned appropriately, then conditional IDS will play the revealing action in context set 2 and suffer regret $\Omega(\sqrt{nk})$.

Surrogate IDS

- Construct a partition $\{\Theta_k\}_{k=1}^K$ over Θ such that for any $\mathcal{E}, \mathcal{E}' \in \Theta_k$ and any $k \in [K]$, we have

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^\ell) \leq \varepsilon\,,$$

  where $\varepsilon > 0$ is the distortion tolerance.
- Construct the surrogate environment $\widetilde{\mathcal{E}}_\ell^* \in \Theta$ based on $\{\Theta_k\}_{k=1}^K$ that needs less information to learn.

Regret Bound of `Surrogate-IDS`

`Surrogate-IDS` minimizes

$$\pi^{\ell}_{\text{s-IDS}} = \operatorname*{argmin}_{\pi \in \Pi} \frac{(\mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi^*}(s^{\ell}_1) - V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] - \varepsilon)^2}{\mathbb{I}^{\pi}_{\ell}(\widetilde{\mathcal{E}}^*_{\ell}; \mathcal{H}_{\ell,H})},$$

for some parameters $\varepsilon > 0$ the will be chosen later.

Regret Bound of `Surrogate-IDS`

`Surrogate-IDS` minimizes

$$\pi^{\ell}_{\text{s-IDS}} = \operatorname*{argmin}_{\pi} \frac{(\mathbb{E}_{\ell}[V^{\mathcal{E}}_{1,\pi^*}(s^{\ell}_1) - V^{\mathcal{E}}_{1,\pi}(s^{\ell}_1)] - \varepsilon)^2}{\mathbb{I}^{\pi}_{\ell}(\widetilde{\mathcal{E}}^*_{\ell}; \mathcal{H}_{\ell,H})},$$

for some parameters $\varepsilon > 0$.

Theorem. A generic regret bound for `surrogate-IDS` is

$$\mathfrak{BR}_L(\pi_{\text{IDS}}) \leq \sqrt{\mathbb{E}[\Gamma^*]\mathbb{I}(\zeta; \mathcal{D}_{L+1})\, L}\,.$$

- For tabular MDPs,

$$\mathbb{E}[\Gamma^*] \lesssim SAH^3, \mathbb{I}(\zeta; \mathcal{D}_{L+1}) \lesssim SAH\,.$$

- For linear MDPs,

$$\mathbb{E}[\Gamma^*] \lesssim dH^3, \mathbb{I}(\zeta; \mathcal{D}_{L+1}) \lesssim dH\,.$$

Some notations

- For a random variable χ we define:

  $$\mathbb{I}_\ell^\pi(\chi;\mathcal{H}_{\ell,h}) = D_{\mathrm{KL}}(\mathbb{P}_{\ell,\pi}((\chi,\mathcal{H}_{\ell,h}) \in \cdot)||\mathbb{P}_{\ell,\pi}(\chi \in \cdot)\otimes\mathbb{P}_{\ell,\pi}(\mathcal{H}_{\ell,h} \in \cdot))\,,$$

  where $\mathbb{P}_{\ell,\pi}$ is the law of χ and the history induced by policy π interacting with a sample from the posterior distribution of $\mathcal{E}$ given $\mathcal{D}_\ell$.