Dynamic Regret of Online Markov Decision Processes

Authors

Peng Zhao, Long-Fei Li, Zhi-Hua Zhou

Contact

{zhaop, lilf, zhouzh}@lamda.nju.edu.cn

ICML International Conference On Machine Learning

LAMDA Learning And Mining from DatA

南京大學 NANJING UNIVERSITY 1902

Online Markov Decision Processes

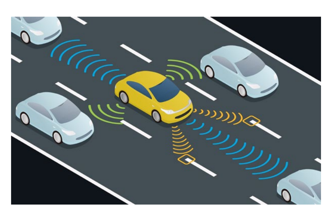
At each round $t = 1,2,\dots,T$:

1. the learner observes the current state x_t, decides a policy $\pi_t: X \times A \to [0,1]$, draws and executes an action a_t from $\pi_t(\cdot|x_t)$;
2. the environment simultaneously picks a loss function $\ell_t: X \times A \to [0,1]$;
3. the learner suffers loss $\ell_t(x_t, a_t)$, observes function ℓ_t and transits the next state x_{t+1} according to the transition kernel $P(\cdot|x_t, a_t)$.

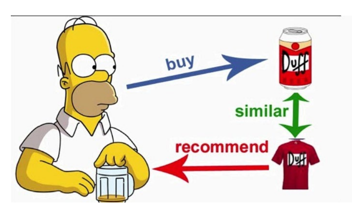
Regret: to learn as well as the best fixed policy

$$\text{Regret}_T = \sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \min_{\pi\in\Pi} \sum_{t=1}^{T} \ell_t(x_t, \pi(x_t)) = \sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \sum_{t=1}^{T} \ell_t(x_t, \pi^*(x_t))$$

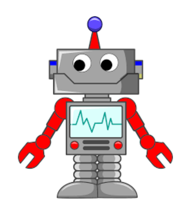
Non-stationary Environments: online MDPs for real-world applications



autonomous driving



online recommendations



robots

Performance Measure: Dynamic Regret

Dynamic Regret : competing with ***any*** policies $\pi_1^c, \dots, \pi_T^c$

$$\text{D-Regret}(\pi_1^c, \cdots, \pi_T^c) = \sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \sum_{t=1}^{T} \ell_t(x_t, \pi_t^c(x_t))$$

adaptive to non-stationarity of environments

universal guarantee against any compared policy sequence

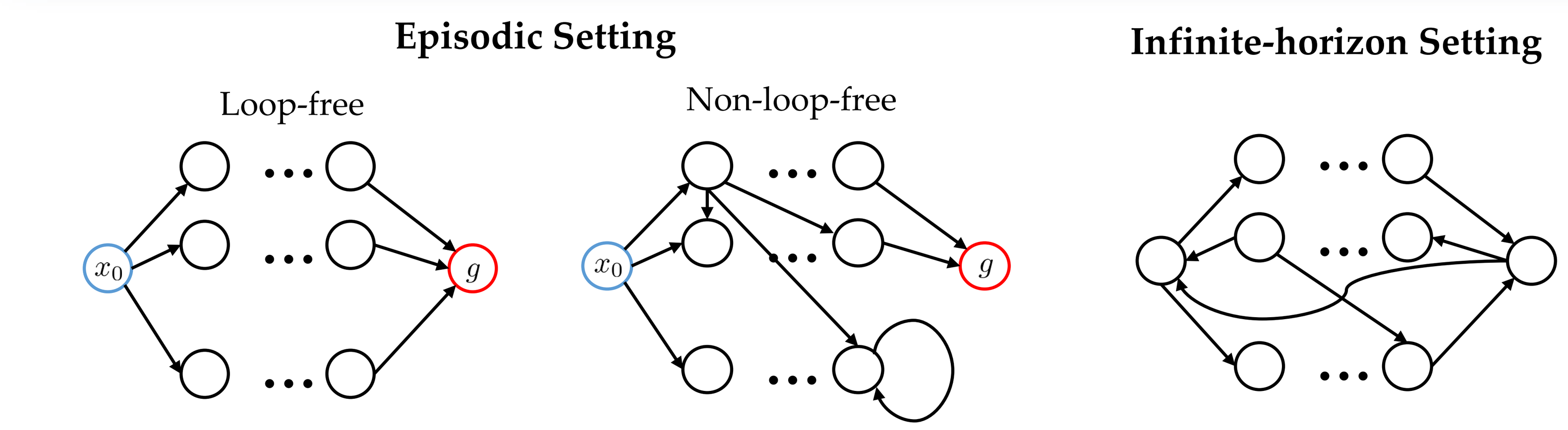
specialize → Static regret, by setting $\pi_1^c = \cdots = \pi_T^c = \pi^*$

$$\text{Regret}_T = \sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \sum_{t=1}^{T} \ell_t(x_t, \pi^*(x_t))$$

specialize → Worst-case dynamic regret, by setting $\pi_t^c = \pi_t^* \in \arg\min_{\pi\in\Pi} \ell_t(x_t, \pi(x_t))$

$$\text{D-Regret}_T^* = \sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \sum_{t=1}^{T} \ell_t(x_t, \pi_t^*(x_t))$$

Our Results



MDP Model	Ours Result (dynamic regret)	Previous Work (static regret)
Episodic loop-free SSP (Section 2)	$\tilde{O}(H\sqrt{K(1+P_T)})$ [Theorem 1]	$\tilde{O}(H\sqrt{K})$ (Zimin & Neu, 2013)
Episodic SSP (Section 3)	$\tilde{O}(\sqrt{B_K(H_* + \bar{P}_K)} + \bar{P}_K)$ [Theorem 3]	$\tilde{O}(\sqrt{H^{\pi^*}DK})$ (Chen et al., 2021a)
Infinite-horizon MDPs (Section 4)	$\tilde{O}(\sqrt{\tau T(1+\tau P_T)} + \tau^2 P_T)$ [Theorem 6]	$\tilde{O}(\sqrt{\tau T})$ (Zimin & Neu, 2013)

- *Our obtained dynamic regret bounds immediately recover the best known static regret.*
- *The dynamic regret for episodic (loop-free) SSP are proved to be minimax optimal.*
- *All our results are achieved by parameter-free algorithms.*

Episodic Loop-free SSP

O-REPS [Zimin & Neu, 2013]: $q_{k+1} = \arg\min_{q\in\Delta(M)} \eta \langle q, \ell_k\rangle + D_\psi(q, q_k)$

Dynamic regret of O-REPS:

$$\sum_{k=1}^{K} \left\langle q_k - q^{\pi_k^c}, \ell_k \right\rangle \le \eta T + \frac{1}{\eta}\left(H \log\frac{|X||A|}{H} + \bar{P}_T \log\frac{1}{\alpha}\right) \text{ with } \bar{P}_T = \sum_{k=2}^{K} \left\| q^{\pi_k^c} - q^{\pi_{k-1}^c} \right\|_1$$

Key challenge: how to deal with the unknown path length $\bar{P}_T$?

Main idea: online ensemble with meta-base two-layer structure.

Step size pool: $\eta_1 \quad \eta_2 \quad \dots \quad \eta_N$

Base-algorithm: $q_{k+1,i} = \arg\min_{q\in\Delta(M)} \eta_i \langle q, \ell_k\rangle + D_\psi(q, q_{k,i})$

Meta-algorithm: $p_{k+1,i} \propto \exp(-\varepsilon \sum_{s=1}^{k} h_{s,i})$, where $h_{s,i} = \langle q_{s,i}, \ell_s\rangle, \forall i \in [N]$.

Dynamic regret decomposition (for any base-learner i):

$$\text{D-Regret}_T = \sum_{k=1}^{K}\langle q_k, \ell_k\rangle - \sum_{k=1}^{K}\langle q^{\pi_k^c}, \ell_k\rangle = \underbrace{\sum_{k=1}^{K}\langle q_k - q_{k,i}, \ell_k\rangle}_{\texttt{meta-regret}} + \underbrace{\sum_{k=1}^{K}\langle q_{k,i} - q^{\pi_k^c}, \ell_k\rangle}_{\texttt{base-regret}}$$

where q_k is the final policy at episode k; $q_{k,i}$ is the policy of the i-th base-learner, $\forall i \in [N]$.

Base-regret. $\sum_{k=1}^{K} \langle q_{k,i^*} - q^{\pi_k^c}, \ell_k\rangle \le \eta_{i^*} T + \frac{H\log(|X||A|/H) + 2\bar{P}_T \log T}{\eta_{i^*}} \le \tilde{O}(\sqrt{T(H+\bar{P}_T)})$

Meta-regret. $\sum_{k=1}^{K}\langle p_k, h_k\rangle - h_{k,i} \le \frac{\log N}{\varepsilon} + \varepsilon \sum_{k=1}^{K} \|h_k\|_\infty^2 \le \sqrt{HT\log N}$

Lower bound: $\mathbb{E}[\text{D-Regret}_K] \ge \Omega(\sqrt{T(H+\bar{P}_T)\log|X||A|})$

Path length of policies : $P_T = \sum_{k=2}^{K}\sum_{l=0}^{H-1} \|\pi_{k,l}^c - \pi_{k-1,l}^c\|_{1,\infty}$

Relationship between P_T and $\bar{P}_t$: $\bar{P}_T \le HP_T$

Episodic SSP

Challenge 1: simultaneously deal with two uncertainties:

- ♠ Unknown horizon length and unknown path length of $\pi_1^c, \dots, \pi_K^c$.

Solution 1: group-wise scheduling :

- ♠ Horizon pool $\mathcal{H} = \{H_1, \dots, H_G\}$, step size grid $\mathcal{E}_i = \{\eta_{i,1}, \dots, \eta_{i,N_i}\}$ for each H_i.

Challenge 2: $\mathbb{E}[\text{D-Regret}_K] = \mathbb{E}\left[(L_K - L_K^{i^*})\right] + \mathbb{E}\left[(L_K^{i^*} - L_K^c)\right]$

Static regret: $\text{base-regret} \le H_{i^*}/\eta_{i^*} + \eta_{i^*} L_K^c \le \tilde{O}(\sqrt{H_{i^*} L_K^c}) \le \tilde{O}(\sqrt{H^{\pi^*}DK})$

$$L_K^{i^*} \le L_K^c + \tilde{O}(\sqrt{H^{\pi^*}DK}) \le DK + \tilde{O}(\sqrt{H^{\pi^*}DK}) = \tilde{O}(DK)$$

$$\text{meta-regret} \le 1/\varepsilon_{i^*} + \boxed{\varepsilon_{i^*} H_{i^*} L_K^{i^*}} \le 1/\varepsilon_{i^*} + \varepsilon_{i^*} H_{i^*} DK = \tilde{O}(\sqrt{H^{\pi^*}DK})$$

Dynamic regret: $L_K^{i^*} \le L_K^c + \text{base-regret}$, but $L_K^c \le DK$ does not holds!

Solution 2: add correction term in both base and meta level.

Base algo. $q_{k+1}^{i,j} = \arg\min_{q\in\Delta(M,H_i,\alpha)} \eta\langle q, \ell_k + a_k\rangle + D_\psi(q, q_k^{i,j})$ where $a_k = 32\eta\ell_k^2$

Meta algo. $p_{k+1} = \arg\min_{p\in\Delta_N} \langle p, h_k + b_k\rangle + D_{\bar{\psi}}(p, p_k)$ where $h_k^{i,j} = \langle q_k^{i,j}, \ell_k\rangle, b_k = 32\varepsilon h_k^2$

weighted entropy: $\bar{\psi}(p) = \sum_{i=1}^{G}\sum_{j=1}^{N_i} \frac{1}{\varepsilon_{i,j}} p_{i,j}\log p_{i,j}$, with $\varepsilon_{i,j} = \frac{\eta_{i,j}}{2H_i}$

$$\text{base-regret} \le (H_{i^*} + \bar{P}_T)/\eta_{i^*,j^*} + \eta_{i^*,j^*} L_K^c - \boxed{\eta_{i^*,j^*}\sum_{k=1}^{K}\langle q_k^{i^*,j^*}, \ell_k^2\rangle}$$

the negative term is crucial to cancel the term in meta-regret

Dynamic regret bound: $\mathbb{E}[\text{D-Regret}_K] \le \tilde{O}\left(\sqrt{(H_* + \bar{P}_K)(H_* + \bar{P}_K + L_K^c)}\right)$

Lower bound: $\mathbb{E}[\text{D-Regret}_K] \ge \Omega\left(\sqrt{DH_*K(1 + \bar{P}_K/H_*)}\right)$ and $\bar{P}_K \ge cP_K, \forall\, c > 0$.

Infinite-horizon MDPs

Reduction to the switching-cost expert problem:

$$\mathbb{E}[\text{D-Regret}_T] \le \sum_{t=1}^{T}\langle q_t - q^{\pi_t^c}, \ell_t\rangle + (\tau+1)\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1 + (\tau+1)^2 P_T + 4(\tau+1)$$

where τ is the mixing time and P_T is the path length defined as $P_T = \sum_{t=2}^{T} ||\pi_t^c - \pi_{t-1}^c||_{1,\infty}$.

Main difficulty: switching-cost in the meta-base structure.

$$\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1 = \sum_{t=2}^{T}\left\|\sum_{i=1}^{N} p_{t,i}q_{t,i} - \sum_{i=1}^{N} p_{t-1,i}q_{t-1,i}\right\|_1 \le \sum_{t=2}^{T}\sum_{i=1}^{N} p_{t,i}\|q_{t,i} - q_{t-1,i}\|_1 + \sum_{t=2}^{T}\|p_t - p_{t-1}\|_1$$

Solution: Add a correction term to penalize unstable base-learners.

Surrogate loss: $h_{t,i} = \langle q_{t,i}, \ell_t\rangle + (\tau+1)\|q_{t,i} - q_{t-1,i}\|_1$

Dynamic regret decomposition (for any base-learner i):

$$\sum_{t=1}^{T}\langle q_t - q^{\pi_t^c}, \ell_t\rangle + (\tau+1)\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1$$

$$= \underbrace{\sum_{t=1}^{T}(\langle p_t, h_t\rangle - h_{t,i}) + (\tau+1)\sum_{t=2}^{T}\|p_t - p_{t-1}\|_1}_{\text{meta-regret}} + \underbrace{\sum_{t=1}^{T}\left\langle q_{t,i} - q^{\pi_t^c}, \ell_t\right\rangle + (\tau+1)\sum_{t=2}^{T}\|q_{t,i} - q_{t-1,i}\|_1}_{\text{base-regret}}$$

Base-regret regarding the best learner i^*: Define $\bar{P}_T = \sum_{t=2}^{T}\left\|q^{\pi_t^c} - q^{\pi_{t-1}^c}\right\|_1$

$$\text{base-regret} \le \eta_{i^*}T + \frac{H\log(|X||A|) + 2\bar{P}_T\log T}{\eta_{i^*}} + (\tau+1)\eta_{i^*}T \le \tilde{O}\left(\sqrt{\tau T(1+\bar{P}_T)}\right)$$

Meta-regret. $\text{meta-regret} \le \frac{\log N}{\varepsilon} + 2\varepsilon(2\tau+3)^2 T \le (2\tau+3)\sqrt{2T\log N}$

Relationship between P_T and $\bar{P}_T$: $\bar{P}_T \le (\tau+2)P_T$