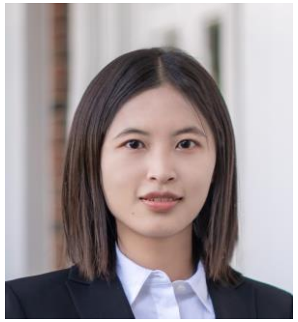
ICML
International Conference
On Machine Learning

NVIDIA

Reinforcement Learning with Segment Feedback


Yihan Du
UIUC


Anna Winnick
Stanford University


Gal Dalal
Nvidia


Shie Mannor
Technion/Nvidia

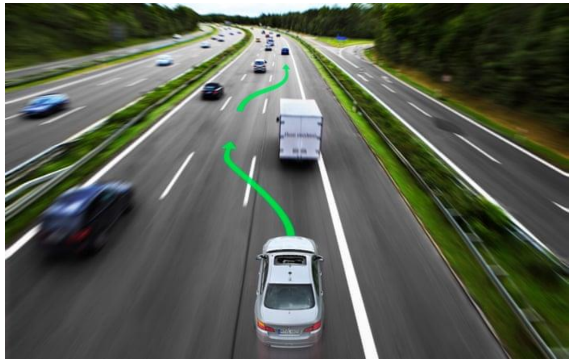

R. Srikant
UIUC

Speaker: Yihan Du
duyihan1996@gmail.com

ICML 2025

Motivation

- Reinforcement learning (RL) [Sutton & Barto, 2018]:
  - An agent interacts with an unknown environment through time
  - Goal of maximizing the expected cumulative reward
  - Applications: robotics, autonomous driving, …
- Classic RL: observe reward for each state-action pair
- However, in real-world applications, e.g., autonomous driving:
  - It is difficult and costly to collect a reward for each state-action pair
- Prior works – RL with trajectory feedback [Efroni et al., 2021; Chatterji et al., 2021]:
  - Observe a reward signal at the end of each trajectory

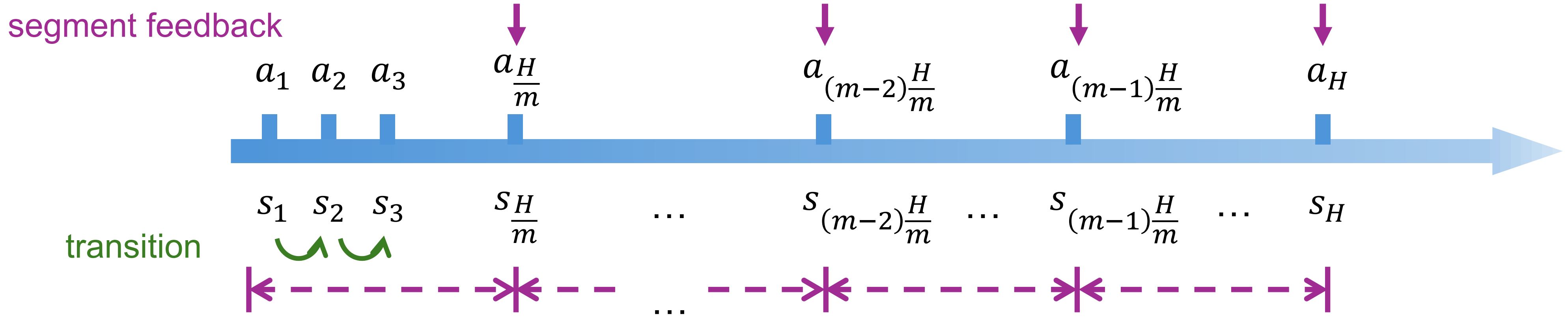


The relationship between feedback frequency and the performance of RL algorithms is still unknown

RL with Segment Feedback

- Episodic Markov decision process (MDP):
  - H: the length of each episode
  - $r(s,a) \in [-r_{max}, r_{max}]$: unknown reward function. Denote $\theta^* := [r(s,a)]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$
  - $p(s'|s,a)$: transition distribution
  - $\pi_h(s)$: policy, specify what action to take in state s at step h
- Value functions: $V_h^\pi(s) = \mathbb{E}[\sum_{t=h}^{H} r(s_t,a_t)|s_h = s,\pi]$. Optimal policy: $\pi^* = \underset{\pi}{\operatorname{argmax}}\, V_h^\pi(s)$ for all $h \in [H]$ and $s \in \mathcal{S}$
- Segment feedback: each episode is equally divided into m segments, observe reward feedback at the end of each segment:
  - Binary feedback y_i: $\Pr[y_i = 1] = \frac{1}{1+\exp(-(\phi^{\tau_i})^\top \theta^*)}$, $\Pr[y_i = 0] = 1 - \Pr[y_i = 1]$ (Thumbs up/down)
  - Sum feedback: $R_i = (\phi^{\tau_i})^\top \theta^* + \sum_{t=(i-1)\cdot\frac{H}{m}+1}^{i\cdot\frac{H}{m}} \varepsilon_t$
- Goal: minimize regret $\mathcal{R}(K) := \sum_{k=1}^{K}(V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1))$

τ_i: the i-th trajectory segment, where $i \in [m]$
$\phi^\tau(s,a)$: the number of times (s,a) is visited in (sub-)trajectory τ



Algorithm for Binary Feedback

Known transition

Algorithm SegBiTS:

- For episode $k = 1, \ldots, K$:
  - $\hat{\theta}_{k-1} \leftarrow \underset{\theta}{\operatorname{argmin}} -\sum_{k'=1}^{k-1} \sum_{i=1}^{m} \log\left(\frac{1}{1+\exp\left(-y_i^{k'} \left(\phi^{\tau_i^{k'}}\right)^\top \theta\right)}\right) + \frac{1}{2}\lambda \|\theta\|_2^2$
  - $\Sigma_{k-1} \leftarrow \sum_{k'=1}^{k-1} \sum_{i=1}^{m} \phi^{\tau_i^{k'}} \left(\phi^{\tau_i^{k'}}\right)^\top + \alpha\lambda I$
  - Sample noise $\xi_k \sim \mathcal{N}(0, \alpha \cdot v(k-1)^2 \cdot \Sigma_{k-1}^{-1})$
  - $\tilde{\theta}_k \leftarrow \hat{\theta}_{k-1} + \xi_k$
  - $\pi^k \leftarrow \underset{\pi}{\operatorname{argmax}} (\phi^\pi)^\top \hat{\theta}_k$
  - Play episode k with policy π^k. Observe trajectory τ^k and binary segment feedback $\{y_i^k\}_{i\in[m]}$

- λ: regularization parameter
- $\alpha := \exp\left(\frac{H r_{max}}{m}\right) + \exp\left(-\frac{H r_{max}}{m}\right) + 2$
- $v(k-1)$: part of the confidence radius for $\hat{\theta}_{k-1}$
- $\phi^\pi(s,a)$: the expected number of times (s,a) is visited in an episode under π

Algorithm for Sum Feedback

Known transition

Algorithm E-LinUCB:

- Let $w^* \in \Delta_\Pi$ and z^* be the optimal solution and optimal value of the optimization

$$\min_{w \in \Delta_\Pi} \left\| \left(\sum_{\pi \in \Pi} w(\pi) \left(\sum_{i=1}^{m} \mathbb{E}_{\tau_i \sim \pi}[\phi^{\tau_i} (\phi^{\tau_i})^T]\right)\right)^{-1} \right\| \quad \text{// E-experimental design}$$

- $K_0 \leftarrow \tilde{O}((z^*)^2 H^4)$
- Round the continuous sampling distribution w^* into K^0 discrete sampling policies $(\pi^1, \dots, \pi^{K_0})$
- Play K_0 episodes with policies $\pi^1, \dots, \pi^{K_0}$. Observe trajectories $\tau^1, \dots, \tau^{K_0}$ and sum feedback $\{R_i^1\}_{i \in [m]}, \dots, \{R_i^{K_0}\}_{i \in [m]}$
- For episode $k = K_0 + 1, \dots, K$:
  - $\hat{\theta}_{k-1} \leftarrow \left(\lambda I + \sum_{k'=1}^{k-1} \sum_{i=1}^{m} \phi^{\tau_i^{k'}} \left(\phi^{\tau_i^{k'}}\right)^\top\right)^{-1} \sum_{k'=1}^{k-1} \sum_{i=1}^{m} \phi^{\tau_i^{k'}} R_i^{k'}$
  - $\Sigma_{k-1} \leftarrow \lambda I + \sum_{k'=1}^{k-1} \sum_{i=1}^{m} \phi^{\tau_i^{k'}} \left(\phi^{\tau_i^{k'}}\right)^\top$
  - $\pi^k \leftarrow \operatorname{argmax}_{\pi \in \Pi}((\phi^\pi)^\top \hat{\theta}_{k-1} + \beta(k-1) \cdot \|\phi^\pi\|_{\Sigma_{k-1}^{-1}})$, where $\beta(k-1)$ is part of the confidence radius for $\hat{\theta}_{k-1}$
  - Play episode k with policy π^k. Observe trajectory τ^k and sum feedback $\{R_i^k\}_{i \in [m]}$

Theoretical Results

Theorem 1. With probability at least $1-\delta$, for any $K>0$, the regret of algorithm SegBiTS is bounded by

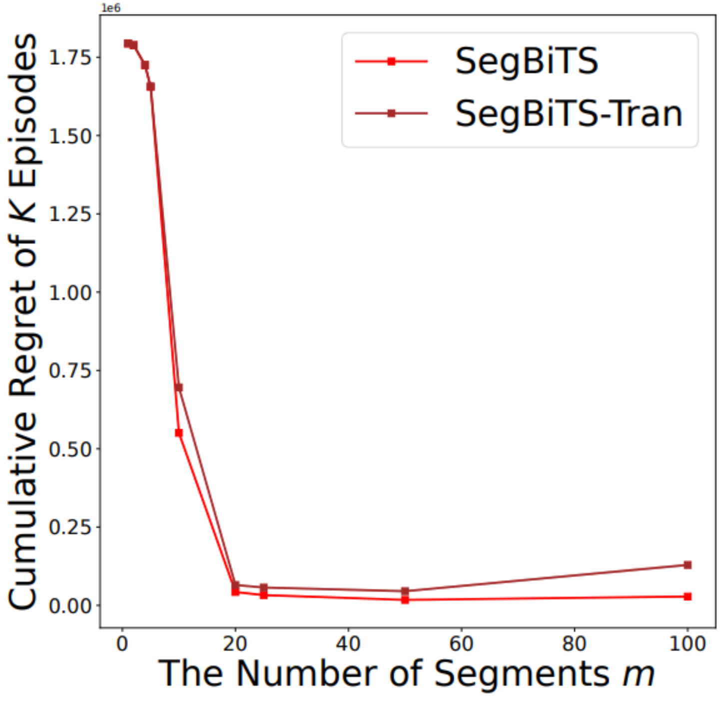
$$\tilde{O}\left(\exp\left(\frac{Hr_{max}}{2m}\right)\nu(K)\sqrt{|\mathcal{S}||\mathcal{A}|}\cdot\left(\sqrt{Km|\mathcal{S}||\mathcal{A}|\max\left\{\frac{H^2}{m\alpha\lambda},1\right\}}+H\sqrt{\frac{K}{\alpha\lambda}}\right)\right)$$

Theorem 2. With probability at least $1-\delta$, for any $K>0$, the regret of algorithm E-LinUCB is bounded by

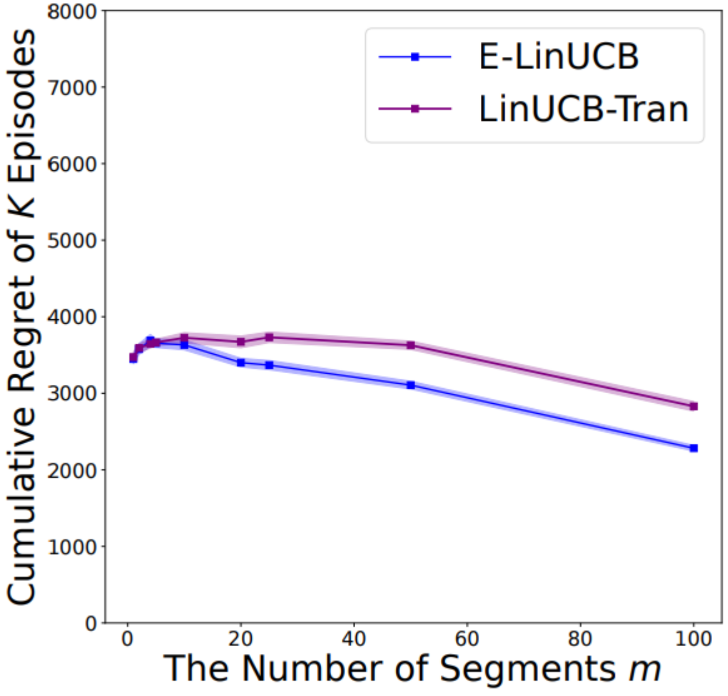
$$\tilde{O}\left(|\mathcal{S}||\mathcal{A}|\sqrt{HK}+(z^*)^2H^5+|\mathcal{S}||\mathcal{A}|H\right)$$

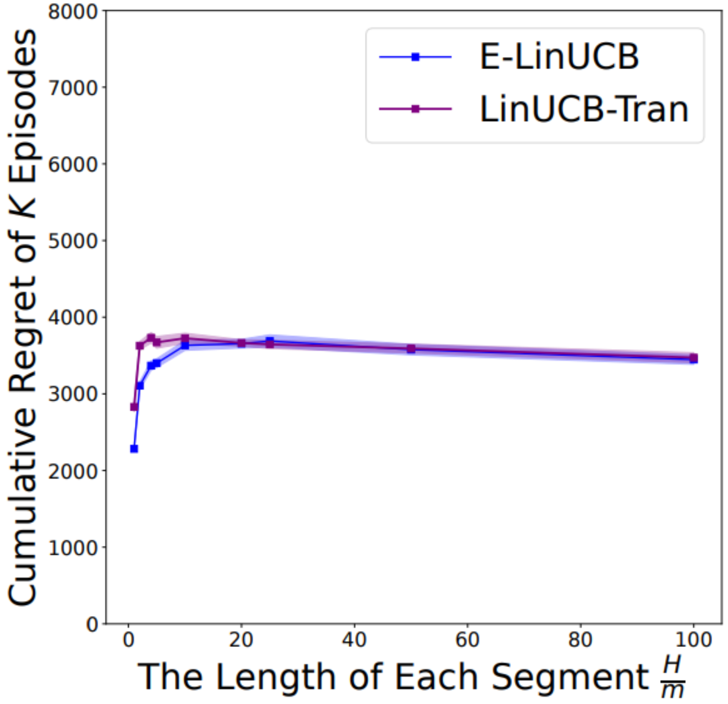
- The influence of the number of segments m on learning performance:
  - Under binary feedback, increasing m significantly helps accelerate learning
  - Under sum feedback, surprisingly, increasing m does not help accelerate learning much
- Lower bounds and extensions to the unknown transition setting are also provided in our paper

Empirical Evaluation



Binary feedback





Sum feedback

Conclusion

- Study a general model called RL with segment feedback
  - Bridge the gap between per-state-action feedback in classic RL and trajectory feedback seamlessly
- Design algorithms SegBiTS and E-LinUCB for binary and sum feedback settings, respectively
- Our theoretical and empirical results exhibit how the number of segments m impacts learning performance:
  - Under binary feedback, increasing m significantly helps accelerate learning
  - Under sum feedback, surprisingly, increasing m does not help accelerate learning much

References

- Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, R. Srikant. Reinforcement Learning with Segment Feedback. ICML, 2025.
- Sutton, R. S. and Barto, A. G. Reinforcement learning: An introduction. MIT press, 2018.
- Efroni, Y., Merlis, N., and Mannor, S. Reinforcement learning with trajectory feedback. AAAI, 2021.
- Chatterji, N., Pacchiano, A., Bartlett, P., and Jordan, M. On the theory of reinforcement learning with once-per-episode feedback. NeurIPS, 2021.

Thank You

Yihan Du

Postdoc
UIUC
duyihan1996@gmail.com