Tempo Adaptation in Non-stationary Reinforcement Learning

Hyunin Lee [1] Yuhao Ding [1] Jongmin Lee [1] Ming Jin[2]
Javad Lavaei[1] Somayeh Sojoudi[1]

[1]UC Berkeley [2]Virginia Tech

Overlooked issue: Time synchronization

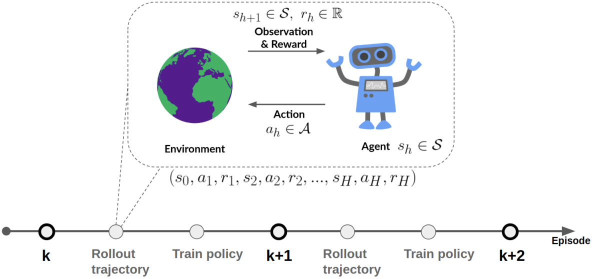

Figure 1: Conventional Non-stationary RL environment

- **Key observation**: In reality, environmental changes occur over wall-clock time (t) rather than episode progress (k).
- Existing works: episode k → collect data & train policy → episode $k+1$.
- In reality: time t_k → spend Δt for collecting data & training → time $t_{k+1} = t_k + \Delta t$.

Remove time synchronization

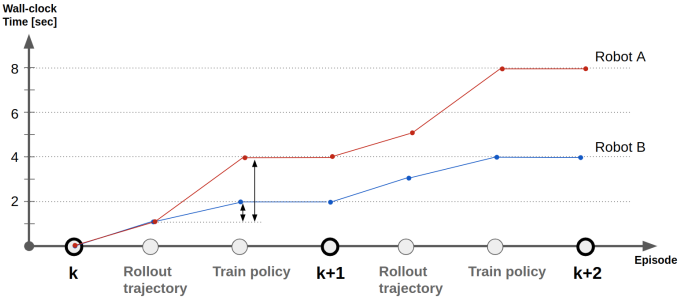

Figure 2: Different training time makes agent encounters different environment

- In time-desynchorzied environment, the agent should choose **when to interact** $(t_1, t_2, \ldots, t_K)$ additional to **how many times to interact** (K)
- The choice of *interaction times* $(t_1, t_2, \ldots, t_K)$ significantly impacts the suboptimality gap of the policy.

Contribution

- We propose a Proactively Synchronizing Tempo (`ProST`) framework that computes suboptimal $\{\mathfrak{t}_1, \mathfrak{t}_2, ..., \mathfrak{t}_K\}(= \{\mathfrak{t}\}_{1:K})$.
- ProST framwork computes suboptimal $\{\mathfrak{t}\}_{1:K}$ by minimizing the upper bound of its performance metric, dynamic regret.
- One interesting property is that we show suboptimal $\{\mathfrak{t}\}_{1:K}$ strikes a balance between the policy training time (**agent tempo**) and how fast the environment changes (**environment tempo**).

ProsT framework

For given $t \in [0, T]$, ProST framework computes K^*, $\{t_1^*, t_2^*, .., t_{K^*}^*\}$, then $\{\pi_{t_1^*}, \pi_{t_2^*}, .., \pi_{t_{K^*}^*}\}$ into two components

- Time optimizer
- Future policy optimizer

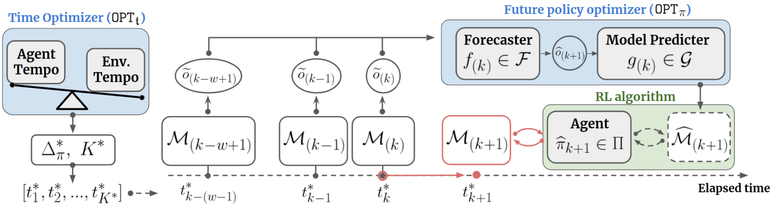

Figure 3: ProST framework

Future policy optimizer

For given $\mathfrak{t}_k$, $\mathfrak{t}_{k+1}$, it computes a near-optimal policy of $\mathfrak{t}_{k+1}$ at time $\mathfrak{t}_k$

Definition (MDP forecaster $g \circ f$)

Consider two function classes $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F} : \mathcal{O}^w \to \mathcal{O}$ and $\mathcal{G} : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R} \times \Delta(\mathcal{S})$, where $w \in \mathbb{N}$. Then, for $f_{(k)} \in \mathcal{F}$ and $g_{(k)} \in G$, we define MDP forecaster at time t_k as $(g \circ f)_{(k)} : \mathcal{O}^w \times \mathcal{S} \times \mathcal{A} \to \mathbb{R} \times \Delta(\mathcal{S})$.

Estimate the future MDP model and optimize.

- At $\mathfrak{t} = \mathfrak{t}_k$
- During $\mathfrak{t} \in (\mathfrak{t}_k, \mathfrak{t}_{k+1})$
  1. $\hat{o}_{(k+1)} = f_{(k)}(\{\tilde{o}\}_{(k-w+1,k)})$
  2. $(\widehat{R}_{(k+1)}(s,a), \widehat{P}_{(k+1)}(\cdot|s,a)) = g_{(k)}(s, a, \hat{o}_{k+1})$
  3. $\widehat{\pi}_{(k+1)} \leftarrow \widehat{\mathcal{M}}_{(k+1)} = \langle \mathcal{S}, \mathcal{A}, H, \widehat{P}_{(k+1)}, \widehat{R}_{(k+1)}, \gamma \rangle$
- At $\mathfrak{t} = \mathfrak{t}_{k+1}$

Time optimizer

Strategy: Δ_π^* is a minimizer of the dynamic regret's upper bound

- Analysis on finite space $|\mathcal{S}|, |\mathcal{A}| < \infty \rightarrow$ `ProST-T`

Theorem (ProST-T dynamic regret $\mathfrak{R}$)

Let $\iota_H^K = \sum_{k=1}^{K-1} \sum_{h=0}^{H-1} \iota_h^{(k+1)}(s_h^{(k+1)}, a_h^{(k+1)})$ *and* $\bar{\iota}_\infty^K := \sum_{k=1}^{K-1} \|\bar{\iota}_\infty^{k+1}\|_\infty$*, where* ι_H^K *is a data-dependent error. For a given* $p \in (0,1)$*, the dynamic regret of the forecasted policies* $\{\widehat{\pi}^{(k+1)}\}_{1:K-1}$ *of ProST-T is upper bounded with probability at least* $1 - p/2$ *as follows:*

$$\mathfrak{R}\left(\{\widehat{\pi}^{(k+1)}\}_{1:K-1}, K)\right) \leq \mathfrak{R}_I + \mathfrak{R}_{II}$$

where $\mathfrak{R}_I = \bar{\iota}_\infty^K/(1-\gamma) - \iota_H^K + C_p \cdot \sqrt{K-1}$, $\mathfrak{R}_{II} = C_{II}[\Delta_\pi] \cdot (K-1)$*, and* $C_p, C_{II}[\Delta_\pi]$ *are some functions of* p, Δ_π*, respectively.*

- $\mathfrak{R}_I \leftarrow$ Forecasting model error $\leftarrow B(\Delta_\pi)$ (rate of environment's change)
- $\mathfrak{R}_{II} \leftarrow$ Policy optimization error $\leftarrow \Delta_\pi$ (rate of agent's adaption)
- $\Delta_{*\pi}$ strikes a balance between $\mathfrak{R}_I$ and $\mathfrak{R}_{II}$

Δ_π bounds for sublinear $\mathfrak{R}_{II}$

Δ_π^* should satisfy sublinear dynamic regret to K

- δ : approximation gap
- τ : entropy regularization parameter
- η : learning rate

Proposition (Δ_π bounds for sublinear $\mathfrak{R}_{II}$)

A total step H is given by MDP. For a number $\epsilon > 0$ such that $H = \Omega\left(\log\left((\widehat{r}_{max} \vee r_{max})/\epsilon\right)\right)$, we choose δ, τ, η to satisfy $\delta = \mathcal{O}(\epsilon)$, $\tau = \Omega\left(\epsilon / \log|\mathcal{A}|\right)$ and $\eta \leq (1-\gamma)/\tau$, where $\widehat{r}_{max}$ and r_{max} are the maximum reward of the forecasted model and the maximum reward of the environment, respectively. Define $\mathbb{N}_{II} := \{n \mid n > \frac{1}{\eta\tau}\log\left(\frac{C_1(\gamma+2)}{\epsilon}\right), n \in \mathbb{N}\}$, where C_1 is a constant. Then $\mathfrak{R}_{II} \leq 4\epsilon(K-1)$ for all $\Delta_\pi \in \mathbb{N}_{II}$.

$\mathfrak{R}_I \leftarrow$ Forecasting model error $\leftarrow B(\Delta_\pi)$

SW-LSE : Sliding window regularized LSE

Theorem (Dynamic regret $\mathfrak{R}_I$ when f = SW-LSE)

For given $p \in (0,1)$, if the exploration bonus constant β and regularization parameter λ satisfy $\beta = \Omega(|\mathcal{S}|H\sqrt{\log(H/p)})$, $\lambda \geq 1$, then the $\mathfrak{R}_I$ is bounded with probability $1-p$,

$$\mathfrak{R}_I \leq C_I[B(\Delta_\pi)] \cdot w + C_k \cdot \sqrt{\frac{1}{w}\log\left(1 + \frac{H}{\lambda}w\right)}$$

where $C_I[B(\Delta_\pi)] = (1/(1-\gamma) + H) \cdot B_r(\Delta_\pi) + (1 + H\hat{r}_{max})\gamma/(1-\gamma) \cdot B_p(\Delta_\pi)$, and C_k is a constant on the order of $\mathcal{O}(K)$.

Δ_π bounds for sublinear $\mathfrak{R}_I$

Proposition (Δ_π bounds for sublinear $\mathfrak{R}_I$)

Denote $B(1)$ as the environment tempo when $\Delta_\pi = 1$, which is a summation over all time steps. Assume that the environment satisfies $B_r(1) + B_p(1)\hat{r}_{max}/(1-\gamma) = o(K)$ and we choose $w = \mathcal{O}((K-1)^{2/3}/(C_I[B(\Delta_\pi)])^{2/3})$. Define the set $\mathbb{N}_I$ to be $\{n \mid n < K,\ n \in \mathbb{N}\}$. Then $\mathfrak{R}_I$ is upper-bounded as $\mathfrak{R}_I = \mathcal{O}\left(C_I[B(\Delta_\pi)]^{1/3}(K-1)^{2/3}\sqrt{\log((K-1)/C_I[B(\Delta_\pi)])}\right)$ and also satisfies a sublinear upper bound, provided that $\Delta_\pi \in \mathbb{N}_I$.

Δ_π^* strikes a balance between $\mathfrak{R}_I$ and $\mathfrak{R}_{II}$

- $\mathfrak{R}_I$ upperbound is increasing on a interval $\mathbb{N}_I \cap \mathbb{N}_{II}$
- $\mathfrak{R}_{II}$ upperbound is decreasing on a interval $\mathbb{N}_I \cap \mathbb{N}_{II}$

Theorem (Suboptimal tempo Δ_π^*)

Let $k_{Env} = (\alpha_r \vee \alpha_p)^2 C_I[B(1)]$, $k_{Agent} = \log(1/(1-\eta\tau))C_1(K-1)(\gamma+2)$. *Consider three cases:* **case1**: $\alpha_r \vee \alpha_p = 0$, **case2**: $\alpha_r \vee \alpha_p = 1$, **case3**: $0 < \alpha_r \vee \alpha_p < 1$ *or* $\alpha_r \vee \alpha_p > 1$. *Then* Δ_π^* *depends on the environment's drifting constants as follows:*

- *Case1:* $\Delta_\pi^* = T$.
- *Case2:* $\Delta_\pi^* = \log_{1-\eta\gamma}(k_{Env}/k_{Agent}) + 1$.
- *Case3:* $\Delta_\pi^* = \exp\left(-W\left[-\frac{\log(1-\eta\tau)}{\max(\alpha_r,\alpha_p)-1}\right]\right)$, *provided that the parameters are chosen so that* $k_{Agent} = (1-\eta\tau)k_{Env}$.

Performance

Benchmark methods

- MBPO : state of the art model-based policy optimization.
- Pro-OLS : policy optimization algorithm that predicts future V.
- ONPG : adaptive algorithm that fine-tunes the policy on current data.
- FTRL : adaptive algorithm that maximizes the performance on all previous data.

Table 1: Average reward returns

Speed	$B(G)$	**Swimmer-v2**					**Halfcheetah-v2**					**Hopper-v2**				
		Pro-OLS	ONPG	FTML	MBPO	ProST-G	Pro-OLS	ONPG	FTML	MBPO	ProST-G	Pro-OLS	ONPG	FTML	MBPO	ProST-G
1	16.14	-0.40	-0.26	-0.08	-0.08	**0.57**	-83.79	-85.33	-85.17	-24.89	**-19.69**	**98.38**	95.39	97.18	92.88	92.77
2	32.15	0.20	-0.12	0.14	-0.01	**1.04**	-83.79	-85.63	-86.46	-22.19	**-20.21**	98.78	97.34	**99.02**	96.55	98.13
3	47.86	-0.13	0.05	-0.15	-0.64	**1.52**	-83.27	-85.97	-86.26	-21.65	**-21.04**	97.70	98.18	98.60	95.08	**100.42**
4	63.14	-0.22	-0.09	-0.11	-0.04	**2.01**	-82.92	-84.37	-85.11	-21.40	**-19.55**	98.89	97.43	97.94	97.86	**100.68**
5	77.88	-0.23	-0.42	-0.27	0.10	**2.81**	-84.73	-85.42	-87.02	**-20.50**	-20.52	97.63	99.64	99.40	96.86	**102.48**
A	8.34	1.46	2.10	**2.37**	-0.08	0.57	-76.67	-85.38	-83.83	-40.67	**83.74**	104.72	**118.97**	115.21	100.29	111.36
B	4.68	**1.79**	-0.72	-1.20	0.19	0.20	-80.46	-86.96	-85.59	-29.28	**76.56**	80.83	**131.23**	110.09	100.29	127.74

*Whole training procedure is in Appendix

Ablation study

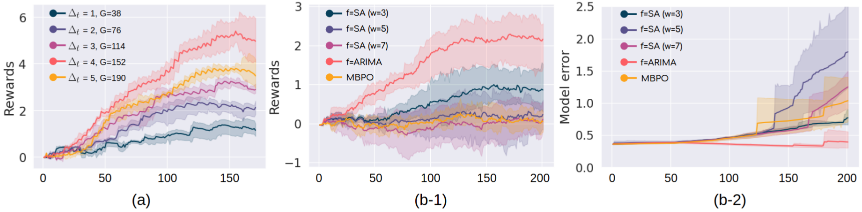

Figure 4: (a) Optimal Δ_π^*. (b-1) Different forecaster f (ARIMA, SA). (b-2) The Mean squared Error (MSE) model loss of four `ProST-G` with different forecasters(ARIMA and three SA) and the MBPO. x-axis are all episodes.