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

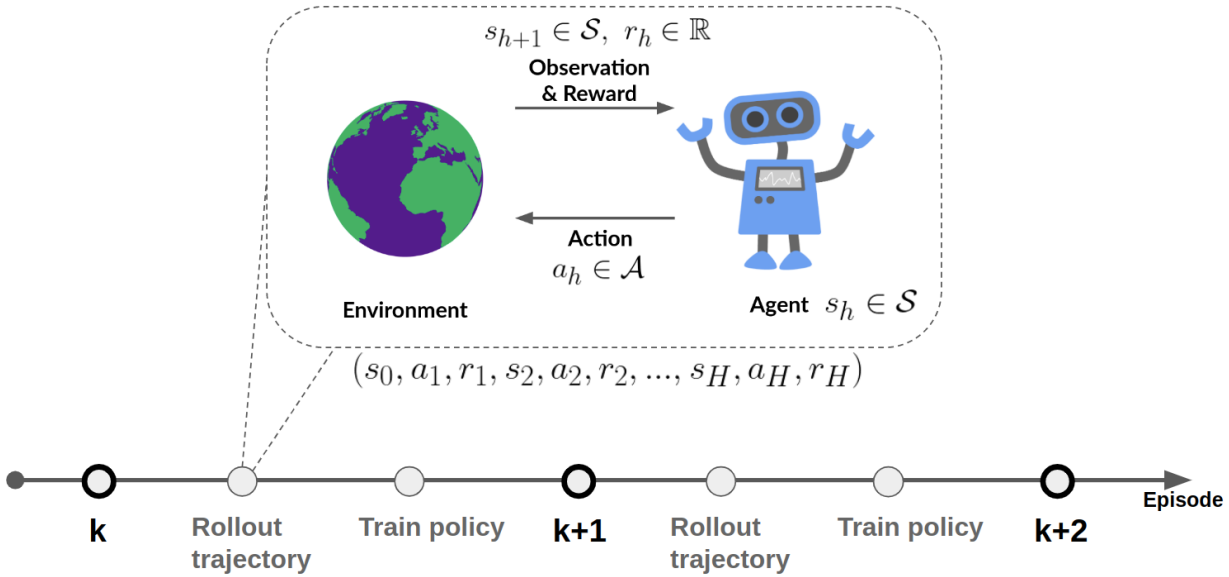

Figure 1: Conventional Non-stationary RL environment

Key observation: In reality, environmental changes occur over wall-clock time (t) rather than episode progress (k).

Existing works: episode $k \rightarrow$ collect data & train policy $\rightarrow$ episode $k+1$.

In reality: time $t_k \rightarrow$ spend Δt for collecting data & training $\rightarrow$ time $t_{k+1} = t_k + \Delta t$.

Remove time synchronization

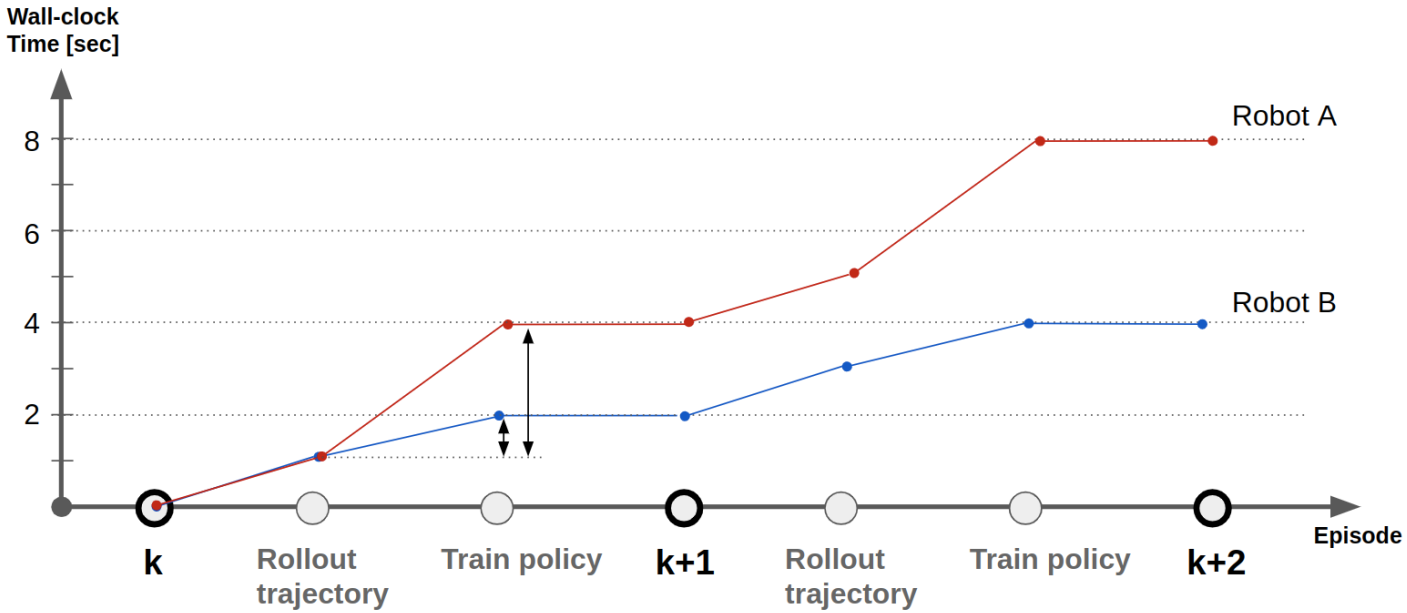

Figure 2: Di erent training time makes agent encounters di erent environment

In time-desynchorzied environment, the agent should choose **when to interact** $(t_1; t_2; \cdots; t_K)$ additional to **how many times to interact** (K)

The choice of *interaction times* $(t_1; t_2; \cdots; t_K)$ significantly impacts the suboptimality gap of the policy.

Contribution

We propose a Proactively Synchronizing Tempo (ProST) framework that computes suboptimal $\{t_1, t_2, \ldots, t_K\} (= \{t\}_{1:K})$.

ProST framwork computes suboptimal $\{t\}_{1:K}$ by minimizing the upper bound of its performance metric, dynamic regret.

One interesting property is that we show suboptimal $\{t\}_{1:K}$ strikes a balance between the policy training time (**agent tempo**) and how fast the environment changes (**environment tempo**).

ProST framework

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

Time optimizer

Future policy optimizer

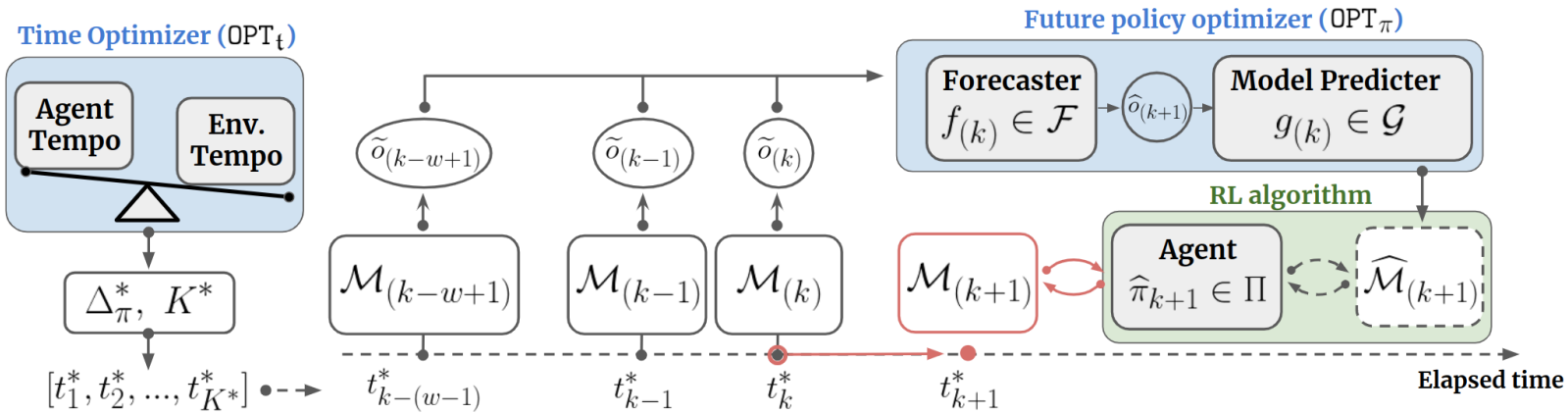

Figure 3: ProST framework

Future policy optimizer

For given t_k, t_{k+1}, it computes a near-optimal policy of t_{k+1} at time t_k

Definition (MDP forecaster $g \circ f$)

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

Estimate the future MDP model and optimize.

At $t = t_k$

During $t \in (t_k, t_{k+1})$

$$\hat{o}_{(k+1)} = f_{(k)}(\{o\}_{(k-w+1:k)})$$
$$(\widehat{R}_{(k+1)}(s,a), \widehat{P}_{(k+1)}(\cdot|s,a)) = g_{(k)}(s,a,\hat{o}_{k+1})$$
$$\widehat{\pi}_{(k+1)} \leftarrow \widehat{\mathcal{M}}_{(k+1)} = \langle \mathcal{S}, \mathcal{A}, H, \widehat{P}_{(k+1)}, \widehat{R}_{(k+1)}, \gamma \rangle$$

At $t = t_{k+1}$

Time optimizer

Strategy: $\Delta^{\square}$ is a minimizer of the dynamic regret's upper bound

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

Theorem (ProST-T dynamic regret R)

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

$$R(\{\hat{} \overset{(k+1)}{}\}_{1 \leq K-1}; K)) \leq R_I + R_{II}$$

where $R_I = \overset{K}{\infty} / (1 - \quad) - \overset{K}{H} + C_p \sqrt{K-1}$; $R_{II} = C_{II}[\Delta\ ] \cdot (K-1)$, *and* $C_p; C_{II}[\Delta\ ]$ *are some functions of* p, $\Delta\ $, *respectively.*

$R_I \leftarrow$ Forecasting model error $\leftarrow B(\Delta\ )$ (rate of environment's change)

$R_{II} \leftarrow$ Policy optimization error $\leftarrow \Delta\ $ (rate of agent's adaption)

$\Delta_{\square}$ strikes a balance between R_I and R_{II}

Δ bounds for sublinear R_{II}

$\Delta^{\square}$ should satisfy sublinear dynamic regret to K

: approximation gap

: entropy regularization parameter

: learning rate

Proposition (Δ bounds for sublinear R_{II})

A total step H is given by MDP. For a number >0 such that $H = \Omega(\log((\hat{r}_{max} \square r_{max})/\ ))$, we choose ; ; to satisfy $= O(\ )$; $= \Omega(\ /\log|A|)$ and $\leq (1-\ )/\ $, where $\hat{r}_{max}$ and r_{max} are the maximum reward of the forecasted model and the maximum reward of the environment, respectively. De ne $N_{II}=\{n \mid n > \frac{1}{\ }\log(\frac{C_1(\ +2)}{\ });n \in N\}$, where C_1 is a constant. Then $R_{II} \leq 4\ (K-1)$ for all $\Delta\ \in N_{II}$.

$R_l \leftarrow$ Forecasting model error $\leftarrow B(\Delta\ )$

SW-LSE : Sliding window regularized LSE

Theorem (Dynamic regret R_l when f = SW-LSE)

For given $p \in (0;1)$, if the exploration bonus constant and regularization parameter satisfy $= \Omega(|S|H\sqrt{\log(H/p)})$; ≥ 1, then the R_l is bounded with probability $1-p$,

$$R_l \leq C_l[B(\Delta\ )]\,w + C_k \sqrt{\frac{}{\frac{1}{w}\log(1+\frac{H}{-}w)}}$$

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

Δ bounds for sublinear R_l

Proposition (Δ bounds for sublinear R_l)

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

$\Delta^{\square}$ strikes a balance between R_I and R_{II}

R_I upperbound is increasing on a interval $N_I \cap N_{II}$

R_{II} upperbound is decreasing on a interval $N_I \cap N_{II}$

Theorem (Suboptimal tempo Δ^*)

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

- *Case1:* $\Delta^{\square} = T$.
- *Case2:* $\Delta^{\square} = \log_{1-\ \ }(k_{Env}/k_{Agent}) + 1$.
- *Case3:* $\Delta^{\square} = \exp(-W[-\frac{\log(1-\ \ )}{\max(\ _r;\ _p)-1}])$, *provided that the parameters are chosen so that* $k_{Agent} = (1-\ \ )k_{Env}$.

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 . (b-1) Di erent forecaster f (ARIMA, SA). (b-2) The Mean squared Error (MSE) model loss of four ProST-G with di erent forecasters(ARIMA and three SA) and the MBPO. x-axis are all episodes.