Timing as an Action: Learning When to Observe and Act

Helen Zhou [1] Audrey Huang [2] Kamyar Azizzadenesheli [3] David Childers [1] Zachary C. Lipton [1]

[1]CMU [2]UIUC [3]NVIDIA

Motivation

In many real-world settings, observations are expensive, forcing agents to commit to courses of action for designated periods of time.

Consider the setting where a patient with a chronic illness visits a doctor, who prescribes them a medication and schedules follow-up appointments. Importantly,

1. The doctor must choose not only which treatment (**action**) to recommend but also how long (**delay**) to recommend it for.
2. The doctor doesn't observe the patient's intermediate state or benefit of medication until the next appointment (**no observations** of state or reward until **after** the delay).
3. There is some **cost** to each appointment (observation and action cost).

Timing-as-an-action: Setting

We introduce the *timing-as-an-action* Markov decision process:

- Infinite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{K}, P, R, \gamma, C, s_0)$, with *delay space* $\mathcal{K} = \{1, 2, ..., K\}$.
- Agent maximizes expected $\gamma \in [0, 1)$-discounted sum of per-period primitive rewards $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ starting from s_0.
- Per interaction, agents (1) incur a fixed known cost $C \in \mathbb{R}_{\geq 0}$, (2) observe the state s and the *aggregate* reward since last observation $g = -C + \sum_{j=0}^{k-1} \gamma^j r_j$, and (3) choose actions and delays according to $\pi : \mathcal{S} \to \Delta(\mathcal{A} \times \mathcal{K})$.
- State evolves with Markov transition matrix $\mathbf{P}_a$, with entries $P(s'|s, a)$ for k periods.

Reformulation as an MDP

By collapsing the partially observable problem to observation periods, obtain a fully observed variable-duration MDP with actions (a, k), transitions $P(s'|s, a, k)$ defined by $\mathbf{P}_{a,k} = \mathbf{P}_a^k$ for all $(a, k) \in \mathcal{A} \times \mathcal{K}$, and rewards $g \sim G(s, a, k)$ the distribution over aggregated rewards induced by R and P with expected value

$$\mathbb{E}[G(s, a, k)] = -C + \sum_{j=0}^{k-1} \gamma^j \mathbb{E}[R(s_j, a)|s, a].$$

When a policy π interacts continuously with M, it observes a trajectory $(s_0, a_0, g_0, s_1, a_1, g_1, \ldots)$, and its state-action value function of policy π is:

$$Q^\pi(s, a, k) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^{\left(\sum_{\tau'=0}^{\tau-1} k_{\tau'}\right)} g_\tau | \pi, s_0 = s, a_0 = a, k_0 = k\right]$$

The optimal policy can be shown to satisfy a modified Bellman equation

$$Q(s, a, k) = \mathbb{E}[G(s, a, k)] + \gamma^k \mathbb{E}_{s' \sim P(\cdot|s,a,k)}[\max_{a',k'} Q(s', a', k')] \tag{1}$$

This representation allows RL with a simple *model-free* baseline:

- Apply Q-learning in reformulated MDP with update

$$\widehat{Q}(s, a, k) \leftarrow g + \gamma^k \max_{a',k'} \widehat{Q}(s', a', k'). \tag{2}$$

Model-based methods can further exploit the special structure of the delay action.

Model-based Algorithms

For each step, given data $\{(s_i, a_i, k_i, g_i, s_i')\}_{i=1}^N$

- Estimate $\widehat{P}$ by maximum likelihood

$$\widehat{P} = \operatorname{argmax}_{p \in \mathcal{P}} \sum_{i=1}^{N} \log p_{a_i,k_i}(s_i'|s_i), \tag{3}$$

  - **Timing-naive**: $\mathcal{P} = \{P : \mathcal{S} \times \mathcal{A} \times \mathcal{K} \to \Delta(\mathcal{S})\}$ is the set of all valid transitions
  - **Timing-aware**: $\mathcal{P} = \mathcal{P}_1 = \{p : p_{a,k}(\cdot|s) = [\mathbf{p}_{a,1}]^k(\cdot|s), \forall (s, a, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{K}\}$, is the 1-step transitions iterated k times.

- Estimate $\widehat{R}$ by least squares

$$\widehat{R} = \operatorname{argmin}_{R' \in \mathcal{R}} \frac{1}{N} \sum_{i=1}^{N} (\mathcal{G}_{R',\widehat{P}}(s_i, a_i, k_i) - g_i)^2, \tag{4}$$

where $\mathcal{G}_{R',P'}(s, a, k) = -C + \sum_{\tau=0}^{k-1} \gamma^\tau \mathbb{E}_{s' \sim P'_{a,k}(\cdot|s)}[R'(s', a)]$ is structured reward.

- Update $\widehat{Q}$ by value iteration in estimated MDP

Guarantees

In the generative tabular setting with n samples from $S \times A \times K$ states, the model based approach satisfies

$$\left\|Q^* - Q^{\pi_{\widehat{Q}}}\right\|_\infty \leq \frac{2}{1-\gamma} \left\|\mathcal{G}_{R,P} - \mathcal{G}_{\widehat{R},\widehat{P}}\right\|_\infty + \frac{2\gamma}{(1-\gamma)^2} \max_{s,a,k} \left\|P(\cdot|s, a, k) - \widehat{P}(\cdot|s, a, k)\right\|_1.$$

where w.p. $\geq 1 - \delta$

$$\left\|\mathcal{G}_{\widehat{R},\widehat{P}} - \mathcal{G}_{R,P}\right\|_\infty \lesssim \frac{1}{(1-\gamma)} \left(SAK\varepsilon_{\widehat{P}}\right)^{1/2} + \left(\frac{G_{\max}^2 S^2 A^2 K}{n}\right)^{1/2} + \left(\frac{1}{(1-\gamma)^2} \frac{G_{\max}^2 S^2 A^2 K}{n} \varepsilon_{\widehat{P}}\right)^{1/4},$$

Here $\varepsilon_{\widehat{P}} = \max_{s,a,k} \|\widehat{P}_{a,k}(\cdot|s) - P_{a,k}(\cdot|s)\|_1$ is the transition estimation error.

For **Timing Naive** $\varepsilon_{\widehat{P}} \lesssim S\sqrt{\frac{AK \log(1/\delta)}{n}}$, for **Timing Aware** $\varepsilon_{\widehat{P}} \lesssim S\sqrt{\frac{A \log(K/\delta)}{n}}$

Total error for timing-aware is $\tilde{O}(SA^{3/4}K^{1/2}n^{-1/2})$ and for timing-naive is $\tilde{O}(SA^{3/4}Kn^{-1/2})$

Estimation Performance

Timing-aware estimation improves sample efficiency and allows extrapolating across delay lengths.

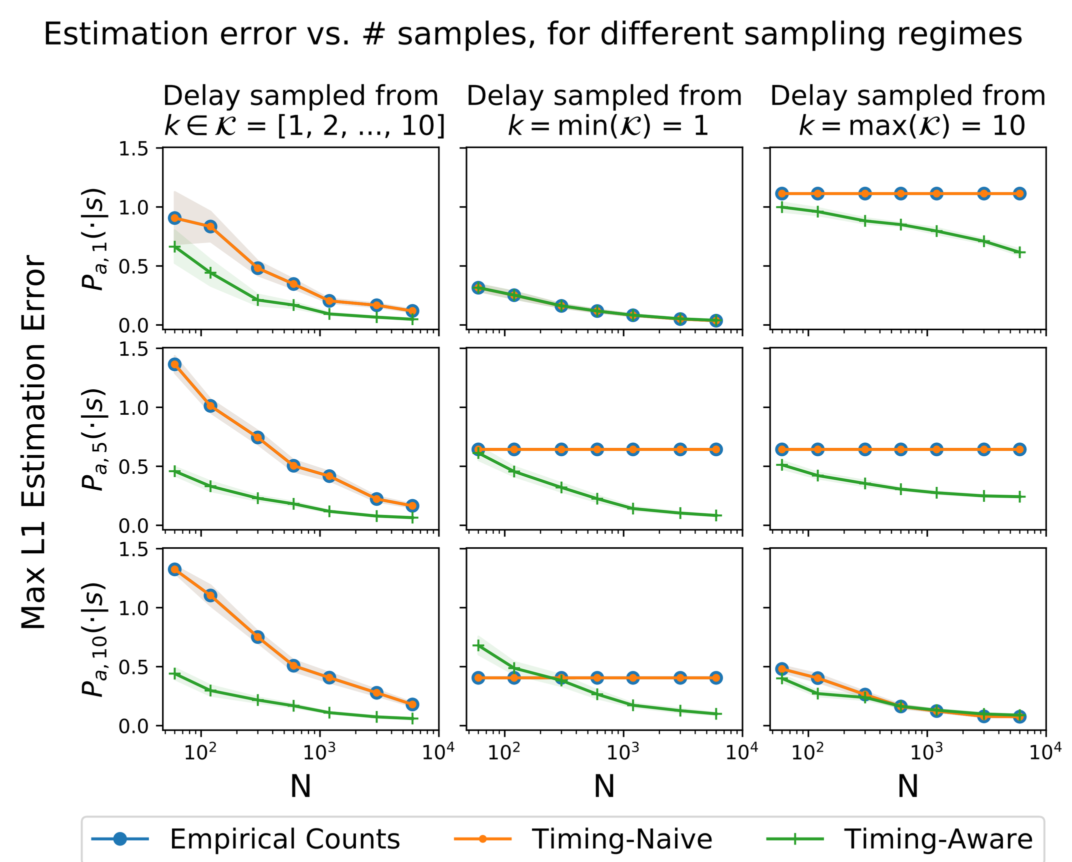

Figure 1. Estimation error $\max_{a,s} \|P_{a,k}(\cdot|s) - \widehat{P}_{a,k}(\cdot|s)\|_1$ (with 95% CI) for $\widehat{P}_{a,1}$, $\widehat{P}_{a,5}$, and $\widehat{P}_{a,10}$ vs. # of samples N generated from three sampling regimes: (a) generative setting, (b) sampling $k = \min(\mathcal{K})$, and (c) sampling $k = \max(\mathcal{K})$.

RL Experiments

We test the above algorithms in 3 standard tabular RL environments, augmented with action cost C and choice of delay k.

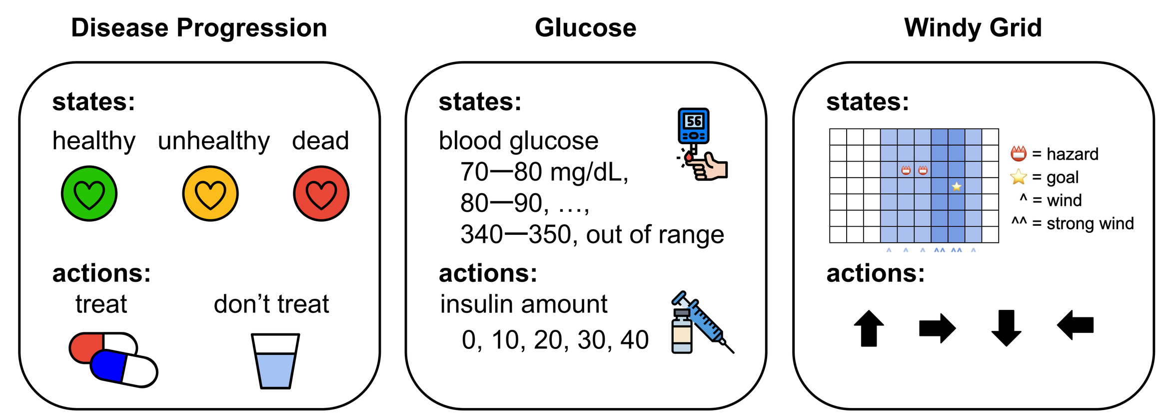

Figure 2. Summary of disease, glucose, and windy grid environments.

Evaluation: Algorithms are evaluated in the online setting by average cumulative reward.

Sample via $\epsilon-$greedy exploration: w.p. $1 - \epsilon$ execute $(a, k) := \arg\max_{a,k} \widehat{Q}(s, a, k)$

else w.p. ϵ explore: choose a uniformly over $\mathcal{A}$ and $k = 1$.

RL Results

Simulations confirm the performance gains from model-based and timing-aware methods.

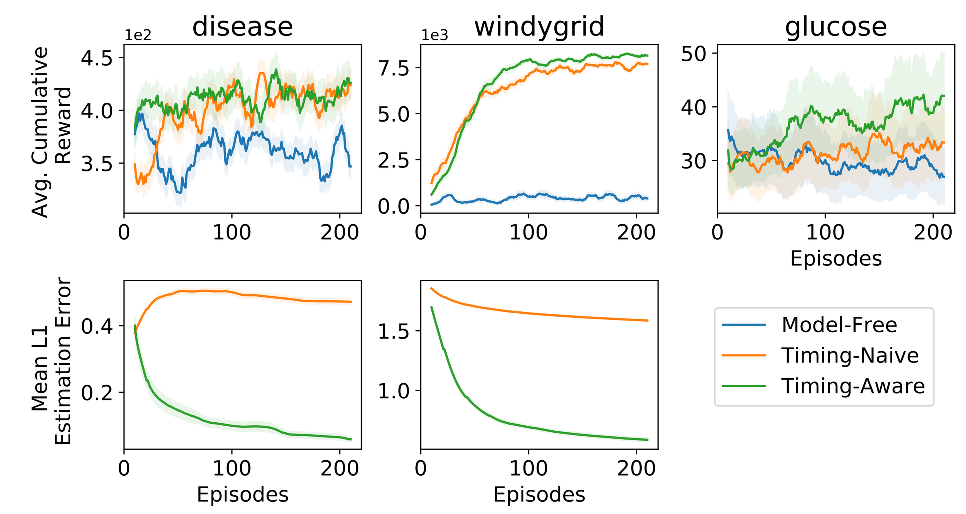

Figure 3. Average cumulative reward and mean L1 error ($\|\widehat{P}_{a,k}(\cdot|s) - P_{a,k}(\cdot|s)\|_1$ averaged over all s, a, k) across 50 trials, smoothed with a running average over 20 episodes. Shaded regions are std. err.

Table 1. Final average cumulative reward in each setting after 200 episodes (in hundreds).

	Disease Progression	Windy Grid	Glucose
Timing-Aware	**4.26 (4.00–4.53)**	**81.5 (80.3–82.6)**	**0.420 (0.287–0.554)**
Timing-Naive	4.24 (4.00–4.47)	76.9 (75.0–78.8)	0.334 (0.224–0.443)
Model-Free	3.47 (3.28–3.65)	3.96 (1.83–6.10)	0.270 (0.183–0.356)

Discussion

- Timing-as-an-action poses interesting theoretical and practical challenges for bringing RL into real-world settings with costly observations and actions
- Timing-aware model-based method leverages the structure of timing-as-an-action to obtain sample complexity advantages over model-free and timing-naive model-based.
- Estimation using the timing-aware model-based approach is more sample-efficient than timing-naive, which can translate into improvements in the RL setting.