

---

# When to Sense and Control?

## A Time-adaptive Approach for Continuous-Time RL

---

Lenart Treven\*, Bhavya Sukhija, Yarden As, Florian Dörfler, Andreas Krause  
ETH Zurich, Switzerland

### Abstract

Reinforcement learning (RL) excels in optimizing policies for discrete-time Markov decision processes (MDP). However, various systems are inherently continuous in time, making discrete-time MDPs an inexact modeling choice. In many applications, such as greenhouse control or medical treatments, each interaction (measurement or switching of action) involves manual intervention and thus is inherently costly. Therefore, we generally prefer a time-adaptive approach with fewer interactions with the system. In this work, we formalize an RL framework, *Time-adaptive Control & Sensing (TACOS)*, that tackles this challenge by optimizing over policies that besides control predict the duration of its application. Our formulation results in an extended MDP that any standard RL algorithm can solve. We demonstrate that state-of-the-art RL algorithms trained on TACOS drastically reduce the interaction amount over their discrete-time counterpart while retaining the same or improved performance, and exhibiting robustness over discretization frequency. Finally, we propose OTACOS, an efficient model-based algorithm for our setting. We show that OTACOS enjoys sublinear regret for systems with sufficiently smooth dynamics and empirically results in further sample-efficiency gains.

## 1 Introduction

Nearly all state-of-the-art RL algorithms (Schulman et al., 2017; Haarnoja et al., 2018; Lillicrap et al., 2015; Schulman et al., 2015) were developed for discrete-time MDPs. Nevertheless, continuous-time systems are ubiquitous in nature, ranging from robotics, biology, medicine, environment and sustainability etc. (cf. Spong et al., 2006; Jones et al., 2009; Lenhart and Workman, 2007; Panetta and Fister, 2003; Turchetta et al., 2022). Such systems can be naturally modeled with stochastic differential equations (SDEs), but computational approaches necessitate discretization. Furthermore, in many applications, obtaining measurements and switching actions is expensive. For instance, consider a greenhouse of fruits or medical treatment recommendations. In both cases, each measurement (crop inspection, medical exam) or switching of actions (climate control, treatment adjustment) typically involves costly human intervention. Hence, minimizing such interactions with the underlying system is desirable. This underlying challenge is rarely addressed in the RL literature.

In practice, a time-equidistant discretization frequency is set, often manually, adjusted to the underlying system’s characteristic time scale. This is challenging, however, especially for unknown/uncertain systems, and systems with multiple dominant time scales (Engquist et al., 2007). Therefore, for many real-world applications having a global frequency of control is inadequate and wasteful. For example, in medicine, patient monitoring often requires higher frequency interaction during the onset of illness and lower frequency interactions as the patient recovers (Kaandorp and Koole, 2007).

In this work, we address this limitation of standard RL methods and propose a novel RL framework, **Time-adaptive Control & Sensing (TACOS)**. TACOS reduces a general continuous-time RL problem with underlying SDE dynamics to an equivalent discrete-time MDP, that can be solved with any

---

\*Correspondence to [lenart.treven@inf.ethz.ch](mailto:lenart.treven@inf.ethz.ch)

RL algorithm, including standard policy gradient methods like PPO and SAC (Schulman et al., 2017; Haarnoja et al., 2018). We summarize our contributions below.

### Contributions

1. We reformulate the problem of time-adaptive continuous time RL to an equivalent discrete-time MDP that can be solved with standard RL algorithms.
2. Using our formulation, we extend standard policy gradient techniques (Haarnoja et al. (2018) and Schulman et al. (2017)) to the time-adaptive setting. Our empirical results on standard RL benchmarks (Freeman et al., 2021) show that TACOS outperforms its discrete-time counterpart in terms of policy performance, computational cost, and sample efficiency.
3. To further improve sample efficiency, we propose a model-based RL algorithm, OTACOS. OTACOS uses well-calibrated probabilistic models to capture epistemic uncertainty and, similar to Curi et al. (2020) and Treven et al. (2023), leverages the principle of optimism in the face of uncertainty to guide exploration during learning. We theoretically prove that OTACOS suffers no regret and empirically demonstrate its sample efficiency.

## 2 Problem statement

We consider a general nonlinear continuous time dynamical system with continuous state  $\mathcal{X} \subset \mathbb{R}^{d_x}$  and action  $\mathcal{U} \subset \mathbb{R}^{d_u}$  space. The underlying dynamics are governed by a (controllable) SDE:

$$d\mathbf{x}_t = \mathbf{f}^*(\mathbf{x}_t, \mathbf{u}_t)dt + \mathbf{g}^*(\mathbf{x}_t, \mathbf{u}_t)d\mathbf{B}_t. \quad (1)$$

Here  $\mathbf{x}_t \in \mathcal{X}$  is the state at time  $t$ ,  $\mathbf{u}_t \in \mathcal{U}$  the control input,  $\mathbf{f}^*$ ,  $\mathbf{g}^*$  are unknown drift and diffusion functions and  $\mathbf{B}_t$  is a standard Brownian motion in  $\mathbb{R}^{d_B}$ . Our goal is to find a control policy  $\pi_{\mathcal{U}} : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{U}$  which maximizes an unknown reward  $b^*(\mathbf{x}_t, \mathbf{u}_t)$  over a fixed horizon  $\mathcal{T} \stackrel{\text{def}}{=} [0, T]$ , i.e.,

$$\max_{\pi \in \Pi} \mathbb{E} \left[ \int_{t \in \mathcal{T}} b^*(\mathbf{x}_t, \pi_{\mathcal{U}}(\mathbf{x}_t, t))dt \right],$$

where the expectation is taken w.r.t. the policy and stochastic dynamics and  $\Pi$  is the class of policies<sup>2</sup> over which we search.

In practice, we can only measure the system state and execute control policies in discrete points in time. In this work, we focus on problems where state measurement and control are synchronized in time. We refer to these synchronized time points as *interactions* in the following parts of this paper. Synchronizing state measurement and control contrasts standard time-adaptive approaches such as event-triggered control (Heemels et al., 2021), where the state is measured arbitrarily high frequency and control inputs are changed only so often to ensure stability. It is also in contrast to the complementary setting, where control inputs are changing at an arbitrarily high frequency but measurements are collected adaptively in time (Treven et al., 2023). An adaptive control approach as Heemels et al. (2021) is very important for many real-world applications but similarly, an adaptive measurement strategy is crucial for efficient learning in RL (Treven et al., 2023). Our approach treats both of these requirements jointly.

We consider two different scenarios for continuous-time control: (i) Penalizing interactions with some cost, (ii) bounded number of interactions, i.e., hard constraint on control/measurement steps.

**Interaction cost** We consider the setting where every interaction we take has an inherent cost  $c(\mathbf{x}_t, \mathbf{u}_t) > 0$ . Note that we consider this cost structure for its simplicity and TACOS works for more general cost functions that depend on the duration of application for the action  $\mathbf{u}_t$  or the previous action  $\mathbf{u}_{t-1}$  and thus captures many practical real-world settings. We define this task more formally below

$$\max_{\pi \in \Pi, \pi_{\mathcal{T}}} \mathbb{E} \left[ \sum_{i=0}^{K-1} \int_{t_{i-1}}^{t_i} b^*(\mathbf{x}_t, \pi_{\mathcal{U}}(\mathbf{x}_{t_{i-1}}, t_{i-1}))dt - c(\mathbf{x}_{t_{i-1}}, \pi_{\mathcal{U}}(\mathbf{x}_{t_{i-1}}, t_{i-1})) \right], \quad (2)$$

$$t_i = \pi_{\mathcal{T}}(\mathbf{x}_{t_{i-1}}, t_{i-1}) + t_{i-1}, t_0 = 0, t_K = T, \forall (\mathbf{x}, t) \in \mathcal{X} \times \mathcal{T}; \pi_{\mathcal{T}}(\mathbf{x}, t) \in [t_{\min}, t_{\max}].$$

Here  $t_{\min} > 0$  is the minimal duration for which we have to apply the control,  $t_{\max} \in [t_{\min}, T]$  the maximum duration, and  $\pi_{\mathcal{T}}$  is a policy that predicts the duration of applying the action.

<sup>2</sup>We assume that  $\Pi$  is the set of  $L_{\pi}$ -Lipschitz policies

**Bounded number of interactions** In this setting, the number of interactions with the system is limited by a known amount  $K$ . Intuitively, this represents a scenario where we have a finite budget for the inputs that we can apply and have to decide on the best strategy to space these  $K$  inputs over the full horizon. A formal definition of this task is given below

$$\max_{\pi \in \Pi, \pi_{\mathcal{T}}} \mathbb{E} \left[ \sum_{i=0}^{K-1} \int_{t_{i-1}}^{t_i} b^*(\mathbf{x}_t, \pi_{\mathcal{U}}(\mathbf{x}_{t_{i-1}}, t_{i-1}, i-1)) dt \right], \quad (3)$$

$$t_i = \pi_{\mathcal{T}}(\mathbf{x}_{t_{i-1}}, t_{i-1}, i-1) + t_{i-1}, \quad t_0 = 0, t_K = T, \forall (\mathbf{x}, t, i) : \pi_{\mathcal{T}}(\mathbf{x}, t, i) \in [t_{\min}, t_{\max}].$$

In the absence of the transition costs or the bound on the number of interactions, intuitively the policy would propose to interact with the system as frequently as possible, i.e., every  $t_{\min}$  seconds. The additional costs/constraints ensure that we do not converge to this trivial (but unrealistic) solution.

### 3 TACOS: Time Adaptive Control or Sensing

In the following, we reformulate the continuous-time problem as an equivalent discrete-time MDP. We first denote the state and running reward flows of Equation (1). The state flow by applying action  $\mathbf{u}_k$  for  $t_k$  time reads:

$$\mathbf{x}_{k+1} = \Xi(\mathbf{x}_k, \mathbf{u}_k, t_k),$$

$$\Xi(\mathbf{x}, \mathbf{u}, t) \stackrel{\text{def}}{=} \mathbf{x} + \int_0^t \mathbf{f}^*(\mathbf{x}_s, \mathbf{u}) ds + \int_0^t \mathbf{g}^*(\mathbf{x}_s, \mathbf{u}) d\mathbf{B}_s.$$

We assume that every time we interact with the system, we also obtain the integrated reward and define the reward flow as

$$\Xi_{b^*}(\mathbf{x}, \mathbf{u}, t) \stackrel{\text{def}}{=} \int_0^t b^*(\Xi(\mathbf{x}, \mathbf{u}, s), \mathbf{u}) ds. \quad (4)$$

Due to the stochasticity of  $(\mathbf{B}_t)_{t \in \mathcal{T}}$ , the state flow  $\Xi(\mathbf{x}, \mathbf{u}, t)$  and the reward flow  $\Xi_{b^*}(\mathbf{x}, \mathbf{u}, t)$  are stochastic. For ease of notation, we denote

$$\Phi_{\mathbf{f}^*}(\mathbf{x}_k, \mathbf{u}_k, t_k) \stackrel{\text{def}}{=} \mathbb{E}[\Xi(\mathbf{x}_k, \mathbf{u}_k, t_k)], \quad \Phi_{b^*}(\mathbf{x}_k, \mathbf{u}_k, t_k) \stackrel{\text{def}}{=} \mathbb{E}[\Xi_{b^*}(\mathbf{x}_k, \mathbf{u}_k, t_k)]$$

$$\mathbf{w}_k^{\mathbf{x}} \stackrel{\text{def}}{=} \Xi(\mathbf{x}_k, \mathbf{u}_k, t_k) - \Phi_{\mathbf{f}^*}(\mathbf{x}_k, \mathbf{u}_k, t_k), \quad \mathbf{w}_k^{b^*} \stackrel{\text{def}}{=} \Xi_{b^*}(\mathbf{x}_k, \mathbf{u}_k, t_k) - \Phi_{b^*}(\mathbf{x}_k, \mathbf{u}_k, t_k),$$

and the concatenated state and reward flow function, and noise as:

$$\Phi^*(\mathbf{x}_k, \mathbf{u}_k, t_k) = \begin{pmatrix} \Phi_{\mathbf{f}^*}(\mathbf{x}_k, \mathbf{u}_k, t_k) \\ \Phi_{b^*}(\mathbf{x}_k, \mathbf{u}_k, t_k) \end{pmatrix}, \quad \mathbf{w}_k = \begin{pmatrix} \mathbf{w}_k^{\mathbf{x}} \\ \mathbf{w}_k^{b^*} \end{pmatrix}. \quad (5)$$

In this work, we search for policies that return the next control we apply and also the time for how long to apply the control.

#### 3.1 Reformulation of Interaction Cost setting to Discrete-time MDPs

We convert the problem with interaction costs to a standard MDP which any RL algorithm for continuous state-action spaces can solve. To this end, we restrict ourselves to a policy class:

$$\Pi_{IC} = \{ \pi : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{U} \times \mathcal{T} \mid \pi_{\mathcal{T}}(\cdot, t) \in [t_{\min}, t_{\max}], \pi \text{ is } L_{\pi} - \text{Lipschitz} \}.$$

For simplicity, we denote by  $\pi_{\mathcal{T}}$  the component of the policy that predicts the duration of applying the action and with  $\pi_{\mathcal{U}}$  the component that predicts the action value. The policies we consider map state  $\mathbf{x}$  and time-to-go  $t$  to control  $\mathbf{u}$  and the time  $\tau$  for how long we apply the action  $\mathbf{u}$ . We define the augmented state  $\mathbf{s} = (\mathbf{x}, b, t)$ , where  $\mathbf{x}$  is the state,  $b$  integrated reward and  $t$  time-to-go. With the introduced notation we arrive at a discrete-time MDP problem formulation

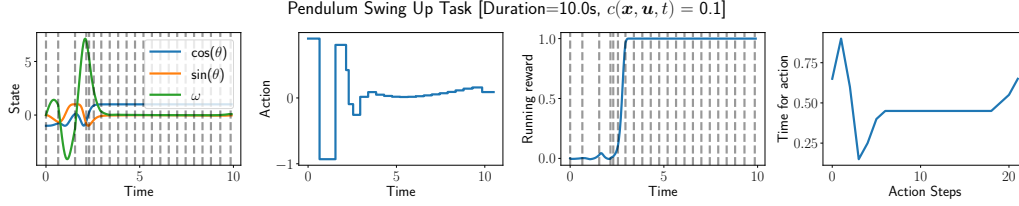
$$\max_{\pi \in \Pi_{IC}} V_{\pi, \Phi^*}(\mathbf{x}_0, T) = \max_{\pi \in \Pi_{IC}} \mathbb{E} \left[ \sum_{k=0}^{K-1} r(\mathbf{s}_k, \pi(\mathbf{s}_k)) \right] \quad (6)$$

$$\text{s.t. } \mathbf{s}_{k+1} = \Psi_{\Phi^*}(\mathbf{s}_k, \pi(\mathbf{s}_k), \mathbf{w}_k), \quad \mathbf{s}_0 = (\mathbf{x}_0, 0, T), \quad \sum_{k=0}^{K-1} \pi_{\mathcal{T}}(\mathbf{x}_k, t_k) = T,$$

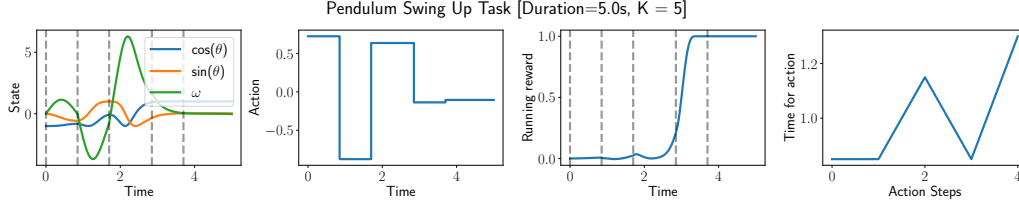
where we have

$$\Psi_{\Phi^*}(\mathbf{s}_k, \pi(\mathbf{s}_k), \mathbf{w}_k) = (\Phi^*(\mathbf{x}_k, \pi(\mathbf{x}_k, t_k)) + \mathbf{w}_k, t_k - \pi_{\mathcal{T}}(\mathbf{x}_k, t_k))$$

$$r(\mathbf{s}_k, \pi(\mathbf{s}_k)) = \Xi_{b^*}(\mathbf{x}_k, \pi(\mathbf{x}_k, t_k)) - c(\mathbf{x}_k, \pi_{\mathcal{U}}(\mathbf{x}_k, t_k)).$$



(a) We add a constant switch cost of 0.1 and significantly reduce the number of interactions from 200 to 24. Initially, the policy applies maximal bang-bang torque for longer times, until the pendulum reaches the top. On the top, we measure and change the controller at a higher frequency in order to keep the pendulum stable, at the position with the highest reward.



(b) We set a tight bound of  $K = 5$  for the number of interactions and observe that we can still solve the task.

Figure 1: Experiment on the Pendulum environment for the average cost and a bounded number of switches setting.

### 3.2 Reformulation of Bounded Number of Interactions to Discrete-time MDPs

The second setting is similar to the one studied by Ni and Jang (2022). In this case, we consider the following class of policies:

$$\Pi_{BI} = \{ \pi : \mathcal{X} \times \mathcal{T} \times \mathbb{N} \rightarrow \mathcal{U} \times \mathcal{T} \mid \forall k \in [K] : \pi(\cdot, \cdot, k) \text{ is } L_\pi\text{-Lipschitz} \}.$$

For an augmented state  $\mathbf{s} = (\mathbf{x}, b, t, k)$ , our policies map states  $\mathbf{x}$ , time-to-go  $t$ , number of past interactions  $k$  to a controller  $\mathbf{u}$  and the time duration  $\tau$  for applying the action. Here the optimal control problem reads

$$\begin{aligned} \max_{\pi \in \Pi_{BI}} V_{\pi, \Phi^*}(\mathbf{x}_0, T) &= \max_{\pi \in \Pi_{BI}} \mathbb{E} \left[ \sum_{k=0}^{K-1} r(\mathbf{s}_k, \pi(\mathbf{s}_k)) \right] \\ \text{s.t. } \mathbf{s}_{k+1} &= \Psi_{\Phi^*}(\mathbf{s}_k, \pi(\mathbf{s}_k), \mathbf{w}_k), \quad \mathbf{s}_0 = (\mathbf{x}_0, 0, T, 0), \end{aligned} \quad (7)$$

where,

$$\begin{aligned} \Psi_{\Phi^*}(\mathbf{s}_k, \pi(\mathbf{s}_k), \mathbf{w}_k) &= (\Phi^*(\mathbf{x}_k, \pi(\mathbf{x}_k, t_k, k)) + \mathbf{w}_k, t_k - \pi_{\mathcal{T}}(\mathbf{x}_k, t_k, k), k + 1) \\ r(\mathbf{s}_k, \pi(\mathbf{s}_k)) &= \Xi_{b^*}(\mathbf{x}_k, \pi(\mathbf{x}_k, t_k, k)). \end{aligned}$$

In the following, we provide a simple proposition which shows that our reformulated problem is equivalent to its continuous-time counterpart from Section 2.

**Proposition 1.** *The problem in Equation (2) and 3 are equivalent to Equation (6) and 7, respectively.*

Figure 1 depicts the influence of interaction cost and  $K$  on the controller’s performance for the pendulum environment.

## 4 TACoS with Model-free RL Algorithms

We now illustrate the performance of TACoS on several well-studied robotic RL tasks. We consider the RC car (Kabzan et al., 2020), Greenhouse (Tap, 2000), Pendulum, Reacher, Halfcheetah and Humanoid environments from Brax (Freeman et al., 2021). Thus our experiments range from environments necessitating time-adaptive control like the Greenhouse, a realistic and highly dynamic race car simulation, and a very high dimensional RL task like the Humanoid.<sup>3</sup>

<sup>3</sup> $\mathcal{X} \subset \mathbb{R}^{244}, \mathcal{U} \subset \mathbb{R}^{17}$ . We provide our implementation at <https://github.com/lasgroup/TaCoS>.

We investigate both the bounded number of interactions and interaction cost settings in our experiments. In particular, we study how the bound  $K$  affects the performance of TACOS and compare it to the standard equidistant baseline. We further study the interplay between the stochasticity of the environments (magnitude of  $g^*$ ) and interaction costs and the influence of  $t_{\min}$  on TACOS. For all experiments in this section, we combine SAC with TACOS (SAC-TACOS).

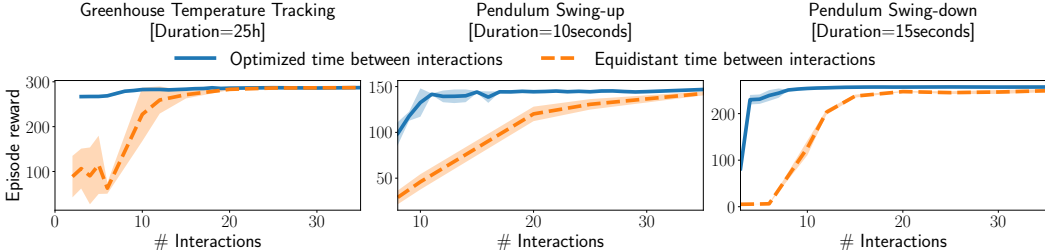


Figure 2: We study the effects of the bound on interactions  $K$  on the performance of the agent. TACOS performs significantly better than equidistant discretization, especially for small values of  $K$ .

**How does the bound on the number of interactions  $K$  affect TACOS?** We analyze the bounded number of interactions setting (cf. Section 3.2) of TACOS, studying the relationship between the number of interactions and the achieved episode reward. We compare our algorithm with the standard equidistant time discretization approach which splits the whole horizon  $T$  into  $T/K$  discrete time steps at which an interaction takes place. We evaluate the two methods in the greenhouse and pendulum environments. For the pendulum, we consider the swing-up and swing-down tasks. The results are reported in Figure 2. The time-adaptive approach performs significantly better than the standard equidistant time discretization. This is particularly the case for the greenhouse and pendulum swing-down tasks. Both tasks involve driving the system to a stable equilibrium and thus, while high-frequency interaction might be necessary at the initial stages, a fairly low interaction frequency can be maintained when the system has reached the equilibrium state. This demonstrates the practical benefits of time-adaptive control.

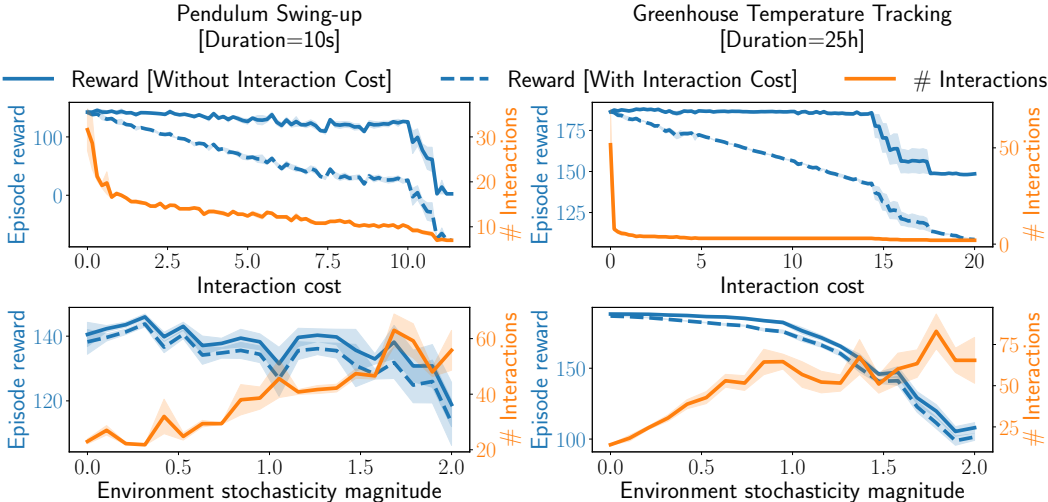


Figure 3: Effect of interaction cost (first row) and environment stochasticity (second row) on the number of interactions and episode reward for the Pendulum and Greenhouse tasks.

**How does the interaction cost magnitude influence TACOS?** We investigate the setting from Section 3.1 with interaction costs. In our experiments, we always pick a constant cost, i.e.,  $c(x, u) = C$ . We study the influence of  $C$  on the episode reward and on the number of interactions that the policy has with the system within an episode. We again evaluate this on the greenhouse and pendulum environment. For the pendulum, we consider the swing-up task. The results are presented in the first row of Figure 3. Noticeably, increasing  $C$  reduces the number of interactions. The decrease is

drastic for the greenhouse environment since it can be controlled with considerably fewer interactions without having any effect on the performance. Generally, we observe that decreasing the number of interactions, that is, increasing  $C$ , also results in a slight decline in episode reward.

**How does environment stochasticity influence the number of interactions?** We analyze the influence of the environment’s stochasticity, i.e., the magnitude of the diffusion term  $g^*$ , on the episode reward and number of interactions on TACOS. Intuitively, the more stochastic the environment, the more interactions we would require to stabilize the system. We again evaluate our method on the greenhouse and pendulum swing-up tasks. The results are reported in the second row of Figure 3. The results verify our intuition that more stochasticity in the environment generally leads to more interactions. However, we observe that the policy is still able to achieve high rewards for a wide range of magnitude of  $g^*$ . This showcases the robustness and adaptability of TACOS to stochastic environments.

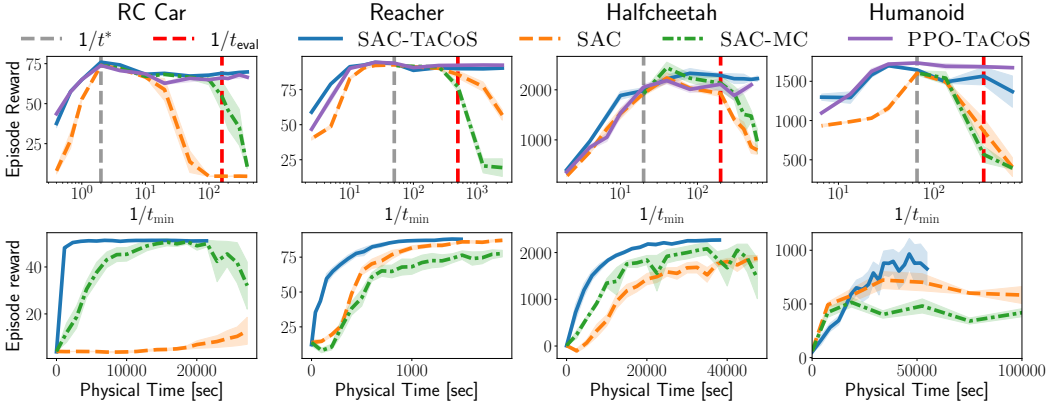


Figure 4: We compare the performance of TACOS in combination with SAC and PPO with the standard SAC algorithm and SAC with more compute (SAC-MC) over a range of values for  $t_{\min}$  (first row). In the second row, we plot the episode reward versus the physical time in seconds spent in the environment for SAC-TACoS, SAC, and SAC-MC for a specific evaluation frequency  $1/t_{\text{eval}}$ . We exclude PPO-TACoS in this plot as it, being on-policy, requires significantly more samples than the off-policy methods. While all methods perform equally well for standard discretization (denoted with  $1/t^*$ ), our method is robust to interaction frequency and does not suffer a performance drop when we decrease  $t_{\min}$ .

**How does  $t_{\min}$  influence TACOS?** As highlighted in Section 1, picking the right discretization for interactions is a challenging task. We show that TACOS can naturally alleviate this issue and adaptively pick the frequency of interaction while also being more computationally and data-efficient. Moreover, we show that TACOS is robust to the choice of  $t_{\min}$ , which represents the minimal duration an action has to be applied, i.e., its inverse is the highest frequency at which we can control the system. In this experiment, we consider SAC-TACoS and compare it to the standard SAC algorithm. TACOS adaptively picks the number of interactions and therefore during an episode of time  $T$ , it effectively collects less data than the standard discrete-time RL algorithm.<sup>4</sup> This makes comparison to the discrete-time setting challenging since environment interactions and physical time on the environment are not linearly related for TACOS as opposed to the standard discrete-time setting. Nevertheless, to be fair to the discrete-time method, we give SAC more physical time on the system for all environments, effectively resulting in the collection of more data for learning. Since the standard SAC algorithm updates the policy relative to the data amount, we consider a version of SAC, SAC-MC (SAC more compute), which leverages the additional data it collects to perform more gradient updates. This version essentially performs more policy updates than SAC-TACoS and thus is computationally more expensive. Furthermore, to demonstrate the generality of our framework, we also combine TACOS with PPO (PPO-TACoS).

We report the performance after convergence across different  $t_{\min}$  in the first row of Figure 4. From our experiment, we conclude that SAC-TACoS and PPO-TACoS are robust to the choice of  $t_{\min}$

<sup>4</sup>A standard RL algorithm would collect  $T/t_{\min}$  data points per episode.

and perform equally well when  $t_{\min}$  is decreased, i.e., frequency is increased. This is in contrast to the standard RL methods, which have a significant drop in performance at high frequencies. This observation is also made in prior work (Hafner et al., 2019). Crucially, this highlights the sensitivity of the standard RL methods to the frequency of interaction. In the second row of Figure 4 we show the learning curve of the methods for a specific frequency  $1/t_{\text{eval}}$ . From the curve, we conclude that SAC-TACOS achieves higher rewards with significantly less physical time on the environment. We believe this is because our method explores more efficiently (akin to Dabney et al., 2020; Eberhard et al., 2022), and also learns a much stronger/continuous-time representation of the underlying MDP.

Interestingly, at the default frequency used in the benchmarks  $1/t^*$ , all methods perform similarly. However, slightly decreasing the frequency already leads to a drastic drop in performance for all methods. Intuitively, decreasing the frequency prevents us from performing the necessary fine-grained control and obtaining the highest performance.

While we have access to the optimal frequency  $1/t^*$  for these benchmarks, for a general and unknown system it is very difficult to estimate this frequency. Furthermore, as we observe in our experiments, picking a very high frequency is also not an option when using standard RL algorithms. We believe this is where TACOS excels as it adaptively picks the frequency of interaction, thereby relieving the problem designer of this decision.

## 5 Efficient Exploration for TACOS via Model-Based RL

In this section, we propose a novel model-based RL algorithm for TACOS called **Optimistic TACOS** (OTACOS). We analyze the episodic setting, where we interact with the system in episodes  $n = 1, \dots, N$ . In episode  $n$ , we execute the policy  $\pi_n$ , collect measurements and integrated rewards  $(\mathbf{x}_{n,0}, b_{n,0}), \dots, (\mathbf{x}_{n,k_n}, b_{n,k_n})$ , and prepare the data  $\mathcal{D}_n = \{(\mathbf{z}_{n,1}, \mathbf{y}_{n,1}), \dots, (\mathbf{z}_{n,k_n}, \mathbf{y}_{n,k_n})\}$ , where  $\mathbf{z}_{n,i} = (\mathbf{x}_{n,i-1}, \mathbf{u}_{n,i-1}, t_{n,i-1})$  and  $\mathbf{y}_{n,i} = (\mathbf{x}_{n,i}, b_{n,i})$ . From the dataset  $\mathcal{D}_{1:n} \stackrel{\text{def}}{=} \cup_{i \leq n} \mathcal{D}_i$  we build a model  $\mathcal{M}_n$  for the unknown function  $\Phi^*$  such that it is well-calibrated in the sense of the following definition.

**Definition 1** (Well-calibrated statistical model of  $\Phi^*$ , Rothfuss et al. (2023)). *Let  $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{U} \times \mathcal{T}$ . We assume  $\Phi^* \in \bigcap_{n \geq 0} \mathcal{M}_n$  with probability at least  $1 - \delta$ , where statistical model  $\mathcal{M}_n$  is defined as*

$$\mathcal{M}_n \stackrel{\text{def}}{=} \{ \mathbf{f} : \mathcal{Z} \rightarrow \mathbb{R}^{d_x+1} \mid \forall \mathbf{z} \in \mathcal{Z}, \forall j \in \{1, \dots, d_x + 1\} : |\mu_{n,j}(\mathbf{z}) - f_j(\mathbf{z})| \leq \beta_n(\delta) \sigma_{n,j}(\mathbf{z}) \},$$

Here,  $\mu_{n,j}$  and  $\sigma_{n,j}$  denote the  $j$ -th element in the vector-valued mean and standard deviation functions  $\boldsymbol{\mu}_n$  and  $\boldsymbol{\sigma}_n$  respectively, and  $\beta_n(\delta) \in \mathbb{R}_{\geq 0}$  is a scalar function that depends on the confidence level  $\delta \in (0, 1]$  and which is monotonically increasing in  $n$ .

Similar to model-based RL algorithms for the discrete-time setting (Kakade et al., 2020; Curi et al., 2020; Sukhija et al., 2024), we follow the principle of optimism in the face of uncertainty and select the policy  $\pi_n$  for both settings of TACOS (cf. Sections 3.1 and 3.2) by solving:

$$\pi_n \stackrel{\text{def}}{=} \operatorname{argmax}_{\pi \in \Pi_{\square}} \max_{\Phi \in \mathcal{M}_{n-1}} V_{\pi, \Phi}(\mathbf{x}_0, T), \quad (8)$$

where  $\square \in \{IC, BI\}$  is the appropriate policy class from Section 3. Running OTACOS for  $N$  episodes, we measure the performance via the *regret*:

$$R_N = \sum_{n=1}^N (V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) - V_{\pi_n, \Phi^*}(\mathbf{x}_0, T)).$$

Here  $\pi^*$  is the optimal policy from the class of policies we optimize over. Any kind of regret bound requires certain assumptions on the regularity of the underlying dynamics (1).

**Assumption 1** (Dynamics model). *Given any norm  $\|\cdot\|$ , we assume that the drift  $\mathbf{f}^*$ , and diffusion  $\mathbf{g}^*$  are  $L_{\mathbf{f}^*}$  and  $L_{\mathbf{g}^*}$ -Lipschitz continuous, respectively, with respect to the induced metric. We further assume  $\sup_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{g}^*(\mathbf{z})\|_F \leq A$ .*

Assumption 1 ensures the existence of the SDE (1) solution under policy  $\pi_n$ . To provide bounds on the performance of OTACOS for settings Sections 3.1 and 3.2 we also need some assumptions on the noise and reward model.

**Assumption 2** (Reward and noise model for Section 3.1 Setting). *Given any norm  $\|\cdot\|$ , we assume that running reward  $b$  is  $L_b$ -Lipschitz continuous, with respect to the induced metric. We further assume boundedness of the reward  $0 \leq b^*(\mathbf{x}, \mathbf{u}) \leq B$ , and interaction cost  $0 \leq c(\mathbf{x}, \mathbf{u}) \leq C$ . The dynamics noise is independent and follows:  $\mathbf{w}_k^x \sim \mathcal{N}(0, \sigma^2(\mathbf{x}_k, \mathbf{u}_k, t_k)I_{d_x})$ .*

**Assumption 3** (Reward and noise model for Section 3.2 Setting). *Given any norm  $\|\cdot\|$ , we assume that the running reward  $b$  is  $L_b$ -Lipschitz continuous, w.r.t. to the induced metric.*

Finally, we assume that we learn a well-calibrated model of the unknown flow  $\Phi^*$ .

**Assumption 4** (Well calibration assumption). *Our learned model is an all-time-calibrated statistical model of  $\Phi^*$ , i.e., there exists an increasing sequence of  $(\beta_n(\delta))_{n \geq 0}$  such that our model satisfies the well-calibration condition, cf., Definition 1.*

Analogous assumptions are made for model-based RL algorithms in the discrete-time setting (Curi et al., 2020; Sukhija et al., 2024). This calibration assumption is satisfied if  $\Phi^*$  can be represented with Gaussian Process (GP) (Williams and Rasmussen, 2006; Kirschner and Krause, 2018) models.

**Theorem 2.** *Consider the setting from Section 3.1 and let Assumption 1, 2, and Assumption 4 hold. Then we have with probability at least  $1 - \delta$ :*

$$R_N \leq \mathcal{O}\left(\beta_{N-1} T^{3/2} \sqrt{N \mathcal{I}_N}\right)$$

Now consider, the setting with a bounded number of switches  $K$ , and let Assumption 1, 3, and Assumption 4 hold. Then, we get with probability at least  $1 - \delta$

$$R_N \leq \mathcal{O}\left(\beta_{N-1}^K K e^{D(L_{f^*} + L_{g^*}^2)(1+L_\pi)TK} \sqrt{N \mathcal{I}_N}\right),$$

where  $D$  is a constant. Here, with  $\mathcal{I}_N$  we denote the model-complexity after observing  $N$  points (Curi et al., 2020), which quantifies the difficulty of learning  $\Phi^*$ . For GPs, it behaves similar to the maximum information gain  $\gamma_N$  (Srinivas et al., 2009), i.e., implying sublinear regret for several common kernels (Vakili et al., 2021).

As a proof of concept, we evaluate OTACoS on the pendulum and RC car environment for the interaction cost setting.<sup>5</sup> As baselines, we adapt common model-based RL methods such as PETS (Chua et al., 2018) and planning with the mean to TACoS. We call them PETS-TACoS and MEAN-TACoS, respectively. The result is reported in Figure 5. From the figure, we conclude that OTACoS is more sample efficient than other model-based baselines and SAC-TACoS (SAC-TACoS requires circa 6000 episodes for the pendulum and 2000 for the RC car).

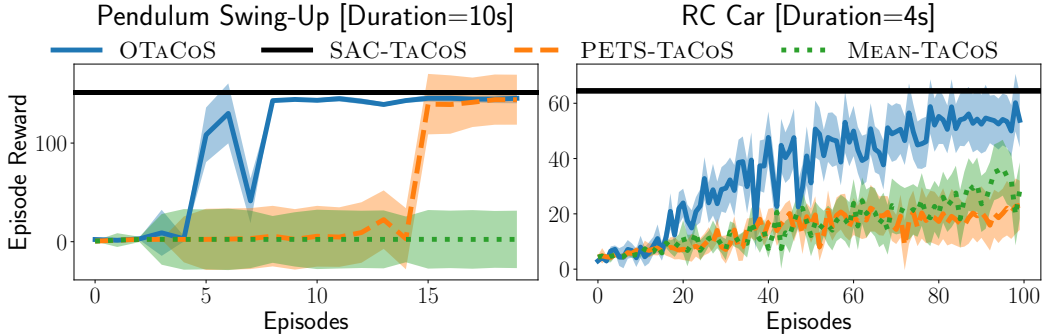


Figure 5: We run OTACoS on the pendulum and RC car environment. We report the achieved reward averaged over five different seeds with one standard error.

## 6 Related Work

Similar to this work, Holt et al. (2023); Ni and Jang (2022); Karimi (2023) consider continuous-time deterministic dynamical systems where the measurements or control input changes can only

<sup>5</sup>The code is available at <https://github.com/lasgroup/model-based-rl>.



happen at discrete time steps. Moreover, Holt et al. (2023) proposes a similar problem as ours from Section 3.1, where they specify a cost on the number of interactions. However, their solution is based on a heuristic, where a measurement is taken when the variance of the potential reward surpasses a prespecified threshold. On the contrary, we directly tackle this problem at hand and propose a general framework for time-adaptive control that does not rely on any heuristics. Karimi (2023) adapt SAC (Haarnoja et al., 2018) to include a regularization term, which effectively adds a cost for every discrete interaction. Ni and Jang (2022) induce a soft-constraint on the duration  $\tau$  of each action in the environment. However, all the aforementioned works propose heuristic techniques to minimize interactions, whereas we formalize the problem systematically for the more general case of SDEs and show that it has an underlying MDP structure that any RL algorithm can leverage. In addition, we propose a no-regret model-based RL algorithm for this setting and analyze its sample complexity.

Temporal abstractions are considered also in the framework of options (Sutton et al., 1999; Mankowitz et al., 2014; Mann and Mannor, 2014; Harb et al., 2018). However, a key difference to TACOS is that in the options framework, the agent measures the state even between the controller switches.

**Learning to repeat actions** Several works observe that repeating actions in the discrete-time MDPs problems such as Atari (Mnih et al., 2013; Braylan et al., 2015) or Cartpole (Hafner et al., 2019) significantly increase the speed of learning. However, the action repeat is fixed through the entire rollout and treated as a hyperparameter. Durugkar et al. (2016); Vezhnevets et al. (2016); Srinivas et al. (2017); Sharma et al. (2017); Lee et al. (2020); Grigsby et al. (2021); Chen et al. (2021); Nam et al. (2021); Yu et al. (2021); Biedenkapp et al. (2021); Krale et al. (2023) automate the selection of action repeat, and show superior performance over the fixed number setting. Dabney et al. (2020) empirically show that repeating the actions helps with the exploration, effectively having a similar effect that colored noise exploration has over the standard white noise exploration (Eberhard et al., 2022).

**Continuous-time RL** Following the seminal work of Doya (2000) and the advances in Neural ODEs of Chen et al. (2018), continuous-time RL has regained interest (Cranmer et al., 2020; Greydanus et al., 2019; Yildiz et al., 2021; Lutter et al., 2021). Moreover, modeling in continuous-time is found to be particularly useful when learning from different data sources where each source is collected at a different frequency (Burns et al., 2023; Zheng et al., 2023). An important line of work exists for modeling continuous dynamics for the case when states and actions are discrete, called Markov Jump Processes (Kallianpur and Sundar, 2014; Berger, 1993; Huang et al., 2019; Seifner and Sanchez, 2023). Another line of work that is close to ours is event and self-Triggered Control (Astrom and Bernhardsson, 2002; Anta and Tabuada, 2010; Heemels et al., 2012, 2021), where they model continuous-time control systems by implementing changes to the input only when stability is at risk, ensuring efficient and timely interventions. Treven et al. (2023) propose a no-regret continuous-time model-based RL algorithm, which akin to OTACOS, performs optimistic exploration. They study the problem where controls can be executed continuously in time and propose adaptive measurement selection strategies. Similarly, we propose a novel model-based RL algorithm, OTACOS, based on the principle of optimism in the face of uncertainty. We show that OTACOS has no regret for sufficiently smooth dynamics and has considerable sample-efficiency gains over its model-free counterpart.

## 7 Conclusion and discussion

We study the problem of time-adaptive RL for continuous-time systems with continuous state and action spaces. We investigate two practical settings where each interaction has an inherent cost and where we have a hard constraint on the number of interactions. We propose a novel RL framework, TACOS, and show that both of these settings result in extended MDPs which can be solved with standard RL algorithms. In our experiments, we show that combining standard RL algorithms with TACOS results in a significant reduction in the number of interactions without having any effect on the performance for the interaction cost setting. Furthermore, for the second setting, TACOS achieves considerably better control performance despite having a small budget for the number of interactions. Moreover, we show that TACOS improves robustness to a large range of interaction frequencies, and generally improves sample complexity of learning. Finally, we propose, OTACOS, a no-regret model-based RL algorithm for TACOS and show that it has further sample efficiency gains.

## Acknowledgments and Disclosure of Funding

This project has received funding from the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40 180545, the Microsoft Swiss Joint Research Center, grant of the Hasler foundation (grant no. 21039) and the SNSF Postdoc Mobility Fellowship 211086.

## References

- Anta, A. and Tabuada, P. (2010). To sample or not to sample: Self-triggered control for nonlinear systems. *IEEE Transactions on automatic control*, 55(9):2030–2042.
- Astrom, K. J. and Bernhardsson, B. M. (2002). Comparison of riemann and lebesgue sampling for first order stochastic systems. In *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, volume 2, pages 2011–2016. IEEE.
- Berger, M. A. (1993). *Markov Jump Processes*, pages 121–138. Springer New York, New York, NY.
- Biedenkapp, A., Rajan, R., Hutter, F., and Lindauer, M. (2021). Temporal: Learning when to act. In *International Conference on Machine Learning*, pages 914–924. PMLR.
- Bobkov, S. G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28.
- Braylan, A., Hollenbeck, M., Meyerson, E., and Miikkulainen, R. (2015). Frame skip is a powerful parameter for learning to play atari. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.
- Burns, K., Yu, T., Finn, C., and Hausman, K. (2023). Offline reinforcement learning at multiple frequencies. In *Conference on Robot Learning*, pages 2041–2051. PMLR.
- Chen, C., Tang, H., Hao, J., Liu, W., and Meng, Z. (2021). Addressing action oscillations through learning policy inertia. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7020–7027.
- Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.
- Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *NeurIPS*.
- Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. (2020). Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*.
- Curi, S., Berkenkamp, F., and Krause, A. (2020). Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170.
- Dabney, W., Ostrovski, G., and Barreto, A. (2020). Temporally-extended  $\{\epsilon\}$ -greedy exploration. *arXiv preprint arXiv:2006.01782*.
- Djellout, H., Guillin, A., and Wu, L. (2004). Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *The Annals of Probability*, 32(3):2702–2732.
- Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245.
- Durugkar, I. P., Rosenbaum, C., Dernbach, S., and Mahadevan, S. (2016). Deep reinforcement learning with macro-actions. *arXiv preprint arXiv:1606.04615*.
- Eberhard, O., Hollenstein, J., Pinneri, C., and Martius, G. (2022). Pink noise is all you need: Colored noise exploration in deep reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

- Engquist, B., Li, X., Ren, W., Vanden-Eijnden, E., et al. (2007). Heterogeneous multiscale methods: a review. *Communications in Computational Physics*, 2(3):367–450.
- Freeman, C. D., Frey, E., Raichuk, A., Girgin, S., Mordatch, I., and Bachem, O. (2021). Brax - a differentiable physics engine for large scale rigid body simulation.
- Greydanus, S., Dzamba, M., and Yosinski, J. (2019). Hamiltonian neural networks. *Advances in neural information processing systems*, 32.
- Grigsby, J., Yoo, J. Y., and Qi, Y. (2021). Towards automatic actor-critic solutions to continuous control. *arXiv preprint arXiv:2106.08918*.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.
- Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.
- Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. (2018). When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Heemels, W., Johansson, K. H., and Tabuada, P. (2021). Event-triggered and self-triggered control. In *Encyclopedia of Systems and Control*, pages 724–730. Springer.
- Heemels, W. P., Johansson, K. H., and Tabuada, P. (2012). An introduction to event-triggered and self-triggered control. In *2012 IEEE 51st IEEE conference on decision and control (CDC)*, pages 3270–3285. IEEE.
- Holt, S., Hüyük, A., and van der Schaar, M. (2023). Active observing in continuous-time control. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Huang, Y., Kavitha, V., and Zhu, Q. (2019). Continuous-time markov decision processes with controlled observations. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 32–39. IEEE.
- Jones, D. S., Plank, M., and Sleeman, B. D. (2009). *Differential equations and mathematical biology*. CRC press.
- Kaandorp, G. C. and Koole, G. (2007). Optimal outpatient appointment scheduling. *Health care management science*, 10:217–229.
- Kabzan, J., Valls, M. I., Reijgwart, V. J., Hendriks, H. F., Ehmke, C., Prajapat, M., Bühler, A., Gosala, N., Gupta, M., Sivanesan, R., et al. (2020). Amz driverless: The full autonomous racing system. *Journal of Field Robotics*, 37(7):1267–1294.
- Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. (2020). Information theoretic regret bounds for online nonlinear control. *NeurIPS*, 33:15312–15325.
- Kallianpur, G. and Sundar, P. (2014). 266Jump Markov Processes. In *Stochastic Analysis and Diffusion Processes*. Oxford University Press.
- Karimi, A. (2023). Decision frequency adaptation in reinforcement learning using continuous options with open-loop policies.
- Kirschner, J. and Krause, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR.
- Krale, M., Simão, T. D., and Jansen, N. (2023). Act-then-measure: reinforcement learning for partially observable environments with active measuring. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 33, pages 212–220.
- Lee, J., Lee, B.-J., and Kim, K.-E. (2020). Reinforcement learning for control with multiple frequencies. *Advances in Neural Information Processing Systems*, 33:3254–3264.

- Lenhart, S. and Workman, J. T. (2007). *Optimal control applied to biological models*. CRC press.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
- Lutter, M., Mannor, S., Peters, J., Fox, D., and Garg, A. (2021). Value iteration in continuous actions, states and time. *arXiv preprint arXiv:2105.04682*.
- Mankowitz, D. J., Mann, T. A., and Mannor, S. (2014). Time regularized interrupting options. In *International Conference on Machine Learning*.
- Mann, T. and Mannor, S. (2014). Scaling up approximate value iteration with options: Better policies with fewer iterations. In *International conference on machine learning*, pages 127–135. PMLR.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
- Nam, H. A., Fleming, S., and Brunskill, E. (2021). Reinforcement learning with state observation costs in action-contingent noiselessly observable markov decision processes. *Advances in Neural Information Processing Systems*, 34:15650–15666.
- Ni, T. and Jang, E. (2022). Continuous control on time. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*.
- Panetta, J. C. and Fister, K. R. (2003). Optimal control applied to competing chemotherapeutic cell-kill strategies. *SIAM Journal on Applied Mathematics*, 63(6):1954–1971.
- Rothfuss, J., Sukhija, B., Birchler, T., Kassraie, P., and Krause, A. (2023). Hallucinated adversarial control for conservative offline policy evaluation. In *Uncertainty in Artificial Intelligence*, pages 1774–1784. PMLR.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Seifner, P. and Sanchez, R. J. (2023). Neural markov jump processes. *arXiv preprint arXiv:2305.19744*.
- Sharma, S., Srinivas, A., and Ravindran, B. (2017). Learning to repeat: Fine grained action repetition for deep reinforcement learning. *arXiv preprint arXiv:1702.06054*.
- Spong, M. W., Hutchinson, S., Vidyasagar, M., et al. (2006). *Robot modeling and control*, volume 3. Wiley New York.
- Srinivas, A., Sharma, S., and Ravindran, B. (2017). Dynamic action repetition for deep reinforcement learning. In *Proc. AAAI*.
- Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.
- Sukhija, B., Treven, L., Sancaktar, C., Blaes, S., Coros, S., and Krause, A. (2024). Optimistic active exploration of dynamical systems. *NeurIPS*.
- Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.
- Tap, F. (2000). *Economics-based optimal control of greenhouse tomato crop production*. Wageningen University and Research.
- Treven, L., Hübotter, J., Sukhija, B., Dörfler, F., and Krause, A. (2023). Efficient exploration in continuous-time model-based reinforcement learning.

- Turchetta, M., Corinzia, L., Sussex, S., Burton, A., Herrera, J., Athanasiadis, I., Buhmann, J. M., and Krause, A. (2022). Learning long-term crop management strategies with cyclegym. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11396–11409. Curran Associates, Inc.
- Vakili, S., Khezeli, K., and Picheny, V. (2021). On information gain and regret bounds in gaussian process bandits. In *AISTATS*.
- Vezhnevets, A., Mnih, V., Osindero, S., Graves, A., Vinyals, O., Agapiou, J., and kavukcuoglu, k. (2016). Strategic attentive writer for learning macro-actions. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.
- Yildiz, C., Heinonen, M., and Lähdesmäki, H. (2021). Continuous-time model-based reinforcement learning. In *International Conference on Machine Learning*, pages 12009–12018. PMLR.
- Yu, H., Xu, W., and Zhang, H. (2021). Taac: Temporally abstract actor-critic for continuous control. *Advances in Neural Information Processing Systems*, 34:29021–29033.
- Zheng, Q., Henaff, M., Amos, B., and Grover, A. (2023). Semi-supervised offline reinforcement learning with action-free trajectories. In *International conference on machine learning*, pages 42339–42362. PMLR.

## **Contents of Appendix**

<b>A</b>	<b>Extended Theory</b>	<b>15</b>
A.1	Transition Cost setting . . . . .	15
A.2	Bounded number of transition . . . . .	18
<b>B</b>	<b>Additional Experiments</b>	<b>24</b>

## A Extended Theory

In this section, we prove Theorem 2 for OTACOS. We separate the section into two parts; proof for the transaction cost setting (Appendix A.1) and the proof for the bounded number of switches setting (Appendix A.2).

We start with the definitions of model complexity and sub-Gaussian random vector that we will use extensively in this section.

**Definition 2** (Model Complexity). *We define the model complexity as is defined by Curi et al. (2020).*

$$\mathcal{I}_N := \max_{\mathcal{D}_1, \dots, \mathcal{D}_N} \sum_{n=1}^N \sum_{(\mathbf{x}, \mathbf{u}, t) \in \mathcal{D}_n} \|\boldsymbol{\sigma}_n(\mathbf{x}, \mathbf{u}, t)\|_2^2. \quad (9)$$

**Definition 3.** *A random variable  $x \in \mathbb{R}$  is said to be sub-Gaussian with variance proxy  $\sigma^2$  if  $\mathbb{E}[x] = 0$  and we have:*

$$\mathbb{E}[e^{tx}] \leq e^{\frac{\sigma^2 t^2}{2}}, \quad \forall t \in \mathbb{R}$$

*A random vector  $\mathbf{x} \in \mathbb{R}^d$  is said to be sub Gaussian with variance proxy  $\sigma^2$  if for any  $\mathbf{e} \in \mathbb{R}^d$ ,  $\|\mathbf{e}\|_2 = 1$  the random variable  $\mathbf{x}^\top \mathbf{e}$  is  $\sigma^2$  sub Gaussian. We write  $\mathbf{x} \sim \text{subG}(\sigma^2)$ .*

In the following, we will be distinguishing between the state of the augmented MDP  $\mathbf{s}$  and the true state of the dynamical system  $\mathbf{x}$ . The augmented state at time step  $k$  includes the true state of the system,  $\mathbf{x}_k$ , the integrated reward  $b_k$  between  $k - 1$  and  $k$ , and the time to left to go  $t_k$ , i.e.,  $\mathbf{s}_k = [\mathbf{x}_k^\top, b_k, t_k]^\top$ .

### A.1 Transition Cost setting

We prove our regret bound for the transition cost case in the following. We start with the difference lemma which adapts Sukhija et al. (2024, Lemma 2) to our setting.

**Lemma 3** (Difference lemma). *Define  $V_{\pi_n, \Phi}(\mathbf{x}, \tau)$  as*

$$\mathbb{E}_{\pi, \Phi} \left[ \sum_{k \geq 0}^{K(\tau)-1} r(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)) \middle| \mathbf{x}_0 = \mathbf{x} \right]; \text{ where } \sum_{k=0}^{K(\tau)-1} \pi_{\mathcal{T}}(\mathbf{x}_k, t_k) = \tau$$

*that is the total reward starting with time to go  $\tau$  and state  $\mathbf{x}$  for the policy  $\boldsymbol{\pi}$  and dynamics  $\Phi$ . Here the expectation w.r.t.  $\boldsymbol{\pi}$ ,  $\Phi$  represents the expectation of the underlying trajectory induced by the policy  $\boldsymbol{\pi}$  on the dynamics  $\Phi$ . Then we have for all  $\boldsymbol{\pi}$ ,  $\Phi'$ ,  $\Phi^*$ ,  $\mathbf{x}_0$ ,  $T < 0$ ;*

$$V_{\pi, \Phi'}(\mathbf{x}_0, T) - V_{\pi, \Phi^*}(\mathbf{x}_0, T) = \mathbb{E}_{\pi, \Phi^*} \left[ \sum_{k \geq 0} V_{\pi, \Phi'}(\hat{\mathbf{x}}_{k+1}, t_{k+1}) - V_{\pi, \Phi'}(\mathbf{x}_{k+1}, t_{k+1}) \right], \quad (10)$$

*where  $\hat{\mathbf{x}}_{k+1}$  is the state of  $\hat{\mathbf{s}}_{k+1} = \Psi_{\Phi'}(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k), \mathbf{w}_k)$  and  $\mathbf{x}_{k+1}$  is the state of  $\mathbf{s}_{k+1} = \Psi_{\Phi^*}(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k), \mathbf{w}_k)$ .*

*Proof.*

$$\begin{aligned} V_{\pi, \Phi^*}(\mathbf{x}_0, T) &= \mathbb{E}_{\pi, \Phi^*} \left[ \sum_{k \geq 0} r(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)) \right] \\ &= \mathbb{E}_{\pi, \Phi^*} \left[ r(\mathbf{s}_0, \boldsymbol{\pi}(\mathbf{s}_0)) + \sum_{k \geq 1} r(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)) \right] \\ &= \mathbb{E}_{\pi, \Phi^*} [r(\mathbf{s}_0, \boldsymbol{\pi}(\mathbf{s}_0)) + V_{\pi, \Phi^*}(\mathbf{x}_1, t_1)] \\ &= \mathbb{E}_{\pi, \Phi^*} [r(\mathbf{s}_0, \boldsymbol{\pi}(\mathbf{s}_0)) + V_{\pi, \Phi'}(\hat{\mathbf{x}}_1, t_1) - V_{\pi, \Phi'}(\mathbf{x}_0, T)] + \\ &\quad + \mathbb{E}_{\pi, \Phi^*} [V_{\pi, \Phi}(\mathbf{x}_0, T) - V_{\pi, \Phi'}(\hat{\mathbf{x}}_1, t_1) + V_{\pi, \Phi^*}(\mathbf{x}_1, t_1)] \\ &= V_{\pi, \Phi'}(\mathbf{x}_0, T) + \mathbb{E}_{\pi, \Phi^*} [V_{\pi, \Phi}(\mathbf{x}_1, t_1) - V_{\pi, \Phi'}(\hat{\mathbf{x}}_1, t_1)] \end{aligned}$$

$$+ \mathbb{E}_{\pi, \Phi^*} [V_{\pi, \Phi^*}(\mathbf{x}_1, t_1) - V_{\pi, \Phi'}(\mathbf{x}_1, t_1)]$$

Hence we have:

$$\begin{aligned} & V_{\pi, \Phi^*}(\mathbf{x}_0, T) - V_{\pi, \Phi'}(\mathbf{x}_0, T) = \\ & = \mathbb{E}_{\pi, \Phi^*} [V_{\pi, \Phi'}(\mathbf{x}_1, t_1) - V_{\pi, \Phi'}(\hat{\mathbf{x}}_1, t_1)] + \mathbb{E}_{\pi, \Phi^*} [V_{\pi, \Phi^*}(\mathbf{x}_1, t_1) - V_{\pi, \Phi'}(\mathbf{x}_1, t_1)] \end{aligned}$$

By repeating the step inductively the result follows.  $\square$

In the following, we leverage the result above to bound the regret of our optimistic planner w.r.t. the difference in value functions.

**Lemma 4** (Per episode regret bound). *Let Assumption 4 hold, then we have with probability at least  $1 - \delta$  for all  $n \geq 0$ .*

$$V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) \leq \mathbb{E}_{\pi_n, \Phi^*} \left[ \sum_{k \geq 0} V_{\pi_n, \Phi_n}(\hat{\mathbf{x}}_{n, k+1}, t_{n, k+1}) - V_{\pi_n, \Phi_n}(\mathbf{x}_{n, k+1}, t_{n, k+1}) \right]. \quad (11)$$

*Proof.* Since we choose the policy optimistically, we get

$$V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) - V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) \leq V_{\pi_n, \Phi_n}(\mathbf{x}_0, T) - V_{\pi_n, \Phi^*}(\mathbf{x}_0, T).$$

Applying Lemma 3 the result follows.  $\square$

Now we derive an upper and lower bound on our value function.

**Lemma 5** (Objective upper bound). *Let  $\pi$  be any policy from the class  $\Pi_{TC}$  and consider any  $T > 0$ , then we have:*

$$-\frac{C}{t_{\min}} T \leq V_{\pi, \Psi^*}(\mathbf{x}_0, T) \leq BT.$$

*Proof.* Since running reward is bounded  $0 \leq b^*(\mathbf{x}, \mathbf{u}) \leq B$ , the number of steps  $K$  we can do in an episode is bounded with  $0 \leq K \leq \frac{T}{t_{\min}}$ , and switch cost is bounded  $0 \leq c(\mathbf{x}, \mathbf{u}) \leq C$  we have:

$$-\frac{C}{t_{\min}} T \leq V_{\pi, \Psi^*}(\mathbf{x}_0, T) \leq BT.$$

$\square$

A key lemma we use to bound the difference in value functions is the following from Kakade et al. (2020).

**Lemma 6** (Absolute expectation Difference Under Two Gaussians (Lemma C.2. Kakade et al. (2020))). *Let  $\mathbf{z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2 \mathbf{I})$  and  $\mathbf{z}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \sigma^2 \mathbf{I})$ , and for any (appropriately measurable) positive function  $g$ , it holds that:*

$$\mathbb{E}[g(\mathbf{z}_1)] - \mathbb{E}[g(\mathbf{z}_2)] \leq \min \left\{ \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|}{\sigma^2}, 1 \right\} \sqrt{\mathbb{E}[g^2(\mathbf{z}_1)]}$$

Furthermore, due to Assumption 4 we can also bound the distance between the next state prediction by the true system  $\Phi^*$  and the optimistic system  $\Phi_n$ .

**Lemma 7.** *Let Assumption 4 hold, then we have the following for all  $n \geq 0$ .*

$$\|\mathbf{x}_{n, k+1} - \hat{\mathbf{x}}_{n, k+1}\| \leq 2\sqrt{d_{\mathbf{x}}}\beta_{n-1} \|\boldsymbol{\sigma}_{n-1}(\mathbf{x}_{n, k}, \boldsymbol{\pi}_n(\mathbf{x}_{n, k}, t_{n, k}))\|$$

*Proof.*

$$\begin{aligned} \|\mathbf{x}_{n, k+1} - \hat{\mathbf{x}}_{n, k+1}\| &= \|\Phi^*(\mathbf{x}_k, \boldsymbol{\pi}_n(\mathbf{x}_{n, k}, t_{n, k})) + \mathbf{w}_{n, k} - (\Phi_n(\mathbf{x}_{n, k}, \boldsymbol{\pi}_n(\mathbf{x}_{n, k}, t_{n, k})) + \mathbf{w}_{n, k})\| \\ &= \|\Phi^*(\mathbf{x}_k, \boldsymbol{\pi}_n(\mathbf{x}_{n, k}, t_{n, k})) - \Phi_n(\mathbf{x}_{n, k}, \boldsymbol{\pi}_n(\mathbf{x}_{n, k}, t_{n, k}))\| \\ &\leq 2\sqrt{d_{\mathbf{x}}}\beta_{n-1} \|\boldsymbol{\sigma}_{n-1}(\mathbf{x}_{n, k}, \boldsymbol{\pi}_n(\mathbf{x}_{n, k}, t_{n, k}))\|, \end{aligned}$$

where the last inequality follows from the fact that  $\Phi^*, \Phi_n \in \mathcal{M}_{n-1}$   $\square$



Next, we relate the regret at each episode to the model epistemic uncertainty using Lemma 3 and Lemma 7.

**Corollary 8.** *Let Assumption 1 – 2 and Assumption 4 hold, then we have for all  $n \geq 0$  with probability at least  $1 - \delta$ .*

$$V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) \leq \frac{2\sqrt{d_{\mathbf{x}}}\beta_{n-1}T}{\sigma} \left( B + \frac{C}{t_{\min}} \right) \mathbb{E} \left[ \sum_{k \geq 0} \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}))\| \right] \quad (12)$$

*Proof.* From Lemma 4 we have:

$$V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) \leq \mathbb{E} \left[ \sum_{k \geq 0} V_{\pi_n, \Phi_n}(\mathbf{x}_{n,k+1}, t_{n,k+1}) - V_{\pi_n, \Phi_n}(\hat{\mathbf{x}}_{n,k+1}, t_{n,k+1}) \right].$$

Lemma 6 can be applied to positive function  $g$ . We hence make a transformation and apply it to  $g(\cdot) = V_{\pi_n, \Phi_n}(\cdot, t_{n,k+1}) + \frac{C}{t_{\min}}T$ , which is positive due to Lemma 5. Moreover,  $\forall \mathbf{x} \in \mathcal{X}$ ;

$$g(\cdot) = V_{\pi_n, \Phi_n}(\cdot, t_{n,k+1}) + \frac{C}{t_{\min}}T \leq Bt_{n,k+1} + \frac{C}{t_{\min}}T \leq T\left(B + \frac{C}{t_{\min}}\right).$$

Applying Lemma 6 we obtain:

$$V_{\pi_n, \Phi_n}(\mathbf{x}_{n,k+1}, t_{n,k+1}) - V_{\pi_n, \Phi_n}(\hat{\mathbf{x}}_{n,k+1}, t_{n,k+1}) \leq \frac{T}{\sigma} \left( B + \frac{C}{t_{\min}} \right) \mathbb{E} [\|\mathbf{x}_{n,k+1} - \hat{\mathbf{x}}_{n,k+1}\|]$$

Finally, applying Lemma 7 we arrive at:

$$V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) \leq \frac{2\sqrt{d_{\mathbf{x}}}\beta_{n-1}T}{\sigma} \left( B + \frac{C}{t_{\min}} \right) \mathbb{E} \left[ \sum_{k \geq 0} \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}))\| \right]$$

□

Now we can prove our regret bound for the transition cost case.

**Theorem 9.** *Let Assumption 1 – 2 and Assumption 4 hold, then we have for all  $n \geq 0$  with probability at least  $1 - \delta$ .*

$$\begin{aligned} R_N &= \sum_{n=1}^N V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) \\ &\leq \frac{2\sqrt{d_{\mathbf{x}}}\beta_{N-1}T^{3/2}}{\sigma^2 t_{\min}} \left( B + \frac{C}{t_{\min}} \right) \sqrt{N\mathcal{I}_N} \end{aligned}$$

*Proof.* We compute:

$$\begin{aligned} R_N &= \sum_{n=1}^N V_{\pi_n, \Phi^*}(\mathbf{x}_0, T) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T) \\ &\leq \frac{2\sqrt{d_{\mathbf{x}}T}}{\sigma} \left( B + \frac{C}{t_{\min}} \right) \sum_{n=1}^N \beta_{n-1} \mathbb{E} \left[ \sum_{k \geq 0} \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}))\| \right] \\ &\leq \frac{2\sqrt{d_{\mathbf{x}}}\beta_{N-1}T}{\sigma} \left( B + \frac{C}{t_{\min}} \right) \mathbb{E} \left[ \sum_{n=1}^N \sum_{k \geq 0} \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}))\| \right] \\ &\leq \frac{2\sqrt{d_{\mathbf{x}}}\beta_{N-1}T}{\sigma} \left( B + \frac{C}{t_{\min}} \right) \sqrt{\frac{TN}{t_{\min}}} \mathbb{E} \left[ \sqrt{\sum_{n=1}^N \sum_{k \geq 0} \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}))\|^2} \right] \end{aligned}$$

$$\leq \frac{2\sqrt{d_{\mathbf{x}}}\beta_{N-1}T^{3/2}}{\sigma\sqrt{t_{\min}}} \left( B + \frac{C}{t_{\min}} \right) \sqrt{N\mathcal{I}_N}$$

Here the first inequality follows because of Corollary 8, the second inequality follows due to the monotonicity of sequence  $(\beta_n)_{n \geq 0}$ , the third inequality follows by Cauchy–Schwarz and the last one by maximizing the term in expectation.  $\square$

Our regret  $R_N$  is sublinear if  $\beta_{N-1}\sqrt{N\mathcal{I}_N}$  is sublinear. For general well-calibrated models this is tough to verify. However, for Gaussian process dynamics,  $\mathcal{I}_N$  is equal to (up to constant factors) the maximum information gain  $\gamma_N$  (Srinivas et al., 2009) (c.f., Curi et al. (2020, Lemma 17)). The maximum information gain is sublinear for a rich class of kernels (Vakili et al., 2021), i.e., yielding sublinear regret for OTACOS (see Sukhija et al. (2024, Theorem 2) for more detail).

## A.2 Bounded number of transition

We overload the notation in this section and add number of switches to the value function, such that we have  $V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, 0) = V_{\pi_n, \Phi^*}(\mathbf{x}_0, T)$

**Lemma 10** (Per episode regret bound). *We have:*

$$\begin{aligned} V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, 0) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T, 0) &\leq \\ &\leq \mathbb{E} \left[ \sum_{k=0}^{K-1} V_{\pi_n, \Phi_n}(\mathbf{x}_{n,k+1}, t_{n,k+1}, k+1) - V_{\pi_n, \Phi_n}(\widehat{\mathbf{x}}_{n,k+1}, t_{n,k+1}, k+1) \right], \end{aligned}$$

where  $\widehat{\mathbf{x}}_{n,k+1}$  is the state of one step hallucinated component  $\widehat{\mathbf{s}}_{n,k+1} = \Psi_{\Phi_n}(\mathbf{s}_{n,k}, \pi_n(\mathbf{s}_{n,k}), \mathbf{w}_{n,k})$  and  $\mathbf{x}_{n,k+1}$  is the state of  $\mathbf{s}_{n,k+1} = \Psi_{\Phi^*}(\mathbf{s}_{n,k}, \pi_n(\mathbf{s}_{n,k}), \mathbf{w}_{n,k})$ .

*Proof.*

$$\begin{aligned} V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, 0) &= \mathbb{E} \left[ \sum_{k \geq 0} r(\mathbf{s}_{n,k}, \pi_n(\mathbf{s}_{n,k})) \right] = \mathbb{E} \left[ r(\mathbf{s}_{n,0}, \pi_n(\mathbf{s}_{n,0})) + \sum_{k \geq 1} r(\mathbf{s}_{n,k}, \pi_n(\mathbf{s}_{n,k})) \right] \\ &= \mathbb{E} [r(\mathbf{s}_{n,k}, \pi_n(\mathbf{s}_{n,0})) + V_{\pi_n, \Phi^*}(\mathbf{x}_{n,1}, t_{n,1}, 1)] \\ &= \mathbb{E} [r(\mathbf{s}_{n,k}, \pi_n(\mathbf{s}_{n,0})) + V_{\pi_n, \Phi_n}(\mathbf{x}_{n,1}, t_{n,1}, 1) - V_{\pi_n, \Phi_n}(\mathbf{x}_0, T, 0)] + \\ &\quad + \mathbb{E} [V_{\pi_n, \Phi_n}(\mathbf{x}_0, T, 0) - V_{\pi_n, \Phi_n}(\mathbf{x}_{n,1}, t_{n,1}, 1) + V_{\pi_n, \Phi^*}(\mathbf{x}_{n,1}, t_{n,1}, 1)] \\ &= V_{\pi_n, \Phi_n}(\mathbf{x}_0, T, 0) + \mathbb{E} [V_{\pi_n, \Phi_n}(\widehat{\mathbf{x}}_{n,1}, t_{n,1}, 1) - V_{\pi_n, \Phi_n}(\mathbf{x}_{n,1}, t_{n,1}, 1)] \\ &\quad + \mathbb{E} [V_{\pi_n, \Phi^*}(\mathbf{x}_{n,1}, t_{n,1}, 1) - V_{\pi_n, \Phi_n}(\mathbf{x}_{n,1}, t_{n,1}, 1)] \end{aligned}$$

Hence we have:

$$\begin{aligned} V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, 0) - V_{\pi_n, \Phi_n}(\mathbf{x}_0, T, 0) &= \\ &= \mathbb{E} [V_{\pi_n, \Phi_n}(\widehat{\mathbf{x}}_{n,1}, t_{n,1}, 1) - V_{\pi_n, \Phi_n}(\mathbf{x}_{n,1}, t_{n,1}, 1)] + \mathbb{E} [V_{\pi_n, \Phi^*}(\mathbf{x}_{n,1}, t_{n,1}, 1) - V_{\pi_n, \Phi_n}(\mathbf{x}_{n,1}, t_{n,1}, 1)] \end{aligned}$$

Repeating the step inductively the result follows and using  $V_{\pi_n, \Phi^*}(\mathbf{x}_{n,K}, t_{n,K}, K) = 0$  we prove the lemma.  $\square$

### A.2.1 Subgaussianity of the noise

In principle, we could assume that the noise  $\mathbf{w}_k$  is Gaussian and then with the same analysis obtain the regret bound. However, stochastic flows are in many cases not exactly Gaussian but only sub-Gaussian. For such noise we need can not apply Lemma 6 and need to resort to different analysis. First we show that under mild assumptions on the SDE dynamics functions  $\mathbf{f}^*$  and  $\mathbf{g}^*$  the resulting noise  $\mathbf{w}_k$  is sub-Gaussian.

To derive this result we will follow the work of Djellout et al. (2004). We present the results in quite informal way, for more rigorous statements we refer the reader to Djellout et al. (2004).

**Definition 4** (Wasserstein distance). *Let  $(\mathcal{E}, d_{\mathcal{E}})$  be a metric space and let  $\mu, \nu$  be two probability measures on  $\mathcal{E}$ . We define:*

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(x,y) \sim \gamma} [d(x,y)^p]^{\frac{1}{p}}$$

**Definition 5** (Kullback–Leibler divergence). *Let  $(\mathcal{E}, d_{\mathcal{E}})$  be a metric space and let  $\mu, \nu$  be two probability measures on  $\mathcal{E}$ . We define:*

$$H(\nu||\mu) = \begin{cases} \mathbb{E}_{x \sim \nu} \left[ \log \left( \frac{d\nu(x)}{d\mu(x)} \right) \right], & \text{if } \nu \ll \mu \\ +\infty, & \text{else} \end{cases}$$

**Definition 6** ( $L^p$ -transportation cost information inequality). *Let  $(\mathcal{E}, d_{\mathcal{E}})$  be a metric space and let  $\mu$  be a probability measure on  $\mathcal{E}$ . We say that  $\mu$  satisfy the  $L^p$ -transportation cost information inequality, and for short write  $\mu \in T_p(C)$ , if there exists a constant  $C$  such that for any measure  $\nu$  on  $\mathcal{E}$  we have:*

$$W_p(\mu, \nu) \leq \sqrt{2CH(\nu||\mu)}.$$

We now state an important theorem of Bobkov and Götze (1999) that we will use later.

**Theorem 11** (From Bobkov and Götze (1999)). *Let  $(\mathcal{E}, d_{\mathcal{E}})$  be a metric space and let  $\mu$  be a probability measure on  $\mathcal{E}$ . We have that  $\mu \in T_1(C)$  if and only if for any  $\mu$ -integrable and  $L_F$ -Lipschitz function  $F : (\mathcal{E}, d_{\mathcal{E}}) \rightarrow \mathbb{R}$  and for any  $\lambda \in \mathbb{R}$  we have:*

$$\mathbb{E}_{x \sim \mu} \left[ e^{\lambda(F(x) - \mathbb{E}_{x \sim \mu}[F(x)])} \right] \leq e^{\frac{\lambda^2}{2} CL_F^2}$$

Next, we provide a condition under which  $\Xi(\mathbf{x}, \mathbf{u}, t)$  is sub-Gaussian random variable for any  $t \in \mathcal{T}$ .

**Corollary 12** (Adjusted Corollary 4.1 of Djellout et al. (2004)). *Assume*

$$\sup_{\substack{\mathbf{x} \in \mathbb{R}^{d_x} \\ \mathbf{u} \in \mathbb{R}^{d_u}}} \|\mathbf{g}^*(\mathbf{x}, \mathbf{u})\|_F \leq A, \quad \|\mathbf{f}^*(\mathbf{x}, \mathbf{u}) - \mathbf{f}^*(\hat{\mathbf{x}}, \hat{\mathbf{u}})\| \leq L_{f^*} \|(\mathbf{x}, \mathbf{u}) - (\hat{\mathbf{x}}, \hat{\mathbf{u}})\|,$$

and denote the law of  $(\Xi(\mathbf{x}, \mathbf{u}, t))_{t \in \mathcal{T}}$  on the space  $C(\mathcal{T}, \mathbb{R}^{d_{\Xi}})$  (space of continuous functions from  $\mathcal{T}$  to  $\mathbb{R}^{d_{\Xi}}$ ) by  $\mathbb{P}_{\mathbf{x}}$ . Then, there exist a constant  $C = C(A, L_{f^*}, T)$  such that  $\mathbb{P}_{\mathbf{x}} \in T_1(C)$  on the space  $C(\mathcal{T}, \mathbb{R}^{d_{\Xi}})$  equipped with the metric:

$$d(\gamma_1, \gamma_2) = \sup_{t \in [0, T]} \|\gamma_1(t) - \gamma_2(t)\|$$

Lets  $\mathbf{e}$  be a(ny) unit vector in  $\mathbb{R}^{d_{\Xi}}$  and define:

$$\begin{aligned} F_{\mathbf{e}, t} &: C(\mathcal{T}, \mathbb{R}^{d_{\Xi}}) \rightarrow \mathbb{R} \\ F_{\mathbf{e}, t} &: \gamma \mapsto \gamma(t)^\top \mathbf{e} \end{aligned}$$

We have:

$$\begin{aligned} |F_{\mathbf{e}, t}(\gamma_1) - F_{\mathbf{e}, t}(\gamma_2)| &= |(\gamma_1(t) - \gamma_2(t))^\top \mathbf{e}| \\ &\leq \|\gamma_1(t) - \gamma_2(t)\| \|\mathbf{e}\| = \|\gamma_1(t) - \gamma_2(t)\| \\ &\leq \sup_{t \in \mathcal{T}} \|\gamma_1(t) - \gamma_2(t)\| = d(\gamma_1, \gamma_2) \end{aligned}$$

Therefore for any  $\mathbf{e}, t$  the function  $F_{\mathbf{e}, t}$  is 1-Lipschitz. Since we have

$$\mathbb{E}[|F_{\mathbf{e}, t}(\gamma)|] = \int_{C(\mathcal{T}, \mathbb{R}^{d_{\Xi}})} |\gamma(t)| d\mathbb{P}_{\mathbf{x}}(\gamma) = \mathbb{E}[|\Xi(\mathbf{x}, \mathbf{u}, t)^\top \mathbf{e}|] < \infty$$

the function  $F_{\mathbf{e}, t}$  is also  $\mathbb{P}_{\mathbf{x}}$ -integrable. Combining the latter observation with the Theorem 11 we obtain that for any  $\mathbf{e} \in \mathbb{R}^{d_{\Xi}}$  and any  $t \in \mathcal{T}$  we have:

$$\mathbb{E}_{\Xi(\mathbf{x}, \mathbf{u}, t)} \left[ e^{\lambda(\Xi(\mathbf{x}, \mathbf{u}, t)^\top \mathbf{e} - \mathbb{E}[\Xi(\mathbf{x}, \mathbf{u}, t)^\top \mathbf{e}])} \right] = \mathbb{E}_{\gamma \sim \mathbb{P}_{\mathbf{x}}} \left[ e^{\lambda(F_{\mathbf{e}, t}(\gamma) - \mathbb{E}_{\gamma \sim \mathbb{P}_{\mathbf{x}}}[F_{\mathbf{e}, t}(\gamma)])} \right] \leq e^{\frac{\lambda^2}{2} C}$$

Hence under the assumption of Theorem 2 for Bounded number of switches setting we have that for any  $t \in \mathcal{T}$  the random variable  $\Xi(\mathbf{x}, \mathbf{u}, t) - \mathbb{E}[\Xi(\mathbf{x}, \mathbf{u}, t)] \sim \text{subG}(C)$ . The variance proxy  $C$  depends on  $A, L_{f^*}, T$ .

### A.2.2 Lipschitiness of the expected flow $\Phi^*$

To apply analysis for the case when noise  $w_k$  is any sub-Gaussian we also need to show that the dynamics function  $\Phi^*$  is Lipschitz. We first start with some general results.

**Lemma 13.** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $A \subset [n]$  and denote  $B = A^C$ . If we have:*

- $\|f(\mathbf{x}_A, \mathbf{x}_B) - f(\widehat{\mathbf{x}}_A, \mathbf{x}_B)\|_2 \leq L_A \|\mathbf{x}_A - \widehat{\mathbf{x}}_A\|_2$ ,
- $\|f(\mathbf{x}_A, \mathbf{x}_B) - f(\mathbf{x}_A, \widehat{\mathbf{x}}_B)\|_2 \leq L_B \|\mathbf{x}_B - \widehat{\mathbf{x}}_B\|_2$ ,

then  $f$  is  $2(L_A + L_B)$  Lipschitz.

*Proof.* We have:

$$\begin{aligned}
\|f(\mathbf{x}) - f(\widehat{\mathbf{x}})\|_2 &= \|f(\mathbf{x}_A, \mathbf{x}_B) - f(\widehat{\mathbf{x}}_A, \widehat{\mathbf{x}}_B)\|_2 \\
&= \|f(\mathbf{x}_A, \mathbf{x}_B) - f(\widehat{\mathbf{x}}_A, \mathbf{x}_B) + f(\widehat{\mathbf{x}}_A, \mathbf{x}_B) - f(\widehat{\mathbf{x}}_A, \widehat{\mathbf{x}}_B)\|_2 \\
&\leq L_A \|\mathbf{x}_A - \widehat{\mathbf{x}}_A\|_2 + L_B \|\mathbf{x}_B - \widehat{\mathbf{x}}_B\|_2 \\
&\leq (L_A + L_B) (\|\mathbf{x}_A - \widehat{\mathbf{x}}_A\|_2 + \|\mathbf{x}_B - \widehat{\mathbf{x}}_B\|_2) \\
&\leq 2(L_A + L_B) \left\| \begin{pmatrix} \mathbf{x}_A - \widehat{\mathbf{x}}_A \\ \mathbf{x}_B - \widehat{\mathbf{x}}_B \end{pmatrix} \right\|_2 \\
&= 2(L_A + L_B) \|\mathbf{x} - \widehat{\mathbf{x}}\|_2
\end{aligned}$$

□

**Lemma 14** (Lipschitzness of  $\Phi_{f^*}$ ). *There exists a positive constant  $L_{\Phi_{f^*}}$  such that the flow  $\Phi_{f^*}$  is  $L_{\Phi_{f^*}}$ -Lipschitz.*

*Proof.* We will first prove coordinate-wise Lipschitzness. We observe:

1. Lipschitzness in time:

$$\begin{aligned}
\|\Phi_{f^*}(\mathbf{x}, \mathbf{u}, t) - \Phi_{f^*}(\mathbf{x}, \mathbf{u}, \widehat{t})\| &= \left\| \int_0^t \mathbb{E}[f^*(\mathbf{x}_s, \mathbf{u})] ds - \int_0^{\widehat{t}} \mathbb{E}[f^*(\mathbf{x}_s, \mathbf{u})] ds \right\| \\
&\leq \int_{\widehat{t}}^t \mathbb{E}[\|f^*(\mathbf{x}_s, \mathbf{u})\|] ds \leq F |t - \widehat{t}|
\end{aligned}$$

2. Lipschitzness in state  $\mathbf{x}$ : To prove this, consider the  $\delta \mathbf{x}_t = \Xi(\mathbf{x}, \mathbf{u}, t) - \Xi(\widehat{\mathbf{x}}, \mathbf{u}, t)$ , then we have

$$\begin{aligned}
d\delta \mathbf{x}_t &= (f^*(\mathbf{x}_t, \mathbf{u}) - f^*(\widehat{\mathbf{x}}_t, \mathbf{u}))dt + (g^*(\mathbf{x}_t, \mathbf{u}) - g^*(\widehat{\mathbf{x}}_t, \mathbf{u}))d\mathbf{B}_t \\
&= \delta \mathbf{f}_t^* dt + \delta \mathbf{g}_t^* d\mathbf{B}_t.
\end{aligned}$$

Note that  $\|\delta \mathbf{f}_t^*\| \leq L_{f^*} \|\delta \mathbf{x}_t\|$  and  $\|\delta \mathbf{g}_t^*\| \leq L_{g^*} \|\delta \mathbf{x}_t\|$  since both functions are Lipschitz. Define  $\mathbf{y}_t = \delta \mathbf{x}_t^\top \delta \mathbf{x}_t$  and use Ito's Lemma to get

$$d\mathbf{y}_t = 2\delta \mathbf{x}_t^\top (\delta \mathbf{f}_t^* dt + \delta \mathbf{g}_t^* d\mathbf{B}_t) + \text{tr}(\delta \mathbf{g}_t^* (\delta \mathbf{g}_t^*)^\top) dt$$

Moreover,

$$\begin{aligned}
\mathbb{E}[\mathbf{y}_t] &= \int_0^t 2\mathbb{E}[\delta \mathbf{x}_s^\top \delta \mathbf{f}_s^*] + \mathbb{E}[\text{tr}(\delta \mathbf{g}_s^* (\delta \mathbf{g}_s^*)^\top)] ds \\
&\leq \int_0^t 2\mathbb{E}[\|\delta \mathbf{x}_s\| \|\delta \mathbf{f}_s^*\|] + \mathbb{E}[\|\delta \mathbf{g}_s^*\|^2] ds \\
&\leq \int_0^t (2L_{f^*} + L_{g^*}^2) \mathbb{E}[\|\delta \mathbf{x}_s\|^2] ds
\end{aligned}$$

Note that  $\mathbf{y}_t = \|\delta \mathbf{x}_t\|^2$ , so we can apply Grönwall's inequality to get

$$\mathbb{E}[\|\delta \mathbf{x}_t\|^2] \leq \|\delta \mathbf{x}_0\|^2 e^{(2L_{f^*} + L_{g^*}^2)t}.$$

Moreover,

$$\|\mathbb{E}[\delta \mathbf{x}_t]\| \leq \sqrt{\mathbb{E}[\|\delta \mathbf{x}_t\|^2]} \leq \|\delta \mathbf{x}_0\| e^{\frac{2L_{\mathbf{f}^*} + L_{\mathbf{g}^*}^2}{2}t} \leq \|\delta \mathbf{x}_0\| e^{\frac{2L_{\mathbf{f}^*} + L_{\mathbf{g}^*}^2}{2}T}.$$

Hence we have:

$$\|\Phi_{\mathbf{f}^*}(\mathbf{x}, \mathbf{u}, t) - \Phi_{\mathbf{f}^*}(\hat{\mathbf{x}}, \mathbf{u}, t)\| \leq \|\mathbf{x} - \hat{\mathbf{x}}\| e^{\frac{2L_{\mathbf{f}^*} + L_{\mathbf{g}^*}^2}{2}T}.$$

3. Lipschitzness in action  $\mathbf{u}$ : We denote  $\delta \mathbf{x}_t = \Xi(\mathbf{x}, \mathbf{u}, t) - \Xi(\mathbf{x}, \hat{\mathbf{u}}, t)$  and  $\delta \mathbf{u} = \mathbf{u} - \hat{\mathbf{u}}$ . Following the same steps as in the proof of Lipschitzness in state we arrive at:

$$d\mathbf{y}_t = 2\delta \mathbf{x}_t^\top (\delta \mathbf{f}_t^* dt + \delta \mathbf{g}_t^* d\mathbf{B}_t) + \text{tr}(\delta \mathbf{g}_t^* (\delta \mathbf{g}_t^*)^\top) dt$$

Integration yields:

$$\begin{aligned} \mathbb{E}[\mathbf{y}_t] &= \int_0^t 2\mathbb{E}[\delta \mathbf{x}_s^\top \delta \mathbf{f}_s^*] + \mathbb{E}[\text{tr}(\delta \mathbf{g}_s^* (\delta \mathbf{g}_s^*)^\top)] ds \\ &\leq \int_0^t 2\mathbb{E}[\|\delta \mathbf{x}_s\| \|\delta \mathbf{f}_s^*\|] + \mathbb{E}[\|\delta \mathbf{g}_s^*\|^2] ds \\ &\leq \int_0^t 2\mathbb{E}[L_{\mathbf{f}^*} \|\delta \mathbf{x}_s\| (\|\delta \mathbf{x}_s\| + \|\delta \mathbf{u}\|)] + \mathbb{E}\left[2L_{\mathbf{g}^*}^2 (\|\delta \mathbf{x}_s\|^2 + \|\delta \mathbf{u}\|^2)\right] ds \\ &\leq \int_0^t (3L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2) \mathbb{E}[\mathbf{y}_s] + (L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2) \|\delta \mathbf{u}\| ds, \end{aligned}$$

where we used  $(a+b)^2 \leq 2a^2 + 2b^2$  and  $ab \leq \frac{1}{2}(a^2 + b^2)$ . Applying Grönwall's inequality results in:

$$\begin{aligned} \mathbb{E}[\|\delta \mathbf{x}_t\|^2] &\leq \|\delta \mathbf{u}\|^2 (L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2) e^{(3L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2)t} \\ &\leq \|\delta \mathbf{u}\|^2 (L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2) e^{(3L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2)T} \end{aligned}$$

Applying Lemma 13 on 2. and 3. we have that  $\Phi_{\mathbf{f}^*}(\cdot, \cdot, t)$  is  $2\left(e^{\frac{2L_{\mathbf{f}^*} + L_{\mathbf{g}^*}^2}{2}T} + \sqrt{L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2} e^{\frac{3L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2}{2}T}\right)$ -Lipschitz. Applying Lemma 13 on 1. and  $\Phi_{\mathbf{f}^*}(\cdot, \cdot, t)$  and bounding  $2 \leq 4$  we finally obtain that  $\Phi_{\mathbf{f}^*}$  is  $4\left(e^{\frac{2L_{\mathbf{f}^*} + L_{\mathbf{g}^*}^2}{2}T} + \sqrt{L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2} e^{\frac{3L_{\mathbf{f}^*} + 2L_{\mathbf{g}^*}^2}{2}T} + F\right)$ -Lipschitz.  $\square$

**Corollary 15** (Lipschitzness of the  $\Phi_{b^*}$ ). *The cost flow  $\Phi_{b^*}$  is  $\mathcal{O}\left(e^{C_1(L_{\mathbf{f}^*} + L_{\mathbf{g}^*}^2)T}\right)$ -Lipschitz, where  $C_1$  is a constant.*

*Proof.* Same as in the proof of Lemma 14 we first show coordinate-wise Lipschitzness.

1. We first show Lipschitzness in time:

$$\begin{aligned} |\Phi_{b^*}(\mathbf{x}, \mathbf{u}, t) - \Phi_{b^*}(\mathbf{x}, \mathbf{u}, \hat{t})| &= \left| \mathbb{E} \left[ \int_{\hat{t}}^t b^*(\mathbf{x}_s, \mathbf{u}) ds \right] \right| \\ &\leq \mathbb{E} \left[ \int_{\hat{t}}^t |b^*(\mathbf{x}_s, \mathbf{u})| ds \right] \\ &\leq \mathbb{E} [B(t - \hat{t})] = B(t - \hat{t}). \end{aligned}$$

2. To obtain Lipschitzness in state observe:

$$|\Phi_{b^*}(\mathbf{x}, \mathbf{u}, t) - \Phi_{b^*}(\hat{\mathbf{x}}, \mathbf{u}, t)| = \left| \mathbb{E} \left[ \int_0^t b^*(\Xi(\mathbf{x}, \mathbf{u}, s), \mathbf{u}) - b^*(\Xi(\hat{\mathbf{x}}, \mathbf{u}, s), \mathbf{u}) ds \right] \right|$$

$$\begin{aligned}
&\leq \mathbb{E} \left[ \int_0^t |b^*(\Xi(\mathbf{x}, \mathbf{u}, s), \mathbf{u}) - b^*(\Xi(\hat{\mathbf{x}}, \mathbf{u}, s), \mathbf{u})| ds \right] \\
&\leq L_{b^*} \mathbb{E} \left[ \int_0^t \|\Xi(\mathbf{x}, \mathbf{u}, s) - \Xi(\hat{\mathbf{x}}, \mathbf{u}, s)\| ds \right] \\
&\leq L_{b^*} \int_0^t \sqrt{\mathbb{E} \left[ \|\Xi(\mathbf{x}, \mathbf{u}, s) - \Xi(\hat{\mathbf{x}}, \mathbf{u}, s)\|^2 \right]} ds \\
&\leq L_{b^*} \|\mathbf{x} - \hat{\mathbf{x}}\| \int_0^t e^{\frac{2L_{f^*} + L_{g^*}^2}{2}s} ds \\
&= \frac{2L_{b^*}}{2L_{f^*} + L_{g^*}^2} \left( e^{\frac{2L_{f^*} + L_{g^*}^2}{2}t} - 1 \right) \|\mathbf{x} - \hat{\mathbf{x}}\|
\end{aligned}$$

3. Finally, for Lipschitzness in action observe:

$$\begin{aligned}
|\Phi_{b^*}(\mathbf{x}, \mathbf{u}, t) - \Phi_{b^*}(\mathbf{x}, \hat{\mathbf{u}}, t)| &= \left| \mathbb{E} \left[ \int_0^t b^*(\Xi(\mathbf{x}, \mathbf{u}, s), \mathbf{u}) - b^*(\Xi(\mathbf{x}, \hat{\mathbf{u}}, s), \mathbf{u}) ds \right] \right| \\
&\leq \mathbb{E} \left[ \int_0^t |b^*(\Xi(\mathbf{x}, \mathbf{u}, s), \mathbf{u}) - b^*(\Xi(\mathbf{x}, \hat{\mathbf{u}}, s), \mathbf{u})| ds \right] \\
&\leq L_{b^*} \mathbb{E} \left[ \int_0^t \|\Xi(\mathbf{x}, \mathbf{u}, s) - \Xi(\mathbf{x}, \hat{\mathbf{u}}, s)\| + \|\mathbf{u} - \hat{\mathbf{u}}\| ds \right] \\
&\leq L_{b^*} t \|\mathbf{u} - \hat{\mathbf{u}}\| + L_{b^*} \int_0^t \sqrt{\mathbb{E} \left[ \|\Xi(\mathbf{x}, \mathbf{u}, s) - \Xi(\mathbf{x}, \hat{\mathbf{u}}, s)\|^2 \right]} ds \\
&\leq L_{b^*} t \|\mathbf{u} - \hat{\mathbf{u}}\| + L_{b^*} \|\mathbf{u} - \hat{\mathbf{u}}\| \sqrt{L_{f^*} + 2L_{g^*}^2} \int_0^t e^{\frac{3L_{f^*} + 2L_{g^*}^2}{2}s} ds \\
&= L_{b^*} \left( t + \frac{2\sqrt{L_{f^*} + 2L_{g^*}^2}}{3L_{f^*} + 2L_{g^*}^2} \left( e^{\frac{3L_{f^*} + 2L_{g^*}^2}{2}t} - 1 \right) \right) \|\mathbf{u} - \hat{\mathbf{u}}\| \\
&\leq L_{b^*} \left( T + \frac{2\sqrt{L_{f^*} + 2L_{g^*}^2}}{3L_{f^*} + 2L_{g^*}^2} \left( e^{\frac{3L_{f^*} + 2L_{g^*}^2}{2}T} - 1 \right) \right) \|\mathbf{u} - \hat{\mathbf{u}}\|
\end{aligned}$$

Applying Lemma 13 result follows.  $\square$

**Corollary 16** (Lipschitzness of  $\Phi^*$ ). *The unknown function  $\Phi^*$  is  $L_{\Phi} = L_{\Phi_f} + L_{\Phi_b} = \mathcal{O}\left(e^{D(L_{f^*} + L_{g^*}^2)T}\right)$ -Lipschitz, where  $D$  is constant.*

### A.2.3 Regret bound

**Lemma 17** (Per episode regret bound (general sub-Gaussian noise)). *Consider the setting with a bounded number of switches  $K$ , and let Assumption 1, 3, and Assumption 4 hold. Then, we get with probability at least  $1 - \delta$ :*

$$\begin{aligned}
&V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, K) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T, K) \leq \\
&\leq \mathcal{O} \left( L_{\sigma}^{K-1} \beta_{n-1}^K e^{C(L_{f^*} + L_{g^*}^2)(1+L_{\pi})TK} \mathbb{E} \left[ \sum_{k=0}^K \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}, k))\|_2 \right] \right)
\end{aligned}$$

*Proof.* Applying Lemma 5 of Curi et al. (2020) the result follows.  $\square$

**Theorem 18.** *Consider the setting with a bounded number of switches  $K$ , and let Assumption 1, 3, and Assumption 4 hold. Then, we get with probability at least  $1 - \delta$ :*

$$R_N = \sum_{n=1}^N V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, K) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T, K)$$

$$\leq \mathcal{O} \left( L_{\sigma}^{K-1} \beta_{N-1}^K \sqrt{K} e^{C(L_{f^*} + L_{g^*}^2)(1+L_{\pi})TK} \sqrt{N\mathcal{I}_N} \right)$$

*Proof.* We apply Lemma 17 and Cauchy-Schwarz:

$$\begin{aligned} R_N &= \sum_{n=1}^N V_{\pi_n, \Phi^*}(\mathbf{x}_0, T, K) - V_{\pi^*, \Phi^*}(\mathbf{x}_0, T, K) \\ &\leq \sum_{n=1}^N \mathcal{O} \left( L_{\sigma}^{K-1} \beta_{n-1}^K e^{C(L_{f^*} + L_{g^*}^2)(1+L_{\pi})TK} \mathbb{E} \left[ \sum_{k=0}^K \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}, k))\|_2 \right] \right) \\ &\leq \mathcal{O} \left( L_{\sigma}^{K-1} \beta_{N-1}^K e^{C(L_{f^*} + L_{g^*}^2)(1+L_{\pi})TK} \right) \mathbb{E} \left[ \sum_{n=1}^N \sum_{k=0}^K \|\sigma_{n-1}(\mathbf{x}_{n,k}, \pi_n(\mathbf{x}_{n,k}, t_{n,k}, k))\|_2 \right] \\ &\leq \mathcal{O} \left( L_{\sigma}^{K-1} \beta_{N-1}^K e^{C(L_{f^*} + L_{g^*}^2)(1+L_{\pi})TK} \right) \sqrt{K} \sqrt{N\mathcal{I}_N} \end{aligned}$$

Here we first applied Lemma 17. Then we used the monotonicity of  $(\beta_n)_{n \geq 0}$  sequence. In the last step we first applied maximum over the collected data, then Cauchy-Schwarz inequality and finally the definition of model complexity.  $\square$

## B Additional Experiments

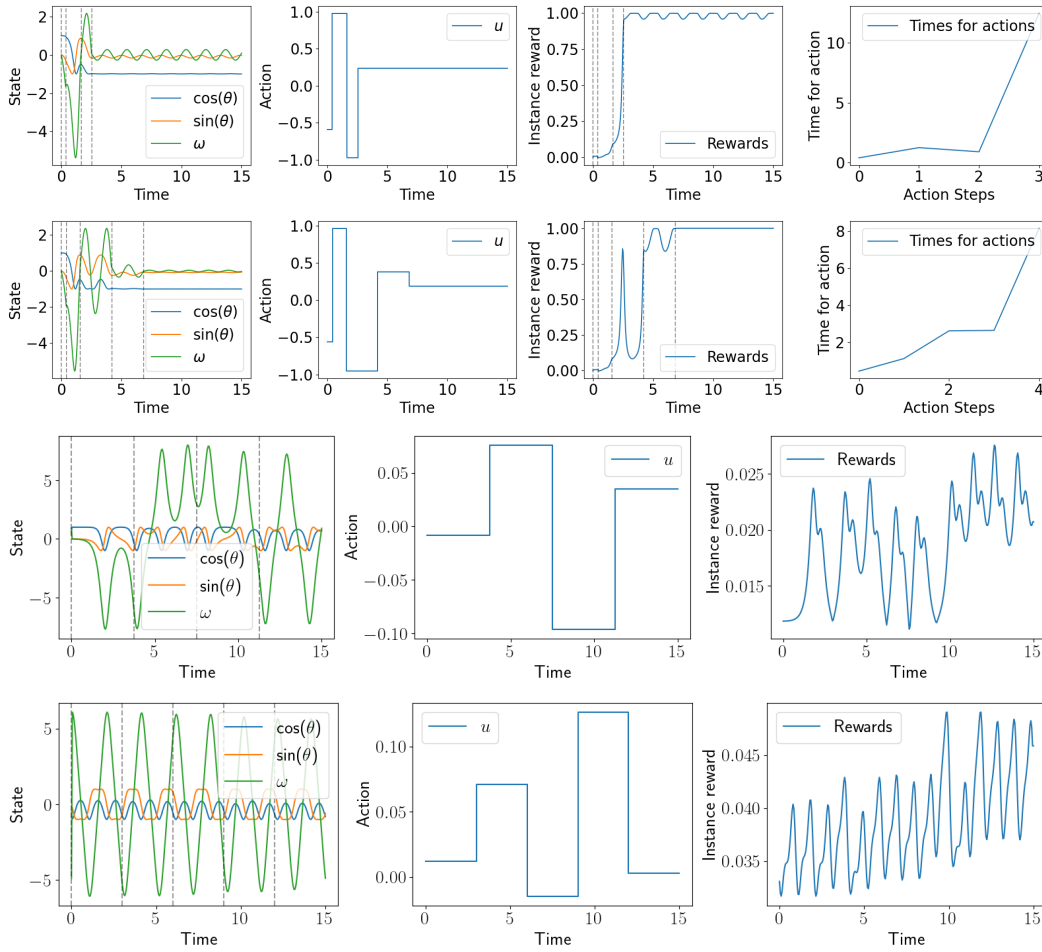


Figure 6: Pendulum swing-down task. Row 1: 4 interactions, optimized interaction times, Row 2: 5 interactions, optimized interaction times, Row 3: 4 equidistant interactions, Row 4: 4 equidistant interactions.



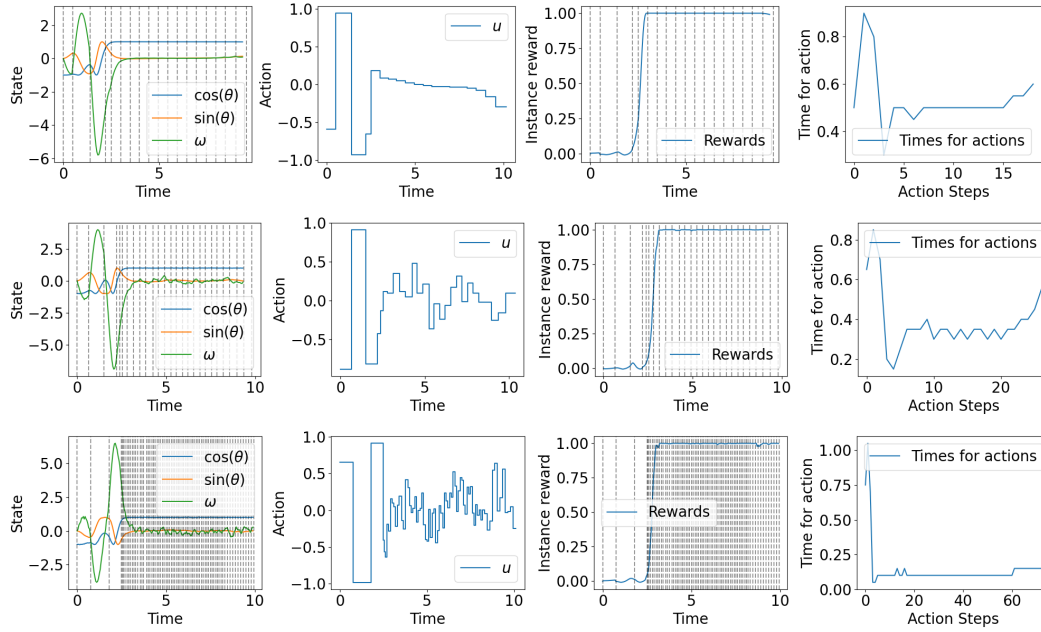


Figure 7: When stochasticity of the environments increases, we need more interactions at the unstable equilibrium (Pendulum on top). The stochasticity scale goes from 0.1 to 0.5 to 1.0 from top to bottom row respectively.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? paper's contributions and scope?

Answer: [Yes]

Justification: We highlight the problem setting, algorithm, and theoretical and empirical results in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: In Section 5 we highlight the assumptions of our work, which also correspond to the limitations of our theoretical analysis and also the setting for which our algorithm yields theoretical guarantees.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are accompanied by the relevant assumptions that are listed in Section 5 and we provide all proofs in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the code with all hyperparameters we used to run the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code with all the hyperparameters to run the experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the experimental details either, minor part in the main paper, and the major part in the accompanying code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experiments are run with 5 seeds and mean performance with standard error is reported in all our plots.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the details of compute in the readme.txt file as part of the enclosed code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics [https://neurips.cc/public/EthicsGuidelines?](https://neurips.cc/public/EthicsGuidelines)

Answer: [Yes]

Justification: Our research conforms, in every respect, to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper proposes a method to improve exploration in reinforcement learning in the nonepisodic setting, and is not tied to specific applications. As such, it shares the many potential societal consequences that are associated with reinforcement learning and automation as a whole, spanning from environmental impact to concerns on ethics and alignment.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release high-risk data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all creators whose code we used in our experiments in Section 4.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code is documented where applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.