No-Regret Safe Learning for Online Nonlinear Control with Control Barrier Functions

Wenhao Luo[1], Wen Sun[2], and Ashish Kapoor[3]

***Abstract*— Reinforcement Learning (RL) and continuous nonlinear control has been successfully deployed in multiple domains of complicated sequential decision making tasks. However, given the exploration nature of the learning process and the presence of model uncertainty, it is challenging to apply them on safety-critical control tasks due to the lack of safety guarantee. On the other hand, while combining control-theoretical approaches with learning algorithms have shown promise in safe RL applications, the sample efficiency of safe data collection process for control is not well addressed. In this paper, we propose a *provably* sample efficient episodic safe learning framework for online control tasks that leverages safe exploration and exploitation in an unknown, nonlinear dynamical system. In particular, the framework 1) extends control barrier functions (CBFs) in a stochastic setting to achieve provable safety under uncertainty during model learning and 2) integrates an optimism-based exploration strategy to efficiently guide the safe exploration process with learned dynamics for *near optimal* control performance. We provide formal analysis on the episodic regret bound against the optimal controller and probabilistic safety with theoretical guarantee. Simulation results are provided to demonstrate the effectiveness and efficiency of the proposed algorithm.**

I. Introduction

The control of safety-critical system such robotic systems is a difficult challenge under uncertainty and lack of complete information in the real world applications. While Reinforcement Learning (RL) algorithms that seek for long-term reward maximization has achieved significant results in many continuous control tasks [1], [2], it has not yet been widely applied to safety-critical control tasks as the rigorous safety requirements may be easily violated by intermediate policies during policy learning. Safe RL approaches [3]–[6] with constraints satisfaction have been proposed to encode safety consideration in a modified optimality criterion or in the constrained policy exploration process with external knowledge, e.g. an accurate probabilistic system model [6], [7]. However, the effectiveness in preventing risky behaviors relies on the sufficient period of policy learning where the unsafe situations could happen in the early learning stage.

Very recently, integrating data-driven learning-based approach with model-based safe control approaches has received significant attention to achieve model uncertainty reduction while ensuring provable safety [8]–[15]. The process often involves safe policy exploration with data collection from a nominal dynamics model and iteratively reduce learned model uncertainty over time to expand certified safety region of the system's state space [8], [9], [11], [13], [14]. However, such exhaustive data collection for safe learning could suffer from poor scalability and low efficiency for primary task. For example, instead of densely sampling over the space, it may be more beneficial to guide the safe exploration process towards task-prescribed policy optimization. Recent work [12] incorporates the safe learning using Gaussian Process (GP) and CBF into a model-free RL framework (RL-CBF) so that the guided exploration process will not only learn model uncertainty impacting safe behaviors but also optimizing the policy performance. Nevertheless, there is no theoretical guarantee on the learning performance in terms of sample efficiency or the control performance for the primary task.

In this paper, we propose a *provably correct* method that handles both sample efficient safe learning and online nonlinear control task in partially unknown system dynamics. In particular, we develop an Optimism-based Safe Learning for Control framework that integrates 1) stochastic discrete-time control barrier functions (CBF) to ensure forward invariant safety under uncertainty, and 2) an optimism-based exploration strategy that enjoys a formally provable regret bound. We provide rigorous proofs on the guaranteed exploration efficiency, policy optimization performance, and safety during exploration at all times.

Our **main contributions** are: 1) a provably sample efficient episodic online learning framework that integrates safe model-based nonlinear control approaches with optimism-based exploration strategy to simultaneously achieve safe learning and policy optimization for online nonlinear control tasks, and 2) rigorous theoretical analysis of guaranteed safety under learned uncertainty and near-optimal online learning and policy performance with proved regret bound.

II. Preliminaries

A. Dynamical System and Stochastic Control

Consider the following partially *unknown* discrete-time control-affine system dynamics with state $x \in \mathcal{X} \subset \mathbb{R}^n$ and control input $u \in \mathcal{U} \subset \mathbb{R}^m$ for a discrete time index $h \in \mathbb{N}$

$$x_{h+1} = \hat{f}(x_h, u_h) + d(x_h, u_h) + \varepsilon_h, \quad \varepsilon_h \sim \mathcal{N}(0, \Sigma_\sigma) \quad (1)$$

where $\hat{f} : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}^n$ is the known nominal discrete dynamics affine in the control input as $\hat{f}(x_h, u_h) = \hat{F}(x_h) + \hat{G}(x_h)u_h$. We assume $\hat{F} : \mathbb{R}^n \mapsto \mathbb{R}^n$, $\hat{G} : \mathbb{R}^n \mapsto \mathbb{R}^{n \times m}$ are locally Lipschitz continuous and the relative degrees of the nominal model and the actual system are the same, which

[1]The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. wenhaol@cs.cmu.edu

[2]The authors are with the Computer Science Department, Cornell University, Ithaca, New York 14853, USA. ws455@cornell.edu

[3]The authors are with the Microsoft Corporation, Redmond, Washington 98033, USA. akapoor@microsoft.com

are common assumptions as in [13], [14]. $d : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}^n$ denotes the unmodelled part of the system dynamics which is unknown, and ε_h is i.i.d noise sampled from a known Multivariate Gaussian distribution with the covariance matrix $\Sigma_\sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, i.e., $\sigma_1, \ldots, \sigma_n$ are known to the learner. For notation simplicity, we denote the stochastic state transition as $P(\cdot|x, u)$.

In particular, we assume that $d(x, u)$ is modelled by the following nonlinear model, $d(x, u) := W^\star \phi(x, u)$ where $\phi : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}^r$ is a known nonlinear feature mapping and the linear mapping $W^\star \in \mathbb{R}^{n \times r}$ is the unknown system parameters that need to be learned.

The control task is described by a cost function. Given an immediate cost function $c : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}^+$, the primary task-prescribed objective can be defined as

$$\min_{\pi \in \Pi} J^\pi(x_0; c, W^\star) = \mathbb{E}\Big[\sum_{h=0}^{H-1} c(x_h, u_h) | \pi, x_0, W^\star \Big] \quad (2)$$

where $\mathbf{x}_0 \in \mathcal{X}$ is a given starting state and Π is some set of pre-defined feasible controllers. Each controller (or a policy) is a mapping $\pi \in \Pi : \mathcal{X} \mapsto \mathcal{U}$. We denote $J^\pi(x; c, W)$ as the expected total cost of a policy π under cost function c, initial state x_0, and the dynamical system in Eq. 1 whose $d(x, u)$ is parameterized by W. In order to achieve optimal task performance, we need to learn the unmodelled $d(x, u)$ by taking samples to approximate to the true linear mapping $W^\star$ and enforce safety constraints at all times.

B. Discrete-time Control Barrier Functions For Gaussian Dynamical Systems

Consider a stochastic Gaussian discrete-time dynamics described in Eq. 1. A desired safety set $x \in \mathcal{S} \subset \mathcal{X}$ can be denoted by the following safety function $h^s : \mathbb{R}^n \mapsto \mathbb{R}$

$$\mathcal{S} = \{x \in \mathbb{R}^n : h^s(x) \geq 0\} \quad (3)$$

Formally, a safety condition is forward invariant if $x_{h=0} \in \mathcal{S}$ implies $x_h \in \mathcal{S}$ for all $h > 0$ with some designed controller $u \in \mathcal{U}$. Control barrier functions (CBF) [16] are often used to derive such designed controllers that enforce the forward invariance of a set of the system state space.

Definition 1. *[Discrete-time Control Barrier Function under Known Gaussian Dynamics] Assume $h^s(\cdot)$ is L-Lipschitz continuous when $x \in \mathcal{X}$ is bounded. Given $\delta \in (0, 1)$ and horizon H, let $\mathcal{S}$ be the 0-superlevel set of $h^s : \mathbb{R}^n \to \mathbb{R}$ which is a continuously differentiable function. We call $h^s(\cdot)$ a stochastic discrete-time control barrier function (CBF) for dynamical system Eq. 1 if there exists a $\eta \in (0, 1)$, such that for all time steps $h = 0, \ldots, H-1$, given any $x \in \mathcal{S}$:*

$$\sup_{u \in \mathcal{U}} \Bigg[h^s\Big(\hat{f}(x, u) + d(x, u)\Big) - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta}\right)} - h^s(x) \Bigg] \geq -\eta h^s(x) \quad (4)$$

where $\bar{\sigma} = \max\{\sigma_1, \ldots, \sigma_n\}$.

Proposition 2 (Forward Invariant with High Probability)**.** *Consider a control barrier function $h^s(\cdot)$ in Definition 1. Given $x_0 \in \mathcal{S}$, consider any policy $\pi : \mathcal{X} \to \mathcal{U}$ such that at any state x, this policy outputs an action $u = \pi(x)$ that satisfies the constraint Eq. 4. Then executing π to generate a trajectory starting at x_0: $\tau = \{x_0, u_0, \ldots, x_{H-1}, u_{H-1}\}$, with probability at least $1 - \delta$ we have $h(x_h) \geq 0$ for all $h \in [H]$, i.e., all states on the trajectory belong to the safe set $\mathcal{S}$.*

C. Learning Objective

If we had knew the unmodelled dynamics $d(\cdot)$, i.e., the whole stochastic Gaussian dynamical system in Eq. 1 is known, then the safe nonlinear control problem can be modeled as follows:

$$\min_{\pi \in \Pi} J^\pi(x_0; c), \quad (5)$$

$$s.t., \forall x \in \mathcal{X}, \quad h^s\Big(\hat{f}(x, \pi(x)) + d(x, \pi(x))\Big) - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta}\right)} - h^s(x) \geq -\eta h^s(x) \quad (6)$$

In our episodic finite horizon learning framework, we start with some initialization $\overline{W}_0$ which is used to parameterize $d_0(x, u) := \overline{W}_0 \phi(x, u)$ (we will discuss conditions on $\overline{W}_0$ in Section III-A that can ensure safety during the entire learning process). At every episode t, the learner will propose a policy $\pi_t \in \Pi$ (probably based on the current guess $d_t(x, u)$ with $\overline{W}_t$), execute π_t in the real system to generate one trajectory $\{x_0^t, u_0^t, \ldots, x_{H-1}^t, u_{H-1}^t\}$ for H time steps; the learner then incrementally updates model parameter to $\overline{W}_{t+1}$ using observations from all of the past trajectories, and move to the next episode $t + 1$ starting from the same initial state $x \leftarrow x_0$. The ideal goal of the learner is to ensure that π_t is safe (with high probability) in terms of satisfying CBF constraint Eq. 4, and also optimize the cost function over episodes:

$$\text{Regret}_T := \sum_{t=0}^{T-1} \sum_{h=0}^{H-1} c(x_h^t, u_h^t) - \sum_{t=0}^{T-1} J^{\pi^\star}(x_0; c) = o(T), \quad (7)$$

Namely, comparing to the best policy $\pi^\star$ (i.e., the optimal solution of the constrained optimization program in Eq. 5 if assuming perfect model information), the cumulative regret grows sublinearly with respect to the number of episodes T. To that end, the goal in this paper is to minimize the cumulative regret in Eq. 7 subject to safety constraint in Eq. 6 at all times in each episode. Next, we will discuss how to enforce such safety constraint with d learned online and provide the episodic safe learning algorithm to achieve bounded regret in Eq. 7 with rigorous analysis.

III. Algorithm and Analysis

A. Approximate Safety Guarantee

We compute $\overline{W}_0$ (the initialization parameters of $d(x, u)$) via ridge linear regression under known feature mapping

$\phi : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}^r$:

$$\overline{W}_0 = \arg\min_W \sum_{i=1}^{N} \left\| W\phi(x_i, u_i) - (x_i' - \hat{f}(x_i, u_i)) \right\|_2^2 + \lambda \|W\|_F^2 \tag{8}$$

where λ is a regularizer parameter and $\|W\|_F$ is the Frobenius norm of the model parameter matrix $W \in \mathbb{R}^{n \times r}$. Denote the initial empirical regularized covariance matrix as

$$V_0 = \sum_{i=1}^{N} \phi(x_i, u_i)\phi(x_i, u_i)^\top + \lambda I \tag{9}$$

The following assumption states that we will have $d_0(x, u) = \overline{W}_0\phi(x, u)$ as a reasonable good estimate of $d(x, u) = W^\star\phi(x, u)$ for all $x, u \in \mathcal{X} \times \mathcal{U}$ (note that however we cannot guarantee $\overline{W}_0$ will be close to $W^\star$ in terms of ℓ_2 norm).

Assumption 3. *After deriving $\overline{W}_0, V_0$ in Eq. 8, 9 from the initial data $(x_i, u_i, x_i')_{i=1}^N$, we can build the initial confidence ball describing the uncertain region of $W^\star$ as follows:*

$$\textit{Ball}_0 = \left\{ W : \left\| (W - \overline{W}_0)V_0 \right\|_2 \le \beta, \quad \|W\|_2 \le \|W^\star\|_2 \right\} \tag{10}$$

where β is the confidence radius as $\beta := \sqrt{\lambda}\|W^\star\|_2 + \sigma\sqrt{8n\ln(5) + 8d\ln(1 + N/\lambda) + 8\ln(1/\delta)}$.

Then in the following we can show that for any $\widetilde{W} \in \text{Ball}_0$, it's prediction $\widetilde{d}(x, u) = \widetilde{W}\phi(x, u)$ for any given x, u is close to the true prediction $d(x, u) = W^\star\phi(x, u)$ from $W^\star$.

Lemma 4. *Assume the condition in Assumption 3 holds. For all $\widetilde{W} \in \textit{Ball}_0$, we have:*

$$\forall x, u \in \mathcal{X} \times \mathcal{U} : \ \left\| \left(\widetilde{W} - W^\star \right) \phi(x, u) \right\|_2 \le \mathcal{O}(\epsilon).$$

This ensures that when we control our dynamical system using CBF with any model $\widetilde{W} \in \text{Ball}_0$, we can ensure safety update to $\mathcal{O}(\epsilon)$.

Theorem 5 (Policy for Approximate Safety Guarantee with Learned Dynamics)**.** *Assume the conditions in Assumption 3 hold. Consider any $\widetilde{W} \in \textit{Ball}_0$, and define any policy $\pi_s : \mathcal{X} \mapsto \mathcal{U}$ that satisfies the CBF constraint parameterized by $\widetilde{W}$, i.e.,*

$$\forall x \in \mathcal{X} : \pi_s(x) \in \mathcal{U}_x := \left\{ u : h^s\left(\hat{f}(x, u) + \widetilde{W}\phi(x, u) \right) - L\bar{\sigma}\sqrt{2n\ln\left(\frac{Hn}{\delta}\right)} \ge (1 - \eta)h^s(x) \right\} \tag{11}$$

Then with probability at least $1 - \delta$, starting at any safe initial state $h^s(x_0) \ge 0$, π_s generates a safe trajectory $\{x_0, u_0, \ldots, x_{H-1}, u_{H-1}\}$, such that for all time steps $h \in [H]$, $h^s(x_h) \ge -\mathcal{O}(\frac{L\epsilon}{\eta})$, where L is the Lipschitz constant of $h^s(\cdot)$ under bounded $x \in \mathcal{X}$.

Thus the initialization step narrows down the search region for $W^\star$ and we have $W^\star \in \text{Ball}_0$ with probability at least $1 - \delta$. Later on, when we improve our model during iterative learning, as long as we restrict our model $\widetilde{W}$ to Ball_0, we guarantee that any policy that satisfies the CBF constraint under $\widetilde{W}$ (Eq. 11) is guaranteed to be approximately safe in the sense of Theorem 5. Now we move to iterative learning where we aim to search for a policy using an optimism-based algorithm that performs as good as the best benchmark $\pi^\star$ in the sense of minimizing regret defined in Eq. 7 and subject to Eq.11.

B. Optimism-based Safe Learning for Control with Regret Analysis

To achieve no-regret performance for efficient safe learning for control, we leverage the LC^3 algorithm developed in [17] for strategic exploration. Here we modify the policy selection step in LC^3 to take our CBF constraint Eq. 5 into consideration and thus ensures approximate safety (i.e., Theorem 5). Meanwhile, similar to LC^3, we also need to leverage the principle of optimism in the face of uncertainty to achieve small regret and with safety guarantee. We propose the framework of Optimism-based Safe Learning for Control (Algorithm 1) that seeks to minimize the cumulative regret for optimal online control performance with safety guarantee.

Algorithm 1 Optimism-based Safe Learning for Control

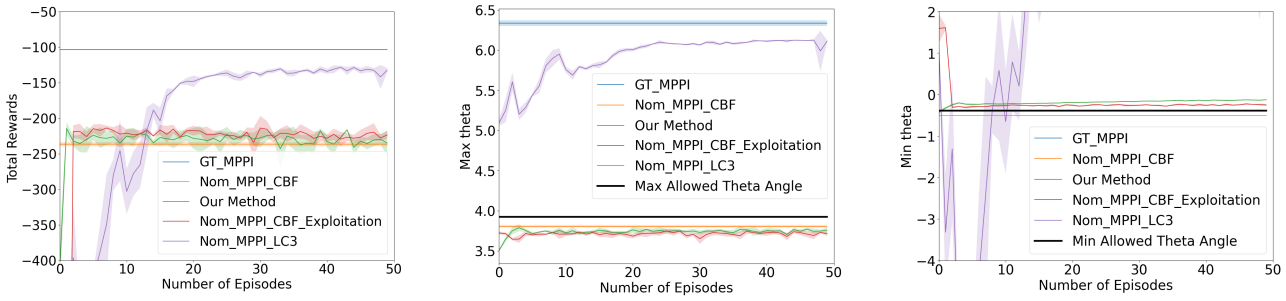
Input: CBF h^s, cost function c, initial data $(x_i, u_i, x_i')_{i=1}^N$, initial confidence region Ball_0 with $\overline{W}_0, \Sigma_0$, number of training episodes T, horizon H, regularizer λ, initial state x_0

Output: a sequence of policies for $t = 0, ..., T$

1: **for** $t = 0, \ldots, T$ **do**
2: $x_0^t \leftarrow x_0$
3: Sample $\widetilde{W}_t \sim \mathcal{N}(\overline{W}_t, \Sigma_t^{-1})$ `# Thompson Sampling for Exploration`
4: $\pi_s^t \leftarrow \arg\min_{\pi \in \Pi_{\widetilde{W}}} J^\pi(x_0^t; c, \widetilde{W}_t)$ `# Safe MPC Planning`
5: Execute π_s^t to sample a trajectory $\tau^t := \{x_h^t, u_h^t, c_h^t, x_{h+1}^t\}_{h=0}^{H-1}$ `# Execution and Data Collection`
6: $\overline{W}_{t+1}, \Sigma_{t+1} \leftarrow$ Update Ball_{t+1} `# Model Update`
Return a sequence of policies for $t = 0, ..., T$

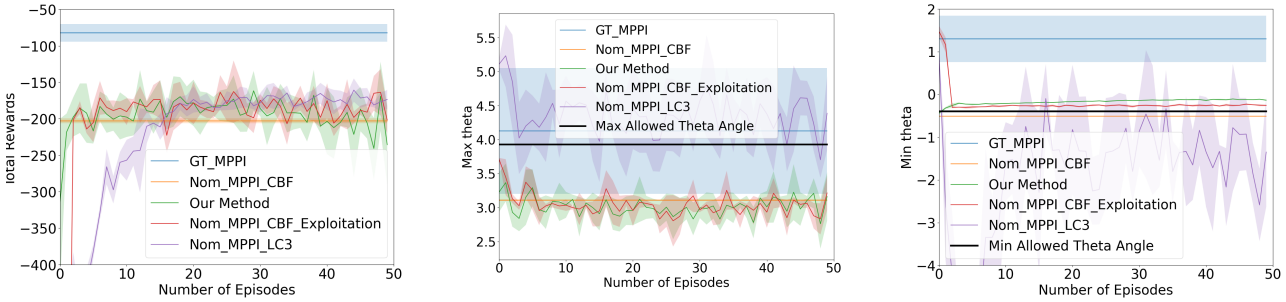
LC^3 [17] shows that with probability $1 - \delta$, for all t, $W^\star \in \{W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \le \beta_t\}$. In our Theorem we prove that with probability at least $1-\delta$, $W^\star \in \text{Ball}_0$. Hence it is not hard to see that with probability at least $1-2\delta$, we have $W^\star \in \text{Ball}_0 \cap \left\{ W : \|(W - \overline{W}_t)\Sigma_t^{1/2}\|_2 \le \beta_t \right\} := \text{Ball}_t$ in our case.

Then we consider the safety constraint. Given any model $\widetilde{W} \in \text{Ball}_t$, we constrain our policy class Π based on the CBF constraint under $\widetilde{W}$ (Eq. 11), i.e., we denote $\Pi_{\widetilde{W}}$ as



(a) (b) (c)

Fig. 1: Performance curves of (a) cumulative rewards, (b) maximum theta angle, and (c) minimum theta angle in Inverted Pendulum environment testing under the same initial condition.



(a) (b) (c)

Fig. 2: Performance curves of (a) cumulative rewards, (b) maximum theta angle, and (c) minimum theta angle in Inverted Pendulum environment with different initial conditions.

follows:

$$\Pi_{\widetilde{W}} = \Big\{ \pi_s \in \Pi : \forall x \in \mathcal{X}, \pi_s(x) \in \Big\{ u : h^s\left(\hat{f}(x,u) + \widetilde{W}\phi(x,u)\right) - L\bar{\sigma}\sqrt{2n \ln\left(\frac{Hn}{\delta}t\right)} \geq (1-\eta)h^s(x) \Big\} \Big\} \tag{12}$$

With this now we select our model and policy optimistically at each episode t, i.e.,

$$\left(W_t, \pi^t\right) := \underset{\widetilde{W} \in \text{Ball}_t}{\arg\min} \underset{\pi \in \Pi_{\widetilde{W}}}{\arg\min} J^{\pi}(x_0^t; c, \widetilde{W}). \tag{13}$$

Then given Eq. 13 and conditioned on the high probability event that $W^\star \in \text{Ball}_t$, and $\pi^\star \in \Pi_{W^\star}$ by definition of $\pi^\star$, we can easily show optimism in the sense that:

$$J^{\pi^t}\left(x_0^t; c, W_t\right) \leq J^{\pi^\star}(x_0; c, W^\star).$$

The optimism allows us to prove the following main statement.

Theorem 6 (Main Result)**.** *Set $\lambda = \bar{\sigma}^2/\|W^\star\|_2^2$. Our algorithm learns a sequence of policies $\pi^0, \ldots, \pi^{T-1}$ in T episodes, such that in expectation, we have:*

$$\mathbb{E}\left[Regret_T\right] \leq \widetilde{\mathcal{O}}\left(H\sqrt{Hr(r+n+H)T}\right).$$

Also with probability at least $1 - O(\delta)$, we have that for all $t \in [T], h^s \in [H]$, $h(x_h^t) \geq -\mathcal{O}(L\epsilon/\eta)$.

With the setup of the optimism, the proof for the regret bound part of the above theorem is mainly following the proof of the main theorem of LC^3 from [17]. The proof for the safety part of the above theorem comes from the fact that $W^t \in \text{Ball}_0$ via the selection rule and the definition of Ball_t, and based on Theorem 5, we know that each trajectory is approximately safe with high probability.

IV. Results

We use the inverted pendulum modified from the OpenAI gym environment [18] with additive disturbance of $0.05\cos(\theta_t - 3)$ on state update to demonstrate the learning performance for control task. The pendulum has ground truth mass $m = 1$ and length $l = 1$, and is controlled by the limited torque input $u \in [-15, 15]$. The standard cost function $c = \theta^2 + 0.1\dot{\theta}^2 + 0.001$ is used to learn the optimal policy keeping the pendulum upright (i.e. $\theta = 0$). The control barrier functions $h_1^s = \theta + 1/8\pi \geq 0$ and $h_2^s = 5/4\pi - \theta \geq 0$ are designed to describe the safety constraint $\theta \in [-1/8\pi, +5/4\pi]$ radians. We define the true

system dynamics as $\theta_{t+1} = \theta_t + \dot{\theta}_{t+1}\Delta t + 0.05\cos(\theta_t - 3)$ and $\dot{\theta}_{t+1} = \dot{\theta}_t + \frac{3g}{2l}\sin\theta_t \Delta t + \frac{3}{ml^2}u\Delta t$.

To describe the partially known system dynamics, we assume a nominal model as $\theta_{t+1} = \theta_t + \dot{\theta}_{t+1}\Delta t$ and $\dot{\theta}_{t+1} = \dot{\theta}_t + \frac{3g}{2l'}\sin\theta_t \Delta t + \frac{3}{m'l'^2}u\Delta t$ with incorrect model parameters $m' = 1.8, l' = 1.8$ available to the learner (hence 80% error in model parameters). Using the same and different initial conditions respectively, Figure 1 and Figure 2 compare the cumulative rewards, maximum and minimum theta angle achieved during testing after each training episode by using (1) MPPI [19] with ground-truth dynamics model (GT-MPPI), (2) MPPI with nominal dynamics model and CBF (Nom-MPPI-CBF), (3) our method of optimism-based safe learning (Algorithm 1), (4) our method with exploitation only, i.e. replace Line 3 in Algorithm 1 by $\widetilde{W}_t \leftarrow \overline{W}_t$ (Nom-MPPI-CBF-Exploitation), and (5) unconstrained Lower Confidence-based Continuous Control algorithm (LC3) [17]. The last three learning-based algorithms are trained for 50 episodes with 20 testing trials after each training episode averaged from four random seeds. It is observed that our method quickly increased cumulative reward in early stage while satisfying the safety constraints as learning process evolves, and our method using exploration behavior (our method) is able to increase reward even faster than our method using exploitation behavior (Nom-MPPI-CBF-Exploitation), empirically implying sample efficiency. In contrast, GT-MPPI and LC3 severely violate angle limitation due to lack of safety consideration, and safe MPPI using CBF with nominal model (Nom-MPPI-CBF) still violates safety constraints with lower cumulative rewards due to the inaccurate nominal model with large error.

V. Conclusion

In this paper, we address the problem of episodic safe learning for online nonlinear control tasks. Unlike previous safe learning and control approaches that exhaustively expanding safety region or optimizing policy performance without efficiency guarantee, we propose an optimism-based online safe learning framework that simultaneously achieve sample efficient learning for safe behaviors and nonlinear control optimization with bounded regret guarantee. Future work include real-world implementations to solve more complex robotic control task with uncertainty and extensions to sample efficient robotic dynamics learning with higher relative degrees system.

References

[1] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *International Conference on Machine Learning*, 2016, pp. 1329–1338.

[2] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[3] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[4] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 22–31.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[6] T. M. Moldovan and P. Abbeel, "Safe exploration in markov decision processes," *arXiv preprint arXiv:1205.4810*, 2012.

[7] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration in finite markov decision processes with gaussian processes," *Advances in Neural Information Processing Systems*, vol. 29, pp. 4312–4320, 2016.

[8] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in neural information processing systems*, 2017, pp. 908–918.

[9] L. Wang, E. A. Theodorou, and M. Egerstedt, "Safe learning of quadrotor dynamics using barrier certificates," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2460–2465.

[10] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2018.

[11] M. J. Khojasteh, V. Dhiman, M. Franceschetti, and N. Atanasov, "Probabilistic safety constraints for learned high relative degree system dynamics," in *Learning for Dynamics and Control*, 2020, pp. 781–792.

[12] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3387–3395.

[13] A. Taylor, A. Singletary, Y. Yue, and A. Ames, "Learning for safety-critical control with control barrier functions," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 708–717.

[14] J. Choi, F. Castañeda, C. Tomlin, and K. Sreenath, "Reinforcement Learning for Safety-Critical Control under Model Uncertainty, using Control Lyapunov Functions and Control Barrier Functions," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[15] M. Ohnishi, L. Wang, G. Notomista, and M. Egerstedt, "Barrier-certified adaptive reinforcement learning with applications to brushbot navigation," *IEEE Transactions on robotics*, vol. 35, no. 5, pp. 1186–1205, 2019.

[16] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *18th European Control Conference (ECC)*. IEEE, 2019, pp. 3420–3431.

[17] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun, "Information theoretic regret bounds for online nonlinear control," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[18] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[19] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic mpc for model-based reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1714–1721.