TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Han, M., Zhang, L., Liu, C., Zhou, Z., Wang, J., & Pan, W. (2026). Reinforcement learning for control with probabilistic stability guarantee: A finite-sample approach. *Automatica*, *188*, Article 112964.
https://doi.org/10.1016/j.automatica.2026.112964

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Automatica 188 (2026) 112964

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief Paper

Reinforcement learning for control with probabilistic stability guarantee: A finite-sample approach☆

Minghao Han [a], Lixian Zhang [b,*], Chenliang Liu [c], Zhipeng Zhou [d], Jun Wang [e], Wei Pan [f]

[a] Department of Control Science and Engineering, Harbin Institute of Technology, China
[b] State Key Laboratory of Robotics and Systems (HIT), Harbin Institute of Technology, China
[c] School of Automation, Central South University, China
[d] Department of Cognitive Robotics, Delft University of Technology, Netherlands
[e] Department of Computer Science, University College London, UK
[f] Department of Computer Science, University of Manchester, UK

ARTICLE INFO

Article history:
Received 18 December 2024
Received in revised form 23 November 2025
Accepted 9 February 2026

Keywords:
Reinforcement learning
Finite sample
Probabilistic bound
Nonlinear control
Lyapunov's method

ABSTRACT

This paper presents a novel approach to reinforcement learning (RL) for control systems that provides probabilistic stability guarantees using finite data. Leveraging Lyapunov's method, we propose a probabilistic stability theorem that ensures mean square stability using only a finite number of sampled trajectories. The probability of stability increases with the number and length of trajectories, converging to certainty as data size grows. Additionally, we derive a policy gradient theorem for stabilizing policy learning and develop an RL algorithm, L-REINFORCE, that extends the classical REINFORCE algorithm to stabilization problems. The effectiveness of L-REINFORCE is demonstrated through simulations on a Cartpole task, where it outperforms the baseline in ensuring stability. This work bridges a critical gap between RL and control theory, enabling stability analysis and controller design in a model-free framework with finite data.

© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Reinforcement learning (RL) has achieved superior performance on some complicated control tasks that are challenging for traditional control engineering methods (Kaufmann et al., 2023; Shi et al., 2025; Xia et al., 2025). An optimal controller can be learned through trial and error (Zhao et al., 2023). However, without using a mathematical model of the system, securing the stability of the closed-loop system still remains a challenging issue for the sample-based RL methods.

The most effective and general approach for the stability analysis of a dynamical system is Lyapunov's method (Lyapunov, 1892). As a basic tool in control theory, the construction/learning of the Lyapunov function is not trivial, and many works are devoted to this problem. In Perkins and Barto (2002), the RL agent controlled the switch between designed controllers using Lyapunov domain knowledge so that any policy is safe and reliable. Petridis and Petridis (2006) proposes a straightforward approach for constructing the Lyapunov functions for nonlinear systems using neural networks. Richards et al. (2018) proposed a learning-based approach for constructing Lyapunov neural networks with the maximized region of attraction. However, these approaches explicitly require a model of the system dynamics. In the well-established control theory, the mathematical model of the dynamics plays an essential role in stability analysis and controller design. However, in model-free learning methods, as the dynamic model is unknown, the Lyapunov condition has to be verified by trying out *all* possible consecutive data pairs in the state space, which makes it impractical to directly exploit Lyapunov's method in a model-free framework. In Berkenkamp et al. (2017), local stability was analyzed by validating the "energy decreasing" condition on discretized points in the subset of state space with the help of a learned model, meaning that only a finite number of inequalities need to be checked. Nevertheless, the discretization technique may become infeasible as the dimension and space of interest increase, limiting its application to rather simple and low-dimensional systems.

More recently, the concept of control with stability guarantee has been explored in the context of model-free RL. In Chow et al. (2018), a Lyapunov-based approach for solving constrained

☆ This work was supported in part by National Natural Science Foundation of China (62225305, 62527807, 62503507), in part by the State Key Laboratory of Robotics and Systems (HIT) under Grant SKLRS-2025-KF-05. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Jorge I. Poveda under the direction of Editor Maria Elena Valcher.

* Corresponding author.
E-mail addresses: ratsos777@gmail.com (M. Han), lixianzhang@hit.edu.cn (L. Zhang), lcliang@csu.edu.cn (C. Liu), zhouzhipeng113@mails.ucas.ac.cn (Z. Zhou), jun.wang@cs.ucl.ac.uk (J. Wang), wei.pan2@newcastle.ac.uk (W. Pan).

https://doi.org/10.1016/j.automatica.2026.112964
0005-1098/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Markov decision processes was proposed with a novel way of constructing the Lyapunov function through linear programming. It should be noted that even though Lyapunov-based methods were adopted, the stability of the system was not addressed. In Postoyan et al. (2017), an initial result was proposed for the stability analysis of deterministic nonlinear systems with an optimal controller for infinite-horizon discounted cost, based on the assumption that the discount is sufficiently close to 1. However, in practice, it is rather difficult to guarantee the optimality of the learned policy unless certain assumptions on the system dynamics are made (Jiang & Jiang, 2015). In Huang et al. (2024), a data-driven output tracking design for unknown linear discrete-time systems was investigated to achieve zero tracking error and user-specified convergence rate. It was shown in Han et al. (2020) that the stability of Markov decision process (MDP) can be analyzed in a model-free and data-driven manner with the Lyapunov method, which is further incorporated with RL for stabilization tasks. Chang et al. (2019) proposed an algorithmic framework to learn control policies and neural network Lyapunov functions for the general class of deterministic nonlinear systems. Chang and Gao (2021) further integrated the neural Lyapunov method proposed in Chang et al. (2019) with model-free RL to learn stabilizing controllers based on data. Ganai et al. (2023) proposed a novel model-free method to accelerate Learning from Observations based on the framework in Chang et al. (2019) for the stabilization control problems in robotics. Dawson et al. (2022) developed a model-based learning approach to synthesize robust feedback controllers based on a robust extension to the theory of control Lyapunov barrier functions. These papers represent an important stream of work in improving the reliability of learning-based control for complex nonlinear systems. However, the results above relied on an infinite amount of data to establish the stability guarantee, which is not practical in reality. As a result, how to analyze stability based on a finite amount of data remains an open question.

On the topic of finite sample-based RL for MDPs, many efforts have been made, and significant progress has been achieved. Kearns and Singh (1998) investigated the convergence rate of Q learning on MDPs with respect to the number of samples. Lazaric et al. (2012) derived the performance bound and conducted the finite-sample analysis for least-squares temporal-difference learning of control policies. Yang et al. (2018) analyzed the finite-sample performance of the batch actor-critic algorithm. The finite sample analysis of many more classical RL algorithms is referred to Chen et al. (2021), J. Zhang et al. (2021), S. Zhang et al. (2021), Zou et al. (2019). Nonetheless, these results typically characterize the performance of the control policy in terms of the discounted sum of rewards given finite samples, while the notion of stability of the closed-loop system is missing and unaddressed.

Another line of work focuses on the finite sample analysis of stochastic approximation (SA) (Robbins & Monro, 1951), which forms the basis of many machine learning problems. Chen et al. (2022) studied a nonlinear stochastic approximation algorithm under Markovian noise and established its finite-sample convergence bounds under various step sizes. The theoretical results were further used to derive the finite-sample bounds of the Q-learning algorithm with linear function approximation (Chen et al., 2022). A rich literature can be found on finite-sample analysis of the SA algorithm with application to RL of MDPs (Chen et al., 2020; Dalal et al., 2018; Huang & Haskell, 2020). However, these works focused on the convergence of the learning process with respect to the number of samples, and the convergence of the states was not considered. The finite-sample analysis for the identification of nonlinear stochastic systems is also of interest for statistical learning (Tsiamis & Pappas, 2019; Tsiamis et al., 2023), though the topic of stability of the system was not a focus. The stability of Markov chains, along with the necessary and sufficient conditions, has been long investigated (Khasminskii, 2012; Meyn & Tweedie, 2012). Nevertheless, a finite-sample-based stability analysis for the MDPs is still missing.

In this paper, we show that the mean square stability of the system can be analyzed based on a finite number of samples without knowing the model of the system. Furthermore, we show how the analysis enables the development of an RL algorithm that learns a stabilizing control policy. The contributions of this paper are summarized as follows:

1. Instead of using infinite sample pairs, given a number of M trajectories of T steps in length, a stability theorem is proposed to provide a probabilistic stability guarantee for the system. The probability is an increasing function of the number M and length T of sampled trajectories, and it converges to 1 as M and T grow.
2. As an independent interest, we also derive the policy gradient theorem for learning a stabilizing policy with sample pairs and the corresponding algorithm. We further reveal that the classic REINFORCE algorithm (Williams, 1992) is a special case of the proposed algorithm for the stabilization problem.

The paper is organized as follows: In Section 2, the problem statement and the preliminary results are given. In Section 3, we propose the probabilistic stability guarantee when only a finite number of samples are accessible and the probabilistic bound is derived. In Section 4, based on the theoretical results, the policy gradient is derived and a model-free RL algorithm (L-REINFORCE) is given. In Section 5, the vanilla version of L-REINFORCE is tested on a simulated Cartpole stabilization task to demonstrate its effectiveness.

2. Preliminaries

2.1. Problem statement

In this work, we focus on the control of stochastic and nonlinear systems characterized by the following Markov decision process (MDP),

$$s_{t+1} \sim P(s_{t+1}|s_t, a_t), \forall t \in \mathbb{Z}_+ \tag{1}$$

where $s \in \mathcal{S} \subseteq \mathbb{R}^n$ denotes the state and $a \in \mathcal{A}$ denotes the action. The state space $\mathcal{S}$ is continuous and the action space $\mathcal{A}$ is finite. $P(s_{t+1}|s_t, a_t)$ denotes the state transition probability density function. The control policy $\pi(a|s)$ is the probability of taking the action $a \in \mathcal{A}$ given the state $s \in \mathcal{S}$. The distribution of the initial state $s_1 \in \mathcal{S}$, is characterized by the probability density function $\rho(s)$.

The mean square stability (MSS) of the stochastic systems described in (1) is defined as follows.

Definition 1 (*Bolzern et al., 2010*)**.** The stochastic system (1) is said to be mean square stable (MSS) if there exists a positive constant b such that $\lim_{t\to\infty} \mathbb{E}_{s_t} \|s_t\|_2^2 = 0$ holds for any initial condition $s_1 \in \{s_1 | \|s_1\|_2^2 \leq b\}$. If b is arbitrarily large then the stochastic system is globally mean square stable (GMSS).

The aim of this paper is twofold. First, we aim to conduct a finite sample stability analysis for the system (1) and derive the probabilistic lower bound. This probabilistic lower bound goes to 1 as the amount of data approaches infinity.

Second, we aim to propose an RL algorithm to learn a stabilizing control policy. The goal of the RL algorithm is to find a control policy π such that the system in (1) is MSS with a probability

lower bounded by δ. The mathematical formulation is given as follows,

$$\text{find } \pi, \text{ s.t. } \mathbb{P}\left(\lim_{t\to\infty} \mathbb{E}_{s_t} \|s_t\|_2^2 = 0\right) \geq \delta \tag{2}$$

Before proceeding, some notations are to be defined. We introduce $c(s) := \min(\|s\|_2^2, \bar{c}), \bar{c} > 0$ to denote the clipped norm of state. The closed-loop transition probability density function is denoted as $P_\pi(s'|s) := \mathbb{E}_{a\sim\pi} P(s'|s,a)$. We also introduce the closed-loop state distribution at a certain instant t as $P(s|\rho,\pi,t)$, which could be defined iteratively: $P(s'|\rho,\pi,t+1) = \int_{\mathcal{S}} P_\pi(s'|s)P(s|\rho,\pi,t)\mathrm{d}s, \forall t \in \mathbb{Z}_{[1,\infty)}$ and $P(s|\rho,\pi,1) = \rho(s)$.

2.2. Sample-based stability analysis

In this subsection, we will introduce the results of stability analysis based on an infinite amount of data. First, the following assumption is commonly exploited in RL literature (Sutton et al., 2009).

Assumption 1. There exists a unique stationary distribution $q_\pi(s) = \lim_{t\to\infty} P(s|\rho,\pi,t)$, for the MDP in (1) under the policy π and initial state distribution ρ.

A further assumption on the initial state distributions is needed for the sample-based stability analysis theorem (Han et al., 2020).

Assumption 2. There exists a positive constant b such that for any $s \in \{s|c(s) \leq b\}$, the initial state probability density function $\rho(s) > 0$.

Remark 1. Assumption 1 essentially assumes that there exists a stationary state distribution q_π, which is commonly assumed in the RL literature (Chen et al., 2021; Sutton et al., 2009; J. Zhang et al., 2021; S. Zhang et al., 2021; Zou et al., 2019). In this work, the assumption is made on the pair consisting of the MDP and the policy π. We do not require that Assumption 1 hold for any control policy. Assumption 2 requires that there exists a sublevel set specified by b, where the probability density function of the initial state distribution $\rho(s) > 0$ for any state. According to Han et al. (2020), this assumption is to ensure that the states in the region of attraction are accessible in a data-based stability theorem. If $\rho(s) > 0$ for any $s \in \mathcal{S}$, then b can be an arbitrary positive value.

Based on the above assumptions, one can exploit Lyapunov's method to prove the sample-based stability theorem. In this paper, we construct the Lyapunov function using the following parameterization,

$$L(s) = (f_\phi(s) - f_\phi(0))^2 + \sigma c(s) \tag{3}$$

where $f_\phi(s)$ is a fully connected neural network (NN) with ReLU activation function. ϕ denotes the parameters of the network and σ is a small positive constant. We introduce the following sample-based stability analysis lemma from (Han et al., 2020).

Lemma 1 (*Han et al., 2020*)**.** *Assume that Assumptions 1–2 hold for system (1). The stochastic system (1) is mean square stable if there exists a function $L : \mathcal{S} \to \mathbb{R}_+$ and positive constants α_1, α_2 and α_3, such that*

$$\alpha_1 c(s) \leq L(s) \leq \alpha_2 c(s) \tag{4}$$

$$\mathbb{E}_{s\sim\mu_\pi}(\mathbb{E}_{s'\sim P_\pi} L(s') - L(s) + \alpha_3 c(s)) \leq 0 \tag{5}$$

where

$$\mu_\pi(s) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P(s|\rho,\pi,t) \tag{6}$$

is the infinite sampling distribution.

It can be found that in Lemma 1, the infinite number of energy-decreasing conditions are replaced by only a single sample-based inequality (5). However, the validation of stability through a sample-based approach comes with a cost: it theoretically requires a tremendous, if not infinite, number of samples to thoroughly estimate the state distributions at instants from 0 to infinity, which is impractical.

3. Finite-sample stability analysis

In this section, we will show that a finite number of samples should be informative enough to guarantee stability with a certain probability. More specifically, the probabilistic stability bound will be given by closing the gap between infinite and finite-sample guarantees.

To estimate the μ_π in Lemma 1, an infinite number of trajectories of infinite time steps are needed, whereas in practice only M trajectories of T time steps are accessible. Thus, in this section, we will first introduce the finite-time sampling distribution (FSD) $\mu_\pi^T := \frac{1}{T}\sum_{t=1}^{T} P(s|\rho,\pi,t)$, as an intermediate to study the effect of the finite sample-based estimation. Apparently, $\lim_{T\to\infty} \mu_\pi^T = \mu_\pi$.

The general idea of exploiting μ_π^T is: we first derive the deviation of $\mathbb{E}_{\mu_\pi^T} \Delta L(s)$ from $\mathbb{E}_{\mu_\pi} \Delta L(s)$ with respect to T, where

$$\Delta L(s) := \mathbb{E}_{s'\sim P_\pi} L(s') - L(s) + \alpha_3 c(s) \tag{7}$$

then we study the effect of estimating $\mathbb{E}_{\mu_\pi^T} \Delta L(s)$ with sample average and derive the probabilistic bound. Finally, the above effects are unified to propose the finite sample-based stability guarantee.

Now, we first close the first gap in terms of deviation between $\mathbb{E}_{\mu_\pi^T} \Delta L(s)$ and $\mathbb{E}_{\mu_\pi} \Delta L(s)$. To quantitatively analyze this effect, we introduce the following assumption.

Assumption 3. There exists a constant $\gamma \in (0,1)$ such that for any π

$$\sum_{t=1}^{T} \|P(s|\pi,\rho,t) - q_\pi(s)\|_1 \leq 2T^\gamma, \forall\, T \in \mathbb{Z}_+ \tag{8}$$

where $\|P(s|\pi,\rho,t) - q_\pi(s)\|_1$ denotes the L_1-distance between probability measures P and q_π.

Remark 2. It should be noted that the Assumption 3 can be satisfied for ergodic MDPs. q_π denotes the stationary state distribution, it naturally follows that $\sum_{t=1}^{T} \|P(s|\pi,\rho,t) - q_\pi(s)\|_1 \leq 2T^{\gamma(T)} \leq 2T$, where $\gamma(T) \in [0,1]$ without any further assumption. The assumption proposed replaces this time-varying $\gamma(T)$ with a constant. Many existing methods assume that the state distribution converges to q_π exponentially with respect to time (Chen et al., 2022; Zou et al., 2019), and this assumption holds for irreducible and aperiodic Markov chains (Levin & Peres, 2017). In our derivation, we do not need the state distribution to uniformly converge to q_π at exponential speed, and therefore relaxed the assumption on the convergence rate. Nevertheless, Assumption 3 allows us to give the quantitative bound for the deviation between $\mathbb{E}_{\mu_\pi} \Delta L(s)$ and $\mathbb{E}_{\mu_\pi^T} \Delta L(s)$ with respect to T.

Based on Assumption 3, we introduce the following Lemma.

Lemma 2. *Assume that Assumptions 1 and 3 hold for system (1). If there exist positive constants α_1, α_2 such that (4) hold, then*

$$\left|\mathbb{E}_{\mu_\pi} \Delta L(s) - \mathbb{E}_{\mu_\pi^T} \Delta L(s)\right| \leq 2\bar{c}(\alpha_3+\alpha_2)T^{\gamma-1} \tag{9}$$

Proof. First of all, we can quantify the variation of $\mathbb{E}_{\mu_\pi} \Delta L(s) - \mathbb{E}_{\mu_\pi^T} \Delta L(s)$ with respect to T.

$$\begin{aligned}
&\mathbb{E}_{\mu_\pi} \Delta L(s) - \mathbb{E}_{\mu_\pi^T} \Delta L(s)\\
=&\int_{\mathcal{S}} \left(\mu_\pi(s) - \frac{1}{T}\sum_{t=1}^{T} P(s|\rho,\pi,t)\right) \Delta L(s)\mathrm{d}s &(10)\\
=&\frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{S}} \left(\mu_\pi(s) - P(s|\rho,\pi,t)\right) \Delta L(s)\mathrm{d}s &(11)
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
&\left|\mathbb{E}_{\mu_\pi} \Delta L(s) - \mathbb{E}_{\mu_\pi^T} \Delta L(s)\right|\\
\leq&\frac{1}{T}\sum_{t=1}^{T} \|\mu_\pi(s) - P(s|\rho,\pi,t)\|_1 \, \|\Delta L(s)\|_\infty
\end{aligned} \tag{12}$$

where $\|\mu_\pi(s) - P(s|\rho,\pi,t)\|_1$ is L_1-distance between μ_π and $P(s|\rho,\pi,t)$, which is bounded by 2. $\|\Delta L(s)\|_\infty$ is the upper bound of ΔL, according to (4),

$$\Delta L(s) = \mathbb{E}_{s'\sim P_\pi} L(s') - L(s) + \alpha_3 c(s) \leq \alpha_2\bar{c} - 0 + \alpha_3\bar{c}$$

Thus we have

$$\|\Delta L(s)\|_\infty \leq (\alpha_3+\alpha_2)\bar{c} \tag{13}$$

According to Assumption 1, the sequence $\{P(s|\rho,\pi,t), t\in\mathbb{Z}_+\}$ converges to $q_\pi(s)$ as t approaches ∞, then by the Abelian theorem, the sequence $\{\frac{1}{T}\sum_{t=1}^{T} P(s|\rho,\pi,t), T\in\mathbb{Z}_+\}$ also converges to $q_\pi(s)$. Therefore, $\mu_\pi(s) = q_\pi(s)$. At last, one obtains the bound in (9),

$$\begin{aligned}
&\left|\mathbb{E}_{\mu_\pi} \Delta L(s) - \mathbb{E}_{\mu_\pi^T} \Delta L(s)\right|\\
\leq&\frac{1}{T}\sum_{t=1}^{T} \|\mu_\pi(s) - P(s|\rho,\pi,t)\|_1 (\alpha_3+\alpha_2)\bar{c} &(14)\\
\leq&2T^{\gamma-1}(\alpha_3+\alpha_2)\bar{c} &(15)
\end{aligned}$$

where (15) is obtained by using Assumption 3. □

In the following, we will derive the probabilistic bound on estimating $\mathbb{E}_{\mu_\pi^T} \Delta L(s)$ with M trajectories of T steps. It is worth mentioning that since M trajectories are independent of each other, each trajectory as a whole is applicable for the estimation of $\Delta L(s)$ under μ_π^T. This will be demonstrated in the following lemma, where increasing M is desirable for the reduction of estimation deviation, while T does not affect the probabilistic bound.

Lemma 3. *Let M denote the number of trajectories and T denote the length of trajectories. If there exist positive constants α_1, α_2 such that (4) hold, then $\forall\beta\geq 0$, the probability that*

$$\frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M} \Delta L(s_{t,m}) \leq \mathbb{E}_{\mu_\pi^T} \Delta L(s) - \beta \tag{16}$$

is upper bounded by

$$\exp\left(-\frac{2M\beta^2}{(2\alpha_2+\alpha_3)^2\bar{c}^2}\right) \tag{17}$$

where

$$\Delta L(s_{t,m}) \coloneqq L(s_{t+1,m}) - L(s_{t,m}) + \alpha_3 c\left(s_{t,m}\right) \tag{18}$$

and $s_{t,m}$ denotes the sampled state in the mth trajectory at time t.

Proof. The high-level plan is to apply Hoeffding's inequality to establish the probabilistic bound. First, Hoeffding's inequality holds under the condition that random variables follow the same distribution and are independent of each other. On the right hand side of (16),

$$\begin{aligned}
&\mathbb{E}_{\mu_\pi^T} \Delta L(s)\\
=&\int_{\mathcal{S}} \mu_\pi^T(s)\Delta L(s)ds &(19)\\
=&\int_{\mathcal{S}} \frac{1}{T}\sum_{t=1}^{T} P(s|\pi,\rho,t)\Big(\mathbb{E}_{s'\sim P_\pi} L(s')\\
&- L(s) + \alpha_3 c(s)\Big)ds &(20)\\
=&\frac{1}{T}\sum_{t=1}^{T}\Big(\int_{\mathcal{S}} P(s|\pi,\rho,t+1)L(s)ds\\
&+ \int_{\mathcal{S}} P(s|\pi,\rho,t)(-L(s)+\alpha_3 c(s))ds\Big) &(21)
\end{aligned}$$

Let us denote the probability that the inequality (16) holds as δ,

$$\begin{aligned}
\delta \coloneqq&\mathbb{P}\Bigg(\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\Big(\frac{L(s_{t+1,m})}{T} - \frac{L(s_{t,m})}{T}\\
&+ \frac{\alpha_3 c(s_{t,m})}{T}\Big) - \mathbb{E}_{\mu_\pi^T} \Delta L(s) \leq -\beta\Bigg)
\end{aligned} \tag{22}$$

which contains a number of M random sequences of variables $L(s_{t+1,m}) \in [0, \alpha_2\bar{c}]$ and $\alpha_3 c(s_{t,m}) - L(s_{t,m}) \in [-\alpha_2\bar{c}, \alpha_3\bar{c}]$. Therefore, by Hoeffding's inequality, the following inequality holds $\forall\beta\geq 0$,

$$\begin{aligned}
\delta \leq& \exp\left(-\frac{2M^2\beta^2}{M(2\alpha_2+\alpha_3)^2\bar{c}^2}\right) &(23)\\
=& \exp\left(-\frac{2M\beta^2}{(2\alpha_2+\alpha_3)^2\bar{c}^2}\right) &(24)
\end{aligned}$$

which concludes the proof. □

In Lemma 3, β is a positive constant that quantifies the deviation between the sample average of $\Delta L(s)$ and its expectation. As β increases, the probability of such a deviation occurring becomes smaller, which is reflected by the decreasing upper bound. Now, the finite sample estimation of $\Delta L(s)$ and $\mathbb{E}_{\mu_\pi} \Delta L(s)$ are connected with $\mathbb{E}_{\mu_\pi^T} \Delta L(s)$ respectively by Lemmas 2 and 3, we will unify them to derive the desired probabilistic stability guarantee.

Theorem 1. *Assume that Assumptions 1–3 hold for system (1). If there exists a function $L: \mathcal{S}\to\mathbb{R}_+$ and positive constants α_1 and α_2, such that (4) holds, and for a number of M trajectories with T time steps there exists a positive constants ϵ and α_3 such that*

$$T \geq \left(\frac{b_1}{\epsilon}\right)^{\frac{1}{1-\gamma}} \tag{25}$$

and

$$\frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M} \Delta L(s_{t,m}) \leq -\epsilon, \tag{26}$$

then the system (1) controlled by π is mean square stable with a probability of at least

$$1-\exp\left(-2M\left(\frac{\epsilon-2T^{\gamma-1}b_1}{b_2}\right)^2\right) \tag{27}$$

where $b_1=(\alpha_3+\alpha_2)\bar{c}$ and $b_2=(2\alpha_2+\alpha_3)\bar{c}$. If the desired confidence of stability guarantee is at least δ, the associated overall sample complexity is at least $\mathcal{O}(\log(\frac{1}{1-\delta}))$. To achieve a confidence δ, M and T have to satisfy $M(\epsilon-T^{\gamma-1}b_1)^2\geq 2\bar{c}^2(2\alpha_2+\alpha_3)^2\log(\frac{1}{1-\delta})$.

Proof. We will use $\Delta L_{M,T}(s)$ to represent the sample average on the left-hand side of (26) for improved readability. Essentially, $\Delta L_{M,T}(s)$ is the unbiased estimation of $\mathbb{E}_{\mu_\pi^T}\Delta L(s)$, of which the probabilistic bound is derived in Lemma 3. However, in Lemma 2 we show that $\mathbb{E}_{\mu_\pi^T}\Delta L(s)$ is not an unbiased estimation of $\mathbb{E}_{\mu_\pi}\Delta L(s)$ since T is finite. Now based on these lemmas, we need to investigate the effect of estimating $\mathbb{E}_{\mu_\pi}\Delta L(s)$ in (5) with $\Delta L_{M,T}(s)$.

Let ω denote the bound in (9), $\omega=2\bar{c}(\alpha_3+\alpha_2)T^{\gamma-1}$. According to (26), $L_{M,T}(s)+\epsilon$ is semi-negative definite, then

$$\begin{aligned}
&\mathbb{P}(\mathbb{E}_{\mu_\pi}\Delta L(s)\leq 0)\\
\geq&\mathbb{P}\left(\mathbb{E}_{\mu_\pi}\Delta L(s)<\Delta L_{M,T}(s)+\epsilon\right) &(28)\\
=&\mathbb{P}\left(\left(\Delta L_{M,T}(s)-\mathbb{E}_{\mu_\pi^T}\Delta L(s)\right)\right.\\
&\left.+\left(\mathbb{E}_{\mu_\pi^T}\Delta L(s)-\mathbb{E}_{\mu_\pi}\Delta L(s)\right)>-\epsilon\right) &(29)\\
\geq&\mathbb{P}\left(\Delta L_{M,T}(s)-\mathbb{E}_{\mu_\pi^T}\Delta L(s)-\omega>-\epsilon\right) &(30)\\
=&1-\mathbb{P}\left(\Delta L_{M,T}(s)-\mathbb{E}_{\mu_\pi^T}\Delta L(s)\leq-(\epsilon-\omega)\right) &(31)
\end{aligned}$$

Then we apply the probabilistic bound derived in Lemma 3, by substituting β in (17) by

$$(\epsilon-\omega)=\epsilon-\frac{1}{T}\sum_{t=1}^{T}\|\mu_\pi(s)-P(s|\rho,\pi,t)\|_1 b_1 \tag{32}$$

It follows that

$$\begin{aligned}
&\mathbb{P}(\mathbb{E}_{\mu_\pi}\Delta L(s)\leq 0)\\
\geq&1-\mathbb{P}(\Delta L_{M,T}(s)-\mathbb{E}_{\mu_\pi^T}\Delta L(s)\leq-(\epsilon-\omega)) &(33)\\
\geq&1-\exp\left(-\frac{2M(\epsilon-\omega)^2}{(2\alpha_2+\alpha_3)^2\bar{c}^2}\right) &(34)
\end{aligned}$$

Substitute (32) into (34) and utilize Assumption 3, one has

$$\begin{aligned}
&\mathbb{P}(\mathbb{E}_{\mu_\pi}\Delta L(s)\leq 0)\\
\geq&1-\exp\left(-\frac{2M(\epsilon-T^{\gamma-1}2b_1)^2}{(2\alpha_2+\alpha_3)^2\bar{c}^2}\right) &(35)
\end{aligned}$$

The desired bound for guaranteeing mean square stability in probability can be obtained in (27). □

4. Reinforcement learning with probabilistic stability guarantee

Based on the theoretical results in the previous section, one can judge whether the system is stable given several finite-length trajectories by evaluating (26). The theoretical results in the stability theorems using Lyapunov's method do not, however, give a prescription for determining the Lyapunov function and controller. To translate the theorem into practical algorithms, the high-level plan is to parameterize $L(s)$ with (3) and the controller $\pi(a|s)$ with an arbitrary NN $\pi_\theta(a|s)$. Then ϕ and θ will be updated separately and iteratively using stochastic gradient descent algorithms until system (1) is stabilized such that (26) is satisfied. We use τ to denote a trajectory ($\tau=\{s_1,a_1,s_2\ldots s_T\}$), and $\tau\sim\pi$ is the shorthand for indicating that the distribution over trajectories depends on π: $P(\tau)=\rho(s_1)\prod_{t=1}^{T}P(s_{t+1}|s_t,a_t)\pi(a_t|s_t)$.

4.1. Policy gradient

In this subsection, we will focus on how to learn the controller in an iterative manner, repeatedly estimating the policy gradient of the target function with samples and updating θ through stochastic gradient descent. $\Delta L(s)$ is temporarily assumed to be given, i.e., ϕ are fixed. In Section 4.2, we will show how the Lyapunov function is selected and learned after θ is updated.

Since the left-hand side of (26) is the unbiased estimate of $\Delta L(s)$ on μ_π^T, the problem can be formulated by

$$\text{find } \theta,\ \text{s.t. } \mathbb{E}_{\mu_{\pi_\theta}^T}\Delta L(s)\leq-\epsilon \tag{36}$$

A straightforward way of solving the constrained optimization problem above would be the first-order method (Bertsekas, 2014) (Chapter 4), also known as gradient descent. At each update step, the gradient of (36) with respect to θ is estimated with samples, and θ updates a small step in the opposite direction of the estimated gradient vector. The gradient of (36) with respect to θ is derived in the following theorem.

Theorem 2. *The gradient of Lyapunov condition (36) is given by the following,*

$$\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=1}^{T}\nabla_\theta\log\pi_\theta(a_t|s_t)l(\tau,t)\right] \tag{37}$$

where

$$l(\tau,t)=\sum_{j=t+1}^{T}c(s_j)+L(s_{T+1}) \tag{38}$$

Proof. In this section, we will show that (37) holds. To proceed, we introduce $P(s_t|\rho,\pi_\theta):=P(s|\rho,\pi_\theta,t)$ to make the derivation more clear. First, the following equalities hold,

$$\begin{aligned}
&\mathbb{E}_{\mu_{\pi_\theta}^T}\Delta L(s)+\epsilon\\
=&\frac{1}{T}\mathbb{E}_{P(\cdot|\rho,\pi_\theta)}L(s_{T+1})-\frac{1}{T}\mathbb{E}_\rho L(s_1)\\
&+\frac{\alpha_3}{T}\sum_{t=1}^{T}\mathbb{E}_{P(\cdot|\rho,\pi_\theta)}c(s_t)+\epsilon
\end{aligned} \tag{39}$$

Take the derivative of the above equation with respect to θ, one has

$$\begin{aligned}
&\nabla_\theta\mathbb{E}_{\mu_{\pi_\theta}^T}\Delta L(s)\\
=&\frac{1}{T}\int_{\mathcal{S}}\nabla_\theta P(s_{T+1}|\rho,\pi_\theta)L(s_{T+1})\\
&+\frac{\alpha_3}{T}\sum_{t=1}^{T}\int_{\mathcal{S}}\nabla_\theta P(s_t|\rho,\pi_\theta)c(s_t)
\end{aligned} \tag{40}$$

For any $t\in\mathbb{Z}_{[2,+\infty)}$ and function $f(s_t)$,

$$\begin{aligned}
&\int_{\mathcal{S}}\nabla_\theta P(s_t|\rho,\pi_\theta)f(s_t)\\
=&\mathbb{E}_{P(s_{t-1}|\rho,\pi_\theta)}\left(\sum_{a_{t-1}}\nabla_\theta\pi_\theta(a_{t-1}|s_{t-1})\right.\\
&\times\int_{\mathcal{S}}P(s_t|s_{t-1},a_{t-1})f(s_t)\\
&\left.+\int_{\mathcal{S}}\nabla_\theta P(s_{t-1}|\rho,\pi_\theta)\mathbb{E}_{P_{\pi_\theta}(s_t|s_{t-1})}f(s_t)\right)
\end{aligned} \tag{41}$$

$$= \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}[f(s_t)\nabla_\theta \log \pi_\theta(a_{t-1}|s_{t-1})] + \int_{\mathcal{S}} \nabla_\theta P(s_{t-1}|\rho,\pi_\theta)\mathbb{E}_{P_{\pi_\theta}(s_t|s_{t-1})} f(s_t) \tag{42}$$

Iterate the above relation to $t=1$, one has

$$\begin{aligned}&\int_{\mathcal{S}} \nabla_\theta P(s_t|\rho,\pi_\theta) f(s_t)\\ &= \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[f(s_t)\sum_{i=1}^{t-1}\nabla_\theta \log \pi_\theta(a_i|s_i)\right]\end{aligned} \tag{43}$$

Substitute the above equation into (40),

$$\begin{aligned}&\nabla_\theta \mathbb{E}_{\mu^T_{\pi_\theta}} \Delta L(s)\\ =&\frac{1}{T}\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)L(s_{T+1})\right]\\ &+\frac{\alpha_3}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[c(s_t)\sum_{t'=1}^{t-1}\nabla_\theta \log \pi_\theta(a_{t'}|s_{t'})\right]\end{aligned} \tag{44}$$

For the term in (44), we move the first sum into the expectation and exchange its order with the second sum, then

$$\begin{aligned}&\nabla_\theta \mathbb{E}_{\mu^T_{\pi_\theta}} \Delta L(s)\\ =&\frac{1}{T}\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)L(s_{T+1})\right]\\ &+\frac{\alpha_3}{T}\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t'=1}^{T-1}\nabla_\theta \log \pi_\theta(a_{t'}|s_{t'})\sum_{t=t'+1}^{T}c(s_t)\right]\end{aligned} \tag{45}$$

$$=\frac{1}{T}\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)l(\tau,t)\right] \tag{46}$$

which concludes the proof. □

Surprisingly, we found that the policy gradient derived in Theorem 2 is very similar to that used in the vanilla policy gradient method, i.e., REINFORCE (Sutton & Barto, 2018), in the classic RL paradigm. In RL, the objective is to minimize a certain objective function $J_\theta = \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}c(s_t)\right]$ and the policy gradient $\nabla_\theta J_\theta$ is given as follows:

$$\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{j=t+1}^{T}c(s_j)\right)\right] \tag{47}$$

Essentially, despite the scale of $\frac{1}{T}$, (37) and (47) are equivalent if one chooses $c(s)$ to be the Lyapunov function and sets $\alpha_3 = 1$. This implies that given the system (1), REINFORCE actually updates the policy towards a solution that can stabilize the system, although it is now aware of under what conditions the solution is guaranteed to be stabilizing. In particular, we can view REINFORCE as a special case of our result, since we prove that many other choices of α_3 and Lyapunov functions are admissible to find a stabilizing solution. The default setting of $c(s)$ as $L(s)$ and $\alpha_3 = 1$ in REINFORCE may not satisfy (26), while we revealed that many other feasible combinations of L and α_3 potentially exist.

In light of this connection with REINFORCE, it is natural to propose a similar learning procedure based on Theorem 2, referred to as Lyapunov-based REINFORCE (L-REINFORCE). L-REINFORCE updates the policy with the policy gradient proposed in (37). Instead of minimizing the sum of costs, L-REINFORCE aims to learn a stochastic policy $\pi_\theta(a|s)$ such that the conditions in Theorem 1 are satisfied.

Table 1
Hyperparameters of L-REINFORCE.

Hyperparameters	Value
Number of trajectories (M)	20
Time steps (T)	250
learning rate (α)	1e−2
Soft replacement (τ)	0.005
α_3 in (26)	0.005
ϵ in (26)	1e−4
σ in (3)	0.01
Upper bound of norm ($\bar{c}$)	1000
Discount(γ)	0.995
Bias ($\bar{b}$)	10
Structure of Lyapunov network $f_\phi(s)$	(64,16,1)
Structure of policy network $\mu_\theta(s)$	(32,3)

Furthermore, to reduce the variance in the estimation of (37) and speed up learning, it is desirable to introduce a baseline function $b(s)$ in (37), and the estimation is still unbiased (Sutton & Barto, 2018):

$$\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\Big(l(\tau,t)-b(s_t)\Big)\right] \tag{48}$$

4.2. Lyapunov function

The Lyapunov function is parameterized using a DNN f_ϕ in (3). In this paper, we choose the value function to be the update target for f as explored in Chow et al. (2019) and leave other possible choices for future work.

In L-REINFORCE, there are two networks: the policy network $\mu_\theta(s)$ and the Lyapunov network $f_\phi(s)$. For the policy network, we use a fully connected MLP with one hidden layer of 32 units activated by ReLU, outputting the logits for softmax. Its expression is as follows,

$$\pi(a|s) = \frac{exp(\mu^a_\theta(s))}{\sum_i exp(\mu^i_\theta(s))} \tag{49}$$

where $\mu^i_\theta(s)$ denotes the i_{th} dimension of $\mu_\theta(s)$. $f_\phi(s)$ is a fully connected MLP with two hidden layers activated by ReLU. f is trained to approximate the value function, $V(s_t) = \sum_{j=t}^{\infty}\gamma^{j-t}\mathbb{E}_{s_j}\left(c(s_j)-\bar{b}\right)$, where $\bar{b}$ is a designed positive constant. The parameter of f is updated through soft replacement, i.e. $\phi_{k+1} \leftarrow (1-\tau)\phi_k + \tau(\phi_k - \alpha\nabla_\phi e^2)$, where e is the temporal difference.

To wrap up, the L-REINFORCE algorithm is summarized in Algorithm 1.

Algorithm 1 L-REINFORCE

repeat
  for 1, 2, …, M **do**
    Collect transition pairs following π_θ for T steps
  end for
  $\theta \leftarrow \theta - \alpha\nabla_\theta \mathbb{E}_{\mu^T_{\pi_\theta}} \Delta L(s)$
  Update ϕ of Lyapunov function/value network to approximate the designed target
until There exists α_3 such that $\mathbb{E}_{\mu^T_{\pi_\theta}} \Delta L(s) < -\epsilon$

5. Simulation results

In this section, we evaluate the performance of the proposed framework. To demonstrate the effectiveness of the proposed method, we consider the stabilization task of a simulated Cartpole (Brockman et al., 2016). The goal is to stabilize the pole vertically at the position $x = 0$. We adopt REINFORCE (Williams, 1992)

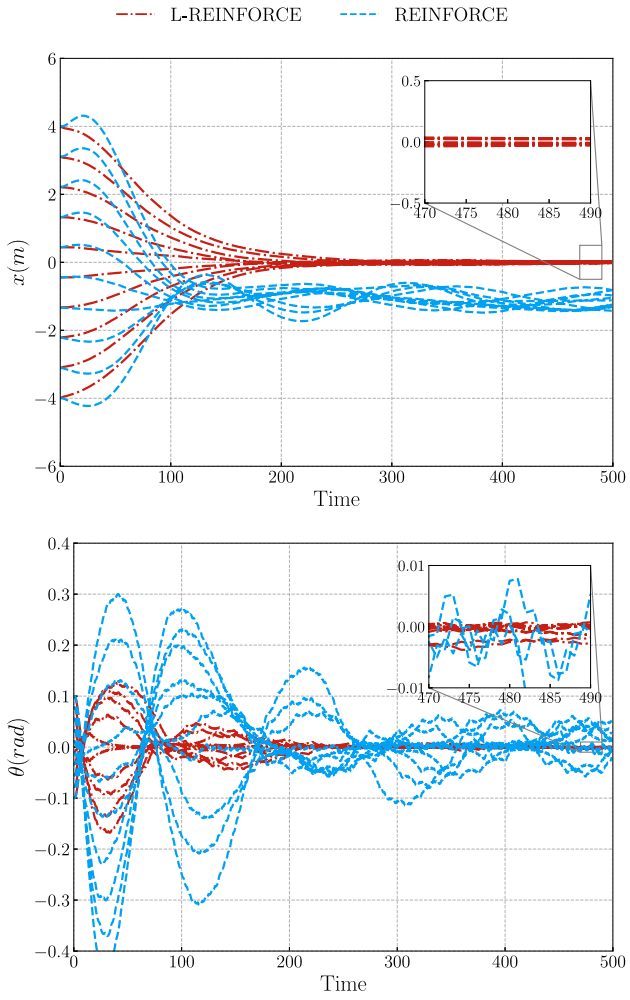


Fig. 1. State trajectories of the controllers trained by L-REINFORCE and REINFORCE. The X-axis denotes the time steps; and the Y-axis denotes the position x in meters and the angle θ in radians, respectively. Zoom-in views are displayed inside the plots.

as the baseline method for comparison. L-REINFORCE and REINFORCE select the action, the horizontal force $F \in \{-10, 0, 10\}(N)$ applied to the cart. The hyperparameters of L-REINFORCE are presented in Table 1. In Table 1, the structures of the neural networks in L-REINFORCE are represented by tuples. $(64, 16, 1)$ represents that the Lyapunov network is composed of three consecutive layers, which contain 64 neurons, 16 neurons, and 1 neuron, respectively. The output of this network is a scalar term. $(32, 3)$ represents that the policy network is composed of two consecutive layers, which contain 32 neurons and 3 neurons, respectively.

The initial horizontal position of the cart x is uniformly distributed on $[-1, 1](m)$, and other states are uniformly distributed on $[-0.05, 0.05]$. During training, the episode ends if the angle θ exceeds the thresholds ± 0.35 rads or the total time step exceeds 500. This condition is not applied during the post-training performance evaluation. During training, REINFORCE uses the 2-norm of the state as the stage cost. The learning rate of REINFORCE is determined to be 1×10^{-2} after tuning. During training, REINFORCE and L-REINFORCE adopt the same values for the number of trajectories M and time steps T for each iteration of policy update.

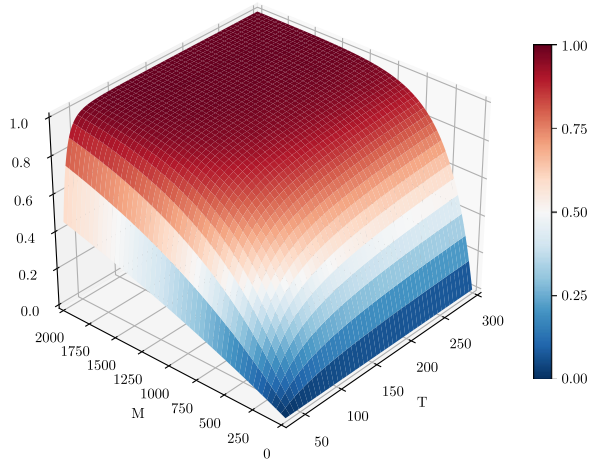


Fig. 2. Visualization of the probabilistic stability bound. The X-axis indicates the length of trajectories T and Y-axis indicates the number of episodes M. The Z-axis indicates the probability of stability and the values are colored differently according to the color bar.

It is important to note that the stability of a system cannot be judged from the cumulative cost (or return) because stability is an inherent property of the system dynamics, and a stable system may not be optimal in terms of the return. Therefore in Fig. 1, we show the transient system behavior under the learned policies of L-REINFORCE and REINFORCE. The systems are initialized at different positions in the space, and their subsequent behaviors are observed. As shown in Fig. 1, starting from different initial states, L-REINFORCE can efficiently stabilize the system. On the other hand, the cartpole controlled by REINFORCE is not stabilized in the x position and oscillates in the angular position.

To let the readers have an intuitive sense of the probabilistic stability bound, the bound for Cartpole under the control of the L-REINFORCE agent is shown in Fig. 2. As shown in Fig. 2, the probability of stability increases sharply as the minimum T requirement (25) is satisfied. Increasing M and T are both helpful for raising the confidence of stability guarantee. The probabilistic stability guarantee measures the reliability of learned policies. By tuning the hyperparameters such as M, T, and α_3, one can achieve the confidence of stability guarantee according to the real needs.

Remark 3. The choice of the clip value $\bar{c}$ may influence both controller performance and sample complexity. Specifically, (1) when $\|s\| > \bar{c}$, the cost saturates and provides limited gradient information for training. A larger $\bar{c}$ gives the controller more informative feedback in these regions. (2) On the other hand, a larger $\bar{c}$ loosens the confidence bound in (27). Consequently, more data is required to achieve the same confidence level. Therefore, $\bar{c}$ indeed acts as a hyperparameter, and its choice reflects a trade-off between informative controller training and the tightness of the probabilistic stability bound.

6. Conclusion and discussion

In this paper, we have investigated the finite sample-based stability analysis and control of stochastic systems characterized by MDP in a model-free manner. Instead of verifying energy decreasing point-wisely on the state space, we proposed a stability theorem where only one sampling-based inequality is to be checked. Furthermore, we showed that with a finite number of trajectories of finite length, it is possible to guarantee stability

with a certain probability, and the probabilistic bound is derived. Finally, we proposed a model-free learning algorithm to learn the controller with a stability guarantee and revealed its connection to REINFORCE.

In the future, an important direction is to extend the theoretical analysis to more efficient algorithms than REINFORCE. On the other hand, the notion of stability in Definition 1 follows the description of MSS given in Benjelloun and Boukas (2002), Bolzern et al. (2010), Zhu et al. (2015). In Teel et al. (2014), a variety of definitions of stability in stochastic hybrid systems have been reviewed, including exponential stability, Lyapunov stability, and Lagrange stability, etc. We believe the extension of the proposed method to investigate other types of stability is another promising direction for future research.

References

Benjelloun, K., & Boukas, E.-K. (2002). Mean square stochastic stability of linear time-delay system with Markovian jumping parameters. *IEEE Transactions on Automatic Control*, *43*(10), 1456–1460.

Berkenkamp, F., Turchetta, M., Schoellig, A., & Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. In *Advances in neural information processing systems* (pp. 908–918).

Bertsekas, D. P. (2014). *Constrained optimization and Lagrange multiplier methods*. Academic Press.

Bolzern, P., Colaneri, P., & De Nicolao, G. (2010). Markov jump linear systems with switching transition rates: Mean square stability with dwell-time. *Automatica*, *46*(6), 1081–1088.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai Gym. arXiv preprint arXiv:1606.01540.

Chang, Y.-C., & Gao, S. (2021). Stabilizing neural control using self-learned almost Lyapunov critics. In *2021 IEEE international conference on robotics and automation* (pp. 1803–1809). IEEE.

Chang, Y.-C., Roohi, N., & Gao, S. (2019). Neural Lyapunov control. *Advances in Neural Information Processing Systems*, *32*.

Chen, Z., Maguluri, S. T., Shakkottai, S., & Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, *33*, 8223–8234.

Chen, Z., Maguluri, S. T., Shakkottai, S., & Shanmugam, K. (2021). Finite-sample analysis of off-policy TD-learning via generalized Bellman operators. *Advances in Neural Information Processing Systems*, *34*, 21440–21452.

Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., & Maguluri, S. T. (2022). Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, *146*, Article 110623.

Chow, Y., Nachum, O., Duenez-Guzman, E., & Ghavamzadeh, M. (2018). A Lyapunov-based approach to safe reinforcement learning. *Advances in Neural Information Processing Systems*, *31*.

Chow, Y., Nachum, O., Faust, A., Ghavamzadeh, M., & Duenez-Guzman, E. (2019). Lyapunov-based safe policy optimization for continuous control. arXiv preprint arXiv:1901.10031.

Dalal, G., Thoppe, G., Szörényi, B., & Mannor, S. (2018). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference on learning theory* (pp. 1199–1233). PMLR.

Dawson, C., Qin, Z., Gao, S., & Fan, C. (2022). Safe nonlinear control using robust neural Lyapunov-barrier functions. In *Conference on robot learning* (pp. 1724–1735). PMLR.

Ganai, M., Hirayama, C., Chang, Y.-C., & Gao, S. (2023). Learning stabilization control from observations by learning Lyapunov-like proxy models. In *2023 IEEE international conference on robotics and automation* (pp. 2913–2920). IEEE.

Han, M., Zhang, L., Wang, J., & Pan, W. (2020). Actor-critic reinforcement learning for control with stability guarantee. *IEEE Robotics and Automation Letters*, *5*(4), 6217–6224.

Huang, C., Chen, C., Xie, K., Lewis, F. L., & Xie, S. (2024). Specified convergence rate guaranteed output tracking of discrete-time systems via reinforcement learning. *Automatica*, *161*, Article 111490.

Huang, W., & Haskell, W. B. (2020). Stochastic approximation for risk-aware Markov decision processes. *IEEE Transactions on Automatic Control*, *66*(3), 1314–1320.

Jiang, Y., & Jiang, Z.-P. (2015). Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, *60*(11), 2917–2929.

Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Champion-level drone racing using deep reinforcement learning. *Nature*, *620*(7976), 982–987.

Kearns, M., & Singh, S. (1998). Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in Neural Information Processing Systems*, *11*.

Khasminskii, R. (2012). *Stochastic stability of differential equations*. Springer.

Lazaric, A., Ghavamzadeh, M., & Munos, R. (2012). Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, *13*(1), 3041–3074.

Levin, D. A., & Peres, Y. (2017). *Markov chains and mixing times*: *vol. 107*, American Mathematical Soc..

Lyapunov, A. M. (1892). *The general problem of the stability of motion (in Russian)* (PhD Dissertation), Univ. Kharkov.

Meyn, S. P., & Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Perkins, T. J., & Barto, A. G. (2002). Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, *3*(Dec), 803–832.

Petridis, V., & Petridis, S. (2006). Construction of neural network based Lyapunov functions. In *The 2006 IEEE international joint conference on neural network proceedings* (pp. 5059–5065). IEEE.

Postoyan, R., Buşoniu, L., Nešić, D., & Daafouz, J. (2017). Stability analysis of discrete-time infinite-horizon optimal control with discounted cost. *IEEE Transactions on Automatic Control*, *62*(6), 2736–2749.

Richards, S. M., Berkenkamp, F., & Krause, A. (2018). The Lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. In *Conference on robot learning* (pp. 466–476).

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Shi, X., Li, Y., Du, C., Chen, C., Zong, G., & Gui, W. (2025). Reinforcement learning-based optimal control for Markov jump systems with completely unknown dynamics. *Automatica*, *171*, Article 111886.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Sutton, R. S., Maei, H. R., & Szepesvári, C. (2009). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems* (pp. 1609–1616).

Teel, A. R., Subbaraman, A., & Sferlazza, A. (2014). Stability analysis for stochastic hybrid systems: A survey. *Automatica*, *50*(10), 2435–2456.

Tsiamis, A., & Pappas, G. J. (2019). Finite sample analysis of stochastic system identification. In *2019 IEEE 58th conference on decision and control* (pp. 3648–3654). IEEE.

Tsiamis, A., Ziemann, I., Matni, N., & Pappas, G. J. (2023). Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, *43*(6), 67–97.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3–4), 229–256.

Xia, K., Wang, J., Zou, Y., Gao, H., & Ding, Z. (2025). Data-driven identifier–actor–critic learning for cooperative spacecraft attitude tracking with orientation constraints. *Automatica*, *173*, Article 112035.

Yang, Z., Zhang, K., Hong, M., & Başar, T. (2018). A finite sample analysis of the actor-critic algorithm. In *2018 IEEE conference on decision and control* (pp. 2759–2764). IEEE.

Zhang, J., Kim, J., O'Donoghue, B., & Boyd, S. (2021). Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 35*, (12), (pp. 10887–10895).

Zhang, S., Zhang, Z., & Maguluri, S. T. (2021). Finite sample analysis of average-reward TD learning and Q-Learning. *Advances in Neural Information Processing Systems*, *34*, 1230–1242.

Zhao, J., Yang, C., Gao, W., Modares, H., Chen, X., & Dai, W. (2023). Linear quadratic tracking control of unknown systems: A two-phase reinforcement learning method. *Automatica*, *148*, Article 110761.

Zhu, J., Yu, X., Zhang, T., Cao, Z., Yang, Y., & Yi, Y. (2015). The mean-square stability probability of H_∞ control of continuous Markovian jump systems. *IEEE Transactions on Automatic Control*, *61*(7), 1918–1924.

Zou, S., Xu, T., & Liang, Y. (2019). Finite-sample analysis for SARSA with linear function approximation. In *Advances in neural information processing systems* (pp. 8665–8675).



Minghao Han received the B.S. degree and Ph.D. degree from the Department of Control Science and Engineering, Harbin Institute of Technology, China, in 2017 and 2022, respectively.

In 2020, he was a visiting researcher at Delft University of Technology, Department of Cognitive Robotics. From 2021 to 2022, he was a visiting researcher at ETH Zürich, Soft Robotics Lab. Since 2023, he has been a Research Fellow at Nanyang Environment & Water Research Institute (NEWRI), Nanyang Technological University. His research interests include control theory, reinforcement learning, and their application in making complex industrial processes and robots more intelligent, safe, and reliable.



Lixian Zhang received the Ph.D. degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2006.

From January 2007 to September 2008, he was a Postdoctoral Fellow in the Department Mechanical Engineering at the Ecole Polytechnique de Montreal, Canada. He was a Visiting Professor at the Process Systems Engineering Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, from February 2012 to March 2013. Since January 2009, he has been with Harbin Institute of Technology, where he is currently a full professor of School of Astronautics. His research interests include nondeterministic switched systems, model predictive control and their applications in multimodal unmanned systems.

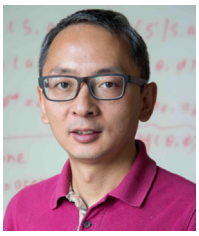
Prof. Zhang serves as Senior Editor for IEEE Control Systems Letters, and served as Associate Editor for IEEE Transactions on Automatic Control, IEEE Transactions on Cybernetics, etc. He was named to the list of Clarivate Highly Cited Researchers in 2016–2021. He is a Fellow of IEEE.



Chenliang Liu received the Ph.D. degree in control science and engineering from the School of Automation, Central South University, in 2024. He is currently an associate professor with the School of Automation, Central South University. From 2023 to 2024, he was a joint training Ph.D. student with the School of Chemistry, Chemical Engineering and Biotechnology at Nanyang Technological University, Singapore. His research interests include artificial intelligence, deep learning, and intelligent manufacturing.



Zhipeng Zhou received the Ph.D. degree in pure mathematics from Academy of Mathematics and Systems Science Chinese Academy of Science, Beijing, in China. He is currently an Algorithm Expert in Alibaba Qwen. His research interests are computer vision, reinforcement learning, and large language model.



Jun Wang received the Ph.D. degree in Computer Science from Delft University of Technology, the Netherlands. He is currently a Chair Professor of Computer Science, University College London, and Founding Director of M.Sc. Web Science and Big Data Analytics. He is also Co-founder and Chief Scientist in MediaGamma Ltd, a UCL start-up company focusing on AI for intelligent audience decision making. His main research interests are in the areas of AI and intelligent systems, including (multiagent) reinforcement learning, deep generative models, and their diverse applications on information retrieval, recommender systems and personalization, data mining, smart cities, bot planning, computational advertising, etc.

Prof. Wang was a recipient of the Beyond Search-Semantic Computing and Internet Economics award by Microsoft Research and also received Yahoo! FREP Faculty award. He has served as an Area Chair in ACM CIKM and ACM SIGIR. His recent service includes co-chair of Artificial Intelligence, Semantics, and Dialog in ACM SIGIR 2018. MediaGamma has received the UCLB One-to-Watch award 2016. He was a technical advisor for startups such as Last. Fm, Passiv Systems, Massive Analytic, Context Scout, and Polecat, and had various projects with Huawei, BT, Microsoft, Yahoo!, Alibaba, Didi, etc.



Wei Pan received his Ph.D. from Imperial College London in 2017. He is currently an Associate Professor in Machine Learning in the Department of Computer Science at the University of Manchester, UK. Previously, he was an Assistant Professor in Robot Dynamics in the Department of Cognitive Robotics at TU Delft, Netherlands, and a Project Leader at DJI, China. He serves as a Senior Editor for IEEE RA-L and ACM TOPML, as well as an Associate Editor or Area Chair for IEEE T-RO, IEEE T-RL, IEEE T-ASE, ICRA, IROS, RSS, CoRL, and L4DC.