Learning-Aided Reach-Avoid Controller Synthesis with Formal Guarantees for Stochastic Systems

Shibo Yan*, Xia Zeng *, Hengjun Zhao*, Zhengfeng Yang†, Zhiming Liu*
*School of Computer and Information Science, Southwest University, Chongqing 400715, China
{xzeng0712, zhaohj2016, zhimingliu88}@swu.edu.cn, yan3425@email.swu.edu.cn
†Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
zfyang@sei.ecnu.edu.cn

***Abstract*—This paper addresses infinite-horizon reach-avoid controller synthesis with formal guarantees for nonlinear stochastic dynamical systems. Learning-based methods handle high-dimensional stochastic control effectively but lack formal safety guarantees, while existing certificate-based approaches suffer from the scalability limitations of interval-based verification. We propose a learning-assisted framework that integrates reinforcement learning, PAC-based polynomial approximation, and sum-of-squares optimization. A learned neural controller is approximated by a polynomial controller, enabling efficient stochastic barrier-like certificate synthesis via semidefinite programming. The framework avoids costly post-verification and directly computes reach-avoid probability lower bounds without the iterative trial process. Experiments demonstrate improved scalability and tighter probabilistic guarantees over state-of-the-art methods.**

***Index Terms*—Stochastic control, Reach-avoid, Formal method, Barrier-like functions, Martingales**

I. Introduction

Controller synthesis under reach-avoid constraints, which simultaneously accounts for safety and reachability, has significant practical relevance in safety-critical applications [1]. Mainstream approaches to designing trustworthy controllers under reach-avoid constraints have relied on the existence of certain certificate functions, such as Lyapunov functions and barrier functions [2]–[4]. In practical problems, uncertainties in system models are often unavoidable. Such problems are commonly modeled using discrete/continuous-time stochastic dynamical systems. For stochastic control, reach-avoid constraints require that the system satisfies the reach-avoid objective with a prescribed probability [5], which are often established using so-called *supermartingale* [6] certificate functions that extend classical certificate functions of deterministic systems [7].

Driven by advances in AI, learning-based methods have gained increasing attention for synthesizing controllers and certificate functions for stochastic systems [8]–[10]. The disturbances in stochastic systems naturally align with data-driven learning techniques such as reinforcement learning (RL) [11] which offer substantial advantages, especially for complex, nonlinear stochastic systems. However, control policies and supermartingale certificate functions obtained through learning-based approaches typically guarantee constraint satisfaction only on the training dataset. Therefore, a post-learning formal verification step remains indispensable to ensure that the synthesized controller rigorously satisfies the desired properties. The representative work [10] proposed to simultaneously learn a control policy and a barrier-like function, termed *reach-avoid supermartingales* (RASM), to guarantee reach-avoid specifications for discrete-time stochastic systems. They subsequently employed interval arithmetic to formally verify the learned certificates, resulting in exponential complexity with the system's dimension due to state space partitioning. Moreover, propagating intervals through learned neural network (NN) controllers requires estimating their Lipschitz constants, further increasing the complexity.

We propose a scalable framework that integrates learning with constraint solving for reach-avoid controller synthesis of stochastic systems. The key idea is to use the RL-generated NN controller as an intermediate artifact to guide the construction of a trustworthy polynomial controller. Specifically, the framework consists of three steps. First, we train an NN controller via RL with reachability and safety objectives. Second, we apply PAC-based polynomial approximation [12] to the NN controller to obtain a polynomial controller. Third, we formulate the synthesis of *stochastic barrier-like certificate functions* (SBCs) for the polynomial-controlled system as a sum-of-squares (SOS) optimization problem, which can be efficiently solved via semi-definite programming (SDP), avoiding interval-based computations and improving the efficiency and scalability of post-verification. Moreover, the optimal lower bound of the reach-avoid probability is incorporated into the SOS objective and thus obtained directly. In contrast, existing methods typically require a prespecified probability value and iteratively synthesize controllers and corresponding SBCs through trial and error to update the probability estimate. Consequently, our approach computes a tighter reach-avoid probability lower bound more efficiently.

In summary, we develop a learning-assisted methodology for reach-avoid controller synthesis in stochastic control systems by integrating learning-based and polynomial-algebraic techniques. We formulate stochastic barrier-like certificate

This work was supported in part by the National Key Research and Development Project, China under Grant 2022YFA1005101, the National Natural Science Foundation of China under Grants (No.62472362, 62272397, 12171159, 92582113). (*Corresponding author: Xia Zeng.*)

2832-6520/26/$31.00 ©2026 IEEE
DOI 10.1109/MEMOCODE70597.2026.00018

synthesis and reach-avoid probability optimization as SOS programs, avoiding the limitations of interval-arithmetic-based verification for neural-network controllers and certificate functions. Experiments show that our approach achieves tighter probability lower bounds and substantially improved scalability over state-of-the-art methods.

II. Preliminaries

Notations. Let $\mathbb{R}$ be the field of real numbers, $\mathbb{R}[\mathbf{x}]$ be the ring of polynomials with coefficients in $\mathbb{R}$ over n-dimensional variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$, $\Sigma[\mathbf{x}] \subset \mathbb{R}[\mathbf{x}]$ be the space of sum-of-squares (SOS) polynomials. Let $\mathbb{N}$ denote the set of nonnegative integers. Let $supp(\mathcal{P})$ denote the support of a distribution $\mathcal{P}$.

A. Problem Formulation

We consider discrete-time stochastic dynamical systems defined by $\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t, \omega_t)$, where $\mathbf{f}(\cdot,\cdot,\cdot) : \Psi \times \mathcal{U} \times \mathcal{N} \to \Psi$ denotes the dynamics function with $\Psi \subset \mathbb{R}^n$, $\mathcal{U} \subset \mathbb{R}^m$, and $\mathcal{N} \subset \mathbb{R}^d$ the system state space, the control action space, the stochastic disturbance space, respectively; $\mathbf{u}_t$ is determined by a control policy $\pi(\mathbf{x}) : \Psi \to \mathcal{U}$, and ω_t is sampled by a specified probability distribution $\mathcal{P}$ with $supp(\mathcal{P}) \subset \mathcal{N}$. A *controlled discrete-time stochastic dynamical system* (CDSDS) is equipped with an initial set $\Theta \subset \Psi$. In this paper, we adopt a 5-tuple $\mathcal{D} \doteq (\mathbf{f}, \pi, \mathcal{P}, \Psi, \Theta)$ to denote such systems. A sequence $(\mathbf{x}_t, \mathbf{u}_t, \omega_t)_{t\in\mathbb{N}}$ with $\mathbf{x}_0 \in \Theta$, $\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t, \omega_t), \forall t \in \mathbb{N}$, and $\mathbf{u}_t, \omega_t$ specified as above constitutes a trajectory of $\mathcal{D}$.

Definition 1 (Reach-Avoid Property)**.** *Let $\delta \in [0,1]$ be a probability threshold. Given a target set $X_g \subseteq \Psi$ and an unsafe set $X_u \subseteq \Psi$, a CDSDS $\mathcal{D}$ satisfies the reach-avoid requirement with probability at least δ if, for any initial state $\mathbf{x}_0 \in \Theta$, $\mathbb{P}_{\mathbf{x}_0}[RA(X_g, X_u)] \geq \delta$, where $RA(X_g, X_u) = \{(\mathbf{x}_t, \mathbf{u}_t, \omega_t)_{t\in\mathbb{N}} \mid \exists t \in \mathbb{N}.\ \mathbf{x}_t \in X_g \wedge (\forall \tau \leq t.\ \mathbf{x}_\tau \notin X_u)\}$ denotes the set of trajectories that reach X_g without reaching X_u.*

Problem 1 (Reach-Avoid Control Problem)**.** *Given a CDSDS $\mathcal{D} = (\mathbf{f}, \pi, \mathcal{P}, \Psi, \Theta)$ where π is not determined, the reach-avoid control problem is to synthesize a control policy $\mathbf{u} = \pi(\mathbf{x})$ which guarantees that $\mathcal{D}$ satisfies the reach-avoid property with probability at least δ as per Definition 1.*

B. Stochastic Barrier-like Certificates

In this paper, we employ the concept of stochastic barrier-like certificates (SBCs) to guarantee reach-avoid specifications for stochastic discrete-time systems.

Definition 2 (Stochastic Barrier-like Certificates [10])**.** *Given a CDSDS $\mathcal{D} = (\mathbf{f}, \pi, \mathcal{P}, \Psi, \Theta)$, let $X_g \subseteq \Psi$ and $X_u \subseteq \Psi$ be the target set and the unsafe set, and let $\delta \in [0,1)$ be the probability threshold. A continuous function $B(\cdot) : \Psi \to \mathbb{R}$ is said to be a stochastic barrier-like certificate (SBC) for reach-avoid requirement with respect to X_g, X_u and δ if it satisfies:*

- *Nonnegativity condition. $B(\mathbf{x}) \geq 0$ for each $\mathbf{x} \in \Psi$;*
- *Initial condition. $B(\mathbf{x}) \leq 1$ for each $\mathbf{x} \in \Theta$;*
- *Safety condition. $B(\mathbf{x}) \geq \frac{1}{1-\delta}$ for each $\mathbf{x} \in X_u$;*
- *Expected decrease condition. There exists $\varepsilon > 0$ such that, for each $\mathbf{x} \in \Psi \backslash X_g$ at which $B(\mathbf{x}) < \frac{1}{1-\delta}$, we have $B(\mathbf{x}) \geq \mathbb{E}_{\omega\sim\mathcal{P}}[B(\mathbf{f}(\mathbf{x}, \pi(\mathbf{x}), \omega))] + \varepsilon$.*

Theorem 1 ([10])**.** *Given a CDSDS $\mathcal{D}$, let $X_g \subseteq \Psi$ and $X_u \subseteq \Psi$ be the target set and the unsafe set, respectively, and let $\delta \in [0,1)$ be the probability threshold. Suppose that there exists an SBC $B(\cdot)$ with respect to X_g, X_u and δ. Then $\mathcal{D}$ satisfies the reach-avoid property as per Definition 1.*

According to Theorem 1, solving Problem 1 can be reduced to designing a control policy $\mathbf{u} = \pi(\mathbf{x})$ for which an SBC can be found. To enable the application of polynomial optimization methods, throughout this paper, we assume that the domain Ψ, the initial set Θ, the goal set X_g, and the unsafe set X_u are all *semi-algebraic sets*. In addition, we assume that $\mathbf{f}$ is a polynomial vector function in $\mathbf{x}, \mathbf{u}, \omega$ for simplicity of presentation, but will also discuss how to relax this restriction to allow non-polynomial dynamics.

III. Verifiable Reach-Avoid Controller Synthesis

In this section, we propose a framework that synthesizes verifiable control policies by integrating reinforcement learning (RL) with the principles of Probably Approximately Correct (PAC) analysis [13].

A. RL-Aided Reference Controller Generation

We employ the widely used Soft Actor-Critic (SAC), an off-policy deep RL algorithm for control tasks with continuous action spaces [14], to train an NN controller by encoding the reach-avoid requirement as optimization objectives. To learn a reach-avoid controller, we define the reward function as

$$r_t = \beta_u\, d_u(\mathbf{x}_t) + \beta_g\big(d_g^w(\mathbf{x}_t) - d_g^w(\mathbf{x}_{t+1})\big) + R_g\, 1_{X_g}(\mathbf{x}_{t+1}) - R_u\, 1_{X_u}(\mathbf{x}_{t+1})$$

where $\beta_u, \beta_g > 0$ are scaling factors, and $R_g, R_u > 0$ are empirical constants. The term $d_u(\mathbf{x}_t)$ denotes the distance from $\mathbf{x}_t$ to the unsafe region X_u; $d_g^w(\mathbf{x}_t)$ represents the dimension-weighted distance from $\mathbf{x}_t$ to the target region X_g, where state dimensions that are more difficult to control are assigned larger weights. The indicator functions $1_{X_g}(\cdot)$, $1_{X_u}(\cdot)$ activate terminal rewards or penalties: R_g is the terminal reward for reaching X_g, and R_u is the terminal penalty for entering X_u. The SAC policy is trained using the above reward function until the termination condition is satisfied. The resulting mean action function $\mu_\theta(\mathbf{x})$ of the trained SAC policy constitutes the synthesized reach-avoid controller, denoted by $\mathbf{u}_{SAC}(\mathbf{x})$.

B. PAC-based Polynomial Controller Synthesis

Using the trained NN controller $\mathbf{u}_{SAC}(\mathbf{x})$ in section III-A, we construct a polynomial controller that is amenable to verification, following the principles of PAC analysis [13]. The approximation error between the PAC-based polynomial controller and the original reference controller $\mathbf{u}_{SAC}(\mathbf{x})$ can be quantified statistically for a prescribed error rate ϵ and confidence level η.

We begin by computing an approximate polynomial $p(\mathbf{x}, \mathbf{c}) \in \mathbb{R}[\mathbf{x}]$ of a prescribed degree d, with $\mathbf{c}$ the indeterminate coefficients. Let $[\mathbf{x}]_d$ denote the vector of all monomials in $\mathbf{x}$ of total degree at most d, arranged according to the graded lexicographic order, i.e., $[\mathbf{x}]_d = (1, x_1, \cdots, x_n, x_1^2, x_1x_2, \cdots, x_{n-1}x_n^{d-1}, x_n^d)^\top$. Let the dimension of $[\mathbf{x}]_d$ be ν, i.e., $\nu = \dim([\mathbf{x}]_d) = \binom{n+d}{d}$. Define $\mathbb{N}_d^n$ as $\{\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n \mid \sum_1^n \alpha_i \leq d\}$ and the monomial $\mathbf{x}^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Thus, $p(\mathbf{x}, \mathbf{c})$ can be written as $p(\mathbf{x}, \mathbf{c}) = \mathbf{c}^T[\mathbf{x}]_d = \sum_{\alpha \in \mathbb{N}_d^n} \mathbf{c}_\alpha \mathbf{x}^\alpha$. We seek a polynomial $p(\mathbf{x}, \mathbf{c})$ that closely approximates $\mathbf{u}_{\text{SAC}}(\mathbf{x})$ by solving

$$\min_{\mathbf{c}} \; e \quad s.t. \; |\mathbf{u}_{\text{SAC}}(\mathbf{x}) - p(\mathbf{x}, \mathbf{c})| \leq e, \;\; \forall \mathbf{x} \in \Psi. \tag{1}$$

Problem (1) is theoretically difficult to solve because it involves infinitely many constraints. The scenario approach of Calafiore et al. [15] overcomes this challenge by replacing the infinite constraint set with a finite sample of K constraints. That is, we determine a polynomial $p(\mathbf{x}, \mathbf{c})$ of degree d by solving the linear program in the variables $\mathbf{c}$ and e obtained from the reformulation of (1) via the scenario optimization approach, i.e., by solving:

$$\min_{\mathbf{c}} \; e \quad s.t. \; |\mathbf{u}_{\text{SAC}}(\mathbf{x}_i) - p(\mathbf{x}_i, \mathbf{c})| \leq e, \;\; \forall \mathbf{x}_i \in \widetilde{\Psi}, \tag{2}$$

where $\widetilde{\Psi}$ is constructed by randomly sampling from Ψ with $|\widetilde{\Psi}| = K$. Moreover, [15] establishes a PAC guarantee relating the solution of the scenario program (2) to that of the original problem (1). For prescribed values of ϵ and η, according to PAC principles [13], there are requirements for the number of sampling points to achieve credible approximation error $\epsilon \geq \frac{2}{K}(\ln \frac{1}{\eta} + \kappa)$ where $\kappa = \nu + 1$.

By solving the linear program (2), we obtain a polynomial $p(\mathbf{x}, \mathbf{c}^*)$ together with the corresponding error parameter e^*. If e^* exceeds a prescribed empirical threshold τ, the approximation error with respect to the reference controller $\mathbf{u}_{\text{SAC}}(\mathbf{x})$ can be reduced by increasing the degree d of the polynomial template and re-solving (2), thereby enlarging the feasible domain for the polynomial coefficients $\mathbf{c}$. In this manner, we ultimately obtain a polynomial $p(\mathbf{x}, \mathbf{c}^*)$ that approximates $\mathbf{u}_{\text{SAC}}(\mathbf{x})$ with a statistical error bound that satisfies the specified tolerance τ at the given confidence level.

IV. Reach-Avoid Verification with Stochastic Barrier-like Certificates

For the controller $p(\mathbf{x})$ obtained via RL and PAC, Theorem 1 implies that an SBC satisfying Definition 2 provides a probabilistic reach-avoid guarantee for the resulting controlled stochastic system. We formulate an SOS optimization problem to synthesize SBCs under $p(\mathbf{x})$. The resulting SDP yields an SBC and its optimized reach-avoid probability bound δ.

The basic idea of SOS relaxation can be explained as follows. Given a basic semi-algebraic set $\mathbb{K}$ defined by $\mathbb{K} = \{\mathbf{x} \in \mathbb{R}^n \,|\, \gamma_1(\mathbf{x}) \geq 0, \ldots, \gamma_l(\mathbf{x}) \geq 0\}$, where $\gamma_j(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$ for $1 \leq j \leq l$, it is not difficult to check that to prove $\forall \mathbf{x}.(\mathbf{x} \in \mathbb{K} \Rightarrow f(\mathbf{x}) \geq 0)$, it is sufficient to show that $f(\mathbf{x}) - \sum_{i=1}^l \sigma_i(\mathbf{x})\gamma_i(\mathbf{x}) \in \Sigma[\mathbf{x}]$ with all $\sigma_i(\mathbf{x}) \in \Sigma[\mathbf{x}]$.

For a prescribed reach-avoid probability threshold δ, we introduce an auxiliary parameter $\hat{\delta}$ and pose the search for an SBC as an optimization problem that maximizes the achievable reach-avoid probability $\hat{\delta}$ according to Theorem 1 as follows:

$$\left.\begin{array}{ll} \max\limits_{\mathbf{b}, \varepsilon, \hat{\delta}} & \hat{\delta} \\ s.t. & \mathbf{x} \in \Psi \Rightarrow B(\mathbf{x}, \mathbf{b}) \geq 0, \\ & \mathbf{x} \in \Theta \Rightarrow 1 - B(\mathbf{x}, \mathbf{b}) \geq 0, \\ & \mathbf{x} \in X_u \Rightarrow B(\mathbf{x}, \mathbf{b}) - \frac{1}{1-\hat{\delta}} \geq 0, \\ & \mathbf{x} \in (\Psi \backslash X_g) \cap (\Psi \backslash X_u) \Rightarrow B(\mathbf{x}, \mathbf{b}) - \\ & \quad \mathbb{E}_{\omega \sim \mathcal{P}}[B(\mathbf{f}(\mathbf{x}, p(\mathbf{x}), \omega), \mathbf{b})] - \varepsilon \geq 0, \\ & 0 \leq \hat{\delta} < 1, \, \varepsilon \geq \varepsilon_c \end{array}\right\} \tag{3}$$

where $B(\mathbf{x}, \mathbf{b})$ is a prescribed polynomial SBC template of a certain degree d_{SBC}, and $\varepsilon_c > 0$ is a prescribed constant for guaranteeing the positivity of ε. If the optimal value $\hat{\delta}^*$ of (3) satisfies $\hat{\delta}^* \geq \delta$, then the closed-loop system under $p(\mathbf{x})$ meets the required reach-avoid specification.

Remark 1. *Noting that the third condition of Definition 2 implies $\mathbf{x} \notin X_u$ for any $\mathbf{x}$ satisfying $B(\mathbf{x}) < \frac{1}{1-\delta}$, we have strengthened the fourth condition of Definition 2 by replacing $B(\mathbf{x}) < \frac{1}{1-\delta}$ with $\mathbf{x} \in \Psi \backslash X_u$ in (3). This technical treatment is intended to simplify the SOS encoding discussed below.*

Suppose that Ψ, Θ, and X_u are basic semi-algebraic sets given by $\Psi := \{\mathbf{x} \in \mathbb{R}^n \mid h_i(\mathbf{x}) \geq 0, \; i = 1, \ldots, m_1\}$, $\Theta := \{\mathbf{x} \in \mathbb{R}^n \mid g_j(\mathbf{x}) \geq 0, \; j = 1, \ldots, m_2\}$, and $X_u := \{\mathbf{x} \in \mathbb{R}^n \mid q_k(\mathbf{x}) \geq 0, \; k = 1, \ldots, m_3\}$. Besides, suppose that $(\Psi \backslash X_g) \cap (\Psi \backslash X_u)$ can be partitioned into subsets, each with the form of $\{\mathbf{x} \in \mathbb{R}^n \,|\, s_\ell(\mathbf{x}) \geq 0, 1 \leq \ell \leq m_4\}$. Then using SOS-relaxation, (3) can be encoded as

$$\left.\begin{array}{ll} \max\limits_{\mathbf{b}, \varepsilon, r} & r \\ s.t. & B(\mathbf{x}, \mathbf{b}) - \sum_{i=1}^{m_1} \sigma_i(\mathbf{x}) h_i(\mathbf{x}) \in \Sigma[\mathbf{x}], \\ & 1 - B(\mathbf{x}, \mathbf{b}) - \sum_{j=1}^{m_2} \phi_j(\mathbf{x}) g_j(\mathbf{x}) \in \Sigma[\mathbf{x}], \\ & B(\mathbf{x}, \mathbf{b}) - r - \sum_{k=1}^{m_3} \kappa_k(\mathbf{x}) q_k(\mathbf{x}) \in \Sigma[\mathbf{x}], \\ & B(\mathbf{x}, \mathbf{b}) - \mathbb{E}_{\omega \sim \mathcal{P}}[B(\mathbf{f}(\mathbf{x}, p(\mathbf{x}), \omega), \mathbf{b})] - \varepsilon \\ & \quad - \sum_{\ell=1}^{m_4} \lambda_\ell(\mathbf{x}) s_\ell(\mathbf{x}) \in \Sigma[\mathbf{x}], \\ & \sigma_i, \phi_j, \kappa_k, \lambda_\ell \in \Sigma[\mathbf{x}], \\ & r \geq 1, \, \varepsilon \geq \varepsilon_c \end{array}\right\} \tag{4}$$

Proposition 1. *Given a CDSDS $\mathcal{D} = (\mathbf{f}, p, \mathcal{P}, \Psi, \Theta)$ with learned policy $p(\mathbf{x})$, a reach-avoid property regarding the target set X_g, unsafe set X_u, and probability threshold δ, and an SBC template $B(\mathbf{x}, \mathbf{b})$, if all involved expressions in $\mathbf{f}, p, \Psi, \Theta, X_g, X_u, B(\mathbf{x}, \mathbf{b})$ are polynomial, then: 1) for appropriate distributions $\mathcal{P}$, (4) is a polynomial optimization problem; 2) if the optimum r^* of (4) satisfies $1 - \frac{1}{r^*} \geq \delta$, then $\mathcal{D}$ satisfies the reach-avoid property with probability δ.*

Proof. To prove 1), it suffices to show that the expectation term in (4), $\mathbb{E}_{\omega \sim \mathcal{P}}[B(\mathbf{f}(\mathbf{x}, p(\mathbf{x}), \omega), \mathbf{b})]$, is polynomial. Since B, $\mathbf{f}$, and p are polynomials by assumption, this expectation reduces to moments $\mathbb{E}_{\omega \sim \mathcal{P}}[\omega^i]$ for i up to a finite bound. Hence, if the moments of $\mathcal{P}$ of different orders admit closed-form expressions, (4) reduces to pure polynomial constraints. This holds for common distributions such as the *triangular*

TABLE I
PERFORMANCE COMPARISON AND EVALUATION ON SEVEN STOCHASTIC CONTROL BENCHMARKS

Environment	Poly.	n_x	NN	$Larac_{\mathrm{SBC}}$ d_c	d_{SBC}	T_{SBC}	P_{SBC}	P_{RASM}	P_s
E_1 2D system [10]	P	2	[2,256,256,1]	4	4	0.372 s	99.26%	93.3%	99.99%
E_2 Inverted pendulum [10]	NP	2	[2,256,256,1]	2	6	3.043 s	98.69%	92.1%	99.02%
E_3 Collision avoidance [16]	P	2	[2,256,256,2]	4	8	2.535 s	97.34%	90.4%	99.97%
E_4 Chain system [17]	P	4	[4,256,256,4]	1	4	1.105 s	95.45%	–	98.19%
E_5 Raychaudhuri system [18]	P	4	[4,256,256,4]	1	4	3.632 s	75.28%	–	89.24%
E_6 Stabilization system [19]	P	5	[5,256,256,2]	1	6	5.410 s	86.44%	–	90.57%
E_7 UAV [20]	NP	6	[6,256,256,2]	1	6	52.181 s	63.99%	–	89.92%

and *normal* distributions. To prove 2), introduce $r = \frac{1}{1-\hat{\delta}}$ in (3). Maximizing $\hat{\delta}$ is then equivalent to maximizing r subject to $r \geq 1$. The constraints of (4) imply those of (3), so the optimum r^* of (4) lower-bounds the optimum $\hat{r}^*$ of (3) with objective r. Thus, by the assumption, $\delta \leq 1 - \frac{1}{r^*} \leq 1 - \frac{1}{\hat{r}^*} = \hat{\delta}^*$. The conclusion follows from Theorem 1. □

Remark 2. *Problem (4) and Proposition 1 can be extended to deal with non-polynomial dynamics* $\mathbf{f}$. *First,* $\mathbf{f}$ *can be over-approximated over* Ψ *through Taylor model [21], that is, we can compute a polynomial* $\mathbf{f}_p(\mathbf{x})$ *and an* n*-dimensional interval* $\mathcal{I}$ *s.t.* $\forall \mathbf{x} \in \Psi.(\mathbf{f}(\mathbf{x}) \in \mathbf{f}_p(\mathbf{x}) + \mathcal{I})$. *Then by introducing auxiliary variables to encode* $\mathcal{I}$ *and strengthening the constraints in (4) involving* $\mathbf{f}$ *to allow all values in* $(\mathbf{f}_p(\mathbf{x}) + \mathcal{I})$, *we can formulate the polynomial optimization problem to synthesize SBCs for non-polynomial systems.*

V. EXPERIMENTS

In this section, we evaluate the proposed method, named $Larac_{SBC}$ (Learning-Aided Reach-Avoid Controller Synthesis via Stochastic Barrier Certificates), on seven nonlinear stochastic control systems that are commonly used as benchmarks in the literature. Our experiments proceed in three parts. First, we compare our approach with the method in [10] on their reported benchmark problems E_1–E_3, with an emphasis on the tightness of the achieved probability guarantees for reach-avoid specifications. Second, we examine the performance of our method on higher-dimensional stochastic control problems E_4–E_7. Third, we analyze the discrepancy between the empirically observed probability of satisfying the reach-avoid property under the synthesized controllers and the lower bounds obtained through our verification procedure.

For each benchmark, we uniformly sample 1000 initial states from the designated initial set. For each initial state, we run $20,000$ Monte Carlo simulations. The minimum of these probabilities over the 1000 initial states is taken as an estimate of the reach-avoid probability of the system.

All experiments are conducted on an Intel Core Ultra 7 265K CPU and an NVIDIA GeForce RTX 4070 Ti SUPER GPU. Table I reports the experimental results, where d_c and d_{SBC} denote the degrees of the polynomial controller and SBC, respectively, and T_{SBC} denotes the SBC synthesis time.

Table I compares the proposed $Larac_{SBC}$ method with the RASM approach of [10]. For all benchmarks E_1–E_3 considered in [10], the controllers synthesized by $Larac_{SBC}$ yield consistently higher reach-avoid probability guarantees. $Larac_{SBC}$ performs post-verification by solving an optimization problem that maximizes the reach-avoid probability subject to the SBC constraints, thus jointly computing the SBC and its optimal probability parameter. In contrast, the RASM method fixes the probability level in advance and iteratively adjusts the controller and barrier-like certificate function to obtain a certifiable probability value.

Table I reports the performance of the $Larac_{SBC}$ method on additional high-dimensional benchmarks, including a non-polynomial stochastic system (UAV) of dimension six. A comparison between the certified reach-avoid probability lower bounds produced by $Larac_{SBC}$ and the simulation-based estimates indicates that the certified bounds closely match the empirical performance. For benchmarks E_4 and E_6, the gap between P_{SBC} and P_s remains within 5%. Even for the more complex case E_7, our method synthesizes a controller that guarantees a reach-avoid probability of at least 63.99%.

In summary, for the synthesis of reach-avoid controllers with formal guarantees for stochastic systems, the proposed framework, integrating learning-based techniques, PAC approximation, and polynomial optimization, is capable of handling higher-dimensional and more complex problems than existing methods. Furthermore, on benchmarks that are tractable for all approaches, our method provides tighter certified lower bounds of reach-avoid probability.

VI. CONCLUSION

This work develops a learning-aided approach for synthesizing reach-avoid controllers with formal guarantees for stochastic systems. An auxiliary controller is first obtained using SAC, followed by a PAC-based approximation method to construct a polynomial controller with explicitly quantified approximation error and an appropriate degree, facilitating formal verification. The SBC conditions are then encoded via SOS relaxation, and a polynomial optimization problem is formulated to maximize the certified lower bound on the reach-avoid probability, yielding both a stochastic barrier certificate and the associated probabilistic guarantee. Compared with the existing RASM method, which relies on interval-analysis-based post-verification, the proposed approach offers better scalability to higher-dimensional stochastic systems while providing high-quality reach-avoid probability guarantees.

REFERENCES

[1] A. Vahidi and A. Eskandarian, "Research advances in intelligent collision avoidance and adaptive cruise control," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 3, pp. 143–153, 2003.

[2] Z. Yang, L. Zhang, X. Zeng, X. Tang, C. Peng, and Z. Zeng, "Hybrid controller synthesis for nonlinear systems subject to reach-avoid constraints," in *Proc. Int. Conf. Comput. Aided Verification (CAV)*. Springer, 2023, pp. 304–325.

[3] C. Dawson, S. Gao, and C. Fan, "Safe control with learned certificates: A survey of neural Lyapunov, barrier, and contraction methods for robotics and control," *IEEE Trans. Robot.*, vol. 39, no. 3, pp. 1749–1767, 2023.

[4] B. Xue, "Reach-avoid controllers synthesis for safety critical systems," *IEEE Trans. Autom. Control*, vol. 69, no. 12, pp. 8892–8899, 2024.

[5] S. Summers and J. Lygeros, "Verification of discrete time stochastic hybrid systems: A stochastic reach-avoid decision problem," *Automatica*, vol. 46, no. 12, pp. 1951–1961, 2010.

[6] D. Williams, *Probability with Martingales*. Cambridge Univ. Press, 1991.

[7] B. Xue, "Sufficient and necessary barrier-like conditions for safety and reach-avoid verification of stochastic discrete-time systems," arXiv:2408.15572, 2024. [Online]. Available: https://arxiv.org/abs/2408.15572

[8] M. Lechner, Đ. Žikelić, K. Chatterjee, and T. A. Henzinger, "Stability verification in stochastic control systems via neural network supermartingales," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, no. 7, 2022, pp. 7326–7336.

[9] K. Chatterjee, T. A. Henzinger, M. Lechner, and Đ. Žikelić, "A learner-verifier framework for neural network controllers and certificates of stochastic systems," in *Proc. Int. Conf. Tools Algorithms Construction Anal. Syst. (TACAS)*. Springer, 2023, pp. 3–25.

[10] Đ. Žikelić, M. Lechner, T. A. Henzinger, and K. Chatterjee, "Learning control policies for stochastic systems with reach-avoid guarantees," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 10. AAAI Press, 2023, pp. 11 926–11 935.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[12] X. Zeng, B. Liu, Z. Zeng, Z. Liu, and Z. Yang, "Safe controller synthesis for nonlinear systems via reinforcement learning and PAC approximation," in *Proc. ACM/IEEE Design Autom. Conf. (DAC)*. New York, NY, USA: ACM, 2024, pp. 283:1–283:6.

[13] G. Calafiore and M. Campi, "The scenario approach to robust control design," *IEEE Trans. Autom. Control*, vol. 51, no. 5, pp. 742–753, 2006.

[14] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, ser. Proc. Mach. Learn. Res., vol. 80. PMLR, 2018, pp. 1861–1870.

[15] M. Campi, S. Garatti, and M. Prandini, "The scenario approach for systems and control design," *Annu. Rev. Control*, vol. 33, no. 2, pp. 149–157, 2009.

[16] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," arXiv:1606.01540, 2016. [Online]. Available: https://arxiv.org/abs/1606.01540

[17] A. Abate, D. Ahmed, A. Edwards, M. Giacobbe, and A. Peruffo, "FOSSIL: A software tool for the formal synthesis of Lyapunov functions and barrier certificates using neural networks," in *Proc. Int. Conf. Hybrid Syst.: Comput. Control (HSCC)*. New York, NY, USA: ACM, 2021, pp. 24:1–24:11.

[18] A. Ferragut and A. Gasull, "Seeking Darboux polynomials," *Acta Appl. Math.*, vol. 139, no. 1, pp. 167–186, 2015.

[19] M. A. Ben Sassi and S. Sankaranarayanan, "Stabilization of polynomial dynamical systems using linear programming based on Bernstein polynomials," arXiv:1501.04578 [math.OC], 2015. [Online]. Available: https://arxiv.org/abs/1501.04578

[20] W. Jin, Z. Wang, Z. Yang, and S. Mou, "Neural certificates for safe control policies," arXiv:2006.08465, 2020. [Online]. Available: https://arxiv.org/abs/2006.08465

[21] X. Chen, E. Abraham, and S. Sankaranarayanan, "Taylor model flowpipe construction for non-linear hybrid systems," in *Proc. IEEE Real-Time Syst. Symp. (RTSS)*. IEEE, 2012, pp. 183–192.