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

Abstract

This paper addresses the design of controllers with formal guarantees for satisfying reach-avoid conditions over an infinite horizon for nonlinear stochastic dynamical systems. Learning-based approaches offer inherent advantages in handling high-dimensional stochastic control problems through efficient data utilization, yet they lack formal guarantees for safety-critical properties. Existing methods attempt to ensure specification satisfaction by jointly synthesizing stochastic barrier-like certificates (SBCs), yet verification remains computationally intractable due to interval-based proofs and the resulting state-space explosion. To overcome these challenges, we propose a learning-assisted controller synthesis framework that integrates reinforcement learning (RL) and PAC-based polynomial approximation with sum-of-squares optimization. The proposed approach fully leverages the strengths of RL to provide a neural network reference controller, and afterwards derives a polynomial controller with high-performance via PAC-approximation, which facilitates efficient SBC synthesis through semidefinite programming, thus eliminating post-verification complexity and improving scalability. Moreover, it enables direct estimation of optimal reach-avoid probability lower bounds, avoiding the iterative trial process required by existing methods. Experiments demonstrate superior scalability and tighter probabilistic guarantees compared with state-of-the-art methods.

Keywords

Stochastic control, Reach-avoid, Formal method, Barrier-like functions, Martingales

ACM Reference Format:

. 2026. Learning-Aided Reach-Avoid Controller Synthesis with Formal Guarantees for Stochastic Systems. In *Proceedings of The Chips to Systems Conference (DAC '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Synthesizing system controllers with formal guarantees under specified requirements is an important research topic. In particular, controller synthesis under reach-avoid constraints, which simultaneously accounts for safety and reachability, has significant practical relevance in safety-critical applications [15].

In recent years, 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 [7, 18, 20]. 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 [13], which are often established using so-called *supermartingale* [17] certificate functions that extend classical certificate functions of deterministic systems [19].

With the rapid advancement of AI technologies, learning-based methods have been increasingly employed to construct controllers for stochastic systems as well as the corresponding certificate functions [5, 11, 16]. The disturbances in stochastic systems naturally align with data-driven learning techniques such as reinforcement learning (RL) [14] 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 [16] proposed to simultanenously 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.

Motivated by these limitations, this paper investigates how to better integrate learning techniques with constraint-solving methods to construct a more scalable approach for synthesizing controllers for stochastic systems w.r.t. reach-avoid constraints. The core idea is that the RL–generated NN controller serves as an intermediate artifact whose purpose is to guide the construction of a trustworthy polynomial controller. Specifically, we propose a framework consisting of three main steps. Firstly, we leverage the strengths of RL to train an NN controller with reachability and safety objectives. Secondly, by incorporating PAC-based polynomial approximation [21] of the NN controller, we obtain a well-performing polynomial controller. Thirdly, to verify the system with the polynomial controller, we transform the synthesis of *stochastic barrier-like certificate functions* (SBCs) into a sum-of-squares (SOS) optimization problem, which enables efficient solving based on semi-definite programming (SDP), thereby avoiding interval-based computations and significantly improving the efficiency and scalability of post-verification. In addition, the optimal lower-bound estimation of the reach-avoid probability is encoded as the optimization objective

DAC '26, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/2026/07
https://doi.org/XXXXXXX.XXXXXXX

in the SOS reformulation, and thus can be directly obtained. Contrarily, existing methods typically require a prespecified probability value as input and iteratively conduct trial-and-error synthesis of controllers and corresponding SBCs, continually updating the probability estimate. As a result, compared with existing approaches, our method is able to compute a tighter lower bound of the reach-avoid probability more efficiently.

In summary, the main contributions of this work are as follows:

- We propose a new methodology for the synthesis of reach-avoid controllers for stochastic control systems, which fully leverages and integrates learning-based techniques with classical polynomial-algebraic methods.
- We transform the construction of stochastic barrier-like certificates as well as the search for optimal reach-avoid probabilities into SOS optimization programs, avoiding the intrinsic limitations of interval-arithmetic-based verification, particularly on NN controllers and certificate functions.
- We conduct comprehensive experimental evaluations and demonstrate that compared with state-of-the-art techniques, the proposed approach provides consistently tighter lower bounds of the reach-avoid probabilities on the same benchmarks, and achieves a significant improvement in the scale of problems it can handle.

2 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 $\mathbf{x} = (x_1, x_1, \ldots, x_n)^\top$, $\mathbb{R}[\mathbf{x}]^n$ be the space of n-dimensional vector polynomials, $\Sigma[\mathbf{x}] \subset \mathbb{R}[\mathbf{x}]$ be the space of sum-of-squares (SOS) polynomials. Let $\mathbb{N}$ denote the set of nonnegative integers. The indicator function of a set A, denoted $1_A(\mathbf{x})$, equals 1 if $\mathbf{x} \in A$ and 0 otherwise. Let $supp(\mathcal{P})$ denote the support of a distribution $\mathcal{P}$.

2.1 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; $\mathbf{x}_t \in \Psi$, $\mathbf{u}_t \in \mathcal{U}$, and $\omega_t \in \mathcal{N}$ denote the system state, the control input, and the stochastic disturbance at time $t \in \mathbb{N}$, 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 $supp(\mathcal{P}) \subset \mathcal{N}$.

In many contexts, a *controlled discrete-time stochastic dynamical system* (CDSDS) is equipped with a domain, the same as the state space Ψ, and an initial set $\Theta \subset \Psi$. In the rest of 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 $\mathcal{D}$. The set of all trajectories of $\mathcal{D}$ forms a probability space with induced probability measure.

The investigated problem of this paper is defined via the following definitions.

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 $\mathcal{D}$ is termed satisfying the reach-avoid requirement with probability at least δ when the system reaches X_g without reaching X_u. Formally, for any initial state $\mathbf{x}_0 \in \Theta$, we have*

$$\mathbb{P}_{\mathbf{x}_0}[RA(X_g, X_u)] \geq \delta$$

with $RA(X_g, X_u) = \{(\mathbf{x}_t, \mathbf{u}_t, \omega_t)_{t\in\mathbb{N}} | \exists t \in \mathbb{N}.\ \mathbf{x}_t \in X_g \wedge (\forall \tau \leq t.\ \mathbf{x}_\tau \notin X_u)\}$ 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, a target set $X_g \subseteq \Psi$, an unsafe set $X_u \subseteq \Psi$, and a specified probability threshold δ, the probabilistic 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.*

2.2 Stochastic Barrier-like Certificates

In this paper, we employ the concept of stochastic barrier-like certificates (SBCs), a class of barrier-like functions also referred to as reach-avoid supermartingales (RASMs) in [16], to guarantee reach-avoid specifications for stochastic discrete-time systems. Drawing on the theory of supermartingale processes [17], [16] introduced RASMs and proved their effectiveness in characterizing the probabilistic reach-avoid properties of stochastic systems.

Definition 2 (Stochastic Barrier-like Certificates[16]). *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 $B(\mathbf{x}) \geq \mathbb{E}_{\omega\sim\mathcal{P}}[B(\mathbf{f}(\mathbf{x}, \pi(\mathbf{x}), \omega))] + \varepsilon$.*

The existence of an SBC for some $\varepsilon > 0$ implies the satisfaction of the reach-avoid property.

Theorem 1 ([16]). *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.*

We give an intuitive explanation of Theorem 1 as follows. The function B is required to be nonnegative on the domain Ψ, bounded above by 1 on the initial set Θ, and bounded below by $\frac{1}{1-\delta}$ on the unsafe set X_u. Consequently, for a trajectory to reach an unsafe state, the value of B must increase by at least a factor of $\frac{1}{1-\delta}$. Moreover, for any state $\mathbf{x} \in \Psi\backslash X_g$ with $B(\mathbf{x}) \leq \frac{1}{1-\delta}$, Theorem 1 requires that the expected value of B decrease by at least some $\varepsilon > 0$, ensuring a strict decreasing trend until the system reaches either the target set X_g or a region where $B(\mathbf{x}) \geq \frac{1}{1-\delta}$. Together, these conditions ensure that the SBC certifies satisfaction of the reach-avoid property with probability at least δ. The proof of Theorem 1 is given in [16].

According to Theorem 1, solving of 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 in the construction of SBCs, 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*, that is, subsets of $\mathbb{R}^n$ represented by polynomial equations and inequalities. 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.

3 Verifiable Reach-Avoid Controller Synthesis via Reinforcement Learning

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 [3]. To circumvent the difficulty of verifying RL-generated deep neural network controllers, we construct polynomial controllers guided by the well-performing RL policies and justified by PAC analysis, which yields probabilistic bounds on the approximation error. The polynomial structure additionally enables numerical solution of the constraints induced by subsequent SBC synthesis, thereby supporting efficient formal verification.

3.1 RL-Aided Reference Controller Generation

We employ the widely used Soft Actor-Critic (SAC) deep RL algorithm [9] to train an NN controller by encoding the reach-avoid requirement as optimization objectives. SAC is an off-policy RL method extensively applied in control tasks with continuous action spaces. It integrates both value-based and policy-based learning approaches, and its primary components include a *stochastic* policy (actor) network $\pi_\theta(\cdot \mid \mathbf{x})$ and two value estimation (critic) networks Q_{ϕ_1} and Q_{ϕ_2}. The actor network $\pi_\theta(\cdot \mid \mathbf{x})$ essentially models a diagonal Gaussian distribution, by taking the current state $\mathbf{x}$ as input and outputs the mean $\mu_\theta(\mathbf{x})$ and the log standard deviation $\log[\sigma_\theta(\mathbf{x})]$ of the stochastic policy. Actions are sampled using the reparameterization trick and subsequently squashed through a tanh function:

$$\mathbf{u}_t = \tanh(\mu_\theta(\mathbf{x}_t) + \sigma_\theta(\mathbf{x}_t) \odot \boldsymbol{\varepsilon}_t),\ \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, I),$$

where $\odot$ denotes the elementwise product and $\mathcal{N}(0, I)$ is the normal Gaussian distribution with the same dimension as $\mathbf{u}$. Finally $\mathbf{u}_t$ is linearly scaled to match the environment's action space. The critic adopts a double-Q architecture. Each critic network takes as input the concatenated state-action vector $(\mathbf{x}_t, \mathbf{u}_t)$ and produces scalar outputs $Q_{\phi_1}(\mathbf{x}_t, \mathbf{u}_t)$ and $Q_{\phi_2}(\mathbf{x}_t, \mathbf{u}_t)$. The minimum of the two estimates is used to suppress overestimation bias.

The design of reward functions is crucial in RL. To learn a reach-avoid controller, we define the reward function as

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

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; $d_g^{\mathbf{w}}(\mathbf{x}_t)$ 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 $1_{X_g}(\cdot)$, $1_{X_u}(\cdot)$ activates 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.

To collect the training data set D, multi-start sampling is conducted over the initial set Θ, and rollouts generate trajectories $(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{u}_t, r_t)$, which are stored in a replay buffer. SAC performs optimization under the maximum-entropy RL framework, with the following objective:

$$\mathcal{J}_\pi = \mathbb{E}_{\mathbf{x}\sim D,\, \mathbf{u}\sim\pi_\theta}\left[\, Q_\phi(\mathbf{x}, \mathbf{u}) - a \log \pi_\theta(\mathbf{u} \mid \mathbf{x}) \,\right],$$

where $\pi_\theta(\mathbf{u} \mid \mathbf{x})$ denotes the probability density evaluated at $\mathbf{u}$ of the diagonal Gaussian distribution generated by $\pi_\theta(\cdot \mid \mathbf{x})$, and $a > 0$ serves as a temperature parameter balancing rewards and exploration entropy. After every certain number of simulation steps, a minibatch is sampled from the replay buffer to alternately update the actor and critic networks. This iterative process continues until the termination condition is satisfied. The resulting mean action function $\mu_\theta(\mathbf{x})$ of the trained actor network constitutes the synthesized reach-avoid controller, denoted by $\mathbf{u}_{SAC}(\mathbf{x})$.

3.2 PAC-based Polynomial Controller Synthesis

Using the trained NN controller $\mathbf{u}_{SAC}(\mathbf{x})$ in Subsection 3.1, we construct a polynomial controller that is amenable to verification, following the principles of PAC analysis [3]. 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.,

$$[\mathbf{x}]_d = (1, x_1, x_2, \cdots, x_n, x_1^2, x_1x_2, \cdots, x_{n-1}x_n^{d-1}, x_n^d)^\top.$$

Let the dimension of $[\mathbf{x}]_d$ be ν, i.e., $\nu = \dim([\mathbf{x}]_d) = \binom{n+d}{d}$. Define $\mathbb{N}^n_d$ as $\{\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n \mid \sum_1^n \alpha_i \le d\}$ 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

$$p(\mathbf{x}, \mathbf{c}) = \mathbf{c}^T[\mathbf{x}]_d = \sum_{\alpha\in\mathbb{N}^n_d} \mathbf{c}_\alpha \mathbf{x}^\alpha.$$

Our objective is to construct a polynomial $p(\mathbf{x}, \mathbf{c})$ that approximates $\mathbf{u}_{SAC}(\mathbf{x})$ as accurately as possible, which amounts to solve

$$\left.\begin{array}{ll} \min\limits_{\mathbf{c}} & e \\ s.t. & |\mathbf{u}_{\text{SAC}}(\mathbf{x}) - p(\mathbf{x}, \mathbf{c})| \le e,\ \ \forall \mathbf{x} \in \Psi. \end{array}\right\} \tag{1}$$

Problem (1) is theoretically difficult to solve because it involves infinitely many constraints. The scenario approach of Calafiore et al. [4] 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:

$$\left.\begin{array}{ll} \min\limits_{\mathbf{c}} & e \\ s.t. & |\mathbf{u}_{\text{SAC}}(\mathbf{x}_i) - p(\mathbf{x}_i, \mathbf{c})| \le e,\ \ \forall \mathbf{x}_i \in \widetilde{\Psi}, \end{array}\right\} \tag{2}$$

where $\widetilde{\Psi}$ is constructed by randomly sampling from Ψ with $|\widetilde{\Psi}| = K$. Moreover, [4] 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 [3], 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 $\mathbf{u}_{SAC}(\mathbf{x})$ 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 $\mathbf{u}_{SAC}(\mathbf{x})$ with a statistical error bound that satisfies the specified tolerance τ at the given confidence level.

Compared with the least-squares-based approximation method which fits $\mathbf{u}_{SAC}(\mathbf{x})$ using fixed polynomial templates and fixed sets of sampled points from Ψ, the PAC-based framework provides a systematic sampling procedure and explicit bounds relating sample size, accuracy, and template complexity, enabling the polynomial structure to be chosen to meet a prescribed accuracy.

4 Reach-Avoid Verification with Stochastic Barrier-like Certificates

For the controller $p(\mathbf{x}, \mathbf{c}^*)$ obtained via the RL and PAC procedures (denoted by $p(\mathbf{x})$ for short in this section), by Theorem 1, the existence of stochastic barrier-like certificates (SBCs) satisfying the constraints of Definition 2 establishes probabilistic reach-avoid guarantees for the resulting closed-loop stochastic system. We will formulate an optimization problem to synthesize SBCs for the system under $p(\mathbf{x})$, by employing sum-of-squares (SOS) relaxations to encode the SBC constraints. The induced SOS program, solvable efficiently as a semidefinite program, yields both the SBC and the corresponding 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

$$\mathbb{K} = \{\mathbf{x} \in \mathbb{R}^n \mid \gamma_1(\mathbf{x}) \geq 0, \ldots, \gamma_l(\mathbf{x}) \geq 0\},$$

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:

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

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 1 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 1 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, X_u can be represented by

$$\begin{cases} \Psi := \{\mathbf{x} \in \mathbb{R}^n \mid h_i(\mathbf{x}) \geq 0, 1 \leq i \leq m_1\}, \\ \Theta := \{\mathbf{x} \in \mathbb{R}^n \mid g_j(\mathbf{x}) \geq 0, 1 \leq j \leq m_2\}, \\ X_u := \{\mathbf{x} \in \mathbb{R}^n \mid q_k(\mathbf{x}) \geq 0, 1 \leq k \leq m_3\}. \end{cases}$$

Besides, suppose that $(\Psi\backslash X_g) \cap (\Psi\backslash X_u)$ can be partitioned into subsets, each with the form of $\{\mathbf{x} \in \mathbb{R}^n \mid s_\ell(\mathbf{x}) \geq 0, 1 \leq \ell \leq m_4\}$. Then using SOS-relaxation, (3) can be encoded as

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

Remark 2. *For ease of presentation, we have assumed $(\Psi\backslash X_g) \cap (\Psi\backslash X_u)$ consists of only one partition of the form $\{\mathbf{x} \in \mathbb{R}^n \mid s_\ell(\mathbf{x}) \geq 0, 1 \leq \ell \leq m_4\}$. For multiple partitions of $(\Psi\backslash X_g) \cap (\Psi\backslash X_u)$, the encoding of (4) can be extended accordingly.*

For the problem (4) we give the following conclusion:

Proposition 2. *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-aoivd property with probability threshold δ.*

Proof. To prove 1), we only need to show the expectation term in (4), i.e., $\mathbb{E}_{\omega\sim\mathcal{P}}[B(\mathbf{f}(\mathbf{x},p(\mathbf{x}),\omega),\mathbf{b})]$, is a polynomial expression. Since $B, \mathbf{f}, p$ are all polynomials from the assumption, the calculation of the expectation is reduced to $\mathbb{E}_{\omega\sim\mathcal{P}}[\omega^i]$ for i under a certain bound. Therefore if the moments of different orders of $\mathcal{P}$ have closed-form solutions, (4) can be reduced to pure polynomial constraints. Such distributions includes common ones like the *triangular* distribution, the *normal* distribution, etc.

To prove 2), introduce a new variable $r = \frac{1}{1-\hat{\delta}}$ in the problem (3). Then maximizing $\hat{\delta}$ in (3) is equivalent to maximizing r with constraint $r \geq 1$. Now it is easy to check that the constraints of (4) implies the constraints of (3), and thus the optimum r^* of (4) gives a lower bound of the optimum $\hat{r}^*$ of (3) with objective r. Therefore from the assumption we obtain that $\delta \leq 1 - \frac{1}{r^*} \leq 1 - \frac{1}{\hat{r}^*} = \hat{\delta}^*$, and the conclusion follows from Theorem 1. □

Remark 3. *Problem (4) and Proposition 2 can be extended to deal with non-polynomial dynamics $\mathbf{f}$. First, $\mathbf{f}$ can be over-approximated over Ψ through Taylor model [6], 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.

5 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 [16] on their reported benchmark problems, including the two-dimensional linear system (E_1), the stochastic inverted pendulum system (E_2), and the collision-avoidance system (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, thereby assessing its capability to synthesize controllers with provable reach-avoid guarantees in more complex settings. This corresponds to the Chain system [1] (E_4), the Raychaudhuri system [8] (E_5), the Stabilization system [12] (E_6), and a classical six-dimensional UAV control system [10] (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, in order to assess the conservativeness of the resulting probabilistic guarantees. This category of comparison is performed on all $E_1 - E_7$. For each benchmark, we simulate the closed-loop system under the controller synthesized by our method. Specifically, we uniformly sample 1000 initial states from the designated initial set. For each initial state, we run 20,000 Monte Carlo simulations to account for the stochastic disturbances, and we estimate the probability of satisfying the reach-avoid specification from the resulting trajectories. The minimum of these probabilities over the 1000 initial states is taken as an estimate of the reach-avoid probability of the system.

We first use an example to show how our method works.

Example 1. *Consider the classic case* Collision Avoidance *from [16]. To demonstrate how the simulated system trajectories enter the unsafe set, we slightly increase the noise in the original equations to 4 times its original value. The dynamics equations are as follows:*

$$x_{t+1} = x_t + 0.2(d_2(d_1 u_t + (1-d_1)\binom{0}{1}) + (1-d_2)\binom{0}{-1}) + 0.2\omega_t,$$

where ω_t follows a triangular distribution, and the specific settings for the parameters d_1, d_2, and the disturbances can be found in [16]. The system state space is $\Psi = \{x \in \mathbb{R}^2 \mid -1 \le x_1, x_2 \le 1\}$, the initial state set is $\Theta = \{x \in \mathbb{R}^2 \mid (-1 \le x_1 \le -0.9 \wedge -0.6 \le x_2 \le 0.6) \vee (0.9 \le x_1 \le 1.0 \wedge -0.6 \le x_2 \le 0.6)\}$, the goal set is $X_g = \{x \in \mathbb{R}^2 \mid -0.2 \le x_1, x_2 \le 0.2\}$, and the unsafe state set is $X_u = \{x \in \mathbb{R}^2 \mid (-0.3 \le x_1 \le 0.3 \wedge 0.7 \le x_2 \le 1) \vee (-0.3 \le x_1 \le 0.3 \wedge -1 \le x_2 \le -0.7)\}$. We design a controller for this stochastic system to ensure that the control behavior satisfies the reach-avoid property with a probability of at least 90%.

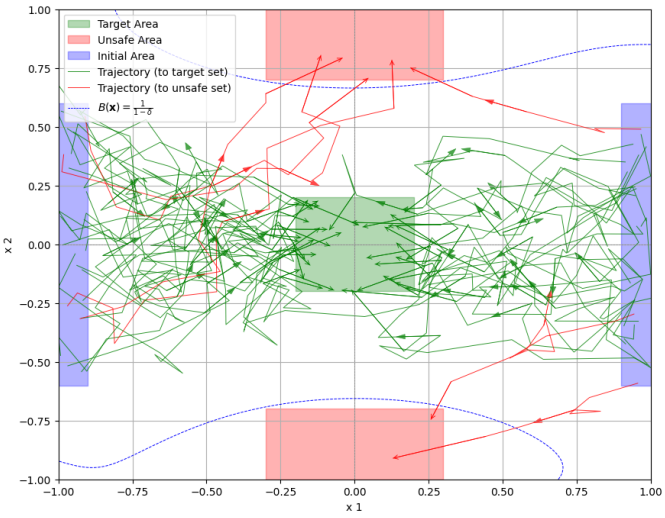

Figure 1: The figure illustrates the behavior of the Collision Avoidance system under the controller designed using our proposed method. It additionally shows the $\frac{1}{1-\delta}$–level set of the SBC (with $\delta = 92.68\%$), which provides a probabilistic guarantee of the reach-avoid property. The green curves with arrows depict clusters of simulated trajectories initialized from states uniformly sampled within the initial set, while the red curves indicate trajectories that enter the unsafe region.

Using the method proposed in this paper, the reference controller obtained through SAC training is a four-layer ReLU network (i.e., [2, 256, 256, 2]). Through PAC approximation, we obtain the following polynomial-form controller candidate $p(\mathbf{x}) = (u_1, u_2)^\top$:

$$u_1 = \underbrace{-2.41x_1 + 0.426x_2 + 0.276x_1^2 - \cdots - 0.453x_2^4 - 0.0691}_{18\ terms},$$

$$u_2 = \underbrace{-0.356x_1 - 3.33x_2 - 0.239x_1^2 - \cdots - 2.79x_2^4 + 0.0243}_{18\ terms}.$$

Furthermore, we verify the existence of an SBC for the system under the controller $p(\mathbf{x})$. By applying an SOS-based reformulation and solving the resulting SDP, we obtain an optimal lower bound of 92.68% for the reach-avoid probability in 3.025s. The resulting SBC (the blue dashed curve in Figure 1) is as follows:

$$B = \underbrace{0.00451 - 0.0709x_1 - 0.0754x_2 + \cdots + 22.2x_1x_2^7 - 61.3x_2^8}_{45\ terms}.$$

Figure 1 provides an intuitive illustration of the system's behavior under $p(\mathbf{x})$. The green curves with arrows represent system trajectories obtained from simulations with randomly sampled initial states. The computed lower bound of 92.68% for the reach-avoid probability is very close to the actual simulation result of 95.12%. These results further demonstrate that the designed controller satisfies the reach-avoid property.

All experiments are conducted on a 3.9 GHz Intel Core Ultra 7 265K CPU and an NVIDIA GeForce RTX 4070 Ti SUPER (16GB

Table 1: Comparison between our approach and [16] on their benchmarks

Environment	Poly.	n_x	$Larac_{SBC}$					P_{RASM}	P_s
			NN	d_c	d_{SBC}	T_{SBC}	P_{SBC}		
E_1 2D system [16]	P	2	[2,256,256,1]	4	4	0.372 s	99.26%	93.3%	99.99%
E_2 Inverted pendulum [16]	NP	2	[2,256,256,1]	2	6	3.043 s	98.69%	92.1%	99.02%
E_3 Collision avoidance [2]	P	2	[2,256,256,2]	4	8	2.535 s	97.34%	90.4%	99.97%

Table 2: Performance of our approach on additional high-dimensional benchmarks

Environment	Poly.	n_x	NN	d_c	d_{SBC}	T_{SBC}	P_{SBC}	P_s
E_4 Chain system [1]	P	4	[4,256,256,4]	1	4	1.105 s	95.45%	98.19%
E_5 Raychaudhuri system [8]	P	4	[4,256,256,4]	1	4	3.632 s	75.28%	89.24%
E_6 Stabilization system [12]	P	5	[5,256,256,2]	1	6	5.410 s	86.44%	90.57%
E_7 UAV [10]	NP	6	[6,256,256,2]	1	6	52.181 s	63.99%	89.92%

VRAM) GPU, under Windows 11 with 32GB RAM, and all the experiment results are presented in Table 1 and Table 2 with the following information:

- The 'Environment' column lists the benchmark sources, the 'Poly.' column identifies each system as polynomial (P) or non-polynomial (NP), and n_x denotes the system's state space dimension.
- For each stochastic control instance, the results of our method are recorded in the '$Larac_{SBC}$' columns. The architectures of the NN controllers trained by RL are listed in the 'NN' column, the degrees of the polynomial controllers obtained via PAC-approximation are listed in the 'd_c' column, the degrees of the synthesized SBCs are listed in the 'd_{SBC}' column, the associated reach-avoid probability lower bounds are listed in the 'P_{SBC}' column, and the time costs of SBC synthesis are listed in the 'T_{SBC}' column.
- The simulation-based reach-avoid probability estimates are reported in the 'P_s' column.

Table 1 compares the proposed $Larac_{SBC}$ method with the RASM approach of [16]. The column 'P_{RASM}' corresponds to the resulting certified probability lower bounds of RASM. We can see that for all benchmarks $E_1 - E_3$ considered in [16], the controllers synthesized by $Larac_{SBC}$ yield reach-avoid probability guarantees that are consistently higher. The superiority of our proposed $Larac_{SBC}$ method arises from its verification strategy. $Larac_{SBC}$ 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 probability value certified by interval-analysis-based verification. Therefore $Larac_{SBC}$ can achieve more accurate probability guarantees efficiently.

Table 2 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, shown in the 'P_{SBC}' and 'P_s'columns respectively, 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 successfully synthesizes a controller that guarantees a reach-avoid probability of at least 64%. The larger gap observed in this case is primarily due to the difficulty, in higher-dimensional systems, of constructing sufficiently expressive SBCs. In contrast, the RASM method is unable to handle the higher-dimensional benchmarks in Table 2, primarily because it relies on interval arithmetic to formally verify the learned certificates. Such approaches inherently depend on state-space partitioning, whose computational complexity grows exponentially with the system dimension.

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.

6 Conclusion

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

References

[1] Alessandro Abate, Daniele Ahmed, Alec Edwards, Mirco Giacobbe, and Andrea Peruffo. 2021. FOSSIL: a software tool for the formal synthesis of lyapunov functions and barrier certificates using neural networks. In *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control (HSCC '21)*. Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. https://doi.org/10.1145/3447928.3456646

[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540* (2016).

[3] G. Calafiore and M. Campi. 2006. The scenario approach to robust control design. *IEEE Transactions on automatic control* 51, 5 (2006), 742–753.

[4] M. Campi, S. Garatti, and M. Prandini. 2009. The scenario approach for systems and control design. *Annual Reviews in Control* 33, 2 (2009), 149–157.

[5] Krishnendu Chatterjee, Thomas A. Henzinger, Mathias Lechner, and Đorđe Žikelić. 2023. A learner-verifier framework for neural network controllers and certificates of stochastic systems. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 3–25.

[6] X. Chen, E. Abraham, and S. Sankaranarayanan. 2012. Taylor model flowpipe construction for non-linear hybrid systems. IEEE, 183–192.

[7] Charles Dawson, Sicun Gao, and Chuchu Fan. 2023. Safe Control With Learned Certificates: A Survey of Neural Lyapunov, Barrier, and Contraction Methods for Robotics and Control. *Trans. Rob.* 39, 3 (2023), 1749–1767. doi:10.1109/TRO.2022.3232542

[8] Antoni Ferragut and Armengol Gasull. 2014. Seeking Darboux Polynomials. *Acta Applicandae Mathematicae* 139 (2014), 167 – 186. https://api.semanticscholar.org/CorpusID:254189880

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 1861–1870.

[10] W. Jin, Z. Wang, Z. Yang, and S. Mou. 2020. Neural Certificates for Safe Control Policies. (2020).

[11] Mathias Lechner, Đorđe Žikelić, Krishnendu Chatterjee, and Thomas A. Henzinger. 2022. Stability verification in stochastic control systems via neural network supermartingales. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 36. 7326–7336.

[12] Mohamed Amin Ben Sassi and Sriram Sankaranarayanan. 2015. Stabilization of polynomial dynamical systems using linear programming based on bernstein polynomials. *arXiv preprint arXiv:1501.04578* (2015).

[13] Sean Summers and John Lygeros. 2010. Verification of discrete time stochastic hybrid systems: A stochastic reach-avoid decision problem. *Automatica* 46, 12 (2010), 1951–1961. doi:10.1016/j.automatica.2010.08.006

[14] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

[15] A. Vahidi and A. Eskandarian. 2003. Research advances in intelligent collision avoidance and adaptive cruise control. *IEEE Transactions on Intelligent Transportation Systems* 4, 3 (2003), 143–153. doi:10.1109/TITS.2003.821292

[16] Đorđe Žikelić, Mathias Lechner, Thomas A. Henzinger, and Krishnendu Chatterjee. 2023. Learning control policies for stochastic systems with reach-avoid guarantees. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1338, 10 pages. doi:10.1609/aaai.v37i10.26407

[17] David Williams. 1991. *Probability with martingales*. Cambridge university press.

[18] Bai Xue. 2024. Reach-Avoid Controllers Synthesis for Safety Critical Systems. *IEEE Trans. Automat. Control* 69, 12 (2024), 8892–8899. doi:10.1109/TAC.2024.3423837

[19] Bai Xue. 2024. Sufficient and necessary barrier-like conditions for safety and reach-avoid verification of stochastic discrete-time systems. *arXiv preprint arXiv:2408.15572* (2024).

[20] Zhengfeng Yang, Li Zhang, Xia Zeng, Xiaochao Tang, Chao Peng, and Zhenbing Zeng. 2023. Hybrid Controller Synthesis for Nonlinear Systems Subject to Reach-Avoid Constraints. In *Computer Aided Verification: 35th International Conference, CAV 2023, Paris, France, July 17–22, 2023, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 304–325. doi:10.1007/978-3-031-37706-8_16

[21] Xia Zeng, Banglong Liu, Zhenbing Zeng, Zhiming Liu, and Zhengfeng Yang. 2024. Safe Controller Synthesis for Nonlinear Systems via Reinforcement Learning and PAC Approximation. In *Proceedings of the 61st ACM/IEEE Design Automation Conference* (San Francisco, CA, USA) *(DAC '24)*. Association for Computing Machinery, New York, NY, USA, Article 283. doi:10.1145/3649329.3657332