Guarantees on Robot System Performance Using Stochastic Simulation Rollouts

Joseph A. Vincent, Aaron O. Feldman, and Mac Schwager, *Member, IEEE*

***Abstract*—In this article, we provide finite-sample performance guarantees for control policies executed on stochastic robotic systems. Given an open- or closed-loop policy and a finite set of trajectory rollouts under the policy, we bound the expected value, value at risk, and conditional value at risk of the trajectory cost, and the probability of failure in a sparse cost setting. The bounds hold, with user-specified probability, for any policy synthesis technique and can be seen as a postdesign safety certification. Generating the bounds only requires sampling simulation rollouts, without assumptions on the distribution or complexity of the underlying stochastic system. We adapt these bounds to also give a constraint satisfaction test to verify the safety of the robot system. We provide a thorough analysis of the bound sensitivity to sim-to-real distribution shifts and provide results for constructing robust bounds that can tolerate some specified amount of distribution shift. Furthermore, we extend our method to apply when selecting the best policy from a set of candidates, requiring a multihypothesis correction. We show the statistical validity of our bounds in the Ant, Half-cheetah, and Swimmer MuJoCo environments and demonstrate our constraint satisfaction test with the Ant. Finally, using the 20-degree-of-freedom MuJoCo Shadow Hand, we show the necessity of the multihypothesis correction.**

***Index Terms*—Motion and path planning, optimization and optimal control, probability and statistical methods, risk-sensitive control.**

I. Introduction

IT IS essential that robots be able to operate safely and successfully under diverse sources of uncertainty, including uncertainty about their own dynamics and state, friction and contact forces, the future motion of other agents, and environment geometry. For example, robot manipulators must interact with objects having uncertain geometries or physical parameters, legged robots must locomote on uncertain terrain, and autonomous vehicles must avoid colliding with other agents whose future trajectories are uncertain. While it is common to have a simulation model reflecting these diverse sources of uncertainty, we rarely have access to closed-form mathematical models, making it difficult to provide rigorous performance and safety guarantees. We address this challenge, presenting statistical performance bounds and safety tests for arbitrary robotic systems given a finite set of trajectory samples from a stochastic simulator.

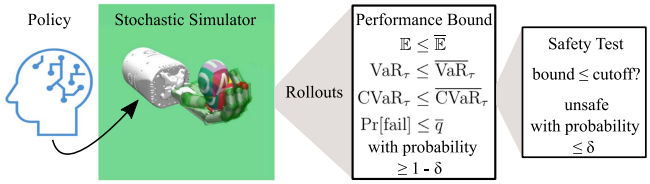

Fig. 1. Overview of our method for bounding performance for a single policy using a stochastic simulator. The policy is executed in simulation n times to collect trajectory rollouts. The cost or constraint function is evaluated for each rollout, and these samples are used to form a distribution-free upper bound on a given performance measure (expected value, value at risk, conditional value at risk, or probability of failure) that is guaranteed to hold with probability at least $1 - \delta$. These probabilistic bounds may also be used to ensure safety by testing constraint satisfaction for the performance measures. The finite-sample bound guarantee ensures that these tests incorrectly accept a policy as safe with at most δ probability. We demonstrate this pipeline for several MuJoCo environments and extend the method to compare multiple policies for manipulating an egg of uncertain mass and friction.

Performance for stochastic robotic systems is typically quantified with an expected trajectory cost, or with risk-sensitive performance measures such as value at risk (VaR) or conditional value at risk (CVaR). These risk-sensitive measures have been widely used in the optimization of financial portfolios [1] and have more recently been adopted for use in robotic control problems [2], [3]. Performance can also be quantified by the probability of task success (e.g., probability that an object is not dropped or a robot does not fall). Symmetrically, safety is often enforced by putting constraints on these performance measures: expected value constraints, VaR constraints, CVaR constraints, or constraints on success probability. We provide probabilistic bounds for all of these performance measures.

Most of the existing approaches for computing or bounding expected value, VaR, CVaR, or success probability: 1) assume a known distribution for the uncertainty (e.g., Gaussian); 2) use a large number of simulation rollouts to approximate uncertainty without formal guarantees; or 3) provide formal guarantees that hold asymptotically as the number of samples approaches

Manuscript received 3 June 2024; accepted 5 August 2024. Date of publication 15 August 2024; date of current version 30 August 2024. This work was supported in part by the NASA University Leadership initiative under Grant #80NSSC20M0163 and in part by the Office of Naval Research under Grant N00014-23-1-2354. The work of Joseph A. Vincent was also supported by a Dwight D. Eisenhower Transportation Fellowship. The work of Aaron O. Feldman was also supported by a National Science Foundation Graduate Research Fellowship under Grant 2146755. This article was recommended for publication by Associate Editor M. R. Dogar and Editor Paolo R. Giordano upon evaluation of the reviewers' comments. *(Joseph A. Vincent and Aaron O. Feldman contributed equally to this work.) (Corresponding author: Joseph A. Vincent.)*

The authors are with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 USA (e-mail: josephav@stanford.edu; aofeldma@stanford.edu; schwager@stanford.edu).

The code associated with this work can be found at https://github.com/StanfordMSL/performance_guarantees.

Digital Object Identifier 10.1109/TRO.2024.3444070

infinity. In contrast, our methods are both distribution-free (they require no knowledge of the underlying probability distribution) and finite-sample (they hold with a finite number of samples). A core insight is that the trajectory cost of the robotic system can be treated as a scalar random variable, and a stochastic simulator can be viewed as an elaborate random number generator, producing independent and identically distributed (IID) samples of the trajectory cost. We can, therefore, apply distribution-free finite-sample statistical analysis tools. We only require two key assumptions.

1) The simulative dynamics model exactly represents the stochastic dynamics encountered during execution, in addition to modeling the roboticist's uncertainty in initial conditions and simulation parameters.
2) Successive simulations are IID, that is, there is no memory or distributional shift between trajectory rollouts.

For the first assumption, there are three uncertainties the simulator may capture: 1) uncertainty over state transitions; 2) uncertainty over initial conditions; and 3) uncertainty over simulation parameters. We assume that these uncertainties are modeled as random variables as is common in, e.g., state estimation and filtering problems.

We derive simple formulas using foundational statistical principles to compute upper bounds on expected value, VaR, and CVaR, as well as an upper bound on the probability of failure in a binary success/failure reward model. These bounds are probabilistic, that is, they are allowed to be wrong δ proportion of the time, for a user-specified error rate δ. We further adapt these bounds to give constraint satisfaction tests for the expected value, VaR, CVaR, and failure probability, with a guaranteed user-specified false positive rate (declaring the system safe, when it is actually unsafe). The probability of the bound being valid $(1-\delta)$ is often referred to as the *confidence* or *coverage* level of the bound. In Section VI, we investigate how the confidence level for each bound changes when there is sim-to-real distribution shift and Assumption 1 is not met. In Section VII, we detail how to retain a desired confidence level in light of such distribution shifts.

An overview of our method for bounding performance and testing constraint satisfaction for a policy is shown in Fig. 1. We also modify the bounds, so they can be used to compare performance among multiple policies, as outlined in Fig. 7. Notably, we show that a multihypothesis correction is required to retain statistical guarantees when comparing multiple policies. We empirically demonstrate the validity of our bounds and constraint satisfaction tests in several MuJoCo [4] environments simulated in Gymnasium [5] and verify our policy comparison bounds in a 20-degree-of-freedom MuJoCo Shadow Hand simulation manipulating an object with uncertain mass and friction.

Our bounds are independent of the system's complexity or dimensionality. They only depend on the number of rollouts and the user-specified error rate. Therefore, our method is appropriate for complex simulation models of high-degree-of-freedom robots, featuring, e.g., discontinuities from contact, uncertainties in friction or reaction forces, fluidic or finite-element simulations for soft robots and deformable objects, and aerodynamic simulations for aerial robots. The bounds also apply regardless of how the simulation is derived, applying also when the simulation itself is learned (e.g., generative world models or multiagent trajectory forecasters). The "simulation" can also be an experimental setup, in which a control policy is repeatedly executed on a physical robot. Furthermore, our bounds are valid for open- or closed-loop policies, as well as deterministic or stochastic policies.

We emphasize that we do not present a new policy optimization or planning technique in this work, but rather present a statistical method for bounding the performance of a given control policy or comparing performance among a set of policies. Our methods can be used as a verification tool to bound the performance or certify the safety of a policy obtained from any upstream optimizer (e.g., a reinforcement learning or optimal control design technique or even a large language model task planner).

In summary, our primary contributions are as follows.

1) Given an open- or closed-loop control policy, we apply probabilistic upper bounds for expected value, VaR, CVaR, and probability of task failure, from a finite set of simulated trajectory rollouts.
2) Similarly, we obtain a probabilistic test to verify the satisfaction of constraints on the expected value, VaR, CVaR, or probability of failure. The test has a user-specified false acceptance rate.
3) We describe a necessary multihypothesis correction to these performance bounds in the case of choosing the best among a finite set of candidate policies.
4) We provide analytic expressions for how the confidence of each bound changes as a function of sim-to-real distribution shift, and we provide simple extensions to each bound to ensure that a desired confidence level holds under a specified amount of sim-to-real distribution shift.

We achieve the above for an arbitrarily complex simulator, learned or model-based, with diverse sources of uncertainty, and with any upstream policy generation or optimization approach. As evidence to this, we demonstrate our approach in several MuJoCo environments including with the 20-degree-of-freedom Shadow Hand.

The rest of this article is organized as follows. We give related work in Section II. In Section III, we introduce our notation and define the problem setting. In Section IV, we present the distribution-free performance bounds and derive constraint satisfaction tests. We empirically validate the bounds and constraint tests in MuJoCo simulations in Section V. In Section VI, we provide expressions for how the confidence level for each bound changes due to sim-to-real distribution shift. In Section VII, we detail how to retain a desired confidence level given some amount of distribution shift. In Section VIII, we introduce the correction required when comparing bounds among multiple policies and demonstrate the validity of this correction in Section IX in simulations with the MuJoCo Shadow Hand manipulating an uncertain object. Finally, Section X concludes this article. We give proofs of theorems and computational details in the Appendix.

II. Related Work

A. Sample-Based Performance Quantification

We build on an emerging literature that uses finite samples from simulation rollouts to produce and optimize for distribution-free guarantees on system performance. Using results from randomized optimization, Akella et al. [6] construct distribution-free bounds on the VaR of a given robustness metric for a robotic system. Their bound on the VaR is a special case of our bound given in Theorem 1. Similar methods are used in [7] to verify safety and robustness of a reinforcement learning policy and in [8] to place probabilistic bounds on the error of a simulated model.

Akella et al. [9] use distribution-free statistics to place bounds on coherent risk measures (such as CVaR) and on the quality of optimization via random search. They use these results to randomly search over policies and choose the one with the least upper bound on risk. In [10], the same bounds are used to find a nonlinear control plan that is better than a specified percentage of plans. As we show in this work, when policy cost is not deterministic, optimizing for the sample-based bound requires a multihypothesis correction to retain validity. This is a crucial step in the policy synthesis process that our article addresses in Section VIII.

Cleaveland et al. [11] perform the verification of closed-loop stochastic systems. Like us, they take a distribution-free finite-sample approach, but with some key differences. While they are primarily interested in verifying closed-loop systems with neural network controllers, we take a broader view to systems, which need not be differentiable, continuous, or defined in closed form. Cleaveland et al. [11] also discuss choosing the least risky controller from a set of controllers but fail to note the need for a multihypothesis correction. Finally, we use a tighter VaR bound not based on the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [12], [13] and take a different approach to constraints; we provide analysis for handling a variety of risk-sensitive constraints, whereas Cleaveland et al. [11] focus on signal-temporal-logic constraints. Finally, none of these works address how bound confidence changes with distribution shift.

B. Risk-Sensitive Control

Our method can be used to certify policies obtained from existing risk-sensitive control techniques. We classify the approaches for risk-sensitive control into three broad categories: parametric, distributionally robust, and sampling-based.

In the parametric category are works imposing distributional and structural assumptions to efficiently quantify risk. Carpin et al. [3] compute risk averse policies (in the sense of CVaR) for finite state and action Markov decision processes (MDPs) by solving a surrogate MDP. Ahmadi et al. [14] synthesize a CVaR-safe controller for linear systems using barrier functions. Lew et al. [15] enforce obstacle avoidance chance constraints assuming a Gaussian dynamics disturbance. While parametric approaches provide efficient mechanisms for quantifying and mitigating risk, we avoid the associated assumptions (i.e., on the dynamics or uncertainty distribution), so that our approach can be applied as a general certification step for arbitrary complex systems.

Instead of assuming a particular uncertainty distribution, some work adopts a distributionally robust approach. The authors of [16] and [17] propose CVaR constrained control where the CVaR is first estimated empirically. Then, based on a known ambiguity set for the disturbance distribution (using the Wasserstein metric) about the empirical CVaR, distributionally robust CVaR constraints are enforced at runtime. Dixit et al. [18] add a constraint on the entropic value at risk in their model-predictive control (MPC) formulation using its dual representation as the worst case expectation within a Kullback–Leibler-divergence-based ambiguity set. In contrast with these works focusing on distribution mismatch, we focus on the setting where our simulator provides samples from the true uncertainty distribution, but our bounds do not require knowledge of that distribution. However, we also provide extensions in Section VII to construct robust bounds that hold even under simulator mismatch. Specifically, we use the simulator and a given robustness tolerance to implicitly define a distributional ambiguity set when constructing the robust bounds.

Sampling-based approaches use repeated draws of empirical performance to estimate the risk without imposing distributional assumptions. In [19], each agent in a team estimates its VaR using the empirical quantile of recently observed rewards. Hiraoka et al. [20] impose a CVaR constraint during policy optimization by rewriting the CVaR as a tail expectation so that it can be approximated from rollouts. During MPC, the authors of [21] and [22] repeatedly rollout controls under the stochastic dynamics and optimize for the sequence minimizing the average associated trajectory cost.

Lew et al. [23] show that, under certain conditions, the sample average solution becomes asymptotically optimal. However, in this work, we are interested in finite-sample performance guarantees. Lew et al. [24] applies the work in [23] to CVaR-constrained trajectory optimization and provides a finite-sample bound on the CVaR constraint. However, they leverage concentration inequalities, which hold uniformly across all controls. Our bounds hold pointwise, necessitating a multihypothesis correction, but when comparing only a modest number of policies, this can yield tighter guarantees. While sampling-based control provides context for our work, we augment such methods with finite-sample distribution-free statistical guarantees.

C. Conformal Prediction

Conformal prediction (CP) is increasingly being used for producing distribution-free guarantees in robotics. Given exchangeable data (a weaker condition than IID) and a scoring function, CP produces a confidence interval on the score of a new data sample [25]. In this work, we adapt analysis tools from CP to obtain our VaR bound and to assess chance constraint satisfaction. Similar to our work, several papers applying CP to robotics have viewed full robot trajectories as one sample. Luo et al. [26] use collected data of unsafe trajectories to augment robotic fault detection systems, achieving a guaranteed false negative rate. Lindemann et al. [27] use CP to augment learned

forecasting models (e.g., predicting pedestrian motion) with a confidence set of possible trajectories. Dixit et al. [28] use adaptive CP to now adapt their confidence sets using online data. As in much of the CP robotics literature, we also view trajectories as the fundamental sample to get IID data. Instead of using offline data, we evaluate performance using a generative stochastic simulator.

D. Concentration Bounds

Key to our approach is the use of sampling-based distribution-free concentration bounds for risk measures as this allows us to produce rigorous performance guarantees when planning with any arbitrarily complex simulator.

Concentration bounds for CVaR were first given by Brown [29], further refined by Wang and Gao [30], and later improved upon by Thomas and Learned-Miller [31]. Each of these results requires bounds on the support of the random variable. Later, Kolla et al. [32] provided concentration bounds in the case the random variable is sub-Gaussian or subexponential (weaker than boundedness by Hoeffding's lemma [33]), which were improved upon in [34]. The CVaR bound we use in this article is from [31], but we express it in a simpler form and give an accompanying proof. We then extend this bound to give a novel bound on the expected value.

Unlike CVaR bounds, VaR bounds place no finite support or sub-Gaussianity restrictions. VaR bounds are often derived via the DKW inequality [32], [35], [36]. Although some authors claim these bounds as contributions, an optimal VaR bound has been available since 2005 [37] (although this optimal VaR bound is derived for continuous random variables, it can be extended for discontinuous random variables using the methods of [38]). In this work, we use a slightly suboptimal VaR bound because of its simple form and derivation, which we present in the Appendix. The VaR bound we use is well established in the statistics community, dating back to 1945 (see [38] and [39]), but its application to bounding policy performance is new. This classical bound is seemingly unknown to many practitioners and is strictly better than the bounds given in [32] and [35].

Concentration bounds on the expected value are more studied than for the VaR or CVaR. Hoeffding's inequality is arguably the most influential bound of this sort [33] although the bounds given later by Anderson are no worse and often better [40]. More modern work has focused on bounds that are not defined in closed form [41] and concentration sequences [42]. The expectation bound we use in this article is equivalent to the Anderson bound [40], but derived and expressed in a slightly different form.

Finally, bounds for the probability of success parameter in the Bernoulli distribution (often called binomial confidence intervals) have been studied extensively. Unlike other concentration bounds, the majority of methods for binomial confidence intervals are not guaranteed to hold with a user-defined probability (e.g., Wald, Wilson, Jeffreys', Agresti–Coull, etc. as described in [43] and [44]). Methods guaranteed to meet the desired confidence level include those by Clopper and Pearson [45], Sterne [46], Crow [47], Eudey [48], [49], and Stevens [50]. Of these approaches, only the bounds from Eudey and Stevens return confidence intervals with exactly the desired coverage, and they do so by inverting randomized hypothesis tests. We show that the confidence interval we derive in Theorem 4 is analogous to the one-sided Clopper–Pearson interval, which is known to be unimprovable among nonrandomized approaches [51].

III. Problem Setting

Here, we introduce notation and formalize the problem of quantifying performance and safety for a stochastic robotic system. For a given fixed time horizon T, we consider either an open-loop policy as a sequence of control actions $(U_0, \ldots, U_{T-1})$ or a closed-loop policy as a mapping (deterministic or stochastic) from state to action $U_t \sim \pi_t(X_t)$. We use the term policy to describe all of these cases and cover all cases with the notation $\mathcal{U}$ to represent the stochastic sequence of control actions obtained by executing the policy in simulation. When executing the policy, the control actions drive state evolution from a random starting state X_0 via stochastic dynamics F_t

$$X_0 \sim \mathcal{X}_0 \tag{1a}$$

$$X_{t+1} \sim F_t(X_t, U_t), \quad t = 0, \ldots, T-1. \tag{1b}$$

We do not assume access to an explicit functional form or distribution for F_t, but instead we assume that we can sample $X_{t+1} \sim F_t(X_t, U_t)$ IID from a simulation of the system. We write the stochastic state trajectory as $\mathcal{X} = (X_0, \ldots, X_T)$, as with the control policy $\mathcal{U}$. The trajectory cost J associated with a policy and a state sequence realization is composed of stagewise costs c_t as

$$J(\mathcal{X}, \mathcal{U}) = c_T(X_T) + \sum_{t=0}^{T-1} c_t(X_t, U_t). \tag{2}$$

As an important alternative case, we also consider the sparse reward setting often seen in reinforcement learning. We let $J = 1$ denote task failure and $J = 0$ denote task success, to keep the interpretation of J as a cost rather than a reward.

In addition to a cost function, we also consider a constraint function g, which may be used to impose trajectory constraints for safety (such as avoiding collisions, or avoiding control input limits) or for task success (such as not dropping an object, not falling down, or attaining a discrete goal). We consider the generated trajectory a success when it satisfies a given set of trajectory constraints

$$g(\mathcal{X}, \mathcal{U}) \leq 0. \tag{3}$$

The aforementioned constraint also applies in the binary success/failure setting, in which case $g = 0$ is task success and $g = 1$ is task failure, again with the convention of lower g being safer. When executing the policy, we assume that cost J can still be defined even when the trajectory violates safety constraints and do not truncate or reject such trajectories. Rather, "soft" safety penalties may be incorporated by modifying the stagewise costs c_t in J, in addition to g, which separately encodes "hard" safety constraints.

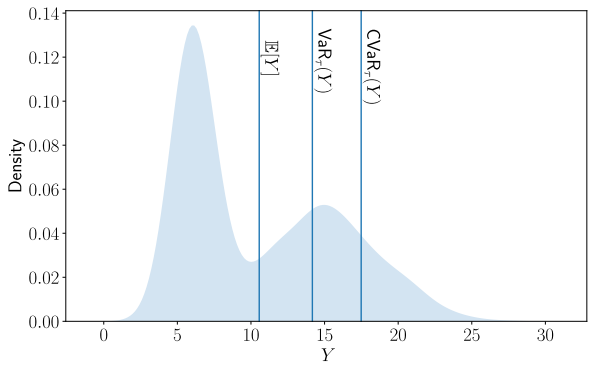

Fig. 2. Visualization of the expected value, VaR_τ, and CVaR_τ for an example distribution. Classic stochastic optimal control and reinforcement learning both seek to minimize the expected value of the total cost distribution. Risk-sensitive stochastic optimal control and reinforcement learning consider other measures of performance, such as VaR or CVaR of the cost.

Though the policy is fixed, the associated cost J and the constraint function g evaluate to random values due to the stochastic dynamics, which generate different state sequence realizations on different runs under the same policy. Therefore, to measure system performance or to impose safety constraints, we must define a summarizing statistic over the cost or constraint functions, which we call a performance measure.

A. Performance Measures

Consider a scalar random variable Y, representing either the value of the cost function J or the constraint function g. We define a performance measure $\mathcal{P}(Y)$ as a summary statistic for the random variable Y. The most common choice of performance measure is the expected value $\mathcal{P}(Y) = \mathbb{E}[Y]$. However, to be more conservative, we often consider alternative measures that are risk sensitive, such as the VaR or CVaR of the trajectory cost. VaR, expected value, and CVaR are defined in the following and visualized in Fig. 2. Alternatively, in the binary success/failure setting, the probability of failure is the natural performance measure to be minimized. We consider finite-sample methods for upper bounding these quantities.

Recall that the cumulative distribution function (CDF), which exists for any scalar random variable, whether continuous or not, is defined as $\text{CDF}(y) := \Pr[Y \leq y]$. We define $\text{VaR}_\tau(Y)$ in terms of the CDF as

Definition 1 (VaR): Given a scalar random variable Y, the VaR of Y at quantile $\tau \in (0, 1)$ is

$$\text{VaR}_\tau(Y) := \inf\{y \mid \text{CDF}(y) \geq \tau\}. \tag{4}$$

If Y has an invertible CDF, then

$$\text{VaR}_\tau(Y) = \{y \mid \text{CDF}(y) = \tau\} = \text{CDF}^{-1}(\tau). \tag{5}$$

In plain words, $\text{VaR}_\tau = y$ ensures that τ proportion of the probability mass of the random variable Y is below the value y. The VaR can be seen as a generalization of the inverse of the CDF and closely relates to the quantile as used in statistics.

Recall that the standard definition of the expected value of a continuous scalar random variable is $\mathbb{E}[Y] = \int_{-\infty}^{\infty} yp(y)\,dy$, where $p(y)$ is the probability density function. In fact, a more general definition of the expected value can be stated in terms of the VaR_τ as follows.

Definition 2 (Expected value): Given scalar random variable Y, the expected value of y is

$$\mathbb{E}[Y] := \int_0^1 \text{VaR}_\tau(Y)\,d\tau. \tag{6}$$

The aforementioned definition applies to any scalar random variable (continuous, discrete, or mixed). In the case of a continuous random variable with the invertible CDF, one can recover the standard definition through a change of variables[1] $\tau = \text{CDF}(y)$.

One criticism of VaR as a risk-sensitive performance measure is that it ignores high-cost outcomes that may lie in the $1 - \tau$ rightmost tail of the distribution. CVaR, defined below, addresses this concern.

Definition 3 (CVaR): Given scalar random variable Y and $\tau \in [0, 1)$, the CVaR of Y at quantile τ is

$$\text{CVaR}_\tau(Y) := \frac{1}{1-\tau}\int_\tau^1 \text{VaR}_\gamma(Y)\,d\gamma. \tag{7}$$

If Y has an invertible CDF, then

$$\text{CVaR}_\tau(Y) = \mathbb{E}[Y \mid Y \geq \text{VaR}_\tau(Y)]. \tag{8}$$

There are several equivalent definitions for $\text{CVaR}_\tau(Y)$. We prefer the one above as it applies to all scalar random variables (continuous, discrete, or mixed), and it highlights the clear relationship between $\text{VaR}_\tau(Y)$ and $\mathbb{E}[Y]$. We can see that $\text{CVaR}_\tau(Y)$ is simply the expected value of Y taken over the top $1 - \tau$ tail of the probability mass (and renormalized by $1 - \tau$ to ensure that it remains a valid expectation). Taking $\tau = 0$ recovers the expected value. In contrast to VaR_τ, CVaR_τ captures high-cost tail events. It, therefore, incorporates the worst case cost situations and is often quite conservative. $\text{CVaR}_\tau(Y)$ is always greater than or equal to both $\mathbb{E}[Y]$ and $\text{VaR}_\tau(Y)$. However, $\mathbb{E}[Y]$ and $\text{VaR}_\tau(Y)$ may lie in either order, depending on τ and the distribution of the random variable Y.

For simple distributions (e.g., Gaussians), $\mathbb{E}[Y]$, $\text{VaR}_\tau(Y)$, and $\text{CVaR}_\tau(Y)$ can be found. However, in more realistic robotics scenarios, distributions are non-Gaussian and usually are unknown, so none of these performance measures can be obtained in closed form. This motivates the need in this article to give finite-sample bounds for these quantities.

We next formalize the probability of failure as a performance measure for problems with binary success/failure cost models.

Definition 4 (Failure probability): Given a Bernoulli random variable Y, we refer to $q = \Pr[Y = 1]$ as the failure probability.

Note that failure probability is actually a special case of expected value, since $\Pr[Y = 1] = \mathbb{E}[Y]$ for a binary random variable Y. However, the binary case allows for significantly tighter bounds than the general bounds on the expected value. We, therefore, treat this case separately.

[1]With the change of variables $\tau = \text{CDF}(y)$, we have $d\tau = p(y)\,dy$ (recalling that $p(y) = dCDF(y)/dy$). The domain of integration $\tau \in (0, 1)$ becomes $y \in (-\infty, \infty)$, and since $\text{VaR}_\tau(Y) = \text{CDF}^{-1}(\tau)$, the integrand becomes $\text{CDF}^{-1}(\text{CDF}(y)) = y$, leading to $\int_{-\infty}^{\infty} yp(y)\,dy$.

B. Cost Performance and Safety Performance

The aforementioned performance measures can be applied to the cost function $\mathcal{P}(J)$ to measure cost performance or to the constraint function $\mathcal{P}(g)$ to measure safety performance. In this article, we present bounds $\overline{P}$ to probabilistically bound the cost performance from a finite set of simulation rollouts. Specifically, we give bounds of the form

$$\Pr[\mathcal{P}(J) \leq \overline{P}] \geq 1 - \delta \tag{9}$$

where δ is a user-defined error rate for the bound.

Similarly, safety is often quantified by putting constraints on a performance measure applied to the constraint function, $\mathcal{P}(g) \leq C$, for some constant C. Specifically, we consider safety as a bound on any of the aforementioned performance measures:

$$\mathbb{E}[g] \leq C_{\mathbb{E}} \tag{10a}$$

$$\text{VaR}_\tau(g) \leq C_{\text{VaR}} \tag{10b}$$

$$\text{CVaR}_\tau(g) \leq C_{\text{CVaR}} \tag{10c}$$

$$\Pr[g = 1] \leq C_q. \tag{10d}$$

where in (10d) we consider the case where g is assumed to be binary with $g = 1$ indicating failure.

The aforementioned formulation in (10) already captures chance constraints (as seen in stochastic optimal control), which require that the trajectory constraint (3) be satisfied with sufficiently high probability. This follows because constraining $\text{VaR}_\tau(g)$ in (10b) is equivalent to imposing a chance constraint on g

$$\text{VaR}_\tau(g) \leq 0 \iff \Pr[g \leq 0] \geq \tau. \tag{11}$$

In fact, chance constraints can also be modeled as a binary success/failure of the type (10d), where $g = 1$ denotes the event that the trajectory fails the constraint (3), and $g = 0$ for a successful trajectory satisfying the constraint.

Note that we cannot directly evaluate whether the constraints in (10) hold, because we cannot compute their left-hand side when we only have access to a simulator. We, therefore, define a test for whether $\overline{P} \leq C$, using the finite-sample bound $\overline{P}$. If $\overline{P}$ passes this test, we declare the constraint satisfied. In the following section, we prove that such a test can be constructed with a user-specified false positive error rate, only concluding that an unsafe policy is safe some small fraction of the time.

IV. Finite-Sample Performance Bounds

In this section, we provide finite-sample upper bounds for the expected value, VaR, CVaR, and failure probability. We also define constraint satisfaction tests based on these bounds. We require access to IID samples of the total cost J or constraint function g under the policy being evaluated, which we assume are obtained from repeatedly executing the policy for a given time horizon T in a stochastic simulator of the robot system. Each bound presented holds probabilistically, with probability at least $1 - \delta$, where the randomness stems from the bound itself being a function of a finite set of random samples. In fact, if no distributional assumptions are made, one can only formulate bounds that hold probabilistically (see [52, Sec. 5]). We refer to δ as the user-specified error rate for the bound and to the guaranteed probability $1 - \delta$ that the bound holds as the confidence level of the bound. All proofs are deferred to the Appendix. An overview of the method is shown in Fig. 1.

As explained previously, our results rest upon two foundational assumptions.

Assumption 1 (Accurate simulation model): The simulative dynamics model in (1) exactly represents the stochastic dynamics encountered during execution, in addition to modeling the roboticist's uncertainty in initial conditions and simulation parameters.

Assumption 2 (IID): Successive simulations are IID, that is, there is no memory or distributional shift between trajectory rollouts.

Of course, these assumptions will never exactly hold in practice, as there is always some sim-to-real gap. However, qualitatively, the closer the simulation model is to the real robotic system, the more reliable the bounds will be. In Section VI, we address the sensitivity of the bounds to some sim-to-real gap, and in Section VII, we show how to construct bounds, which are robust to this gap.

A. Performance Bounds

The bounds make use of the concept of order statistics, defined as follows.

Definition 5 (Order statistics): For a set of n samples $J_{1:n}$ drawn IID, we let $J_{(k)}$ denote the kth order statistic, obtained by arranging the samples in order from smallest to largest and taking the kth element in the sequence.

Definition 6 (Binomial distribution): Let $\text{Bin}(k; m, p)$ denote the Binomial CDF, with m trials, success probability p, evaluated at k successes.

Theorem 1 (VaR bound): Consider $\tau, \delta \in (0, 1)$ and n IID cost samples $J_{1:n}$, and let k be the smallest index such that $\text{Bin}(k - 1; n, \tau) \geq 1 - \delta$. We have the following probabilistic upper bound on $\text{VaR}_\tau(J)$:

$$\overline{\text{VaR}}_\tau := J_{(k)} \tag{12}$$

which has the property

$$\Pr[\text{VaR}_\tau(J) \leq \overline{\text{VaR}}_\tau] \geq 1 - \delta.$$

A feasible value for k exists when $n \geq \lceil \ln(\delta)/\ln(\tau) \rceil$, i.e., n is large enough to ensure $\text{Bin}(n - 1; n, \tau) \geq 1 - \delta$.

The aforementioned VaR bound simply chooses one of the order statistics based on a test involving the cumulative binomial distribution. Its proof (in the Appendix) is inspired by similar analyses in CP [25], [53]. In our experience, this bound is considerably tighter than other VaR bounds in the recent literature (see, e.g., [32] and [35]) and tends to be tighter in practice than the CVaR and $\mathbb{E}$ bounds below. As previously stated, the form of this bound is well known in the statistics literature [38], [39], but its application to bounding policy performance is new. The VaR bound in [9] is a special case of this one, which considers the bound arising from the largest order statistic, $J_{(n)}$.

Before stating the CVaR and $\mathbb{E}$ bounds, we require an additional assumption.

Assumption 3 (Almost sure upper bound): We have an almost sure upper bound J_{ub} such that $\Pr[J \leq J_{\text{ub}}] = 1$.

It may seem circular to require one upper bound in order to produce another upper bound. The idea is to combine the order statistics $J_{(i)}$ with the almost sure upper bound to produce a significantly tighter bound. In fact, any finite sample upper bound on CVaR or $\mathbb{E}$ requires knowledge of such an a priori known bound on the right tail of the distribution of J. Two common choices are an almost sure upper bound (as we assume here) or a sub-Gaussian assumption. Without such a tail bound, one can adversarially construct a distribution that violates any claimed CVaR or $\mathbb{E}$ bound by placing a finite probability mass arbitrarily far to the right in the distribution, pulling both CVaR and $\mathbb{E}$ far enough right to violate the claimed bound. By contrast, VaR ignores the tail, so finite sample bounds on VaR can be constructed without a priori bounds on the tail.

Since J is computed as the sum of T stage costs, it suffices to simply find bounds on the stage cost and multiply these by T to bound the trajectory cost. For some cost functions, bounds may be computed analytically. Otherwise, in practice, one may clip the value of the cost function between some user-defined bounds, bounding the support of the total cost by construction. Computation of support bounds can be done offline and tight support bounds are not needed, but tighter support bounds on J will lead to tighter bounds on $\mathbb{E}[J]$ and $\text{CVaR}_\tau(J)$.

Finally, we define a constant that arises in both CVaR and $\mathbb{E}$ bounds, originating from the application of the DKW bound [12], [13] in their derivation, as discussed in the proofs in the Appendix.

Definition 7 (DKW gap): We define the DKW gap as

$$\epsilon(\delta, n) = \sqrt{\frac{-\ln \delta}{2n}}.$$

Theorem 2 (Expected value bound): Consider $\delta \in (0, 0.5]$, an almost sure upper bound J_{ub}, and n IID cost samples $J_{1:n}$. Let k be the smallest index such that $\frac{k}{n} - \epsilon \geq 0$. We have the following probabilistic upper bound on $\mathbb{E}[J]$:

$$\overline{\mathbb{E}} := \epsilon J_{ub} + \left(\frac{k}{n} - \epsilon\right) J_{(k)} + \frac{1}{n} \sum_{i=k+1}^{n} J_{(i)} \tag{13}$$

which has the property

$$\Pr[\mathbb{E}[J] \leq \overline{\mathbb{E}}] \geq 1 - \delta.$$

We require $n \geq -\frac{1}{2}\ln(\delta)$ samples to ensure $\epsilon \leq 1$. If $k = n$, the summation on the right is ignored. For $n < -\frac{1}{2}\ln(\delta)$, we default to the almost sure bound J_{ub}.

Theorem 3 (CVaR bound): Consider $\tau \in [0, 1)$, $\delta \in (0, 0.5]$, an upper bound J_{ub}, and n IID cost samples $J_{1:n}$. Let k be the smallest index such that $\frac{k}{n} - \epsilon - \tau \geq 0$. We have the following probabilistic upper bound on CVaR_τ:

$$\overline{\text{CVaR}}_\tau := \frac{1}{1-\tau}\left[\epsilon J_{ub} + \left(\frac{k}{n} - \epsilon - \tau\right) J_{(k)} + \frac{1}{n} \sum_{i=k+1}^{n} J_{(i)}\right] \tag{14}$$

which has the property

$$\Pr[\text{CVaR}_\tau(J) \leq \overline{\text{CVaR}}_\tau] \geq 1 - \delta.$$

We require $n \geq -\frac{1}{2}\ln(\delta)/(1-\tau)^2$ samples to ensure $\epsilon \leq 1 - \tau$. If $k = n$, the sum on the right is ignored. For $n < -\frac{1}{2}\ln(\delta)/(1-\tau)^2$, we default to the almost sure bound J_{ub}.

Both the aforementioned CVaR and $\mathbb{E}$ bounds take the form of a sample average over the order statistics, excluding some proportion of the smaller order statistics defined by the DKW gap $\epsilon(\delta, n)$ and including the upper bound J_{ub} and the smallest effective order statistic $J_{(k)}$ with special weightings, also determined by $\epsilon(\delta, n)$. Therefore, both of these bounds can be understood as variations on the sample average that typically serves as a proxy for the expected value. The great advantage of these bounds is that, unlike a sample average, they rigorously upper bound the unknown quantity (CVaR or $\mathbb{E}$) with a user-defined probability $1 - \delta$ without knowing the underlying probability distribution of J.

Notice that setting $\tau = 0$ in the CVaR bound gives the $\mathbb{E}$ bound, which is appealing given that this is also true of CVaR and $\mathbb{E}$ from their definitions in Definitions 2 and 3. We note that the CVaR bound in Theorem 3 is mathematically equivalent to the one derived in [31], but expressed in a different form and derived by different means.

While the bounds described can be performed for any choice of δ and τ, for a given number of samples n, as δ decreases, the actual bound values will increase as we require the bounds to hold with higher confidence. As τ increases, the actual bound values will also increase as we seek to bound a larger measure of the cost distribution (noting that $\text{VaR}_\tau(J)$ and $\text{CVaR}_\tau(J)$ are increasing with respect to τ). Similarly, as δ decreases and/or as τ increases, the minimum number of samples needed to yield (nonvacuous) bounds grows.

Finally, we introduce an upper bound on the probability of failure in a binary cost setting.

Theorem 4 (Failure probability bound): Given $\delta \in (0, 1)$ and n IID Bernoulli samples $J_{1:n}$ (where $J = 1$ denotes failure) with $k = \sum_{i=1}^{n} J_i$ failures, we have the following probabilistic upper bound on the probability of failure, $q := \Pr[J = 1]$:

$$\overline{q} = \max\{q' \in [0, 1] \mid \text{Bin}(k; n, q') \geq \delta\} \tag{15}$$

which has the property

$$\Pr[q \leq \overline{q}] \geq 1 - \delta.$$

B. Constraint Satisfaction Tests

The aforementioned theorems bound performance measures on the random cost J attained by a robotic system under a given control policy. Now, we adapt the above bounds to give a test for constraint satisfaction. Consider any of the aforementioned performance measures applied to the constraint function, $\mathcal{P}(g)$, embedded in a constraint

$$\mathcal{P}(g) \leq C.$$

Using the associated finite-sample bound, which we write generically as $\overline{P}$, we define the associated constraint test as $\overline{P} \leq C$. We have the following result.

Theorem 5 (Constraint test): The test for constraint satisfaction $\overline{P} \leq C$ has a false acceptance rate of no more than δ, i.e.,

$$\Pr[\overline{P} \leq C \mid \mathcal{P}(g) > C] \leq \delta.$$

The ability to provide a false acceptance guarantee for the constraint tests is a powerful yet natural consequence of the bound error rates. Indeed, we could not readily guarantee a false acceptance rate if using Monte Carlo estimates for the performance measures.

We noted in Section III that chance constraints

$$\Pr[g \leq 0] \geq \tau$$

can be modeled using either a binary function for $g \leq 0$ or by constraining $\text{VaR}_\tau(g)$. Using this equivalence, we can test whether a chance constraint holds via two equivalent methods:

1) by forming $\overline{\text{VaR}}_\tau = g_{(k)}$ using Theorem 1 and checking whether $g_{(k)} \leq 0$;
2) by first computing the number of successes/failures to satisfy $g(\mathcal{X}, \mathcal{U}) \leq 0$ to form $\overline{q}$ using Theorem 4 and checking whether $\overline{q} \leq 1 - \tau$.

Both the methods also require the same number of minimum samples n to ever accept: $\text{Bin}(n-1; n, \tau) \geq 1 - \delta$, although this condition is enforced directly in Theorem 1.

It may be the case that one wishes to assess a policy's safety, via a constraint satisfaction test $\mathcal{P}(g) \leq C$, and, if safe, obtain a bound on the policy's cost performance $\mathcal{P}(J)$. To disentangle these assessments, we would first repeatedly rollout the policy recording g_i to assess $\mathcal{P}(g) \leq C$. If the policy passes the safety test, i.e., $\overline{P} \leq C$, we then separately rollout the policy recording J_i to bound $\mathcal{P}(J)$.

V. Bound Evaluation Experiments

A. Performance Bounds

In Fig. 3(a)–(d), we empirically validate the bounds given by Theorems 1–4, respectively. We use the error rate $\delta = 0.2$ for all bounds, and the quantile $\tau = 0.7$ for VaR_τ and CVaR_τ. For a fixed policy $\mathcal{U}$, each plot shows the resulting distribution of total cost as a blue histogram generated using 10 000 simulation rollouts. Since the bounds are sample based, they themselves follow a distribution, shown as the overlaid gray histogram. To compute this bound distribution, we repeatedly (1000 times) generate the sample-based bound using a fresh batch of $n = 100$ sampled policy rollouts. The blue dashed vertical line shows the true performance measure we seek to bound,[2] while the gray dashed vertical line shows the δ (0.2 in this case) quantile of the bound distribution.[3] Since the bounds are sample based, an individual generated bound may be invalid and fall below the true performance measure. We observe this in Fig. 3(a) and (d), where portions of the gray bound histogram lie left of the blue dashed line. Yet, the theorems guarantee that the bounds hold with probability at least $1 - \delta$, so at least $1 - \delta$ of the bound

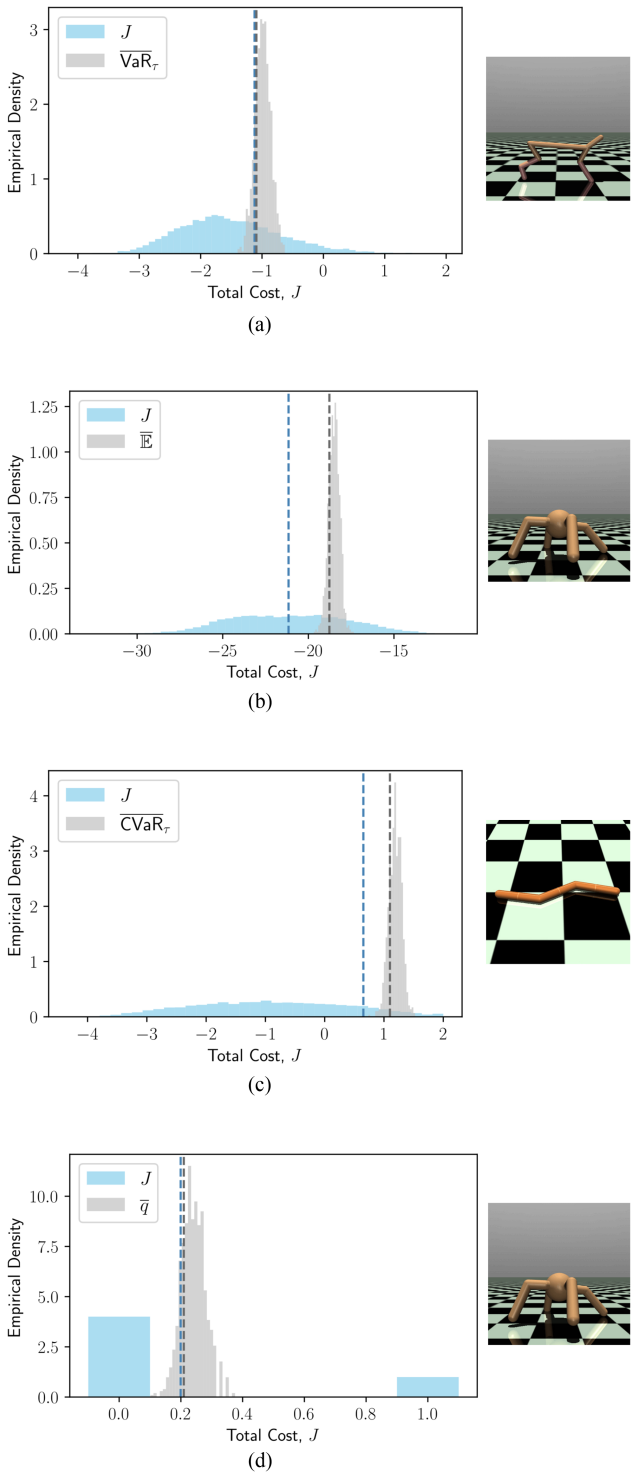

Fig. 3. (a)–(d) Empirical validation of the bounds for VaR_τ, $\mathbb{E}$, CVaR_τ, and q, respectively. Each plot shows, for a single policy, the empirical distribution of total cost J (blue), along with the distribution of the bound (gray). The blue vertical line shows the true measure we seek to bound, and the gray vertical line shows the δ quantile of the bound distribution. Since our theoretical results ensure that the bounds holds with probability $\geq 1 - \delta$, the δ quantile of the bound distribution should exceed the true measure. Thus, visually, our results ensure that the gray line is to the right of the blue line, as validated in each plot. The cost histogram was generated using 10 000 simulations, and the bound histogram was generated by repeatedly computing the bound 1000 separate times. In each case, $n = 100$, $\delta = 0.2$, and $\tau = 0.7$. To demonstrate that the bounds are agnostic to the dynamics, we used the Half Cheetah (a), Ant [(b) and (d)], and Swimmer (c) MuJoCo environments.

[2]Since we do not have access to the true underlying performance measure, we approximate the true measure with a Monte Carlo estimate from the 10 000 simulation rollouts. This is only for visualization purposes.

[3]This theoretical quantile is also approximated for visualization as the empirical quantile of the repeated bound generations.

distribution should exceed the true performance measure. Equivalently, the bound distribution's δ quantile should exceed the true performance measure. Visually, this means that the gray dashed line should lie right of the blue dashed line. We indeed observe this result in all the subfigures, empirically demonstrating that the sample-based bounds hold with probability at least $1-\delta$ as guaranteed by the theorems.

To emphasize that the bounds are agnostic to the form of the dynamics, we use a variety of MuJoCo environments for testing the validity of Theorems 1–4 (Half Cheetah [54], Ant [55], Swimmer [56], and Ant again, respectively). Due to limitations of the MuJoCo simulator, in these experiments, the dynamics of the robot are deterministic. Uncertainty comes from the starting state of the robot, which is randomly initialized using a Gaussian distribution centered about a nominal state. This form of uncertainty could realistically arise when a state estimation algorithm (e.g., Kalman filter) is used to provide a distribution over the starting state. The cost functions used are the negative values of the default rewards in MuJoCo, which encourage forward motion and minimal control input. We use clipping of the stage cost to ensure bounded support when computing the expectation and CVaR bounds. For the sparse cost case in Fig. 3(d), we declared success, setting $J=0$, when the Ant torso remained within the standard height range considered by default in MuJoCo, but still used the continuous cost when optimizing. The open-loop policies considered are obtained as the result of optimizing with the cross-entropy method (CEM) [57]. Thus, the bounds can be viewed as a probabilistic guarantee on the optimizer's solution performance under randomized initial conditions. Further experiment details are in the Appendix.

While the expectation and CVaR bound (see Theorems 2 and 3) are somewhat loose when formed using the relatively small number of 100 samples, the VaR and probability of failure bounds (see Theorems 1 and 4) are quite tight. In our later experiments, we focus on showing results using VaR and failure probability as the relevant statistics.

In the *Bound Comparison* section of the Appendix, we compare the bound distributions (gray) we obtain in Fig. 3 with the bound distributions one would obtain using different bounds from the literature (specifically those from [9], [11], and [32]). We show that the bounds we use are less conservative than others (better estimating the unknown performance measure). Using less conservative bounds allows more accurate understanding of policy performance, so practitioners can better decide whether to deploy a policy or devote more resources toward policy synthesis/improvement.

B. Constraint Satisfaction Tests

In Fig. 4, using our approach from Theorem 5, we show the relationship between the probability of our constraint test holding and the probability that the underlying constraint is actually satisfied. To convey the effect of sample size on the test, we show theoretical curves for $n \in \{10, 50, 100, 500\}$. We empirically validate the theory for the $n=10$ case using simulations of the MuJoCo Ant environment. Specifically, the constraint function $g=0$ is a success if the height of the ant torso

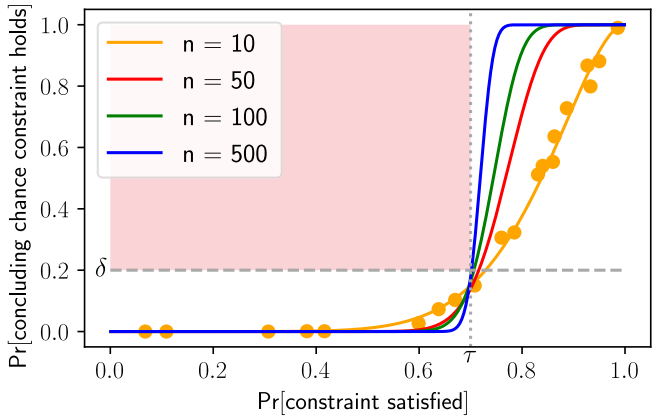

Fig. 4. Visualization of Theorem 5, and empirical validation of the theorem, applied to testing whether a chance constraint holds. Each curve represents the probability of accepting that the chance constraint holds (y-axis) given the true probability of the underlying trajectory constraint being satisfied (x-axis). The validity of the theorem is demonstrated by each curve being below δ when the chance constraint fails to hold, i.e., $\Pr[\text{constraint satisfied}]$ is below τ. Visually, the false acceptance is guaranteed to be below δ so that the curves avoid the region shaded in red in the figure. Here, we use $\delta=0.2$ (horizontal line) and $\tau=0.7$ (vertical line). As the sample size n increases, the curve approaches a step function, i.e., we obtain a perfect discriminator. For $n=10$, we plot empirical results from the Ant environment where the vertical position of the Ant torso always being between $[0.5, 1]$ with probability 0.7 is the chance constraint we seek to assess. The x and y coordinates for each orange dot are separately estimated by averaging over 1000 simulation runs.

remains in the interval [0.5, 1] for an entire rollout (based on the healthy condition specified in MuJoCo), and a failure otherwise. We impose a binary failure probability constraint (10d) requiring that this condition is satisfied with probability at least $\tau=0.7$ (shown by the gray dotted vertical line in Fig. 4). In other words, we have reformulated a chance constraint as a constraint on the failure probability of a binary g. We seek to verify whether or not the provided control policy satisfies this constraint by using the test derived from Theorem 4, checking whether the upper bound on the failure probability is sufficiently low $\overline{q} \leq 1-\tau$. By constructing $\overline{q}$ with user-specified error rate of $\delta=0.2$, we are guaranteed to have a false acceptance rate no greater than $\delta=0.2$ (shown by the dashed gray horizontal line in Fig. 4) by Theorem 5. To validate the test over a range of satisfying and violating policies, we use 20 open-loop policies generated by randomly sampling control actions.

For each policy, we repeatedly (1000 times) collect a fresh set of $n=10$ samples of the constraint function $g_{1:n}$ by executing the control actions from random initial conditions in the Ant environment, and apply Theorem 5 with cutoff $C=1-\tau$ and bound $\overline{P}=\overline{q}$ computed from Theorem 4. We call each such bound computation a trial. We use the empirical fraction of trials for which we concluded that the chance constraint holds (based on the test) to approximate the unknown true probability. This fraction provides the y-coordinate for the associated point in the figure. We obtain the associated x-coordinate, the "ground truth" constraint satisfaction probability $\Pr[g=0]$, through 1000 Monte Carlo simulations. Specifically, we simulated the policy 1000 times and recorded the empirical fraction of the resulting trajectories that were a success, obtaining $g=0$.

The tight agreement between the theoretical and empirical results shows that we can use our method to provide a sampling-based certification method for policy constraint satisfaction. In particular, we observe that both in the empirical and theoretical curves whenever the constraint fails to hold, the acceptance probability is below δ: to the left of the vertical line at $\tau = 0.7$, all curves lie below the horizontal line at $\delta = 0.2$, avoiding the region shaded in red.

Similar figures could have been generated for constraint tests on other performance measures with continuous g, e.g., a CVaR constraint. However, we would not easily be able to generate a corresponding theoretical curve. With Theorem 4, we can compute the theoretical curves using the Binomial distribution and the true probability of success.

VI. Bound Sensitivity to Distribution Shifts

In this section, we give analytical expressions for the effect that changes in cost distributions have on the confidence level of the bounds presented in Section IV. The setting we consider is when the distribution of cost based on our simulator does not match the true distribution of cost when the policy is deployed in the real world. Using samples from the simulator cost distribution, we construct performance bounds with confidence level $1 - \delta_{\text{sim}}$. In the following sections, we give expressions for how this confidence level changes (to $1 - \delta_{\text{true}}$) when the distribution of cost in the real world does not match that of the simulator. We present the sensitivity relationships in this section as corollaries of the bound theorems in Section IV and defer all proofs to the Appendix.

It is important to note the inherent tradeoff between efficiency of a bound and its robustness to distribution shift. If a bound very precisely estimates the unknown parameter, it will be more sensitive to distribution shifts. If one anticipates a certain level of distribution shift, it is not appropriate to use less precise bounds; instead, in Section VII, we detail how to modify the bounds presented in this article to be appropriately robust to distribution shift.

In this article, we measure distribution shift between the simulated cost distribution $\mathcal{D}_{\text{sim}}$ and the true cost distribution $\mathcal{D}_{\text{true}}$ according to the one-sided Kolmogorov–Smirnov (KS) distance for distributions [49]

$$\sup_x \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x). \tag{16}$$

We consider the one-sided KS distance because it captures when a cost CDF shifts downward, relating to a harmful distribution shift of higher cost more often.

Corollary 1 (Sensitivity of VaR bounds): Suppose that we construct $\overline{\text{VaR}}_\tau$ with samples from the simulated cost distribution, $J_{\text{sim}} \sim \mathcal{D}_{\text{sim}}$. Then, suppose that that the true cost distribution, $\mathcal{D}_{\text{true}}$, is close to $\mathcal{D}_{\text{sim}}$ in the one-sided KS distance

$$\sup_x \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha. \tag{17}$$

Then, we have

$$\Pr[\text{VaR}_\tau(J_{\text{true}}) \leq \overline{\text{VaR}}_\tau] \geq 1 - \delta_{\text{true}} \tag{18a}$$

$$\delta_{\text{true}} = 1 - \text{Bin}(k^* - 1; n, \tau + \alpha) \tag{18b}$$

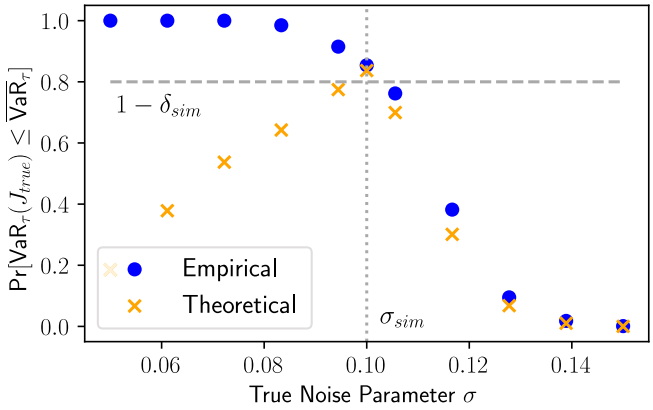

Fig. 5. Confidence level of the VaR bound as a sim-to-real mismatch is varied. The parameter σ controls the standard deviation of the initial state distribution in the Half Cheetah environment. When $\sigma > \sigma_{\text{sim}}$, the true confidence level of the bound degrades. When $\sigma < \sigma_{\text{sim}}$, the true confidence level of the bound strengthens. In blue, we plot the empirical confidence levels estimated by varying σ, using 10 000 simulations to estimate $\text{VaR}_\tau(J_{\text{true}})$, and using 1000 realizations of $\overline{\text{VaR}}_\tau$. In orange, we plot the minimum confidence level guaranteed by (18). The theoretical sensitivity guarantee is valid, always lower than the empirical confidence, but is pessimistic when $\sigma < \sigma_{\text{sim}}$ as even though the one-sided KS distance $\alpha > 0$, the distribution shift actually results in a higher confidence level.

$$k^* = \min\{k \in \{1, \ldots, n\} \mid \text{Bin}(k - 1; n, \tau) \geq 1 - \delta_{\text{sim}}\}. \tag{18c}$$

In Fig. 5, we plot the sensitivity of the VaR bound to distribution shifts imposed by misspecifying a parameter for the Half Cheetah environment, the noise parameter σ over the initial state distribution. The empirical confidence level of the bound $\Pr[\text{VaR}_\tau(J_{\text{true}}) \leq \overline{\text{VaR}}_\tau]$ always exceeds the theoretically predicted $1 - \delta_{\text{true}}$ using Corollary 1. However, the theoretical prediction is pessimistic when σ is decreased from the nominal value σ_{sim} as this distribution shift actually results in increased confidence level. A tighter theoretical bound may be obtained if we assume knowledge of the precise τ' such that $\text{VaR}_\tau(J_{\text{true}}) = \text{VaR}_{\tau'}(J_{\text{sim}})$ and then use τ' instead of the larger $\tau + \alpha$ in (18b).

Corollary 2 (Sensitivity of $\mathbb{E}$ and CVaR bounds): Suppose that we construct $\overline{\text{CVaR}}_\tau$ with samples from the simulated cost distribution, $J_{\text{sim}} \sim \mathcal{D}_{\text{sim}}$. Suppose that the true cost distribution, $\mathcal{D}_{\text{true}}$, is close to $\mathcal{D}_{\text{sim}}$ in the one-sided KS distance

$$\sup_x \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \leq \sqrt{\frac{-\ln(2\delta_{\text{sim}})}{2n}}. \tag{19}$$

Then, we have

$$\Pr[\text{CVaR}_\tau(J_{\text{true}}) \leq \overline{\text{CVaR}}_\tau] \geq 1 - \delta_{\text{true}} \tag{20a}$$

$$\delta_{\text{true}} = e^{-2n(\epsilon - \alpha)^2} \tag{20b}$$

$$\epsilon = \sqrt{\frac{-\ln \delta_{\text{sim}}}{2n}}. \tag{20c}$$

The condition that $\alpha \leq \sqrt{\frac{-\ln(2\delta_{sim})}{2n}}$ is an artifact of the DKW bound holding for $\delta \in (0, 0.5]$. One can remove this condition (at the expense of closed-form expressions for the bounds) by using the KS approach [58] in place of the DKW approach when

constructing the bounds. Finally, since we treat the expected value as a special case of CVaR, the relationships in (20) also hold for the expected value bound.

Corollary 3 (Sensitivity of failure probability bounds): Suppose that we construct $\overline{q}$ by observing k failures out of n samples from a Bernoulli distribution with probability of failure q_{sim}. Then, suppose that the true cost distribution, $\mathcal{D}_{\text{true}}$, is close to $\mathcal{D}_{\text{sim}}$ in the one-sided KS distance

$$\sup_x \ \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \tag{21}$$

i.e., $q_{\text{true}} - q_{\text{sim}} \leq \alpha$. Then, we have

$$\Pr[q_{\text{true}} \leq \overline{q}] \geq 1 - \delta_{\text{true}} \tag{22a}$$

$$\delta_{\text{true}} = \text{Bin}(k^*_\alpha - 1; n, q_{\text{sim}}) \tag{22b}$$

$$k^*_\alpha = \min\{k \in \{0, \ldots, n\} \mid \text{Bin}(k; n, q_{\text{sim}} + \alpha) \geq \delta_{\text{sim}}\}. \tag{22c}$$

VII. Constructing Robust Bounds

Building on the bound sensitivity results, in this section, we show how to construct robust bounds, which will hold with the desired confidence level $(1 - \delta)$ even when there is mismatch between the simulated and real cost distributions. To construct robust bounds, we anticipate the potential cost distribution shift and appropriately increase the bounds, slightly modifying those described in Section IV.

Specifically, we again assume that the true and simulated cost distributions are within α in terms of the one-sided KS distance

$$\sup_x \ \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \tag{23}$$

where $\mathcal{D}_{\text{sim}}$ and $\mathcal{D}_{\text{true}}$ are the simulated and true cost distributions, respectively. Although precisely specifying α is problem dependent and may be challenging, these robust bounds provide a principled mechanism for acting conservatively when the simulator is known to be inaccurate. Practically, α may be chosen based on past experience or can be estimated from Monte Carlo estimates of $\text{CDF}_{\mathcal{D}_{\text{sim}}}$ and $\text{CDF}_{\mathcal{D}_{\text{true}}}$ (using simulated and real-world trajectories, respectively). Each of the expressions we give are presented as corollaries of the bound theorems in Section VI, and all proofs are deferred to the Appendix.

Corollary 4 (Robust VaR bounds): Consider $\tau, \delta \in (0, 1)$ and n IID cost samples from the simulator $J_{1:n} \sim \mathcal{D}_{\text{sim}}$ and assume that $\sup_x \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha$. Then, we have the α-robust VaR bound

$$\Pr[\text{VaR}_\tau(J_{\text{true}}) \leq \overline{\text{VaR}}_\tau(\alpha)] \geq 1 - \delta \tag{24a}$$

$$\overline{\text{VaR}}_\tau(\alpha) = \overline{\text{VaR}}_{\tau+\alpha} \tag{24b}$$

where $\overline{\text{VaR}}_{\tau+\alpha}$ is constructed using $J_{1:n}$ as in Theorem 1 to hold with probability $1 - \delta$.

Corollary 5 (Robust $\mathbb{E}$ and CVaR bounds): Consider $\tau \in [0, 1)$, $\delta \in (0, 0.5]$, an upper bound J_{ub}, and n IID cost samples from the simulator $J_{1:n} \sim \mathcal{D}_{\text{sim}}$ and assume that $\sup_x \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha$. Then, replacing the DKW gap $\epsilon(\delta, n)$ by $\epsilon' = \epsilon(\delta, n) + \alpha$ when applying Theorem 3 with $J_{1:n}$, we have the α-robust CVaR bound

$$\Pr[\text{CVaR}_\tau(J_{\text{true}}) \leq \overline{\text{CVaR}}_\tau(\alpha)] \geq 1 - \delta \tag{25a}$$

$$\overline{\text{CVaR}}_\tau(\alpha) = \frac{1}{1-\tau}\left[\epsilon' J_{ub} + \left(\frac{k}{n} - \epsilon' - \tau\right) J_{(k)} + \frac{1}{n}\sum_{i=k+1}^{n} J_{(i)}\right] \tag{25b}$$

where k is the smallest index such that $\frac{k}{n} - \epsilon' - \tau \geq 0$.

Taking $\tau = 0$ immediately yields a similar robust bound for the expected value.

Corollary 6 (Robust failure probability bounds): Consider $\delta \in (0, 1)$ and n IID Bernoulli samples $J_{1:n}$ obtained in simulation (where $J = 1$ denotes failure) with $k = \sum_{i=1}^n J_i$ failures and assume $\sup_x \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha$ (i.e., $q_{\text{true}} \leq q_{\text{sim}} + \alpha$). Then, we have the α-robust failure probability bound

$$\Pr\left[q_{\text{true}} \leq \overline{q}(\alpha)\right] \geq 1 - \delta \tag{26a}$$

$$\overline{q}(\alpha) = \max\{q' \in [0, 1] \mid \text{Bin}(k; n, q' - \alpha) \geq \delta\}. \tag{26b}$$

In Fig. 6, we plot the confidence level of robust failure probability bounds constructed with different α as we impose a distribution shift by misspecifying the noise parameter σ over the initial state distribution for the Ant environment. Whenever the distribution shift is within the allowed tolerance ($q_{\text{true}} \leq q_{\text{sim}} + \alpha$), as designated by a circle, the confidence level of the bound remains at least $1 - \delta$ confirming Corollary 6.

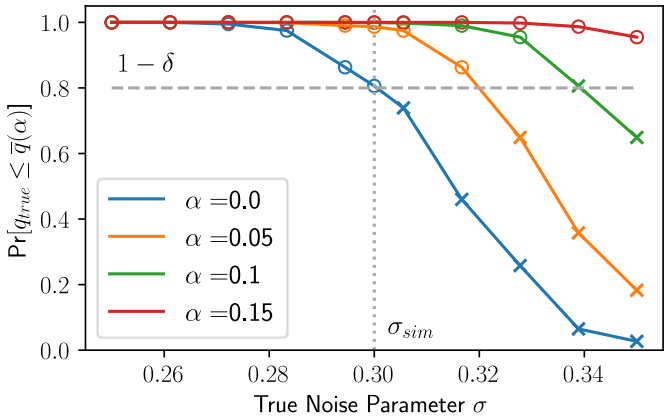

Fig. 6. Confidence level of robust failure probability bounds as a sim-to-real mismatch is varied. Similar to Fig. 5, we plot the empirical confidence level associated with robust bounds constructed with different tolerances $\alpha \in [0, 0.05, 0.1, 0.15]$ as we vary the parameter σ controlling the standard deviation of the initial state distribution in the Ant environment. We observe that even under distribution shift, whenever $q_{\text{true}} \leq q_{\text{sim}} + \alpha$ (denoted by circles), the associated robust bound $\overline{q}(\alpha)$ holds with confidence level at least $1 - \delta$, as expected from Corollary 6. The plotted points are generated by varying σ, using 10 000 simulations to estimate q_{true}, and using 1000 realizations of $\overline{q}(\alpha)$.

VIII. Policy Selection

Thus far, we have presented a method for rigorously assessing the quality of a policy by placing bounds on performance measures of the trajectory cost. A natural extension is to use these bounds when selecting between several candidate policies. In this case, we must take care to apply an appropriate correction

to the bounds for the resulting bound on the chosen policy to remain probabilistically valid. To illustrate the need for a correction, consider a thought experiment where 100 identical policies are considered as candidates for execution. For each of these identical policies, we generate a fresh set of stochastic simulation rollouts and use these to compute a performance bound. We then choose to execute the policy achieving the lowest bound. Of course, this a false choice since all the policies are the same. However, because the rollouts are stochastic, the computed bounds will fluctuate even though all used the same policy. Choosing the lowest bound among the 100 will thus give an artificially low performance bound for the associated policy as the chosen bound is the result of a lucky draw of rollouts. Therefore, we cannot expect the resulting bound to still hold with probability at least $1-\delta$. In other words, since the bounds hold probabilistically, one bound among the 100 is likely to be overly optimistic by chance and may even be lower than the true performance measure, i.e., an invalid bound. To remedy this problem, we explain a statistical correction for comparing bounds among a set of candidate policies.

Consider m (not necessarily independent) policies $\mathcal{U}_{1:m}$, where for each policy $\mathcal{U}_i$, we obtain n IID trajectory rollouts, obtaining trajectory cost samples $J^{(i)}_{1:n}$. Let the performance measure we are interested in be denoted by $\mathcal{P}$ (e.g., $\mathcal{P} = \text{VaR}_\tau$). Then, using the samples $J^{(i)}_{1:n}$ and the results from Section IV, an individual upper bound $\overline{P}^{(i)}$ is computed for each policy, which has corresponding unknown true performance $P^{(i)}$. Now, suppose that we plan to execute the policy with least upper bound, that is,

$$\mathcal{U}^* = \underset{\mathcal{U}_i \in \mathcal{U}_{1:m}}{\arg\min} \quad \overline{P}^{(i)}. \tag{27}$$

Take $\overline{P}^*$ as the individual upper bound associated with $\mathcal{U}^*$ and P^* as the true statistic (e.g., true VaR_τ) associated with $\mathcal{U}^*$. We are interested in understanding with what probability $\overline{P}^*$ upper bounds P^* and formalize this in the following result.

Theorem 6 (Uncorrected confidence level): Given m policies $\mathcal{U}_{1:m}$ with associated unknown true performance $\{P^{(i)}\}_{i=1}^m$ and associated probabilistic bounds $\{\overline{P}^{(i)}\}_{i=1}^m$ individually holding with confidence level $1-\delta$ i.e.,

$$\Pr[P^{(i)} \leq \overline{P}^{(i)}] \geq 1-\delta \quad \forall i \tag{28}$$

let $\overline{P}^*$ be the lowest probabilistic bound and let P^* be the associated true statistic, i.e.,

$$i^* = \underset{i \in \{1,\ldots,m\}}{\arg\min} \ \overline{P}^{(i)} \tag{29a}$$

$$\overline{P}^* = \overline{P}^{(i^*)} \tag{29b}$$

$$P^* = P^{(i^*)}. \tag{29c}$$

Then, $\overline{P}^*$ bounds P^* with probability at least $(1-\delta)^m$, i.e.,

$$\Pr[P^* \leq \overline{P}^*] \geq (1-\delta)^m. \tag{30}$$

Before proceeding, we consider two limiting cases of Theorem 6. Temporarily assume that the individual bounds hold with probability $1-\delta$ exactly, i.e.,

$$\Pr[P^{(i)} \leq \overline{P}^{(i)}] = 1-\delta. \tag{31}$$

Then, we have the following.

1) When each bound is bounding the same statistic (i.e., $P^{(1)} = \cdots = P^{(m)}$), then
$$\Pr[P^* \leq \overline{P}^*] = (1-\delta)^m \tag{32}$$
where this follows as all bounds must hold for the minimum bound to hold in this case.
2) When the cost samples $J^{(i)}_{1:n}$ associated with each policy $\mathcal{U}_i$ are confined to disjoint intervals (i.e., for each $\mathcal{U}_i$, $J \in [J^{(i)}_{lb}, J^{(i)}_{ub}] := D_i$ with $D_i \cap D_j = \emptyset \ \forall i \neq j$), then
$$\Pr[P^* \leq \overline{P}^*] = 1-\delta \tag{33}$$
where this follows as only one bound, the one generated using the policy $\mathcal{U}_i$ having the lowest interval D_i, needs to hold for the minimum bound to hold in this case.

These two limiting cases show the extremes we must consider when making a correction for multiple policies. In the best case (case 2), the resulting bound $\overline{P}^*$ still holds with probability $1-\delta$, i.e., the error rate is unchanged, while in the worst case (case 1), the error rate can increase dramatically for a large set of candidates. Since we want to avoid distributional assumptions, we must consider the worst case scenario, in which case Theorem 6 is tight.

Theorem 6 captures the worst case when comparing multiple policies and selecting the policy with lowest bound. Thus, we can use it to correct for the comparison by first inflating the required probability that each individual bound hold, as described in the following theorem.

Theorem 7 (Multipolicy bound correction): The resulting bound $\overline{P}^*$ obtained when selecting the lowest bound among multiple policies holds with probability at least $1-\delta$ if each individual bound $\overline{P}^{(i)}$ is inflated to satisfy

$$\Pr[P^{(i)} \leq \overline{P}^{(i)}] \geq 1-\bar{\delta} \tag{34a}$$

$$\bar{\delta} = 1-(1-\delta)^{1/m}. \tag{34b}$$

Remark (Multihypothesis connection): This correction is identical to the Šidák correction [59] for testing multiple hypotheses. Although we are only interested in ensuring that the minimizing bound be probabilistically valid, we end up needing to inflate as if we required all bounds to hold due to case 1 (the typical use case of multihypothesis correction).

Remark (Considering shared rollout seeds): In our approach, for each policy, we generate a bound using a fresh set of stochastic simulation rollouts. However, one might consider instead fixing a random seed, e.g., fixing a set of sampled environments, which is then reused when generating the rollouts for each policy. Even with this modification, a multihypothesis correction is still needed. In fact, since under this fixed seed procedure, the rollouts for each policy would no longer be independent, we would have to resort to the weaker Bonferroni correction [59] over the Šidák.

As mentioned previously, introducing the necessity of this bound correction is one of the contributions of this article. In

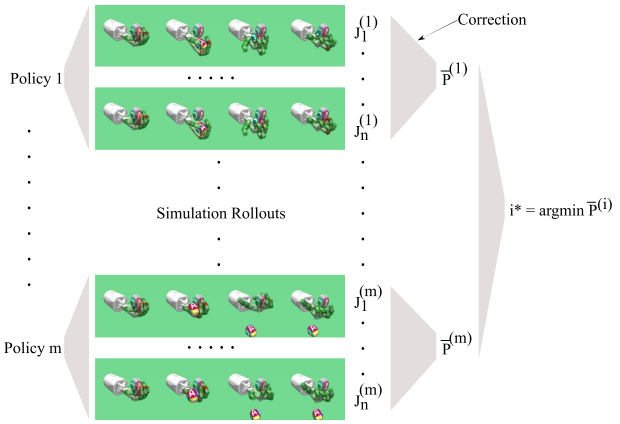

Fig. 7. Overview of our method for comparing performance among a set of policies for manipulating an egg of uncertain mass and friction with the MuJoCo Shadow Hand. Given m candidate policies, each policy is executed in simulation n times, and each associated trajectory cost is recorded. These cost samples are then used to compute a probabilistic upper bound on performance for each policy $\overline{P}^{(i)}$. Finally, the policy achieving the minimum bound is selected, giving a probabilistic guarantee on policy performance. To ensure that this final bound holds with a user-specified probability $(1-\delta)$, we apply a multihypothesis correction to each of the individual bounds.

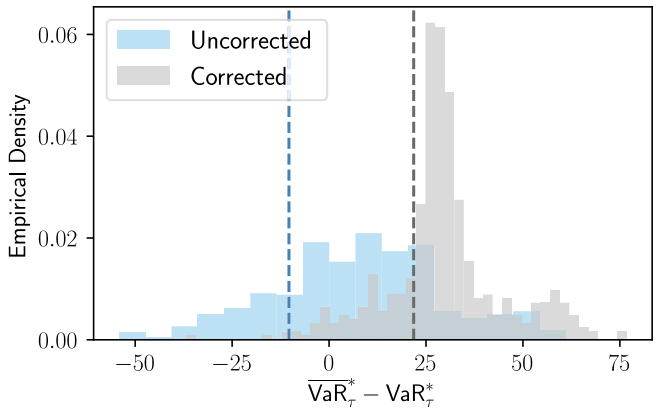

Fig. 8. Validity of the correction given by Theorem 7 for the Shadow Hand environment when manipulating the egg object with uncertain density and friction. Plotted are two distributions, of $\overline{\text{VaR}}^*_\tau - \text{VaR}^*_\tau$. In blue, $\overline{\text{VaR}}^*_\tau$ is chosen without the correction and in gray $\overline{\text{VaR}}^*_\tau$ is chosen with the multihypothesis correction specified in Theorem 7. Each sample in the histogram is generated by selecting the best bound ($\overline{\text{VaR}}^*_\tau$) among 20 precomputed plans based on either the corrected or uncorrected bound and subtracting the true performance measure VaR^*_τ associated with the chosen policy. Each histogram is generated with 500 repetitions. Only for the corrected bound does the δ quantile dashed line lie above 0. Thus, when selecting a policy with multihypothesis correction, the desired error rate is achieved, while this is not the case when using the uncorrected bound.

Fig. 7, we summarize the process for selecting the policy with the lowest performance bound while retaining statistical validity using the multihypothesis correction. In Section IX, we show the necessity of the correction by comparing against the naive uncorrected bound in the setting of object manipulation.

IX. Manipulation Example

In this section, we demonstrate the validity and necessity of the multihypothesis correction in the MuJoCo Shadow Dexterous Hand environment [60] for simulating in-hand manipulation. In realistic settings, the object to be manipulated may have uncertain physical parameters, such as mass and friction, that can only be roughly estimated or inferred from interaction but are not directly observable. In this example, we show how for such a setting, we can use our distribution-free method to select among several candidate control policies for manipulating the object, obtaining an associated performance bound for the chosen policy. This experiment is chosen to emphasize that our approach works with complex dynamics involving contact and the discontinuities therein.

To generate candidate policies, we use a simple sampling-based planner inspired by the approach in [61]. Critical to the success of the planner is not to sample actions for each step of the planning horizon, but to take a spline interpolation of sub-sampled points This reduces the effective size of the action space and enforces smoothness. Using this method, we generated $m = 20$ open-loop plans over a horizon of 100 time steps each designed for manipulating Gymnasium's egg object assuming a nominal density of 1000 kg/m^3, sliding friction coefficient of 1, torsional friction coefficient of 0.005, and rolling friction coefficient of 0.0001. We randomize over density rather than mass as this is more standard in MuJoCo, and the object volume is fixed, meaning that randomizing over density and randomizing over mass are equivalent in this setting.

Given these candidate policies, we applied Theorem 7 to inflate the confidence level and select the bound-minimizing policy. We used VaR_τ with $\tau = 0.7$ as the cost performance measure we wish to bound and specified a bound error rate of $\delta = 0.2$. The policies are evaluated in simulated environments now randomizing the friction and density of the egg object to simulate uncertainty in the true object's physical parameters. The specific uncertainties are as follows:

1) density $\sim$ uniform$[700, 1200]$ kg/m^3;
2) sliding friction coefficient $\sim$ uniform$[0.8, 1.2]$;
3) torsional friction coefficient $\sim$ uniform$[0.004, 0.006]$;
4) rolling friction coefficient $\sim$ uniform$[0.00008, 0.00014]$.

The individual bound for each policy is computed using the total cost obtained in 30 simulated executions of it. The cost is based on aligning the egg with a desired goal position and orientation. The chosen goal orientation requires flipping the egg object from its starting orientation. Fig. 7 provides an overview of the policy selection and bound computation procedure used in this example and shows selected frames from simulations of the first and last of the 20 plans compared.

By repeatedly generating fresh cost samples to run many simulated experiments, we show in Fig. 8 that while our multihypothesis corrected bound is valid with at least the specified probability of $1-\delta$, naively using the uncorrected bound is too optimistic and fails to achieve this confidence level. Specifically, we plot the distribution of $\overline{\text{VaR}}^*_\tau - \text{VaR}^*_\tau$ over 500 repetitions, where in the blue histogram, $\overline{\text{VaR}}^*_\tau$ is chosen without correction, and in the gray histogram, it chosen with the multihypothesis correction. Each histogram sample is generated by selecting

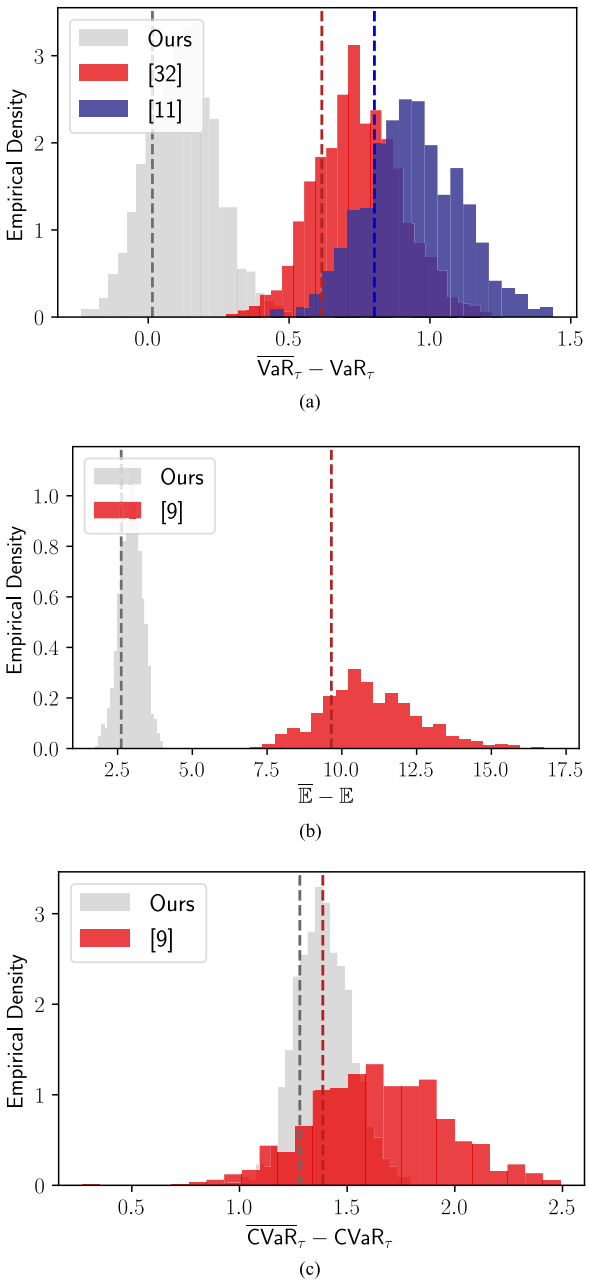

Fig. 9. Distribution of bound offsets for (a) VaR_τ in the Half Cheetah environment, (b) expectation in the Ant environment, and (c) CVaR_τ in the Swimmer environment. All bounds are probabilistically valid, since the vertical dashed lines (the δ quantiles) are to the right of zero, indicating that the generated bounds indeed exceed the true performance measure with probability at least $1-\delta$. In every case, the bounds we use are less conservative than the bounds used in [9], [11], and [32]. As in Fig. 3, the true performance measure is computed using 10 000 rollouts, and each bound histogram was generated by repeatedly computing the given sample-based bound 1000 times, each time using a fresh set of $n = 100$ sampled policy rollouts. We again use $\tau = 0.7$.

the lowest bound ($\overline{\text{VaR}}^*_\tau$) among 20 precomputed open-loop policies and subtracting the true performance measure VaR^*_τ associated with the chosen policy. Since the bound holds when $\overline{\text{VaR}}^*_\tau - \text{VaR}^*_\tau \geq 0$, if the bound holds with probability at least $1-\delta$, the δ quantile of the corresponding $\overline{\text{VaR}}^*_\tau - \text{VaR}^*_\tau$ distribution should be nonnegative. We observe this to be the case when the bound is generated with correction, as the gray dashed line showing the δ quantile of the associated $\overline{\text{VaR}}^*_\tau - \text{VaR}^*_\tau$ distribution lies right of 0. However, the δ quantile when generating the bound without correction, shown as the blue dashed line, lies left of 0. Thus, we observe that the corrected bound holds with probability at least $1-\delta$ as guaranteed by Theorem 7, while the uncorrected bound does not, illustrating the necessity of the multihypothesis correction.

X. Conclusion

In this article, we demonstrate how sampling-based distribution-free bounds can be used to rigorously bound the performance of a control policy applied in a stochastic environment. These bounds can also be used to verify the safety of a policy via constraint tests with a guaranteed false acceptance rate. Furthermore, we provide a thorough analysis of the sensitivity of our bounds to sim-to-real distribution shifts and provide results for constructing robust bounds that can tolerate specified amounts of distribution shift. Finally, we show how to apply these bounds when selecting the best policy from a set of candidates, which requires a multihypothesis correction to retain validity. Because these bounds are distribution-free, they can be applied to complex systems and uncertain environments without requiring knowledge of the underlying problem structure. Rather, our approach only requires simulating policy execution in the stochastic environment and recording the associated cost or constraint value. We empirically demonstrated bound validity in several MuJoCo environments including for the problem of object manipulation, which is high dimensional and has discontinuous dynamics.

In this work, we only studied a few performance measures, but our approach extends to other measures if corresponding bounds are available. Another interesting direction for future work is to use our method to select the best risk-sensitive plans in domain-randomized simulations that are then used as training data for a policy with hopes of better sim-to-real performance. Yet another avenue for future work is to use the sample-based bounds for risk and constraint satisfaction to guide an importance sampling procedure (over open-loop plans) within a stochastic model-based planner such as model predictive path integral (MPPI) [62] or CEM [57].

Appendix
Simulation Details

To create Fig. 3, we perturbed the default starting state in MuJoCo using reset_noise_scale $= 0.1$. The sole exception to this was for the sparse cost case with the Ant where we used reset_noise_scale $= 0.3$ since the default of 0.1 was never enough to push the Ant outside the healthy range of $[0.2, 1]$ when using the optimized action sequence.

To identify the candidate policy, CEM [57] was used. Ten generations of optimization were performed, where in each generation, 100 open-loop policies were generated by sampling the control input at each time step using a Gaussian distribution based on the estimated mean and variance from the previous

generation. Based on a single execution in the random environment, the top performing ten sample plans were then used to fit the Gaussian distribution for the next generation. After the final generation, the top performing plan was selected as the candidate. Plans had a horizon of 20 time steps.

To approximate the cost distribution, the candidate policy was executed 10 000 times. The associated theoretical statistic was then found by taking the empirical average cost for $\mathbb{E}$, the empirical quantile for VaR_τ, or the Monte Carlo approximation for CVaR_τ (see [63]). To compute $\mathbb{E}$ and CVaR_τ, the total costs in the Ant environment were clipped to between $[-2H, 0]$ and between $[-0.325H, 0.1H]$ for the Swimmer environment where $H = \text{˜}20$ was the horizon.

Bound Comparison

Here, we compare the bounds we use for VaR_τ, expected value, and CVaR_τ with those used by other papers in the robotics and statistics literature [9], [11], [32]. Fig. 9 shows the empirical bound distribution generated by each approach using the same experimental parameters as when constructing Fig. 3. In every case, the bounds we use are the least conservative (better estimating the unknown performance measure) while still meeting the desired $1-\delta$ confidence level.

From the derivation of our VaR_τ bound, it is clear that choosing a smaller order statistic for the bound provably violates the $1-\delta$ confidence level. Thus, among VaR_τ bounds constructed as a particular order statistic, ours is unimprovable. Because of this property, our VaR_τ bound will always be less conservative than those used in [11] and [32]. Statements of similar generality cannot be made in the case of the expected value and CVaR_τ bounds, although the empirical results in Fig. 9 provide evidence that the bounds we use in these cases are to be preferred to the bounds used in [9], [11], and [32].

Proofs

Although we do not consider lower bounds or two-sided bounds, we note that lower bounds for each performance measure can be derived in similar fashion to our derivations for the upper bounds. Two-sided bounds can then be constructed by combining lower and upper bounds and adjusting the confidence level (requiring each one-sided bound to hold with probability $1-\delta/2$) using the inclusion–exclusion principle.

A. Proof of Theorem 1

We first provide a lemma that is necessary for our proof.

Lemma 1:

$$\Pr[J_i < \text{VaR}_\tau(J)] \leq \tau. \tag{35}$$

Proof: We may rewrite the strict inequality probability using a left limit approaching $\text{VaR}_\tau(J)$

$$\Pr[J_i < \text{VaR}_\tau(J)] = \lim_{x \to \text{VaR}_\tau(J)^-} \Pr[J_i \leq x] \tag{36}$$

and are guaranteed that the limit exists since CDFs are upper semicontinuous. Since x in the limit satisfies $x < \text{VaR}_\tau(J)$, by definition of VaR as an infimum, we have

$$\Pr[J_i \leq x] < \tau. \tag{37}$$

Since this holds for all x in the limit

$$\lim_{x \to \text{VaR}_\tau(J)^-} \Pr[J_i \leq x] \leq \tau \tag{38}$$

concluding the proof. ■

Now, to prove the main result, from the definition of $\text{VaR}_\tau(J)$, we have

$$\Pr[\text{VaR}_\tau(J) \leq J_{(k)}] = \Pr\left[\sum_{i=1}^{n} \mathbb{1}(J_i < \text{VaR}_\tau(J)) < k\right] \tag{39a}$$

$$= \Pr\left[\sum_{i=1}^{n} \mathbb{1}(J_i < \text{VaR}_\tau(J)) \leq k-1\right]. \tag{39b}$$

Since by Lemma 1, we have

$$\Pr[J_i < \text{VaR}_\tau(J)] \leq \tau \tag{40}$$

the random quantity $\mathbb{1}(J_i < \text{VaR}_\tau(J))$ is Bernoulli where the probability of being 1 is at most τ. Thus

$$\Pr\left[\sum_{i=1}^{n} \mathbb{1}(J_i < \text{VaR}_\tau(J)) \leq k-1\right] \geq \text{Bin}(k-1; n, \tau) \tag{41}$$

since we have the sum of n IID Bernoulli random variables with probability of being 1 at most τ.

Therefore, given values for $\tau, \delta \in (0,1)$ and $J_{1:n}$, choosing

$$\overline{\text{VaR}}_\tau = J_{(k^*)} \tag{42a}$$

$$k^* = \min\{k \mid \text{Bin}(k-1; n, \tau) \geq 1-\delta\} \tag{42b}$$

ensures that

$$\Pr[\text{VaR}_\tau(J) \leq \overline{\text{VaR}}_\tau] \geq 1-\delta. \tag{43}$$

■

Remark: Note that in the case where J has invertible CDF, (40) is tight so that (41) holds exactly. In fact, we can then give a more precise result in this case

$$1-\delta \leq \Pr[\text{VaR}_\tau(J) \leq \overline{\text{VaR}}_\tau] \leq 1-\delta+\text{bin}(k-1; n, \tau) \tag{44}$$

where k is the order statistic chosen for $\overline{\text{VaR}}_\tau$ and $\text{bin}(k-1; n, \tau)$ denotes the Binomial probability mass function evaluated at $k-1$ successes. Thus, the amount of conservatism in the coverage is no more than $\text{bin}(k-1; n, \tau)$.

Remark: Alternatively, one can achieve exactly the desired confidence by randomizing the chosen order statistic [37].

Remark: Our derivation of the VaR bound adapts a very similar result from CP presented below [64].

Theorem 8 (CP conditional result): Given $n+1$ IID samples $S_{1:n+1}$ from some continuous distribution

$$\Pr\left[\Pr[S_{n+1} \leq S_{(k)} \mid S_{1:n}] \geq 1-\epsilon\right] = \text{Bin}(k-1; n, 1-\epsilon). \tag{45}$$

Our proof does not assume a continuous distribution and can be applied to adapt the above result to $\geq \text{Bin}(k-1; n, 1-\epsilon)$ for any distribution.

B. Proof of Theorem 2

The result follows as a special case of Theorem 3 proven in the following. We let $\tau = 0$, noting that $\overline{\text{CVaR}}_\tau = \overline{\mathbb{E}}$ when $\tau = 0$ from the definitions of CVaR_τ and $\mathbb{E}$ in Definitions 2 and 3. ■

C. Proof of Theorem 3

To construct an upper bound for CVaR, we first revisit the definition

$$\text{CVaR}_\tau(Y) := \frac{1}{1-\tau}\int_\tau^1 \text{VaR}_\gamma(Y)\, d\gamma. \tag{46}$$

Note that if we can construct an upper bound on the VaR that holds for all γ, then we can use this to find an upper bound on the CVaR

$$\overline{\text{CVaR}}_\tau(Y) = \frac{1}{1-\tau}\int_\tau^1 \overline{\text{VaR}}_\gamma(Y)\, d\gamma. \tag{47}$$

Note the slight abuse of notation here; in this proof, we require a *simultaneous* VaR bound for which

$$\Pr[\overline{\text{VaR}}_\tau \geq \text{VaR}_\tau \quad \forall\tau] \geq 1-\delta \tag{48}$$

a stronger requirement than we had for the VaR bound presented in Theorem 1. Next, we describe how we obtain a simultaneous VaR bound.

Consider IID samples $Y_1, \ldots, Y_n$ with unknown distribution given by $\text{CDF}(y)$, and let y_{ub} be an associated almost sure upper bound.[4] Let $\widehat{\text{CDF}}(y)$ be the empirical CDF

$$\widehat{\text{CDF}}(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y_i \leq y) \tag{49}$$

and $\epsilon(n,\delta)$ is a constant subtracted from $\widehat{\text{CDF}}(y)$ to obtain a probabilistic lower bound

$$\underline{\text{CDF}}(y) = \begin{cases} \max\{\widehat{\text{CDF}}(y) - \epsilon, 0\}, & \text{if } y < y_{\text{ub}} \\ \underline{\text{CDF}}(y) = 1, & \text{if } y \geq y_{\text{ub}}. \end{cases} \tag{50}$$

By letting

$$\epsilon(n,\delta) = \sqrt{\frac{-\ln\delta}{2n}} \tag{51}$$

(what we call the DKW gap in Definition 7), we know from the DKW bound [12], [13] that

$$\Pr[\underline{\text{CDF}}(y) \leq \text{CDF}(y)\ \forall y] \geq 1-\delta. \tag{52}$$

Let $\overline{\text{VaR}}_\tau$ be the VaR_τ obtained from $\underline{\text{CDF}}(y)$, and note that the DKW bound extends to $\overline{\text{VaR}}_\tau$ to give

$$\Pr[\overline{\text{VaR}}_\tau \geq \text{VaR}_\tau \quad \forall\tau] \geq 1-\delta. \tag{53}$$

[4]We replace our J notation with Y to avoid confusion of j as an index.

To see this, observe that $\underline{\text{CDF}}(y) \leq \text{CDF}(y)$ for all y implies that $\{y \mid \underline{\text{CDF}}(y) \geq \tau\} \subseteq \{y \mid \text{CDF}(y) \geq \tau\}$; therefore, $\inf\{y \mid \underline{\text{CDF}}(y) \geq \tau\} \geq \inf\{y \mid \text{CDF}(y) \geq \tau\}$, since the inf over a subset is greater than or equal to the inf over the larger set. We have $\underline{\text{CDF}}(y) \leq \text{CDF}(y)$ for all y implies that $\overline{\text{VaR}}_\tau \geq \text{VaR}_\tau$ for all τ (from the definition of VaR_τ in Definition 1). The extension of the DKW bound to $\overline{\text{VaR}}_\tau$ follows.

From this bound, we conclude that integrating $\overline{\text{VaR}}_\tau$ over any τ interval gives an upper bound on the integral of VaR_τ over the same interval, which holds with probability at least $1-\delta$. We proceed to analytically integrate $\overline{\text{VaR}}_\tau$ from τ to 1 to compute $\overline{\text{CVaR}}_\tau$ based on the definition of CVaR_τ in Definition 3.

Note that $\overline{\text{VaR}}_\tau$ is a staircase function defined on the domain $\tau \in [0,1]$, which is equal to the smallest order statistic $Y_{(k)}$ such that $(\frac{k}{n} - \epsilon) \geq 0$ over the interval $\tau \in [0, (\frac{k}{n} - \epsilon)]$. It is then equal to $Y_{(k+1)}$ over the next interval $\tau \in ((\frac{k}{n} - \epsilon), (\frac{k+1}{n} - \epsilon)]$, proceeding to $Y_{(n)}$ over $\tau \in ((\frac{n-1}{n} - \epsilon), (1-\epsilon)]$, and finally y_{ub} over the last interval of $\tau \in ((1-\epsilon), 1]$. Notice that all intervals are of length $\frac{1}{n}$, except for the first, which is of length $(\frac{k}{n} - \epsilon)$, and the last, which is of length ϵ.

Integrating this staircase function from a given τ to 1, therefore, evaluates to a sum over order statistics times the length of their respective intervals. The first-order statistic in this sum is the smallest such that its interval appears above the τ quantile, namely, $Y_{(k)}$, where k is the smallest index such that $(\frac{k}{n} - \epsilon) \geq \tau$. The length of this first interval is then $(\frac{k}{n} - \epsilon - \tau)$, and we have for the first term in the sum $(\frac{k}{n} - \epsilon - \tau)Y_{(k)}$, followed by $n-k$ terms of the form $\frac{1}{n}Y_{(i)}$, where $i = k+1, \ldots, n$, and finally the term ϵy_{ub}. Following the definition of CVaR_τ in Definition 3, we normalize the sum by the length of the interval over which we integrate, $\frac{1}{1-\tau}$, to obtain the desired expression.

The bound holds for any $\tau \in [0,1)$ and $\delta \in (0, 0.5]$ (this requirement comes from the DKW inequality). To avoid defaulting to y_{ub}, we require that there is an index $k \leq n$ such that $(\frac{k}{n} - \epsilon - \tau) \geq 0$. Equivalently, $\epsilon \leq 1-\tau$, which implies $n \geq -\frac{1}{2}\ln(\delta)/(1-\tau)^2$.

As noted earlier, this CVaR bound is mathematically equivalent to the one in [31], but our derivation results in a different form for the bound expression. Visualizations of the integral form of CVaR are given in [31, Figs. 3 and 4]. ■

D. Proof of Theorem 4

In this proof, we show that $\Pr[q \leq \overline{q}] \geq 1-\delta$ for $\overline{q}$ chosen, as described in Theorem 4. By the definition of $\overline{q}$, note that

$$q \leq \overline{q} \iff \text{Bin}(k; n, q) \geq \delta \iff k \geq k^* \tag{54}$$

where

$$k^* = \min\{k' \in \{0, \ldots, n\} \mid \text{Bin}(k'; n, q) \geq \delta\}. \tag{55}$$

Then

$$\Pr[q \leq \overline{q}] = \Pr(k \geq k^*) = 1 - \Pr(k \leq k^* - 1) \tag{56a}$$

$$= 1 - \text{Bin}(k^* - 1; n, q) \geq 1 - \delta. \tag{56b}$$

We know $\text{Bin}(k^* - 1; n, q) < \delta$ by construction of k^*. Note that to construct the bound, we do not need knowledge of q; whatever q might be, our rule for choosing $\overline{q}$ is valid. ■

Remark: This is the one-sided Clopper–Pearson bound [45], which is known to be unimprovable among nonrandomized approaches [51]. Note that one could also arrive at this result by employing our VaR bound (see Theorem 1). Going this direction shows that applying CP to samples from a Bernoulli distribution reduces to the Clopper–Pearson result.

E. Proof of Theorem 5

For the test that accepts when $\overline{P} \leq C$, we can upper bound the false acceptance rate as

$$\Pr[\overline{P} \leq C \mid \mathcal{P}(g) > C] \leq \Pr[\overline{P} < \mathcal{P}(g)]. \tag{57}$$

By the bound's guaranteed error rate of δ

$$\Pr[\overline{P} < \mathcal{P}(g)] = 1 - \Pr[\overline{P} \geq \mathcal{P}(g)] \leq \delta. \tag{58}$$

Combining, we bound the test's false acceptance rate

$$\Pr[\overline{P} \leq C \mid \mathcal{P}(g) > C] \leq \delta. \tag{59}$$

■

F. Proof of Theorem 6

As the minimum bound holds whenever all the bounds hold

$$\Pr[P^* \leq \overline{P}^*] \geq \Pr[P^{(i)} \leq \overline{P}^{(i)}\ \forall i]. \tag{60}$$

Even though we do not assume that $\mathcal{U}_i$ are independent, we can assert that the events $P^{(i)} \leq \overline{P}^{(i)}$ and $P^{(j)} \leq \overline{P}^{(j)}$ are independent for $i \neq j$ as bound i is generated with different random samples than bound j. Applying independence yields

$$\Pr[P^{(i)} \leq \overline{P}^{(i)}\ \forall i] = \prod_{i=1}^{m} \Pr[P^{(i)} \leq \overline{P}^{(i)}]. \tag{61}$$

Each bound holds with probability at least $1 - \delta$ so

$$\prod_{i=1}^{m} \Pr[P^{(i)} \leq \overline{P}^{(i)}] \geq (1 - \delta)^m. \tag{62}$$

Combining the steps, we conclude that

$$\Pr[P^* \leq \overline{P}^*] \geq (1 - \delta)^m. \tag{63}$$

G. Proof of Theorem 7

Applying Theorem 6 with $\bar{\delta} = 1 - (1 - \delta)^{1/m}$, we get

$$\Pr[P^* \leq \overline{P}^*] \geq (1 - \bar{\delta})^m = 1 - \delta. \tag{64}$$

■

H. Proof of Corollary 1

Given

$$\sup_x \ \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \tag{65}$$

$$\text{VaR}_\tau(J_{\text{true}}) \leq \text{VaR}_{\tau+\alpha}(J_{\text{sim}}). \tag{66}$$

Furthermore

$$\overline{\text{VaR}}_\tau = J_{(k^*)} \tag{67a}$$

$$k^* = \min\{k \mid \text{Bin}(k - 1; n, \tau) \geq 1 - \delta_{\text{sim}}\}. \tag{67b}$$

Then, utilizing the proof for Theorem 1, we have

$$\Pr[\text{VaR}_\tau(J_{\text{true}}) \leq \overline{\text{VaR}}_\tau] \geq \Pr[\text{VaR}_{\tau+\alpha}(J_{\text{sim}}) \leq k^*] \tag{68a}$$

$$\geq \text{Bin}(k^* - 1; n, \tau + \alpha) \tag{68b}$$

$$\geq 1 - \delta_{\text{true}}. \tag{68c}$$

■

I. Proof of Corollary 2

We are given

$$\sup_x \ \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \leq \sqrt{\frac{-\ln(2\delta_{\text{sim}})}{2n}}. \tag{69}$$

From the proof of Theorem 3, the CVaR bound utilizes a simultaneous lower bound over the CDF of cost. We can interpret the distribution shift by amount α as a shift in the offset of our CDF bound from the empirical CDF

$$\underline{\text{CDF}}(y) = \begin{cases} \max\{\widehat{\text{CDF}}(y) - \epsilon', 0\}, & \text{if } y < y_{\text{ub}} \\ \underline{\text{CDF}}(y) = 1, & \text{if } y \geq y_{\text{ub}}. \end{cases} \tag{70}$$

where $\epsilon' = \epsilon - \alpha$. From the DKW inequality [12], [13], we know that

$$\epsilon' = \sqrt{\frac{-\ln \delta_{\text{true}}}{2n}} \tag{71a}$$

$$\Rightarrow \delta_{\text{true}} = e^{-2n\epsilon'^2} \tag{71b}$$

$$= e^{-2n(\epsilon-\alpha)^2}. \tag{71c}$$

■

J. Proof of Corollary 3

Given k_{sim} observed failures in the simulator out of n trajectories, we have

$$\overline{q} = \max\{q' \in [0, 1] \mid \text{Bin}(k_{\text{sim}}; n, q') \geq \delta_{\text{sim}}\}. \tag{72}$$

Utilizing the proof of Theorem 4, we have

$$\Pr[q_{\text{true}} \leq \overline{q}] \geq \Pr[q_{\text{sim}} + \alpha \leq \overline{q}] \tag{73a}$$

$$= \Pr[k_{\text{sim}} \geq k^*_\alpha] \tag{73b}$$

where

$$k^*_\alpha = \min\{k \in \{0, \ldots, n\} \mid \text{Bin}(k; n, q_{\text{sim}} + \alpha) \geq \delta_{\text{sim}}\}. \tag{74}$$

Thus

$$\Pr[q_{\text{true}} \leq \overline{q}] \geq 1 - \Pr[k_{\text{sim}} \leq k^*_\alpha - 1] \tag{75a}$$

$$= 1 - \text{Bin}(k^*_\alpha - 1; n, q_{\text{sim}}) \tag{75b}$$

$$\geq 1 - \delta_{\text{true}}. \tag{75c}$$

■

K. Proof of Corollary 4

Given

$$\sup_x \quad \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \tag{76}$$

$$\text{VaR}_\tau(J_{\text{true}}) \leq \text{VaR}_{\tau+\alpha}(J_{\text{sim}}). \tag{77}$$

Therefore

$$\Pr[\text{VaR}_\tau(J_{\text{true}}) \leq \overline{\text{VaR}}_\tau(\alpha)] \tag{78a}$$

$$\geq \Pr[\text{VaR}_{\tau+\alpha}(J_{\text{sim}}) \leq \overline{\text{VaR}}_{\tau+\alpha}] \tag{78b}$$

$$\geq 1 - \delta. \tag{78c}$$

■

L. Proof of Corollary 5

Given

$$\sup_x \quad \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \tag{79}$$

we can create a lower bound on $\text{CDF}_{\mathcal{D}_{\text{true}}}$ by lowering a bound on $\text{CDF}_{\mathcal{D}_{\text{sim}}}$ by amount α. That is,

$$\Pr[\underline{\text{CDF}}(y) \leq \text{CDF}_{\mathcal{D}_{\text{true}}}(y) \quad \forall y] \geq 1 - \delta \tag{80}$$

where, following the proof of Theorem 3, and using $\epsilon' = \epsilon + \alpha = \sqrt{-\ln \delta / 2n} + \alpha$

$$\underline{\text{CDF}}(y) = \begin{cases} \max\{\widehat{\text{CDF}}(y) - \epsilon', 0\}, & \text{if } y < y_{\text{ub}} \\ \underline{\text{CDF}}(y) = 1, & \text{if } y \geq y_{\text{ub}}. \end{cases} \tag{81}$$

Following the proof of Theorem 3, the robust bound is formed by replacing ϵ in the original bound equation with ϵ'.

■

M. Proof of Corollary 6

Given

$$\sup_x \quad \text{CDF}_{\mathcal{D}_{\text{sim}}}(x) - \text{CDF}_{\mathcal{D}_{\text{true}}}(x) \leq \alpha \tag{82}$$

we know $q_{\text{true}} \leq q_{\text{sim}} + \alpha$. Then, let

$$k^* = \min\{k' \in \{0, \ldots, n\} \mid \text{Bin}(k'; n, q_{\text{sim}}) \geq \delta\}. \tag{83}$$

With the following implications:

$$k \geq k^* \Rightarrow \text{Bin}(k; n, q_{\text{sim}}) \geq \delta \tag{84a}$$

$$\Rightarrow \text{Bin}(k; n, q_{\text{true}} - \alpha) \geq \delta \Rightarrow q_{\text{true}} \leq \overline{q}(\alpha) \tag{84b}$$

we know that

$$\Pr[q_{\text{true}} \leq \overline{q}(\alpha)] \geq \Pr[k \geq k^*] \tag{85a}$$

$$= 1 - \Pr[k \leq k^* - 1] \geq 1 - \delta. \tag{85b}$$

■

ACKNOWLEDGMENT

The authors would like to acknowledge Taylor Howell for his advice on the sampling-based manipulation policy. This article solely reflects the opinions and conclusions of its authors and not any NASA entity.

REFERENCES

[1] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *J. Risk*, vol. 2, pp. 21–42, 2000.
[2] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes," *Math. Program.*, vol. 125, pp. 235–261, 2010.
[3] S. Carpin, Y.-L. Chow, and M. Pavone, "Risk aversion in finite Markov decision processes using total cost criteria and average value at risk," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 335–342.
[4] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 5026–5033.
[5] R. De Lazcano, K. Andreas, J. J. Tai, S. R. Lee, and J. Terry, "Gymnasium robotics," 2023. [Online]. Available: http://github.com/Farama-Foundation/Gymnasium-Robotics
[6] P. Akella, M. Ahmadi, and A. D. Ames, "A scenario approach to risk-aware safety-critical system verification," 2022, *arXiv:2203.02595*.
[7] H. Krasowski, P. Akella, A. D. Ames, and M. Althoff, "Safe reinforcement learning with probabilistic guarantees satisfying temporal logic specifications in continuous action spaces," in *Proc. IEEE 62nd Conf. Decis. Control*, 2023, pp. 4372–4378.
[8] P. Akella, W. Ubellacker, and A. D. Ames, "Safety-critical controller verification via Sim2Real gap quantification," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 10539–10545.
[9] P. Akella, A. Dixit, M. Ahmadi, J. W. Burdick, and A. D. Ames, "A scenario approach to risk-aware safety-critical system verification," *Artif. Intell.*, vol. 336, 2024, Art. no 104195.
[10] P. Akella, W. Ubellacker, and A. D. Ames, "Probabilistic guarantees for nonlinear safety-critical optimal control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 8120–8126.
[11] M. Cleaveland, L. Lindemann, R. Ivanov, and G. J. Pappas, "Risk verification of stochastic systems with neural network controllers," *Artif. Intell.*, vol. 313, 2022, Art. no. 103782.
[12] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Statist.*, vol. 27, no. 3, pp. 642–669, 1956.
[13] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Ann. Probab.*, vol. 18, no. 3, pp. 1269–1283, 1990.
[14] M. Ahmadi, X. Xiong, and A. D. Ames, "Risk-averse control via CVaR barrier functions: Application to bipedal robot locomotion," *IEEE Control Syst. Lett.*, vol. 6, pp. 878–883, 2022.
[15] T. Lew, R. Bonalli, and M. Pavone, "Chance-constrained sequential convex programming for robust trajectory optimization," in *Proc. Eur. Control Conf.*, 2020, pp. 1871–1878.
[16] A. Hakobyan and I. Yang, "Wasserstein distributionally robust motion control for collision avoidance using conditional value-at-risk," *IEEE Trans. Robot.*, vol. 38, no. 2, pp. 939–957, Apr. 2022.
[17] A. Navsalkar and A. R. Hota, "Data-driven risk-sensitive model predictive control for safe navigation in multi-robot systems," in *2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 1442–1448.
[18] A. Dixit, M. Ahmadi, and J. W. Burdick, "Risk-sensitive motion planning using entropic value-at-risk," in *Proc. Eur. Control Conf.*, 2021, pp. 1726–1732.
[19] E. R. Hunt, C. B. Cullen, and S. Hauert, "Value at risk strategies for robot swarms in hazardous environments," *Proc. SPIE*, vol. 11758, pp. 158–177, 2021.
[20] T. Hiraoka, T. Imagawa, T. Mori, T. Onishi, and Y. Tsuruoka, "Learning robust options by conditional value at risk optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 32, pp. 2619–2629, 2019.
[21] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 4759–4770.
[22] R. Dyro, J. Harrison, A. Sharma, and M. Pavone, "Particle MPC for uncertain and learning-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 7127–7134.
[23] T. Lew, R. Bonalli, and M. Pavone, "Sample average approximation for stochastic programming with equality constraints," 2022, *arXiv:2206.09963*.
[24] T. Lew, R. Bonalli, and M. Pavone, "Risk-averse trajectory optimization via sample average approximation," *IEEE Robot. Autom. Lett.*, vol. 9, no. 2, pp. 1500–1507, Feb. 2024.
[25] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, no. 3, pp. 371–421, 2008.
[26] R. Luo et al., "Sample-efficient safety assurances using conformal prediction," *Int. J. Robot. Res.*, 2023.

[27] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe planning in dynamic environments using conformal prediction," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 5116–5123, Aug. 2023.
[28] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick, "Adaptive conformal prediction for motion planning among dynamic agents," in *Proc. Learn. Dyn. Control Conf.*, 2023, pp. 300–314.
[29] D. B. Brown, "Large deviations bounds for estimating conditional value-at-risk," *Oper. Res. Lett.*, vol. 35, no. 6, pp. 722–730, 2007.
[30] Y. Wang and F. Gao, "Deviation inequalities for an estimator of the conditional value-at-risk," *Oper. Res. Lett.*, vol. 38, no. 3, pp. 236–239, 2010.
[31] P. Thomas and E. Learned-Miller, "Concentration inequalities for conditional value at risk," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6225–6233.
[32] R. K. Kolla, L. Prashanth, S. P. Bhat, and K. Jagannathan, "Concentration bounds for empirical conditional value-at-risk: The unbounded case," *Oper. Res. Lett.*, vol. 47, no. 1, pp. 16–20, 2019.
[33] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
[34] L. Prashanth and S. P. Bhat, "A Wasserstein distance approach for concentration of empirical risk estimates," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 10830–10890, 2022.
[35] B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier, "Qualitative multi-armed bandits: A quantile-based approach," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1660–1668.
[36] S. R. Howard and A. Ramdas, "Sequential estimation of quantiles with applications to A/B testing and best-arm identification," *Bernoulli*, vol. 28, no. 3, pp. 1704–1728, 2022.
[37] R. Zieliński and W. Zieliński, "Best exact nonparametric confidence intervals for quantiles," *Statistics*, vol. 39, no. 1, pp. 67–71, 2005.
[38] H. Scheffe and J. W. Tukey, "Non-parametric estimation. I. Validation of order statistics," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 187–192, 1945.
[39] H. A. David and H. N. Nagaraja, *Order Statistics*. Hoboken, NJ, USA: Wiley, 2004.
[40] T. W. Anderson, "Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function," Dept. Statist., Stanford Univ., Stanford, CA, USA, Tech. Rep. AND ONR 1, 1969.
[41] M. Phan, P. Thomas, and E. Learned-Miller, "Towards practical mean bounds for small samples," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8567–8576.
[42] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon, "Time-uniform, nonparametric, nonasymptotic confidence sequences," *Ann. Statist.*, vol. 49, no. 2, pp. 1055–1080, 2021.
[43] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statist. Sci.*, vol. 16, no. 2, pp. 101–133, 2001.
[44] A. M. Pires and C. Amado, "Interval estimators for a binomial proportion: Comparison of twenty methods," *REVSTAT—Statist. J.*, vol. 6, no. 2, pp. 165–197, 2008.
[45] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.
[46] T. E. Sterne, "Some remarks on confidence or fiducial limits," *Biometrika*, vol. 41, no. 1/2, pp. 275–278, 1954.
[47] E. L. Crow, "Confidence intervals for a proportion," *Biometrika*, vol. 43, nos. 3/4, pp. 423–435, 1956.
[48] M. W. Eudey, "I. on the treatment of discontinuous random variables. II. Statistical model for comparing two methods of diagnosis," Ph.D. dissertation, Dept. Math., Univ. California, Berkeley, CA, USA, 1949.
[49] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, vol. 4. Berlin, Germany: Springer, 2022.
[50] W. L. Stevens, "Fiducial limits of the parameter of a discontinuous distribution," *Biometrika*, vol. 37, no. 1/2, pp. 117–129, 1950.
[51] W. Wang, "Smallest confidence intervals for one binomial proportion," *J. Statist. Plan. Inference*, vol. 136, no. 12, pp. 4293–4306, 2006.
[52] V. Vovk, "Conditional validity of inductive conformal predictors," in *Proc. Asian Conf. Mach. Learn.*, 2012, pp. 475–490.
[53] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2021, *arXiv:2107.07511*.
[54] P. Wawrzyński, "A cat-like robot real-time learning to run," in *Proc. 9th Int. Conf. Adaptive Natural Comput. Algorithms*, 2009, pp. 380–390.
[55] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc Int. Conf. Learn. Representations*, 2016.
[56] R. Coulom, "Reinforcement learning using neural networks, with applications to motor control," Ph.D. dissertation, Dept. Cogn. Sci., Inst. Nat. Polytechnique de Grenoble-INPG, Grenoble, France, 2002.
[57] S. Mannor, R. Y. Rubinstein, and Y. Gat, "The cross entropy method for fast policy search," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 512–519.
[58] Z. Birnbaum and F. H. Tingey, "One-sided confidence contours for probability distribution functions," *Ann. Math. Statist.*, vol. 22, no. 4, pp. 592–596, 1951.
[59] H. Abdi, "The Bonferonni and Šidák corrections for multiple comparisons," *Encyclopedia Meas. Statist.*, vol. 3, pp. 103–107, Jan. 2007.
[60] A. Melnik, L. Lach, M. Plappert, T. Korthals, R. Haschke, and H. Ritter, "Using tactile sensing to improve the sample efficiency and performance of deep deterministic policy gradients for simulated in-hand manipulation tasks," *Front. Robot. AI*, vol. 8, 2021, Art. no. 538773.
[61] T. Howell, N. Gileadi, S. Tunyasuvunakool, K. Zakka, T. Erez, and Y. Tassa, "Predictive sampling: Real-time behaviour synthesis with MuJoCo," 2022, *arXiv:2212.00541*.
[62] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 1433–1440.
[63] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA, USA: SIAM, 2021.
[64] R. Hulsman, "Distribution-free finite-sample guarantees and split conformal prediction," master's thesis, Dept. Statist., Univ. Oxford, 2022.

Joseph A. Vincent received the B.S. degree in aerospace engineering from the University of Kansas, Lawrence, KS, USA, in 2018, and the M.S. degree in aeronautics and astronautics in 2020 from Stanford University, Stanford, CA, USA, where he is currently working toward the Ph.D. degree in aeronautics and astronautics with Stanford University.

His research interests include evaluation of robotic systems using statistical and reachability-based methods.

Aaron O. Feldman received the B.S. degree in information and data sciences from the California Institute of Technology, Pasadena, CA, USA, in 2022. He is currently working toward the Ph.D. degree in aeronautics and astronautics with Stanford University, Stanford, CA, USA.

His current research interests include statistical methods to guarantee safety and improve performance for robotic systems operating under uncertainty.

Mac Schwager (Member, IEEE) received the B.S. degree from Stanford University, Stanford, CA, USA, in 2000, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2005 and 2009, respectively, all in mechanical engineering.

He is currently an Associate Professor of Aeronautics and Astronautics with Stanford University. His research interests include distributed algorithms for control, perception, and learning in groups of robots, and models of cooperation and competition in groups of engineered and natural agents.

Dr. Schwager was the recipient of the NSF CAREER Award in 2014, the DARPA Young Faculty Award in 2018, and a Google Faculty Research Award in 2018, and the IROS Toshio Fukuda Young Professional Award in 2019.