How Generalizable is My Behavior Cloning Policy? A Statistical Approach to Trustworthy Performance Evaluation

Joseph A. Vincent, Haruki Nishimura, *Member, IEEE*, Masha Itkina, Paarth Shah, Mac Schwager, *Member, IEEE*, and Thomas Kollar, *Member, IEEE*

***Abstract*—With the rise of stochastic generative models in robot policy learning, end-to-end visuomotor policies are increasingly successful at solving complex tasks by learning from human demonstrations. Nevertheless, since real-world evaluation costs afford users only a small number of policy rollouts, it remains a challenge to accurately gauge the performance of such policies. This is exacerbated by distribution shifts causing unpredictable changes in performance during deployment. To rigorously evaluate behavior cloning policies, we present a framework that provides a tight lower-bound on robot performance in an arbitrary environment, using a minimal number of experimental policy rollouts. Notably, by applying the standard stochastic ordering to robot performance distributions, we provide a worst-case bound on the *entire distribution* of performance (via bounds on the cumulative distribution function) for a given task. We build upon established results to ensure that the bounds hold with a user-specified confidence level and tightness, and are constructed from as few policy rollouts as possible. In experiments we evaluate policies for visuomotor manipulation in both simulation and hardware. Specifically, we i) empirically validate the guarantees of the bounds in simulated manipulation settings, ii) find the degree to which a learned policy deployed on hardware generalizes to new real-world environments, and iii) rigorously compare two policies tested out-of-distribution. Our data, code, and implementation of confidence bounds are open-source.**

***Index Terms*—Performance evaluation and benchmarking, probability and statistical methods, learning from demonstration, AI-enabled robotics.**

I. Introduction

In THIS letter we focus on evaluating robot policies obtained through the *behavior cloning* (BC) framework of robot learning. BC policies, such as diffusion policies [1], have recently advanced the state-of-the-art (SOTA) in visuomotor policy synthesis, particularly in manipulation [1], [2], [3], [4], [5]. BC is attractive because it avoids the challenges of the sim-to-real gap, which impede the transfer of policies from reinforcement learning and other techniques to real-world robots [6]. In BC, humans give demonstrations, often directly on the robot hardware through teleoperation. Then, a policy is learned based on these demonstrations. This procedure removes the need for a simulation model, avoiding the sim-to-real gap.

However, without an accurate simulation model, evaluation of BC policies relies on real-world tests. In robotics research it is common to evaluate these policies using fewer than 50 policy rollouts, recording the empirical success rate or average reward (e.g., [1], [2], [7]). With such small sample sizes, it can be difficult to interpret the significance of the recorded results. In addition, measuring average performance can be insufficient for applications with safety and reliability requirements.

To address this need, we propose statistical bounds to rigorously evaluate the performance of a BC policy. We quantify performance through a user-specified metric, either a binary success/failure or a continuous reward. Although many BC policies are trained and deployed in hardware, these metrics can also be given for simulated robotics settings. While specifying a suitable reward for policy training can be challenging in practice, we emphasize that we do not use the performance metric for training, only for evaluation. Simple performance metrics (e.g., task success/failure, distance to a goal location) are sufficient for our purposes.

Our proposed method, shown in Fig. 1, is to compute worst-case bounds (useful for determining whether a policy generalizes) on the performance of a policy using a small number of policy rollouts. In Section III, we define an ordering of policy performance distributions that leads to a notion of a worst-case distribution. Our approach can specify the fewest number of policy rollouts required to obtain user-specified *confidence* and *tightness* for a bound on the performance metric. *Confidence* is the probability that the bound holds.[1] *Tightness* quantifies how

[1]Sometimes also referred to as the *coverage* of the bound.

Manuscript received 8 May 2024; accepted 7 August 2024. Date of publication 19 August 2024; date of current version 30 August 2024. This article was recommended for publication by Associate Editor P. Ogren and Editor M. Vincze upon evaluation of the reviewers' comments. The work of Joseph A. Vincent was supported by the Dwight D. Eisenhower Transportation Fellowship. This work was supported in part by the NASA University Leadership initiative under Grant 80NSSC20M0163 to assist the authors with their research and in part by the ONR under Grant N00014-23-1-2354. *(Corresponding author: Joseph A. Vincent.)*

Joseph A. Vincent is with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 USA. The work was primarily done during his internship at Toyota Research Institute (e-mail: josephav@stanford.edu).

Haruki Nishimura, Masha Itkina, Paarth Shah, and Thomas Kollar are with the Toyota Research Institute, Los Altos, CA 94022 USA (e-mail: haruki.nishimura@tri.global; masha.itkina@tri.global; paarth.shah@tri.global; thomas.kollar@tri.global).

Mac Schwager is with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 USA (e-mail: schwager@stanford.edu).

Project Page: https://tri-ml.github.io/stochastic_verification
Letter Repository: https://github.com/TRI-ML/stochastic_verification
Binomial Confidence Intervals: https://github.com/TRI-ML/binomial_cis

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3445635, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3445635

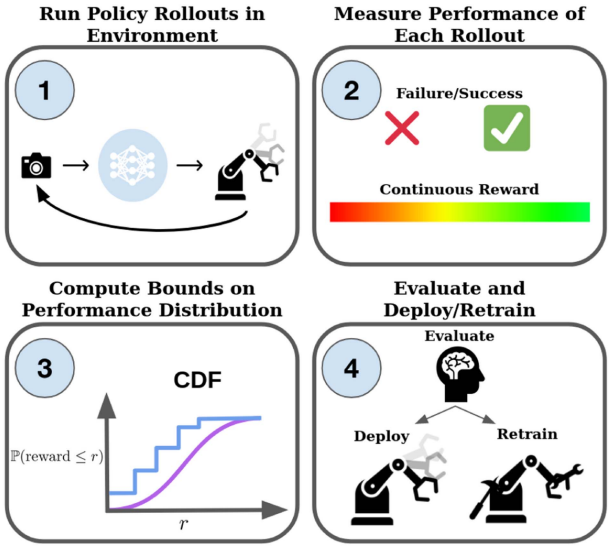

Fig. 1. In our evaluations, first, policy rollouts are collected in the environment of interest. Second, each rollout results in either a binary or continuous performance measurement. Third, an upper confidence bound on the CDF of performance is computed. Finally, based on the bound, the user chooses to deploy or retrain the policy.

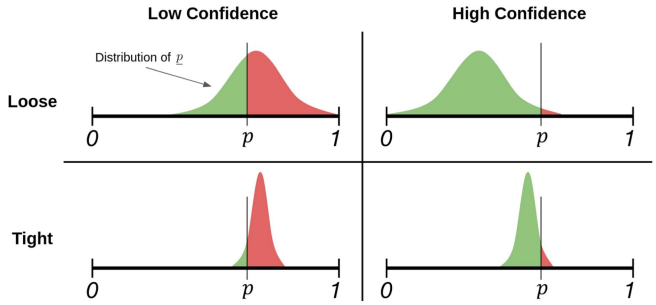

Fig. 2. Hypothetical distributions of a lower confidence bound $\underline{p}$ for an unknown probability of success p. High confidence levels give better chances that the $\underline{p}$ we obtain is lower than p (green shaded region), and tighter bounds give better chances that $\underline{p}$ is close to p.

close the bound tends to be to the true (but unknown) quantity of interest. A bound which is tighter provides a better estimate of the unknown quantity while a bound which is higher confidence provides a better chance that the estimate is conservative. In Fig. 2 we give a visual interpretation of how confidence and tightness relate to the distribution of a lower confidence bound for an unknown probability of success. Achieving higher confidence and tighter bounds on policy performance comes at the cost of more policy rollouts.

The bounds we use require no knowledge of the underlying probability distributions in the policy, training data, robot hardware, or environment physics. The confidence and tightness of these bounds do not depend on the dimensionality of the robot state or observation. We only require performance measurements to be i.i.d. which we further discuss in Section III-B.

We demonstrate how to construct bounds on performance when it is measured by i) a binary success/failure metric, and ii) a continuous metric of reward accumulated over a trajectory. For binary metrics, we seek a lower bound on the unknown probability of task success, utilizing a rarely-used bound which is tighter than the more common Clopper-Pearson bound [8]. Our bound exactly achieves a user-specified confidence level and is as tight as possible. This is enabled by a novel computational method based on mixed monotonic optimization [9]. For continuous metrics, the true distribution of reward is entirely unknown. We propose to use bounds for the cumulative distribution function (CDF) of reward which fully characterizes the reward distribution, allowing flexibility for the user to then analyze performance, e.g., in the tails, at any quantile, or in the mean. For this, we use a tighter version of the Dvoretzky–Kiefer–Wolfowitz (DKW) bound [10], [11].

For either metric, the user specifies the confidence level and the tightness of the bound. Since real-world evaluation is costly, the bounds we employ meet these specifications using the minimum number of policy rollouts. This is in contrast to the more common Clopper-Pearson and DKW bounds.

Finally, in our experiments we apply our framework to investigate the out-of-distribution (OOD) generalization of BC policies, by directly testing in the OOD settings. Because human intuition is often unreliable for OOD performance, we believe this is an application where statistical evaluation methods shine, giving roboticists well-defined confidence levels for the results of their experiments. We demonstrate our bounds in simulation and real-world manipulation settings. We compare our bounds to more common bounds from the literature, showing improved tightness, particularly for few policy rollouts. We also show how our bounds can be used to compare two policies. Finally, in hardware, we bound the performance of a diffusion policy [1] in new environments. An early version of these results appeared in a workshop [12]. Here we substantially expand the analysis and simulation results, add hardware experiments, and add a policy comparison example. Our contributions are as follows:

- A method for bounding the performance of a BC robot policy in an arbitrary environment with minimal policy rollouts, by putting in practice the concept of worst-case bounds on the entire distribution of performance.
- A novel method for computing tightness for binomial confidence intervals, specifically *maximum expected shortage* (the worst-case value for the average degree of underestimation of the true success rate).
- An open-source implementation for computing the binomial bounds described in the letter, which optimally trade-off between confidence, tightness, and sample size: https://github.com/TRI-ML/binomial_cis.

II. Related Work

In this section, we discuss related literature on rigorously evaluating learned robot policies. We first discuss approaches from formal methods, then discuss statistical approaches.

A. Formal Evaluation of Learned Policies

Formal methods applied to robots with learned components can be used to evaluate open-loop policy requirements [13] and closed-loop system requirements [14], [15], [16], [17]. These methods typically have strict limitations on the structure and complexity of the policy and environment model. In settings where these requirements are met, formal methods can evaluate rich closed-loop behaviors by finding certificates such as invariant sets [15], [16], Lyapunov functions [18], [19], and

contraction metrics [20], [21]. However, in this letter we do not assume access to an environment model or policy structure, precluding the use of formal evaluation methods. Next, we discuss statistical evaluation, which is one flavor of black-box validation [22], that that provides probabilistic guarantees.

B. Statistical Evaluation of Learned Policies

When constructing statistical bounds, there are trade-offs between confidence, tightness, and sample size. Related work tends to focus on the trade-off between confidence and sample size, often neglecting tightness. Statistical bounds on the performance of learned policies is explored in [23], [24], [25], [26]. However, these works only bound particular scalar quantities such as the expected value, value-at-risk, and conditional value-at-risk of the performance distribution. We place bounds on the CDF that consider all aspects of the performance distribution. Further, these works do not quantify bound tightness and their bounds can require more policy rollouts than necessary to achieve the specified confidence.

Conformal prediction has been used for uncertainty quantification of robot perception components [27], [28], [29], [30] and policies [31], [32]. This has enabled policies to increase success rates when e.g., avoiding collisions or making language-based plans. However, like the other approaches in this section, conformal prediction i) ignores tails of distributions, ii) does not quantify tightness, and iii) may require more samples than necessary to achieve a desired confidence [33], [34]. In this letter we bound the entire distribution of performance, quantify bound tightness, and ensure that the number of samples used is minimal.

Confidence bounds for unknown success probabilities and CDFs are well studied with the Clopper-Pearson bound being common for success probabilities [8] and with the DKW inequality being common for CDFs [10], [11]. However, these bounds are loose, requiring more samples than necessary. In Sections IV and V, we discuss less common yet well-established bounds which optimally trade-off between confidence, tightness, and number of samples. Common approaches which test for differences in success rates include Fisher's and Barnard's tests [35]. These are the standard methods for performing A/B tests to distinguish success rates. However, while such tests determine which policy performs better, they do not return confidence intervals for the individual success rates (as we do), but rather provide confidence intervals for the odds ratio. The drawback is that the odds ratio lacks information about absolute performance, i.e., pairs of policies with low success rates and with high success rates can have the same odds ratio. Our view is that it is not enough to know which policy performs better, but it is also important to know the performance of each policy individually.

III. Distributional Bounds

A. Bounds

We want to obtain worst-case bounds on the *entire distribution* of performance (via bounds on the CDF), rather than conventional quantities such as expected value. This approach affords a rich understanding of performance, hedging against deploying a policy with some hidden drawbacks (e.g., high expected reward but a long tail of low reward).

Consider two CDFs $F_1(x) = \mathbb{P}(X_1 \leq x)$ and $F_2(x) = \mathbb{P}(X_2 \leq x)$, describing the distribution of performance under two different policies. If $F_2(x) \geq F_1(x)\ \forall x$, then $F_2(x)$ has more probability mass on *lower* performance values. Therefore $F_1(x)$ is preferable to $F_2(x)$, following the standard notion of stochastic ordering [36]. If the inequality does not hold for $\forall x$, we do not claim one distribution is preferable to the other. Thus, we have a partial ordering of performance distributions.

A confidence bound $\overline{F}(x)$ with confidence $1-\alpha$ satisfies

$$\mathbb{P}(F(x) \leq \overline{F}(x)\ \forall x) \geq 1-\alpha. \tag{1}$$

$\overline{F}(x)$ is constructed based on the observed performance of some number of policy rollouts. Since the outcome of policy rollouts is random, $\overline{F}(x)$ is also random, leading to the bound holding probabilistically as in (1). In light of the above discussion on partial ordering of CDFs, $\overline{F}(x)$ is a *worst-case* bound on the distribution of performance with confidence $1-\alpha$. In the case of a binary performance metric, an upper bound on the CDF is simply a lower bound on the success rate.

B. Assumptions

The methods we use in Sections IV and V to construct $\overline{F}(x)$ rely on i.i.d. performance measurements.

Assumption 1: The performance measurements from the policy rollouts are i.i.d. random variables.

That the performance measurements are random variables is ensured by a stochastic policy (or stochastic environment). To be independent, the outcome of one policy rollout must have no influence on the outcome of another. To be identically distributed, there must be no distribution shift between rollouts. For binary metrics, Assumption 1 ensures the performance will follow a Bernoulli distribution with unknown probability of success. For continuous metrics, the assumption allows for the performance to follow any distribution (including discrete and mixed continuous-discrete distributions).

To avoid violating Assumption 1, a testing plan should be developed *before* collecting any policy rollouts, documenting the sampling rule for initial conditions and the number of policy rollouts. Practices which violate Assumption 1 include

- running policy rollouts until enough favorable outcomes are observed (use confidence sequences for this [37]),
- running policy rollouts in a time-varying environment (e.g., lighting changes between rollouts due to sunset).

For further discussion on misuses see [38]. Next, we give bounds for binary and continuous performance metrics.

IV. Confidence Bounds - Binary Metric

We are interested in finding lower confidence bounds on an unknown probability of success using a minimal number of policy rollouts. To determine such bounds we use the approach in [36] to achieve a desired confidence level with minimal policy rollouts. We then introduce a computational method to achieve a desired tightness with minimal policy rollouts.

To find lower bounds on a policy's success rate in a new environment, we treat the result of each policy rollout as a Bernoulli random variable X, where $X=1$ denotes a success and $X=0$ denotes a failure. Then, given the n i.i.d. success/failure measurements, we construct a lower confidence bound $\underline{p}$. The number of observed successes follows a binomial distribution. Before describing the bound, we first introduce optimality criteria for lower confidence bounds.

A. Optimality Criteria

The standard notion of optimality for lower confidence bounds is that of being *uniformly most accurate* (UMA).

Definition 1 (Uniformly Most Accurate, 3.22 of [36]): For test statistic T, a lower confidence bound $\underline{\theta}(T)$ satisfying

$$\mathbb{P}_\theta[\underline{\theta}(T) \leq \theta] \geq 1-\alpha \quad \forall\theta \tag{2}$$

and, amongst all possible lower bounds $\underline{\theta}'(T)$, minimizes

$$\mathbb{P}_\theta[\underline{\theta}'(T) \leq \theta_0] \quad \forall\theta_0 < \theta \tag{3}$$

is a UMA lower confidence bound at confidence level $1-\alpha$.

A UMA lower confidence bound underestimates the unknown parameter θ by as little as possible. *Shortage* (i.e. excess width) is used to quantify the amount of underestimation,

$$\text{shortage} = \max\{\theta - \underline{\theta}, 0\}. \tag{4}$$

Using the Ghosh-Pratt identity [39], [40],

$$\text{ES}(\theta) = \mathbb{E}_\theta[\text{shortage}] = \int_{\theta_0<\theta} \mathbb{P}_\theta[\underline{\theta} \leq \theta_0] d\theta_0. \tag{5}$$

Now, since a UMA lower confidence bound minimizes the integrand of (5), a UMA lower confidence bound also minimizes expected shortage for all values of θ. Although expected shortage is a useful quantity, it depends on the unknown value of θ. To avoid assumptions on θ, we measure tightness according to *maximum expected shortage* (*MES*),

$$\text{MES} = \underset{\theta}{\text{maximize}} \;\; \text{ES}(\theta). \tag{6}$$

MES is the worst-case expected shortage over all possible parameter values θ. The notion of MES is not new [41], but it is not commonly used since the maximization in (6) can be challenging when not concave. Next, we describe a UMA lower bound for an unknown success probability and introduce the first tractable method for computing its MES.

B. Optimal Binomial Bounds

Let $\text{bin}(k,n,p)$ and $\text{Bin}(k,n,p)$ denote the binomial probability mass function and CDF, with k successes, n trials, and success probability p. Let $\lfloor x \rfloor$ denote the floor function. The test statistic $T = U + \sum_{i=1}^{n} X_i$, where $U \sim \mathcal{U}[0,1]$ and X_i observed successes and failures, has probability density

$$f_p(t) = \binom{n}{\lfloor t \rfloor} p^{\lfloor t \rfloor}(1-p)^{n-\lfloor t \rfloor}. \tag{7}$$

Then, the CDF of T is given by

$$F_p(t) = \int_0^t \binom{n}{\lfloor x \rfloor} p^{\lfloor x \rfloor}(1-p)^{n-\lfloor x \rfloor} dx \tag{8a}$$

$$= \text{Bin}(\lfloor t \rfloor - 1, n, p) + (t - \lfloor t \rfloor) \cdot \text{bin}(\lfloor t \rfloor, n, p). \tag{8b}$$

$F_p(t)$ is continuous in both t and p. Thus, by Corollary 3.5.1 of [36], we have the UMA lower confidence bound,

$$\mathbb{P}[\underline{p} \leq p] \geq 1-\alpha \tag{9a}$$

$$\underline{p}(t) = \begin{cases} 0 & \text{if } t < 1-\alpha \\ 1 & \text{if } t > n+1-\alpha \\ p^* & \text{s.t. } F_{p^*}(t) = 1-\alpha \text{ otherwise.} \end{cases} \tag{9b}$$

Note that since the test statistic is defined as a sum of uniform and binomial random variables, given the same number of observed successes, t will be almost surely distinct, resulting in distinct values for $\underline{p}(t)$. Thus, $\underline{p}$ is known as a *randomized confidence bound*, and this property is necessary to construct a UMA confidence bound for an unknown success rate [36].

A bisection search can be used to solve (9b), as $F_p(t)$ is decreasing in p. Taking limits $p \to 0^+$ and $p \to 1^-$, one finds $F_p(t) = 1-\alpha$ has no solution when $t < 1-\alpha$ and $t > n+1-\alpha$. This bound is well-established (see Example 3.5.2 of [36]). The Clopper-Pearson bound is obtained by setting $U = 0$ in the test statistic.

Next, we compute expected shortage. A useful identity is

$$\mathbb{P}[\underline{p} \leq p_0] = \mathbb{P}[t \leq t^*(p_0)] \tag{10}$$

where $t^*(p_0)$ is the unique value that satisfies

$$F_{p_0}(t^*) = 1-\alpha. \tag{11}$$

Then, we can compute expected shortage as

$$\text{ES}(p) = \int_0^p \mathbb{P}_p[\underline{p} \leq p_0] dp_0 = \int_0^p \mathbb{P}_p[t \leq t^*(p_0)] dp_0 \tag{12a}$$

$$= \int_0^p F_p(t^*(p_0)) dp_0. \tag{12b}$$

This integral can be evaluated with standard software.

Next, we find MES via global optimization. Specifically, consider the following reformulation of expected shortage,

$$\text{ES}(p_1, p_2) = \int_0^{p_1} F_{p_2}(t^*(p_0)) dp_0. \tag{13}$$

In this form, expected shortage is increasing in p_1 and decreasing in p_2; this property is known as *mixed monotonicity*. Mixed monotonic functions can be globally optimized [9]. This approach is the first tractable method that certifiably solves the optimization needed to compute MES and is included in our open-source implementation.[2]

Now, given two of the following: i) confidence level $(1-\alpha)$, ii) tightness (MES), iii) number of policy rollouts (n), we can determine the third. For example, since MES is decreasing in n, to determine n given a desired α and MES, one fixes α and computes the smallest n (via a bisection search) which yields an MES below the desired value. In Fig. 3 we show how these quantities vary with one another for our bound and the Clopper-Pearson bound. Since the bound we use is UMA, it always has lower MES. Also novel is the MES computation for Clopper-Pearson. We derive the appropriate form of (12) and compute MES using mixed monotonic programming.[3] Lastly, one can compute a lower confidence bound on the *failure* probability using the same procedure, but with failure counts in the test statistic T rather than success counts.

V. Confidence Bounds - Continuous Metric

When performance is measured with a continuous metric, the policy rollouts result in i.i.d. samples from some unknown distribution of reward. This distribution may be continuous, discrete, or mixed. In this section we describe how to obtain

[2]https://github.com/TRI-ML/binomial_cis

[3]Further details on this derivation for the Clopper-Pearson bound are in the documentation of the linked binomial_cis Github repository.

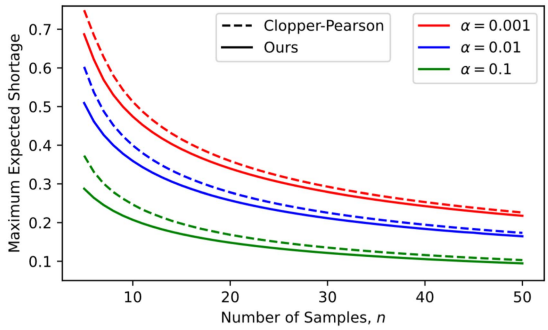

Fig. 3. Variation of MES, n, and α. Our method is always tighter than Clopper-Pearson, and appreciably so at small sample sizes.

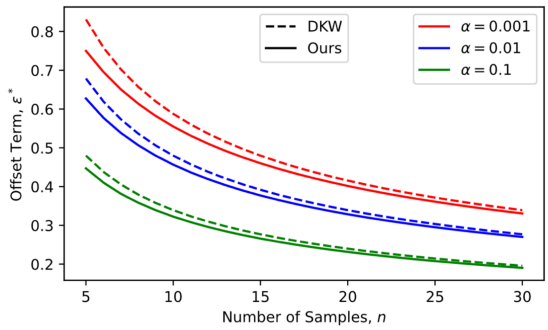

Fig. 4. Variation of ϵ^*, n, and α. Our method is always tighter than the DKW bound, and appreciably so at small sample sizes.

an upper confidence bound on the CDF of this distribution. As in Section IV, the CDF upper bound is constructed in such a way that given user-specified confidence level and tightness, the minimum number of policy rollouts is used to meet these specifications. The results used in this section are not novel (see [42]), but are uncommon in the statistics literature, especially as applied to robotics.

Consider an i.i.d. sample, $X_{1:n}$ from an unknown distribution with CDF $F(x)$. The empirical CDF is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \tag{14}$$

where $\mathbf{1}(x)$ evaluates to 1 if x is true and 0 otherwise. A one-sided Kolmogorov-Smirnov (KS) statistic [36] is

$$D_n^- = \sup_x \; F(x) - F_n(x). \tag{15}$$

By [42], if $F(x)$ continuous, then the distribution of D_n^- is

$$\mathbb{P}(D_n^- \leq \epsilon) = 1 - \epsilon \sum_{k=0}^{\lfloor n(1-\epsilon) \rfloor} w_k \tag{16a}$$

$$w_k = \binom{n}{k} \left(1 - \epsilon - \frac{k}{n}\right)^{n-k} \left(\epsilon + \frac{k}{n}\right)^{k-1}. \tag{16b}$$

Note that there is no dependence on the true $F(x)$. Moreover, if $F(x)$ is not continuous, then (16a) holds with $\geq$ rather than equality. Next, note the equivalence

$$\sup_x \; F(x) - F_n(x) \leq \epsilon \Leftrightarrow F(x) \leq F_n(x) + \epsilon \; \forall x. \tag{17}$$

Then, given i.i.d. samples $X_{1:n}$ from some $F(x)$, we have the following upper confidence bound on the CDF [42],

$$\mathbb{P}(F(x) \leq \overline{F}(x) \; \forall x) \geq 1 - \alpha \tag{18a}$$

$$\overline{F}(x) = F_n(x) + \epsilon^* \tag{18b}$$

$$\text{with } \epsilon^* \text{ chosen s.t. } \mathbb{P}(D_n^- \leq \epsilon^*) = 1 - \alpha. \tag{18c}$$

The tightness of this bound is measured by ϵ^*, the offset from the empirical CDF. The ϵ^* chosen by (18c) is optimal, if made any smaller then (18a) would not hold. ϵ^* is easily computed via a bisection search. Since ϵ^* is decreasing in n, to determine n given a desired α and ϵ^*, one fixes α and computes the smallest n (via a bisection search) which yields an ϵ^* below the desired value. One can also use the DKW inequality [10], [11] to obtain a similar bound, except with $\epsilon^* = \sqrt{-\ln \alpha/(2n)}$. However, the DKW bound is always less tight as shown in Fig. 4.

Fig. 5. Simulation tasks: can, lift, square, tool hang, transport [43].

VI. Experiments

We demonstrate our approach to evaluating BC policies in both simulation and real-world robotic manipulation settings. Performance of BC policies in OOD settings is hard to predict since it is difficult to characterize what aspects of the policy are invariant or sensitive to particular distribution shifts. This motivates the need for rigorous statistical evaluation. Thus, all of our experiments evaluate policies in OOD settings. In the absence of assumptions on the nature of the distribution shift, it is necessary to directly test the policies in the OOD settings.

In simulation, we run many evaluations and show the empirical confidence and tightness of the bounds agree with the theory. In hardware, we i) investigate the degree to which a learned policy generalizes to OOD settings, and ii) compare the OOD generalization of two learned policies. Our approach requires i.i.d. performance scores; a sufficient condition to ensure this is for the initial observations to be i.i.d. from some (potentially unknown) distribution. In simulation, this is easily satisfied. In hardware, we took efforts to eliminate any time-varying conditions that may undermine this assumption. Because our bounds are strictly tighter than those of Clopper-Pearson and DKW, we report the results for our bounds only.

A. Simulation Results

We consider five diffusion policies from [1] for visuomotor manipulation that have shown recent SOTA performance, each trained for a separate task. Fig. 5 shows training environments. After training, we modified each environment as follows,

- *Can:* Changed the color of the can to lime green, and the color of the floor to beige.
- *Lift:* Changed the cube color to blue.
- *Square:* Changed the square color to tan.

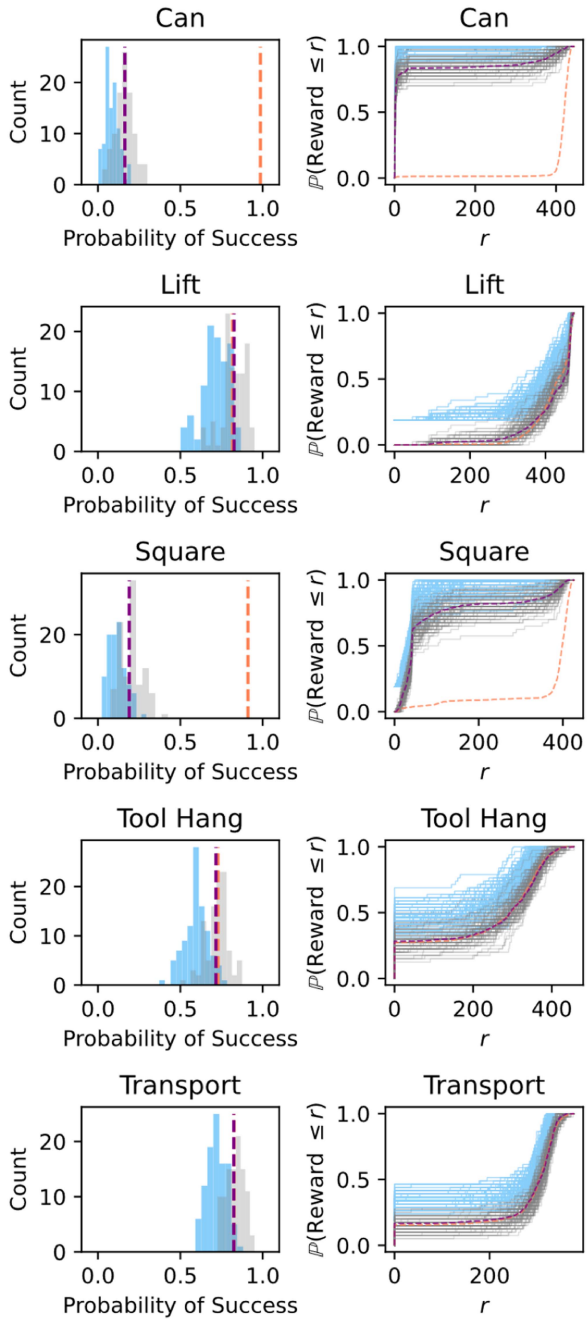

Fig. 6. Visualizations of the confidence bounds for each task. Each plot shows the OOD reward distribution (purple), Monte Carlo estimates of this distribution (gray), our confidence bounds for this distribution (blue), and the ID reward distribution (orange). We use 1000 rollouts to accurately estimate the OOD and ID reward distributions. We observe that the bounds appropriately bound the OOD reward distribution $\sim 95\%$ of the time, while a Monte Carlo estimate is overly optimistic $\sim 50\%$ of the time (since the Monte Carlo estimates are unbiased). From the ID reward distribution we can see that the distribution shifts in the Can and Square environments were harmful while the other shifts were benign.

- *Tool Hang:* Changed the wall color to floral white.
- *Transport:* Changed the color of the cube to lime green and the color of the lid handle to silver.

In Fig. 6 we show our simulation results. To obtain accurate estimates of the true in-distribution (ID) (orange) and OOD (purple) performance distributions, we used 1000 policy rollouts from each environment. We use 4000 rollouts to construct confidence bounds (100 bounds generated from 40 samples each). We also show the Monte Carlo estimates of the OOD performance distribution (gray) one would obtain using 40 samples. We construct 100 bounds to empirically validate the guarantees of the bounds, but in practice one would only construct a single bound. The true ID and OOD performance distributions are typically unknown and are shown only for validating the bounds.

From the figure we see that the distribution shifts for the Can and Square environments were quite harmful to the policy performance, while those for the other tasks were benign. Our bounds tend to accurately reflect these trends. These results reflect the unintuitive nature about when learned policies generalize, e.g., it is not clear why changing the color of the manipulated object is harmful in some cases (Can and Square) but benign in other cases (Lift and Transport). Lastly, the Monte Carlo estimates of the distributions are overly optimistic, making them unfit to be interpreted as bounds. Table I shows good agreement between the empirical confidence and tightness estimates with their theoretical values.

TABLE I
SIMULATION RESULTS OF THE DATA IN FIG. 6

	Confidence			Expected Shortage	
	Theory	Bin.	Cont.	Theory	Bin.
can	0.95	0.97	0.95	0.078	0.079
lift	0.95	0.94	0.97	0.115	0.114
square	0.95	0.90	0.91	0.085	0.077
tool hang	0.95	0.97	0.96	0.127	0.122
transport	0.95	0.96	0.96	0.115	0.105

Theory columns denote the theoretical values for confidence and expected shortage, Bin. columns denote the empirical estimates of these quantities for the binary metric bounds, the Cont. column denotes the empirical confidence level for the continuous metric bounds. Tightness of the continuous bounds is given by ϵ^*, which is not random and thus not listed. In general, we see good agreement between the theoretical values and their empirical estimates (100 trials). With more trials we expect these estimates to converge to their theoretical values.

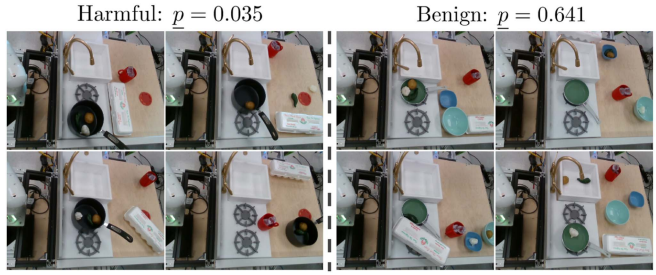

Fig. 7. Snapshots of the environments used in Section VI-B. The robot is trained to pour the ice from the red cup into the sink. Four out of 50 initial conditions are shown for each OOD setting.

B. Generalization of a Single Policy

Here we investigate the degree to which a learned policy generalizes to OOD environments. We test a diffusion policy [1] trained to pour ice from a red cup into a sink. During training there are no other objects on the tabletop. We test the policy in two OOD settings shown in Fig. 7. The first setting is hypothesized to be a harmful distribution shift (another red object in

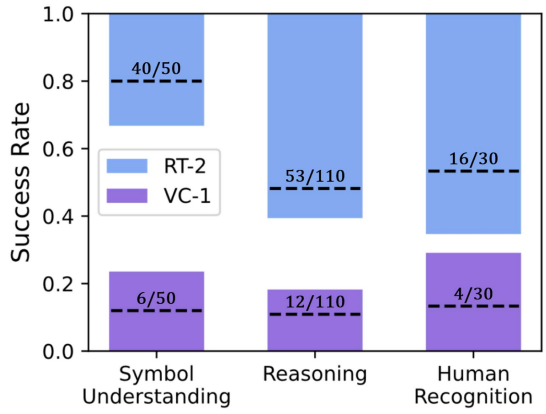

Fig. 8. Confidence intervals for success rates of RT-2 (blue) and VC-1 (purple) policies. For each setting, the intervals are disjoint and we conclude, with 95% confidence, that RT-2 outperforms VC-1. For each policy and setting we also show # successes / # trials and the corresponding empirical success rates (dashed black).

frame) and the second is hypothesized to be a benign distribution shift (no other red objects).

We performed 50 policy rollouts in each setting and found $4/50$ successes in the harmful setting and $38/50$ successes in the benign setting. Applying (9b) at a 95% confidence level we find $\underline{p}_{\text{harmful}} = 0.035$, and $\underline{p}_{\text{benign}} = 0.641$ with MES $= 0.118$. From this, we expect $\underline{p}_{\text{harmful}}$ and $\underline{p}_{\text{benign}}$ to be reasonably close to the true success rates. The low lower-bound for the harmful setting is attributed to frequent grasp failure despite its tendency to move towards the red mug.

The bounds we find support our hypotheses, and also inform us as to the degree of performance we can assert at a 95% confidence level. If a user requires a 60% success rate, then we could claim with 95% confidence that the benign setting is suitable for deployment whereas we could not for the harmful setting. We may then choose to retrain the policy with data from this setting and re-evaluate.

C. Comparing the Generalization of Two Policies

In this experiment we compare two learned visuomotor manipulation policies from [44], where the authors test their RT-2 policy against a VC-1 policy in three settings designed to test emergent capabilities in symbol understanding, reasoning, and human recognition. We compute a lower confidence bound on the success rate of the RT-2 policy ($\underline{p}_{\text{RT-2}}$) and an upper confidence bound on the success rate of the VC-1 policy ($\overline{p}_{\text{VC-1}}$). Note, $\overline{p}_{\text{VC-1}} = 1 - \underline{q}_{\text{VC-1}}$, where $\underline{q}_{\text{VC-1}}$ is a lower confidence bound on the failure rate. Each bound has confidence level 0.975, to ensure a joint confidence of 0.95,

$$\mathbb{P}\left[p_{\text{RT-2}} \in [\underline{p}_{\text{RT-2}}, 1] \cap p_{\text{VC-1}} \in [0, \overline{p}_{\text{VC-1}}]\right] \geq 0.95. \tag{19}$$

Then, if $\overline{p}_{\text{VC-1}} < \underline{p}_{\text{RT-2}}$ (disjoint bounds), we conclude that the RT-2 policy outperforms the VC-1 policy. Furthermore, the chances we come to this conclusion incorrectly are $\leq 5\%$.

In Fig. 8, we show the bounds for each policy in each evaluation setting. We find in each case there is enough evidence to conclude (at a 95% confidence level) that the RT-2 policy outperforms the VC-1 policy. For some settings the gap between the bounds is larger, this may be due to the actual success rates being further apart, or an effect of sample size. With separate bounds for each policy, we can also conclude how well each policy performs individually, an important aspect that standard A/B testing approaches (like Fisher's exact test) do not provide [35]. Finally, this approach easily extends to the case of continuous rewards. A lower bound on the CDF is easily computed since $D_n^+ = \sup_x \; F_n(x) - F(x)$ has the same distribution as D_n^- (16) [42].

VII. Conclusion

The primary contribution of this letter is utilizing uncommon but optimal statistical bounds to efficiently measure policies' OOD generalization from few samples. For binary and continuous performance metrics, we place confidence bounds on the entire distribution (i.e. CDF) of a policy's performance. To this end, a secondary contribution of this letter is the computation of MES for the UMA and Clopper-Pearson lower confidence bounds, which has applicability beyond robotics. The final contribution of our letter is an open-source implementation of the UMA lower confidence bound and associated MES computation, both of which were absent from existing software. In our experiments, we show the validity of our approach as well as how the it can be used to make judgements about policy generalization in OOD settings. Although we focus our discussion and experiments on BC policies, our approach can be applied to any policy.

The bounds presented in this letter are valid for the environment the samples were drawn from. One can extend the analysis in [26] to understand the sensitivity of the bounds' confidence level to distribution shifts, or to construct bounds which are robust to distribution shift. In our letter, the number of policy rollouts is determined by a budget or by constraints on confidence and tightness. However, our ideas can be extended to use confidence sequences where the number of rollouts is not determined beforehand [37], which may save on testing costs in some cases. Another direction is to evaluate performance during a single policy rollout, rather than multiple i.i.d. rollouts. As a final remark, when performing statistical evaluations, researchers should be forthright in their data collection, assumptions/modeling, and interpretation of results.

Acknowledgment

The authors would like to thank Cheng Chi for guidance on diffusion policies. This article solely reflects the opinions and conclusions of its authors and not any NASA entity.

References

[1] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proc. Robot.: Sci. Syst.*, 2023.
[2] E. Jang et al., "BC-Z: Zero-shot task generalization with robotic imitation learning," in *Proc. Conf. Robot Learn.*, 2022, pp. 991–1002.
[3] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," 2024, *arXiv:2401.02117*.
[4] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. Robot.: Sci. Syst.*, 2023.
[5] P. Florence et al., "Implicit behavioral cloning," in *Proc. Conf. Robot Learn.*, 2022, pp. 158–168.
[6] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: A survey," in *2020 IEEE Symp. Ser. Comput. Intell.*, 2020, pp. 737–744.

[7] S. Levine, C. Finn, T. Darrell, and P. Abbeel, " End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.

[8] 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.

[9] B. Matthiesen, C. Hellings, E. A. Jorswieck, and W. Utschick, "Mixed monotonic programming for fast global optimization," *IEEE Trans. Signal Process.*, vol. 68, pp. 2529–2544, 2020.

[10] 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, pp. 642–669, 1956.

[11] P. Massart, "The tight constant in the Dvoretzky-kiefer-wolfowitz inequality," *Ann. Probability*, vol. 18, pp. 1269–1283, 1990.

[12] J. A. Vincent, H. Nishimura, M. Itkina, and M. Schwager, "Full-distribution generalization bounds for imitation learning policies," in *Proc. 1st Workshop Out- Distrib. Generalization Robot. CoRL 2023*, 2023.

[13] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *Proc. Comput. Aided Verification: 29th Int. Conf.*, 2017, pp. 97–117.

[14] H.-D. Tran et al., "NNV: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems," in *Proc. Int. Conf. Comput. Aided Verification*, 2020, pp. 3–17.

[15] J. A. Vincent and M. Schwager, "Reachable polyhedral marching (RPM): An exact analysis tool for deep-learned control systems," 2022, *arXiv:2210.08339*.

[16] S. M. Katz, A. L. Corso, C. A. Strong, and M. J. Kochenderfer, "Verification of image-based neural network controllers using generative models," *J. Aerosp. Inf. Syst.*, vol. 19, no. 9, pp. 574–584, 2022.

[17] N. Rober et al., "Backward reachability analysis of neural feedback loops: Techniques for linear and nonlinear systems," *IEEE Open J. Control Syst.*, vol. 2, pp. 108–124, 2023.

[18] S. M. Richards, F. Berkenkamp, and A. Krause, "The Lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems," in *Proc. Conf. Robot Learn.*, 2018, pp. 466–476.

[19] H. Dai, B. Landry, L. Yang, M. Pavone, and R. Tedrake, "Lyapunov-stable neural-network control," in *Proc. Robot.: Sci. Syst.*, 2021.

[20] S. Singh, S. M. Richards, V. Sindhwani, J.-J. E. Slotine, and M. Pavone, "Learning stabilizable nonlinear dynamics with contraction-based regularization," *Int. J. Robot. Res.*, vol. 40, no. 10/11, pp. 1123–1150, 2021.

[21] D. Sun, S. Jha, and C. Fan, "Learning certified control using contraction metric," in *Proc. Conf. Robot Learn.*, 2021, pp. 1519–1539.

[22] A. Corso, R. Moss, M. Koren, R. Lee, and M. Kochenderfer, "A survey of algorithms for black-box safety validation of cyber-physical systems," *J. Artif. Intell. Res.*, vol. 72, pp. 377–428, 2021.

[23] P. Akella, M. Ahmadi, and A. D. Ames, "A scenario approach to risk-aware safety-critical system verification," 2022, *arXiv:2203.02595*.

[24] P. Akella, A. Dixit, M. Ahmadi, J. W. Burdick, and A. D. Ames, "Sample-based bounds for coherent risk measures: Applications to policy synthesis and verification," *Artif. Intell.*, 2024, Art. no. 104195.

[25] 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.

[26] J. A. Vincent, A. O. Feldman, and M. Schwager, "Guarantees on robot system performance using stochastic simulation rollouts," *IEEE Trans. Robot.*, 2024.

[27] R. Luo et al., "Sample-efficient safety assurances using conformal prediction," in *Proc. Int. Workshop Algorithmic Foundations Robot.*, 2022, pp. 149–169.

[28] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe planning in dynamic environments using conformal prediction," *IEEE Robot. Automat. Lett.*, vol. 8, no. 8, pp. 5116–5123, Aug. 2023.

[29] 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.

[30] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh, "Explore until confident: Efficient exploration for embodied question answering," in *Proc. Robot.: Sci. Syst.*, 2024.

[31] A. Z. Ren et al., "Robots that ask for help: Uncertainty alignment for large language model planners," in *Proc. Conf. Robot Learn.*, 2023, pp. 661–682.

[32] B. Wu, B. D. Lee, B. Bucher, and N. Matni, "Uncertainty aware deployment of pre-trained task conditioned imitation learning policies," in *Proc. 1st Workshop Out- Distrib. Generalization Robot. CoRL 2023*, 2023.

[33] H. Papadopoulos, V. Vovk, and A. Gammerman, " Regression conformal prediction with nearest neighbours," *J. Artif. Intell. Res.*, vol. 40, pp. 815–840, 2011.

[34] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2021, *arXiv:2107.07511*.

[35] M. P. Fay and S. A. Hunsberger, "Practical valid inferences for the two-sample binomial problem," *Statist. Surv.*, vol. 15, pp. 72–110, 2021, doi: 10.1214/21-SS131.

[36] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, vol. 4. Berlin, Germany: Springer, 2022.

[37] 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.

[38] S. Greenland et al., "Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations," *Eur. J. Epidemiol.*, vol. 31, no. 4, pp. 337–350, 2016.

[39] J. K. Ghosh, "On the relation among shortest confidence intervals of different types," *Calcutta Stat. Assoc. Bull.*, vol. 10, no. 4, pp. 147–152, 1961.

[40] J. W. Pratt, "Length of confidence intervals," *J. Amer. Stat. Assoc.*, vol. 56, no. 295, pp. 549–567, 1961.

[41] M. D. d. Edwardes, "The evaluation of confidence sets with application to binomial intervals," *Statistica Sinica*, vol. 8, pp. 393–409, 1998.

[42] Z. Birnbaum and F. H. Tingey, "One-sided confidence contours for probability distribution functions," *Ann. Math. Statist.*, vol. 22, pp. 592–596, 1951.

[43] Y. Zhu et al., "Robosuite: A Modular Simulation Framework and Benchmark for Robot Learning," 2020, *arXiv:2009.12293*.

[44] B. Zitkovich et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. 7th Conf. Robot Learn.*, 2023.