

Conformance Testing for Stochastic Cyber-Physical Systems

Xin Qin¹, Navid Hashemi¹, Lars Lindemann¹, and Jyotirmoy V. Deshmukh¹

¹Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, USA
{xinqin, navidhas, llindema, jdeshmukh}@usc.edu

Abstract—Conformance is defined as a measure of distance between the behaviors of two dynamical systems. The notion of conformance can accelerate system design when models of varying fidelities are available on which analysis and control design can be done more efficiently. Ultimately, conformance can capture distance between design models and their real implementations and thus aid in robust system design. In this paper, we are interested in the conformance of stochastic dynamical systems. We argue that probabilistic reasoning over the distribution of distances between model trajectories is a good measure for *stochastic conformance*. Additionally, we propose the *non-conformance risk* to reason about the risk of stochastic systems not being conformant. We show that both notions have the desirable *transference* property, meaning that conformant systems satisfy similar system specifications, i.e., if the first model satisfies a desirable specification, the second model will satisfy (nearly) the same specification. Lastly, we propose how stochastic conformance and the non-conformance risk can be estimated from data using statistical tools such as conformal prediction. We present empirical evaluations of our method on an F-16 aircraft, an autonomous vehicle, a spacecraft, and Dubin's vehicle.

I. INTRODUCTION

Cyber-physical systems (CPS) are usually designed using a model-based design (MBD) paradigm. Here, the designer models the physical parts and the operating environment of the system and then designs the software used for perception, planning, and low-level control. Such closed-loop systems are then rigorously tested against various operating conditions, where the quality of the designed software is evaluated against *model properties* such as formal design specifications (or other kinds of quantitative objectives). Examples of such property-based analysis techniques include requirement falsification [1]–[5], nondeterministic and statistical verification [6]–[13], and risk analysis [14], [15].

MBD is a fundamentally iterative process in which the designer continuously modifies the software to tune performance or increase safety margins, or change plant models to perform design space exploration [16], e.g., using model abstraction or simplification [17]–[19], or to incorporate new data [20]. Any change to the system model, however, requires repeating the property-based analyses as many times as the number of system properties. The fundamental problem that we consider in this paper is that of conformance [21]–[25]. The notion of conformance is defined w.r.t. the input-output behavior of a model. Typically, model inputs include exogenous disturbances or user-inputs to the model, user-controllable design parameters, and initial operating conditions. For a given input u , let $y = S(u)$ denote the observable behavior of the model S .

Furthermore, let $d(y_1, y_2)$ be a metric defined over the space of the model behaviors. For deterministic models, two models S_1, S_2 are said to be δ -conformant if for all inputs u it holds that $d(y_1, y_2) < \delta$ where $y_1 = S_1(u)$ and $y_2 = S_2(u)$ [22], [23], [25]. This notion of deterministic conformance is useful to reason about worst-case differences between models. However, most CPS applications use components that exhibit stochastic behavior; for example, sensors have measurement noise, actuators can have manufacturing variations, and most physical phenomena are inherently stochastic. The central question that this paper considers is: *What is the notion of conformance between two stochastic CPS models?*

There are some challenges in comparing stochastic CPS models; even if two models are repeatedly excited by the same input, the pair of model behaviors that are observed may be different for every such simulation. Thus, the observable behavior of a stochastic model is more accurately characterized by a distribution over the space of trajectories. A possible way to compare two stochastic models is to use measure-theoretic techniques to compare the distance between the trajectory distributions. A number of divergence measures such as the f-divergences, e.g., the Kullback-Leibler divergence and the total variation distance, or the Wasserstein distance may look like candidate tools to compare the trajectory distributions. However, we argue in this paper that a divergence is not the right notion to use to compare stochastic CPS models. There can be two stochastic models whose output trajectories are very close using any trajectory space metric, but the divergence between their trajectory distributions can be infinite. On the other hand, there can be two trajectory distributions with zero divergence for which the distance between trajectories can be arbitrarily far apart.

This raises an interesting question: how do we then compare two stochastic models? In this paper, we argue that probabilistic bounds derived from the distribution of the distances between model trajectories (excited by the same input) gives us a general definition of conformance that has several advantages, as outlined below. We complement this probabilistic viewpoint further and capture the risk that the distribution of the distances between model trajectories is large leveraging risk measures [26].

First, we show that two stochastic systems that are conformant under our definition inherit the property of *transference* [22]. In simple terms, transference is the property that if the first model has certain logical or quantitative properties, then

the second model also satisfies the same (or nearly same) properties. This property brings several benefits. Consider the scenario where probabilistic guarantees that a model has certain quantitative properties have been established after an extensive and large number of simulations. Ordinarily, if there were any changes made to the model, establishing such probabilistic guarantees would require repeating the extensive simulation-based procedure. However, transference allows us to potentially sample from existing simulations for the first model and sample a small number of simulations from the modified model to establish stochastic conformance between the models, thereby allowing us to establish probabilistic guarantees on the *second* model. We demonstrate examples of such transference w.r.t. quantitative properties arising from quantitative semantics of temporal logic specifications and control-theoretic cost functions.

Next, we show how we can efficiently compute these probabilistic bounds using the notion of *conformal prediction* [27], [28] from statistical learning theory. At a high-level, conformal prediction involves computing quantiles of the empirical distribution of non-conformity scores over a validation dataset to obtain prediction intervals at a given confidence threshold.

The contributions of this paper are summarized as follows:

- We define *stochastic conformance* as a probabilistic bound over the distribution of distances between model trajectories. We also define the *non-conformance risk* to detect systems that are at risk of not being conformant.
- We show that both notions have the desirable transference property, meaning that conformant systems satisfy similar system specifications.
- We show how stochastic conformance and the non-conformance risk can be estimated using statistical tools from risk theory and conformal prediction.

II. PROBLEM STATEMENT AND PRELIMINARIES

Consider the probability space $(\Omega, \mathcal{F}, P)$ where Ω is the sample space, $\mathcal{F}$ is a σ -algebra¹ of Ω , and $P : \mathcal{F} \rightarrow [0, 1]$ is a probability measure. In this paper, our goal is to quantify conformance of stochastic systems, i.e., systems whose inputs and outputs form a probability space with an appropriately defined measure. Let the two stochastic systems be denoted by S_1 and S_2 . The inputs and outputs of stochastic systems are *signals*, i.e., functions from a bounded interval of positive reals known as the *time domain* $\mathbb{T} \subseteq \mathbb{R}_{\geq 0}$ to a metric space, e.g., the standard Euclidean metric. Each stochastic system S_i then describes an input-output relation $S_i : \mathcal{U} \times \Omega \rightarrow \mathcal{Y}$ where $\mathcal{U}$ and $\mathcal{Y}$ denote the sets of all input and output signals. We allow input signals² to be stochastic, and we use the notation $U : \mathbb{T} \times \Omega \rightarrow \mathbb{R}^m$ to denote a stochastic input signals.³ Modeling stochastic systems this way provides great flexibility, and S_i

can e.g., describe the motion of stochastic hybrid systems, Markov chains, and stochastic difference equations.

Assume now that we apply the input signal $U : \mathbb{T} \times \Omega \rightarrow \mathbb{R}^m$ to systems S_1 and S_2 , and let the resulting output signals be denoted by $Y_1 : \mathbb{T} \times \Omega \rightarrow \mathbb{R}^n$ and $Y_2 : \mathbb{T} \times \Omega \rightarrow \mathbb{R}^n$, respectively. We assume that the functions S_1 , S_2 , and U are measurable so that the output signals Y_1 and Y_2 are well-defined stochastic signals. One can hence think of Y_1 and Y_2 to be drawn from the distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, which are functions of the probability space $(\Omega, \mathcal{F}, P)$ as well as the functions S_1 , S_2 , and U . In this paper, we make no restricting assumptions on the functions S_1 , S_2 , and U , and consequently we make no assumptions on the distributions $\mathcal{D}_1$ and $\mathcal{D}_2$.

Informal Problem Statement. Let Y_1 and Y_2 be stochastic output signals of the stochastic systems S_1 and S_2 , respectively, under the stochastic input signal U . How can we quantify closeness of the stochastic systems S_1 and S_2 under U ? To answer this question, we will explore different ways of defining system “closeness” of Y_1 and Y_2 , and we will present algorithms to compute these stochastic notions of closeness. A subsequent problem that we consider is related to transference of properties from one system to another system. Particularly, given a signal temporal logic specification, can we infer guarantees about the satisfaction of the specification of one system from another system if the systems are close under a suitable definition of closeness?

A. Distance Metrics and Risk Measures

To define a general framework for quantifying closeness of stochastic systems, we will use i) different signal metrics to capture the distance between individual realizations $y_1 := Y_1(\cdot, \omega)$ and $y_2 := Y_2(\cdot, \omega)$ of the stochastic signals Y_1 and Y_2 where $\omega \in \Omega$ is a single outcome, and ii) probabilistic reasoning and risk measures to capture stochastic conformance and non-conformance, respectively, under these signal metrics.

We first equip the set of output signals $\mathcal{Y}$ with a function $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ that quantifies distance between signals. A natural choice of d is a signal metric that results in a metric space $(\mathcal{Y}, d)$. We use general signal metrics such as the metric induced by the L_p signal norm for $p \geq 1$. Particularly, define $d_p(y_1, y_2) := \left(\int_{\mathbb{T}} \|y_1(t) - y_2(t)\|^p dt \right)^{1/p}$ so that the L_∞ norm can also be expressed as $d_\infty(y_1, y_2) := \sup_{t \in \mathbb{T}} \|y_1(t) - y_2(t)\|$.

It is now easy to see that a signal metric $d(Y_1, Y_2)$ evaluated over the stochastic signals Y_1 and Y_2 results in a distribution over distances between realizations of Y_1 and Y_2 . To reason over properties of $d(Y_1, Y_2)$, we will use probabilistic reasoning but we will also consider risk measures [26] as introduced next.

A risk measure is a function $R : \mathfrak{F}(\Omega, \mathbb{R}) \rightarrow \mathbb{R}$ that maps the set of real-valued random variables $\mathfrak{F}(\Omega, \mathbb{R})$ to the real numbers. Typically, the input of R indicates a cost. There exist various risk measures that capture different characteristic of the distribution of the cost random variable, such as the mean or the variance. However, we are particularly interested in tail risk measures that capture the right tail of the cost distribution, i.e., the potentially rare but costly outcomes.

¹A σ -algebra on a set Ω is a nonempty collection of subsets of Ω closed under complement, countable unions, and countable intersections.

²Probability spaces over signals are defined by standard notions of cylinder sets [6].

³We will instead of the probability measure P , defined over $(\Omega, \mathcal{F})$, use more generally the notation Prob to be independent of the underlying probability space that we induce, e.g., as a result of transformations via U .

In this paper, we particularly consider the value-at-risk VaR_β and the conditional value-at-risk $CVaR_\beta$ at risk level $\beta \in (0, 1)$. The VaR_β of a random variable $Z : \Omega \rightarrow \mathbb{R}$ is defined as

$$VaR_\beta(Z) := \inf\{\alpha \in \mathbb{R} \mid \text{Prob}(Z \leq \alpha) \geq \beta\},$$

i.e., $VaR_\beta(Z)$ captures the $1 - \beta$ quantile of the distribution of Z from the right. Note that there is an obvious connection between value-at-risk and chance constraints, i.e., it holds that $\text{Prob}(Z \leq \alpha) \geq \beta$ is equivalent to $VaR_\beta(Z) \leq \alpha$. The $CVaR_\beta$ of Z , on the other hand, is defined as

$$CVaR_\beta(Z) := \inf_{\alpha \in \mathbb{R}} (\alpha + (1 - \beta)^{-1} E([Z - \alpha]^+))$$

where $[Z - \alpha]^+ := \max(Z - \alpha, 0)$ and $E(\cdot)$ indicates the expected value. When the function $\text{Prob}(Z \leq \alpha)$ is continuous (in α), it holds that $CVaR_\beta(Z) = E(Z \mid Z \geq VaR_\beta(Z))$, i.e., $CVaR_\beta(Z)$ is the expected value of Z conditioned on the outcomes where Z is greater or equal than $VaR_\beta(Z)$. Finally, note that it holds that $VaR_\beta(Z) \leq CVaR_\beta(Z)$, i.e., $CVaR_\beta$ is more risk sensitive.

For our risk transference results that we present later, we will require that R is *monotone*, *positive homogeneous*, and *subadditive*:

- For two random variables $Z, Z' \in \mathfrak{F}(\Omega, \mathbb{R})$, the risk measure R is monotone if $Z(\omega) \leq Z'(\omega)$ for all $\omega \in \Omega$ implies that $R(Z) \leq R(Z')$.
- For a random variable $Z \in \mathfrak{F}(\Omega, \mathbb{R})$, the risk measure R is positive homogeneous if, for any constant $H \geq 0$, it holds that $R(HZ) = HR(Z)$.
- For two random variables $Z, Z' \in \mathfrak{F}(\Omega, \mathbb{R})$, the risk measure R is subadditive if $R(Z + Z') \leq R(Z) + R(Z')$.

We remark that the VaR_β and the $CVaR_\beta$ satisfies all three properties [26].

B. System specifications

To express specifications, we use Signal Temporal Logic (STL). Let $y : \mathbb{T} \rightarrow \mathbb{R}^n$ be a deterministic signal, e.g., a realization of the stochastic signal Y . The atomic elements of STL are predicates that are functions $\mu : \mathbb{R}^n \rightarrow \{\text{True}, \text{False}\}$. For convenience, the predicate μ is often defined via a predicate function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ as $\mu(y(t)) := \text{True}$ if $h(y(t)) \geq 0$ and $\mu(y(t)) := \text{False}$ otherwise. The syntax of STL is recursively defined as

$$\phi ::= \text{True} \mid \mu \mid \neg\phi' \mid \phi' \wedge \phi'' \mid \phi' U_I \phi'' \quad (1)$$

where ϕ' and ϕ'' are STL formulas. The Boolean operators $\neg$ and $\wedge$ encode negations (“not”) and conjunctions (“and”), respectively. The until operator $\phi' U_I \phi''$ encodes that ϕ' has to be true until ϕ'' becomes true at some future time within the time interval $I \subseteq \mathbb{R}_{\geq 0}$. We derive the operators for disjunction ($\phi' \vee \phi'' := \neg(\neg\phi' \wedge \neg\phi'')$), eventually ($F_I \phi := \top U_I \phi$), and always ($G_I \phi := \neg F_I \neg \phi$).

To determine if a signal y satisfies an STL formula ϕ that is imposed at time t , we can define the semantics as a relation $\models$, i.e., $(y, t) \models \phi$ means that ϕ is satisfied. While the STL semantics are fairly standard [29], we recall them in Appendix A in [30]. Additionally, we can define robust

semantics $\rho^\phi(y, t) \in \mathbb{R}$ that indicate how robustly the formula ϕ is satisfied or violated [31], [32], see Appendix A in [30]. Larger and positive values of $\rho^\phi(y, t)$ hence indicate that the specification is satisfied more robustly. Importantly, it holds that $(y, t) \models \phi$ if $\rho^\phi(y, t) > 0$.

III. CONFORMANCE FOR STOCHASTIC INPUT-OUTPUT SYSTEMS

Our goal is now to quantify closeness of two stochastic systems S_1 and S_2 under the input U . We present our definitions for stochastic conformance and non-conformance risk upfront, and provide motivation for these afterwards.

Definition 1. Let $U : \mathbb{T} \times \Omega \rightarrow \mathbb{R}^m$ be a stochastic input signal, $S_1, S_2 : \mathcal{U} \times \Omega \rightarrow \mathcal{Y}$ be stochastic systems, and $Y_1, Y_2 : \mathbb{T} \times \Omega \rightarrow \mathbb{R}^n$ be stochastic output signals with $Y_1 := S_1(U, \cdot)$ and $Y_2 := S_2(U, \cdot)$. Further, let $\epsilon \in \mathbb{R}$ be a conformance threshold, $\delta \in (0, 1)$ be a failure probability, and $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a signal metric. Then, we say that the systems S_1 and S_2 under the input U are (ϵ, δ) -conformant if

$$\text{Prob}(d(Y_1, Y_2) \leq \epsilon) \geq 1 - \delta. \quad (2)$$

Additionally, let $R : \mathfrak{F}(\Omega, \mathbb{R}) \rightarrow \mathbb{R}$ be a risk measure and $r \in \mathbb{R}$ be a risk threshold. Then, S_1 and S_2 under the input U are at risk of being r -non-conformant if

$$R(d(Y_1, Y_2)) > r. \quad (3)$$

Eq. (2) is referred to as stochastic conformance and Eq. (3) as non-conformance risk. Let us now motivate and discuss these two definitions. While the definition of conformance in equation (2) appears natural at first sight, there are at least two competing ways of defining stochastic conformance. First, as Y_1 and Y_2 are distributions, it would be possible to define conformance as $D(Y_1, Y_2)$ where D is a distance function that measures the difference between two distributions, such as a divergence (Kullback–Leibler or f -divergence). However, our definition provides an intuitive interpretation in the signal space where system specifications are typically defined, while it is usually difficult to provide such an interpretation for divergences between distributions. Additionally, the divergence between Y_1 and Y_2 may be unbounded (or zero) even when equation (2) holds (does not hold).

Proposition 1. There exist stochastic systems S_1 and S_2 and distance metrics d where equation (2): i) is satisfied for $\epsilon > 0$ and $\delta = 0$, i.e., w.p. 1, but where the divergence between the systems is unbounded, and ii) is not satisfied for any given $\epsilon > 0$ and $\delta \in (0, 1)$, but where the divergence between the systems is zero.

Proof. Let us first prove i). For simplicity, consider systems S_1 and S_2 where the stochastic input and output signals are defined over the time domain $\mathbb{T} := \{t_0, \dots, t_T\}$. Further, for all $t \in \mathbb{T}$ let $y_1(t) := 0$ and $y_2(t) := \epsilon$. Clearly, equation (2) is satisfied, e.g., for d_∞ . The distributions D_1 and D_2 (joint distributions of $Y_1(t)$ and $Y_2(t)$, respectively) are Dirac distributions centered at 0 and ϵ , respectively. The Kullback–Leibler divergence between these two distributions is ∞ [33].

Let us now prove ii). Let $\mathbb{T}$ consist of a single time point for simplicity so that Y_1 and Y_2 are random variables defined over a sample space $\mathbb{R}$. Let D_1 and D_2 be the *same* uniform distribution over $[0, a]$. Clearly, the divergence between D_1 and D_1 is zero. We know that the distribution of $Y := Y_1 - Y_2$ has support on $[-a, a]$, and that the probability density function of Y is $p(y) := 1/a - |y|/a^2$. We can now compute $\text{Prob}(|y| \leq \epsilon) = 2\epsilon/a - \epsilon^2/a^2$. Given $\epsilon > 0$ and $\delta \in (0, 1)$, we pick an $\bar{\delta} \in (0, 1)$ such that $\bar{\delta} > \delta$. We then solve the quadratic equation $2\epsilon/a - \epsilon^2/a^2 = 1 - \bar{\delta}$ subject to the constraint that $\epsilon \leq a$. Consequently, we find that $a \geq \epsilon/(1 - \bar{\delta})(1 + \sqrt{\bar{\delta}})$ results in $\text{Prob}(|y| \leq \epsilon) < 1 - \delta$ so that (2) is not satisfied. $\square$

Another way of defining stochastic conformance was presented in [34] where the authors consider a task-specific definition of stochastic conformance where satisfaction probabilities are required to be approximately equal. In other words, two stochastic systems are called c -approximately probabilistically conformant if $|\text{Prob}((Y_1, \tau) \models \phi) - \text{Prob}((Y_2, \tau) \models \phi)| \leq c$. In this definition, it may happen that two systems are c -approximately probabilistically conformant for a small value of c , while the systems produce completely different behaviors and individual realizations y_1 and y_2 are vastly different. Additionally to not being task specific, our definition covers the risk of being r -non-conformant in equation (3).

Finally, we would like to remark that the definition of conformance in equation (2) is related to the definition of non-conformance risk in equation (3). In fact, when the risk measure R is the value-at-risk VaR_β , then we know that

$$\text{VaR}_\beta(d(Y_1, Y_2)) > r \Leftrightarrow \text{Prob}(d(Y_1, Y_2) \leq r) < \beta$$

since $\text{VaR}_\beta(d(Y_1, Y_2)) \leq r$ is equivalent to $\text{Prob}(d(Y_1, Y_2) \leq r) \geq \beta$ according to Section II. Consequently, if $\beta := 1 - \delta$ and $r := \epsilon$ then $\text{VaR}_\beta(d(Y_1, Y_2)) > r$ implies that the systems S_1 and S_2 under U are not (ϵ, δ) -conformant.

The notion of conformance in Definition 1 is useful when the input U describes internal inputs such as system parameters (an unknown mass), exogeneous disturbances from known sources, or initial system conditions. In other words, the distribution U is known, making U a *known unknown*. However, in case of external inputs that could be manipulated (e.g. user inputs that represent rare malicious attacks), the input U may be unknown, making U an *unknown unknown*. We therefore provide an alternative definition of conformance.

Definition 2. Let $U \in \mathcal{U}$ be an unknown deterministic input signal, $S_1, S_2 : \mathcal{U} \times \Omega \rightarrow \mathcal{Y}$ be stochastic systems, and $Y_1, Y_2 : \mathbb{T} \times \Omega \rightarrow \mathbb{R}$ be stochastic output signals with $Y_1 := S_1(U, \cdot)$ and $Y_2 := S_2(U, \cdot)$. Further, let $\epsilon \in \mathbb{R}$ be a conformance threshold, $\delta \in (0, 1)$ be a failure probability, and $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a signal metric. Then, we say that the systems S_1 and S_2 are (ϵ, δ) -conformant if

$$\text{Prob}\left(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \epsilon\right) \geq 1 - \delta. \quad (4)$$

Additionally, let $R : \mathfrak{F}(\Omega, \mathbb{R}) \rightarrow \mathbb{R}$ be a risk measure and $r \in \mathbb{R}$ be a risk threshold. Then, we say that the systems

S_1 and S_2 under the input U are at risk of being r -non-conformant if

$$R\left(\sup_{U \in \mathcal{U}} d(Y_1, Y_2)\right) > r. \quad (5)$$

Based on this definition, note that it will be inherently more difficult to verify Definition 2 compared to Definition 1 due to the sup-operator.

IV. TRANSFERENCE OF SYSTEM PROPERTIES UNDER CONFORMANCE

We expect two systems S_1 and S_2 that are (ϵ, δ) -conformant in the sense of Definitions 1 and 2 to have similar behaviors with respect to satisfying a given system specification. Therefore, we will define the notion of transference with respect to a performance function $C : \mathcal{Y} \rightarrow \mathbb{R}$ that measures how well a signal $y \in \mathcal{Y}$ satisfies this system specification. Towards capturing similarity between S_1 and S_2 with respect to C , the signal metric d has to be chosen carefully.

Definition 3. Let $d : \mathcal{Y} \times \mathcal{Y}$ be a signal metric and $C : \mathcal{Y} \rightarrow \mathbb{R}$ be a performance function. Then, we say that C is Hölder continuous w.r.t. d if there exists constants $H, \gamma > 0$ such that, for any two signals $y_1, y_2 : \mathbb{T} \rightarrow \mathbb{R}^n$, it holds that

$$|C(y_1) - C(y_2)| \leq H d(y_1, y_2)^\gamma \quad (6)$$

A specific example of the performance function C is the robust semantics ρ^ϕ of an STL specification ϕ . In fact, the robust semantics ρ^ϕ are Hölder continuous w.r.t. the sup-norm d_∞ for constants $H = 1$ and $\gamma = 1$ [35, Lemma 2]. For the convenience of the reader, we state the proof of [35, Lemma 2] with the notation used in this paper in Appendix B in [30]. The robust semantics are also Hölder continuous w.r.t. the Skorokhod metric, see Appendix C in [30]. A commonly used performance function in control is $C(y) = \int_0^T y(t)^\top y(t) dt$, and we note that this choice of C is Hölder continuous w.r.t. d_1 as shown in Appendix D in [30]. Finally, note that the Hölder continuity condition in equation (6) implies that, for any constants $c, \epsilon \in \mathbb{R}$, it holds that

$$(C(y_1) \geq c \wedge d(y_1, y_2) \leq \epsilon) \Rightarrow C(y_2) \geq c - H\epsilon^\gamma. \quad (7)$$

A. Transference under stochastic conformance

With the definition of C being Hölder continuous w.r.t. d , we can now derive a stochastic transference result under stochastic conformance as per Definition 1.

Theorem 1. Let the premises in Definitions 1 and 3 hold. Further, let the systems S_1 and S_2 under the input U be (ϵ, δ) -conformant so that equation (2) holds and let C be Hölder continuous w.r.t. d so that equation (6) holds. Then, it holds that

$$\begin{aligned} \text{Prob}(C(Y_1) \geq c) &\geq 1 - \bar{\delta} \Rightarrow \\ \text{Prob}(C(Y_2) \geq c - H\epsilon^\gamma) &\geq 1 - \delta - \bar{\delta}. \end{aligned}$$

Proof. By assumption, it holds that $\text{Prob}(d(Y_1, Y_1) \leq \epsilon) \geq 1 - \delta$ and $\text{Prob}(C(Y_1) \geq c) \geq 1 - \bar{\delta}$ so that we know that

$\text{Prob}(d(Y_1, Y_1) > \epsilon) \leq \delta$ and $\text{Prob}(C(Y_1) < c) \leq \bar{\delta}$. We can now apply the union bound over these two events so that

$$\text{Prob}(d(Y_1, Y_1) > \epsilon \vee C(Y_1) < c) \leq \delta + \bar{\delta}.$$

From here, we can simply see that

$$\text{Prob}(d(Y_1, Y_1) \leq \epsilon \wedge C(Y_1) \geq c) \geq 1 - \delta - \bar{\delta}.$$

Since C is Hölder continuous w.r.t. d , which implies that equation (7) holds, it is easy to conclude that $\text{Prob}(C(Y_2) \geq c - H\epsilon^\gamma) \geq 1 - \delta - \bar{\delta}$. $\square$

Theorem 1 tells us that i) (ϵ, δ) -conformance of systems S_1 and S_2 under U , and ii) Hölder continuity of the performance function C w.r.t. the metric d enables us to derive a probabilistic lower bound for the performance of system S_2 w.r.t. C from the performance of system S_1 .

We can derive a transference result similar to Theorem 1 when we assume that the systems S_1 and S_2 are (ϵ, δ) -conformant in the sense of Definition 2 instead of Definition 1.

Theorem 2. *Let the premises in Definitions 2 and 3 hold. Further, let the systems S_1 and S_2 be (ϵ, δ) -conformant so that equation (4) holds and let C be Hölder continuous w.r.t. d so that equation (6) holds. Then, it holds that*

$$\begin{aligned} \text{Prob}\left(\inf_{U \in \mathcal{U}} C(Y_1) \geq c\right) &\geq 1 - \bar{\delta} \Rightarrow \\ \text{Prob}\left(\inf_{U \in \mathcal{U}} C(Y_2) \geq c - H\epsilon^\gamma\right) &\geq 1 - \delta - \bar{\delta} \end{aligned}$$

Proof. By assumption, it holds that $\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) \leq \epsilon) \geq 1 - \delta$ and $\text{Prob}(\inf_{U \in \mathcal{U}} C(Y_1) \geq c) \geq 1 - \bar{\delta}$ so that we know that $\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) > \epsilon) \leq \delta$ and $\text{Prob}(\inf_{U \in \mathcal{U}} C(Y_1) < c) \leq \bar{\delta}$. We can now apply the union bound over these two events so that

$$\text{Prob}\left(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) > \epsilon \vee \inf_{U \in \mathcal{U}} C(Y_1) < c\right) \leq \delta + \bar{\delta}.$$

From here, we can simply see that

$$\text{Prob}\left(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) \leq \epsilon \wedge \inf_{U \in \mathcal{U}} C(Y_1) \geq c\right) \geq 1 - \delta - \bar{\delta}.$$

This equation tells us that, for each $U \in \mathcal{U}$, we have

$$\text{Prob}(d(Y_1, Y_1) \leq \epsilon \wedge C(Y_1) \geq c) \geq 1 - \delta - \bar{\delta}.$$

Since C is Hölder continuous w.r.t. d , we know that equation (6) holds for each $U \in \mathcal{U}$. Consequently, we can conclude that $\text{Prob}(\inf_{U \in \mathcal{U}} C(Y_2) \geq c - H\epsilon^\gamma) \geq 1 - \delta - \bar{\delta}$. $\square$

B. Transference under non-conformance risk

On the other hand, by considering the notion of r -non-conformance risk, we expect that two systems S_1 and S_2 that are not at risk of being r -non-conformant have a similar risk of violating a specification. Here, we define the risk of violating a specifications by following ideas from [14] as $R(-C(Y_1))$ and $R(-C(Y_2))$.

Theorem 3. *Let the premises in Definitions 1 and 3 hold. Further, let the systems S_1 and S_2 under the input U not be at risk of being r -non-conformant so that equation (3) does not hold (i.e., $R(d(Y_1, Y_2)) \leq r$) and let C be Hölder*

continuous w.r.t. d with $\gamma = 1$ so that equation (6) holds. If the risk measure R is monotone, positive homogeneous, and subadditive, it holds that

$$R(-C(Y_2)) \leq R(-C(Y_1)) + Hr.$$

Proof. We can derive the following chain of inequalities

$$\begin{aligned} R(-C(Y_2)) &\stackrel{(a)}{\leq} R(-C(Y_1) + Hd(Y_1, Y_1)) \\ &\stackrel{(b)}{\leq} R(-C(Y_1)) + R(Hd(Y_1, Y_1)) \\ &\stackrel{(c)}{=} R(-C(Y_1)) + HR(d(Y_1, Y_1)) \\ &\stackrel{(d)}{\leq} R(-C(Y_1)) + Hr \end{aligned}$$

where (a) follows since C is Hölder continuous w.r.t. d and since R is monotone, (b) follows since R is subadditive, and (c) follows since R is positive homogeneous, while the inequality (d) follows since S_1 and S_2 under U are not at risk of being r -non-conformant, i.e., $R(d(Y_1, Y_2)) \leq r$. $\square$

This result implies that the risk of system S_2 w.r.t. the performance function C is upper bounded by the risk of system S_1 w.r.t. C if the systems S_1 and S_2 are not at risk of being r -non-conformant. We remark that a similar result appeared in our prior work [35]. Here, we present these results in the more general context of conformance and extend the result as we use general performance functions C , which additionally requires R to be positive homogeneous. Additionally, we derive a transference result similar to Theorem 3 when we assume that the systems S_1 and S_2 are not at risk of being r -non-conformant in the sense of Definition 2 instead of Definition 1.

Theorem 4. *Let the premises in Definitions 2 and 3 hold. Further, let the systems S_1 and S_2 not be at risk of being r -non-conformant so that equation (5) does not hold (i.e., $R(\sup_{U \in \mathcal{U}} d(Y_1, Y_2)) \leq r$) and let C be Hölder continuous w.r.t. d with $\gamma = 1$ so that equation (6) holds. If the risk measure R is monotone, positive homogeneous, and subadditive, it holds that*

$$R(-\inf_{U \in \mathcal{U}} C(Y_2)) \leq R(-\inf_{U \in \mathcal{U}} C(Y_1)) + Hr.$$

Proof. We can derive the following chain of inequalities

$$\begin{aligned} R(-\inf_{U \in \mathcal{U}} C(Y_2)) &\stackrel{(a)}{\leq} R(-\inf_{U \in \mathcal{U}} C(Y_1) + H \sup_{U \in \mathcal{U}} d(Y_1, Y_1)) \\ &\stackrel{(b)}{\leq} R(-\inf_{U \in \mathcal{U}} C(Y_1)) + R(H \sup_{U \in \mathcal{U}} d(Y_1, Y_1)) \\ &\stackrel{(c)}{=} R(-\inf_{U \in \mathcal{U}} C(Y_1)) + HR(\sup_{U \in \mathcal{U}} d(Y_1, Y_1)) \\ &\stackrel{(d)}{\leq} R(-\inf_{U \in \mathcal{U}} C(Y_1)) + Hr \end{aligned}$$

where (a) follows since $-\inf_{U \in \mathcal{U}} C(Y_2) = \sup_{U \in \mathcal{U}} -C(Y_2)$, since C is Hölder continuous w.r.t. d , and since R is monotone, (b) follows since R is subadditive, and (c) follows since R is positive homogeneous. The inequality (d) follows since S_1 and S_2 are not at risk of being r -non-conformant, i.e., $\sup_{U \in \mathcal{U}} R(d(Y_1, Y_2)) \leq r$. $\square$

V. STATISTICAL ESTIMATION OF STOCHASTIC CONFORMANCE

We propose algorithms to compute stochastic conformance and the non-conformance risk. In practice, note that one will be limited to discrete-time stochastic systems to apply these algorithms.

A. Estimating stochastic conformance

To estimate stochastic conformance, we use conformal prediction which is a statistical tool introduced in [28], [36] to obtain valid uncertainty regions for complex prediction models without making assumptions on the underlying distribution or the prediction model [27], [37]–[40]. Let $Z, Z^{(1)}, \dots, Z^{(k)}$ be $k + 1$ independent and identically distributed random variables modeling a quantity known as the *nonconformity score*. Our goal is to obtain an uncertainty region for Z based on $Z^{(1)}, \dots, Z^{(k)}$, i.e., the random variable Z should be contained within the uncertainty region with high probability. Formally, given a failure probability $\delta \in (0, 1)$, we want to construct a valid uncertainty region over Z (defined in terms of a value $\bar{Z}$) that depends on $Z^{(1)}, \dots, Z^{(k)}$ such that $\text{Prob}(Z \leq \bar{Z}) \geq 1 - \delta$.

By a surprisingly simple quantile argument, see [27, Lemma 1], one can obtain $\bar{Z}$ to be the $(1 - \delta)$ th quantile of the empirical distribution of the values $Z^{(1)}, \dots, Z^{(k)}$ and ∞ . By assuming that $Z^{(1)}, \dots, Z^{(k)}$ are sorted in non-decreasing order, and by adding $Z^{(k+1)} := \infty$, we can equivalently obtain $\bar{Z} := Z^{(p)}$ where $p := \lceil (k + 1)(1 - \delta) \rceil$ with $\lceil \cdot \rceil$ being the ceiling function.

We can now use conformal prediction to estimate stochastic conformance as defined in Definition 1 by setting $Z := d(Y_1, Y_2)$. We therefore assume that we have access to a calibration dataset D_{cal} that consists of realizations $y_1^{(i)}$ and $y_2^{(i)}$ from the stochastic signals $Y_1 \sim \mathcal{D}_1$ and $Y_2 \sim \mathcal{D}_2$, respectively.

Theorem 5. *Let the premises of Definition 1 hold and D_{cal} be a calibration dataset with datapoints $(y_1^{(i)}, y_2^{(i)})$ drawn from $\mathcal{D}_1 \times \mathcal{D}_2$. Further, define $Z^{(i)} := d(y_1^{(i)}, y_2^{(i)})$ for all $i \in \{1, \dots, |D_{\text{cal}}|\}$ and $Z^{(|D_{\text{cal}}|+1)} := \infty$, and assume that the $Z^{(i)}$ are sorted in non-decreasing order. Then, it holds that $\text{Prob}(d(Y_1, Y_2) \leq \bar{Z}) \geq 1 - \delta$ with $\bar{Z}$ defined as $\bar{Z} := Z^{(p)}$ where $p := \lceil (|D_{\text{cal}}| + 1)(1 - \delta) \rceil$. Thus, the systems S_1 and S_2 under the input U are (ϵ, δ) -conformant if $\bar{Z} \leq \epsilon$.*

We see that checking stochastic conformance as defined in Definition 1 is computationally simple when we have a calibration dataset D_{cal} . Checking stochastic conformance as defined in Definition 2, however, is more difficult due to the existence of the sup-operator. To compute this notion of conformance, we make two assumptions: i) the set $\mathcal{U}$ is compact, and ii) for every realization $\omega \in \Omega$, the function $d(Y_1(\cdot, \omega), Y_2(\cdot, \omega))$ is Lipschitz continuous with Lipschitz constant L . While knowledge of the Lipschitz constant L would presume knowledge about the closeness of the systems S_1 and S_2 , it would only provide a conservative over-approximation. We will, however, not need to know the Lipschitz constant L and estimate L instead along with probabilistic guarantees.

Algorithm 1 Conformance Estimation as per Definition 2

Input: Failure probability $\delta \in (0, 1)$ and grid size $\kappa > 0$
Output: $\bar{Z}$ such that $\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \bar{Z} + L\kappa) \geq 1 - \delta$

- 1: Construct κ -net $\bar{\mathcal{U}}$ of $\mathcal{U}$
- 2: **for** $\bar{U} \in \bar{\mathcal{U}}$ **do**
- 3: Obtain calibration set $D_{\text{cal}}^{\bar{U}}$ consisting of realizations $(y_1^{(i)}, y_2^{(i)})$ under $\bar{U}$
- 4: Compute $\bar{Z}_{\bar{U}} := Z^{(p)}$ by applying Theorem 5 but instead using dataset $D_{\text{cal}}^{\bar{U}}$
- 5: $\bar{Z} := \max_{\bar{U} \in \bar{\mathcal{U}}} \bar{Z}_{\bar{U}}$

Our approach is summarized in Algorithms 1 and 2. Algorithm 1 computes $\bar{Z}$ such that $\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \bar{Z} + L\kappa) \geq 1 - \delta$ when L is known and where κ is a gridding parameter, while Algorithm 2 estimates the Lipschitz constant. We present a description of these algorithms upfront and state their theoretical guarantees afterwards.

In line 1 of Algorithm 1, we construct a κ -net $\bar{\mathcal{U}}$ of $\mathcal{U}$, i.e., we construct a finite set $\bar{\mathcal{U}}$ so that for each $U \in \mathcal{U}$ there exists a $\bar{U} \in \bar{\mathcal{U}}$ such that $\bar{d}(U, \bar{U}) \leq \kappa$ where $\bar{d} : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ is a metric. For this purpose, simple gridding strategies can be used as long as the set $\mathcal{U}$ has a convenient representation. Alternatively, randomized algorithms can be used that sample from $\mathcal{U}$ [41]. In lines 2-4, we apply Theorem 5 for each element $\bar{U} \in \bar{\mathcal{U}}$. Therefore, we obtain realizations $(y_1^{(i)}, y_2^{(i)})$ from $\mathcal{D}_1 \times \mathcal{D}_2$ under $\bar{U}$ (line 3). We then compute $\bar{Z}_{\bar{U}}$ so that $\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot)) \leq \bar{Z}_{\bar{U}}) \geq 1 - \delta$ (line 4). Finally, we set $\bar{Z} := \max_{\bar{U} \in \bar{\mathcal{U}}} \bar{Z}_{\bar{U}}$ (line 5).

In Algorithm 2, we compute $\bar{L}$ such that $\text{Prob}(L \leq \bar{L}) \geq 1 - \delta_L$. We uniformly sample control inputs (U', U'') (line 2), obtain realizations (y_1', y_2') from $\mathcal{D}_1 \times \mathcal{D}_2$ under U' and realizations (y_1'', y_2'') from $\mathcal{D}_1 \times \mathcal{D}_2$ under U'' (line 3), and compute the non-conformity score $L^{(i)}$ (line 4). In line 5, we obtain an estimate $\bar{L}$ of the Lipschitz constant L that holds with a probability of $1 - \delta_L$ over the randomness introduced in Algorithm 1.

Theorem 6. *Let the premises of Definition 2 hold. If the Lipschitz constant L of $d(Y_1(\cdot, \omega), Y_2(\cdot, \omega))$ is known uniformly over $\omega \in \Omega$, then, for a gridding parameter $\kappa > 0$, the output $\bar{Z}$ of Algorithm 1 ensures that*

$$\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \bar{Z} + L\kappa) \geq 1 - \delta$$

Thus, the systems S_1 and S_2 are (ϵ, δ) -conformant if $\bar{Z} + L\kappa \leq \epsilon$. Otherwise, let $\delta_L \in (0, 1)$ be a failure probability, then the output $\bar{L}$ of Algorithm 2 ensures that

$$\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \bar{Z} + \bar{L}\kappa) \geq 1 - \delta - \delta_L$$

where Prob is defined over the randomness introduced in Algorithm 2.

Proof. From line 4 of Algorithm 1 we know that $\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot)) \leq \bar{Z}_{\bar{U}}) \geq 1 - \delta$ for each $\bar{U} \in \bar{\mathcal{U}}$.

Algorithm 2 Lipschitz Constant Estimation of L

Input: Failure probabilities $\delta_L \in (0, 1)$, grid size $\kappa > 0$, calibration size $K_L > 0$

Output: $\bar{L}$ such that $\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \bar{L} + \bar{L}\kappa) \geq 1 - \delta - \delta_L$

- 1: **for** i from 1 to K_L **do**
- 2: Sample (U', U'') uniformly from $\mathcal{U} \times \mathcal{U}$
- 3: Obtain realizations (y'_1, y'_2) under U' and (y''_1, y''_2) under U''
- 4: Compute $L^{(i)} := |d(y'_1, y'_2) - d(y''_1, y''_2)| / \bar{d}(U', U'')$
- 5: Compute $\bar{L} := L^{(p)}$ where $p := \lceil (K_L + 1)(1 - \delta'') \rceil$

Due to Lipschitz continuity, we can conclude that for each $U \in \mathcal{U}$ that is such that $\bar{d}(U, \bar{U}) \leq \kappa$ it holds that

$$\text{Prob}(d(Y_1, Y_2) \leq \bar{Z}_{\bar{U}} + L\kappa) \geq 1 - \delta.$$

Since $\bar{\mathcal{U}}$ is a κ -net of $\mathcal{U}$, it follows that $\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_2) \leq \bar{Z} + L\kappa) \geq 1 - \delta$.

For the second part of the proof, note that from line 5 of Algorithm 2 we know that $\text{Prob}(L \leq \bar{L}) \geq 1 - \delta_L$. We can now union bound over this event and $\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot)) \leq \bar{Z}_{\bar{U}}) \geq 1 - \delta$ so that

$$\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot)) \leq \bar{Z}_{\bar{U}} \wedge L \leq \bar{L}) \geq 1 - \delta - \delta_L.$$

The rest of the proof follows as in the first part. $\square$

B. Estimating non-conformance risk

We next briefly summarize how to estimate the value-at-risk and the conditional value-at-risk following standard results such as from [14], [42] and [43], respectively.

Proposition 2. *Let the premises of Definition 1 hold and D_{cal} be a calibration dataset with datapoints $(y_1^{(i)}, y_2^{(i)})$ drawn from $\mathcal{D}_1 \times \mathcal{D}_2$. Let $\beta \in (0, 1)$ be a risk level and $\gamma \in (0, 1)$ be a failure threshold. Define $Z^{(i)} := d(y_1^{(i)}, y_2^{(i)})$ for each $i \in \{1, \dots, |D_{cal}|\}$ and assume that $\text{Prob}(Z \leq \alpha)$ is continuous in α . Then,*

$$\text{Prob}(\underline{VaR}_\beta \leq VaR_\beta(d(Y_1, Y_2)) \leq \overline{VaR}_\beta) \geq 1 - \gamma.$$

where $\underline{VaR}_\beta := \inf \left\{ \alpha \in \mathbb{R} \mid \widehat{\text{Prob}}(Z \leq \alpha) - \sqrt{\frac{\ln(2/\gamma)}{2|D_{cal}|}} \geq \beta \right\}$ and $\overline{VaR}_\beta := \inf \left\{ \alpha \in \mathbb{R} \mid \widehat{\text{Prob}}(Z \leq \alpha) + \sqrt{\frac{\ln(2/\gamma)}{2|D_{cal}|}} \geq \beta \right\}$ with the empirical cumulative distribution function $\widehat{\text{Prob}}(Z \leq \alpha) := \frac{1}{|D_{cal}|} \sum_{i=1}^{|D_{cal}|} \mathbb{I}(Z^{(i)} \leq \alpha)$ and the indicator function $\mathbb{I}$.

For estimating $CVaR_\beta(Z)$, we assume that the random variable $d(Y_1, Y_2)$ has bounded support, i.e., that $\text{Prob}(d(Y_1, Y_2) \in [a, b]) = 1$. Note that $d(Y_1, Y_2)$ is usually bounded from below by $a := 0$ if d is a metric. To obtain an upper bound, we assume that the distance function saturated at b , e.g., by clipping values larger than b to b . In practice, this means that realizations that are far apart already have a large distance and are capped to b .

Proposition 3. *Let the premises of Definition 1 hold and D_{cal} be a calibration dataset with datapoints $(y_1^{(i)}, y_2^{(i)})$ drawn from $\mathcal{D}_1 \times \mathcal{D}_2$. Let $\beta \in (0, 1)$ be a risk level and $\gamma \in (0, 1)$ be a failure threshold. Define $Z^{(i)} := d(y_1^{(i)}, y_2^{(i)})$ for each $i \in \{1, \dots, |D_{cal}|\}$ and assume that $\text{Prob}(d(Y_1, Y_2) \in [a, b]) = 1$. Then, it holds that*

$$\text{Prob}(\underline{CVaR}_\beta \leq CVaR_\beta(d(Y_1, Y_2)) \leq \overline{CVaR}_\beta) \geq 1 - \gamma.$$

where $\overline{CVaR}_\beta := \widehat{CVaR}_\beta + \sqrt{\frac{5 \ln(3/\gamma)}{|D_{cal}|(1-\beta)}}(b - a)$ and $\underline{CVaR}_\beta := \widehat{CVaR}_\beta - \sqrt{\frac{11 \ln(3/\gamma)}{|D_{cal}|(1-\beta)}}(b - a)$ where the empirical estimate of $CVaR_\beta(Z)$ is $\widehat{CVaR}_\beta := \inf_{\alpha \in \mathbb{R}} (\alpha + (|D_{cal}|(1 - \beta))^{-1} \sum_{i=1}^{|D_{cal}|} [Z^i - \alpha]^+)$.

As a consequence of these two lemmas, we know that with a probability of $1 - \gamma$ the systems S_1 and S_2 under the input U are at risk of not being conformant if $\underline{VaR}_\beta \geq \alpha$ or $\overline{CVaR}_\beta \geq \alpha$ based on the risk measure of choice.

VI. CASE STUDIES

We now demonstrate the practicality of stochastic conformance and risk analysis through various case studies. For validation, if we obtain the value $\bar{Z}$ using a conformal prediction procedure for a nonconformity score defined by the random variable Z , i.e., such that $\text{Prob}(Z \leq \bar{Z}) \geq 1 - \delta$. Then, given a test set D_{test} , the validation score is defined as $VS(Z) := |\{z \in D_{test} \mid z \leq \bar{Z}\}| / |D_{test}|$.

A. Dubin's car.

Dubin's car models the motion of a point mass vehicle. The state variables are the x and y position, θ denotes the steering angle and v the longitudinal velocity. While both θ and v are typically assumed to be control inputs, we adapt the case study from [44] where the angular velocity $\omega(t)$ at each time t is assumed to be given so that $\theta(t) := T_s \pi + \sum_{i=1}^t \omega(i) T_s$ where $T_s := 0.1s$. In this example, we assume that $\omega(i) := \frac{\pi}{50T_s}$ for $i \in [1, 25]$, and $\omega(i) := -\frac{\pi}{50T_s}$ for $i \in [26, 50]$. The velocity $v(t)$ is provided by a feedback controller. The dynamics are assumed to have additive white Gaussian noise $\eta^x(t), \eta^y(t) \sim \mathcal{N}(0, 0.005)$. The dynamical equations of motion are as described below:

$$\begin{aligned} x(t+1) &= x(t) + T_s v(t) \cos(\theta(t)) + \eta^x(t) \\ y(t+1) &= y(t) + T_s v(t) \sin(\theta(t)) + \eta^y(t) \end{aligned}$$

The two systems that we compare have two different feedback controllers. The first feedback controller uses the method from [44], [45] and the second controller uses the method from [46]. We plot a set of sampled trajectories in Fig. 1a. This figure also shows the set of initial states $\mathcal{I} := [-1, 0] \times [-1, 0]$. The controller aims to ensure that the system trajectory stays within a series of sets $\mathcal{T}_1$ through $\mathcal{T}_{50}$, the corresponding STL specification is $\phi_{dubin} := \bigwedge_{i=1}^{50} F_{[i-1, i]}([x_i \ y_i] \in \mathcal{T}_i)$. For the experiments that follow, we uniformly sampled initial states from $\mathcal{I}$ and noise η^x, η^y from the described Gaussian distribution.

Effect of calibration set size. In the first experiment, we wish to benchmark the effect of the size of the calibration set D_{cal}

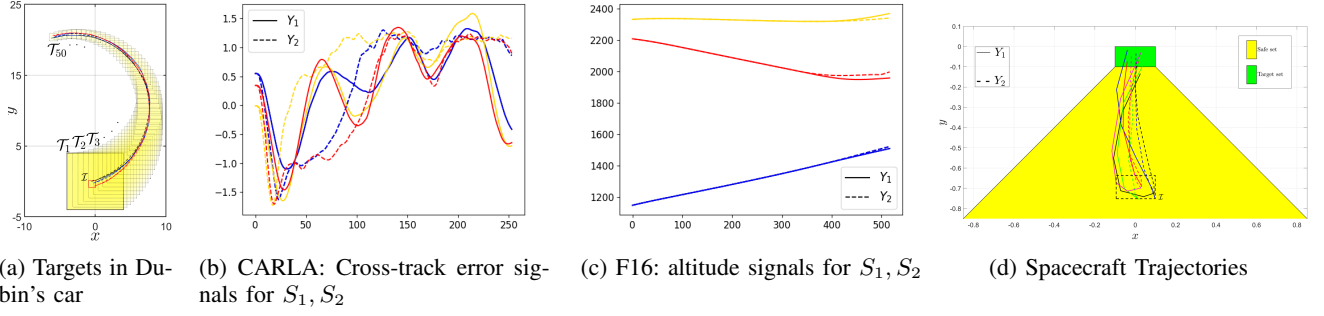


Fig. 1: The solid lines refer to Y_1 and the dashed lines refer to Y_2 ; in each of the displayed plots, the initial condition for each pair of realizations is the same.

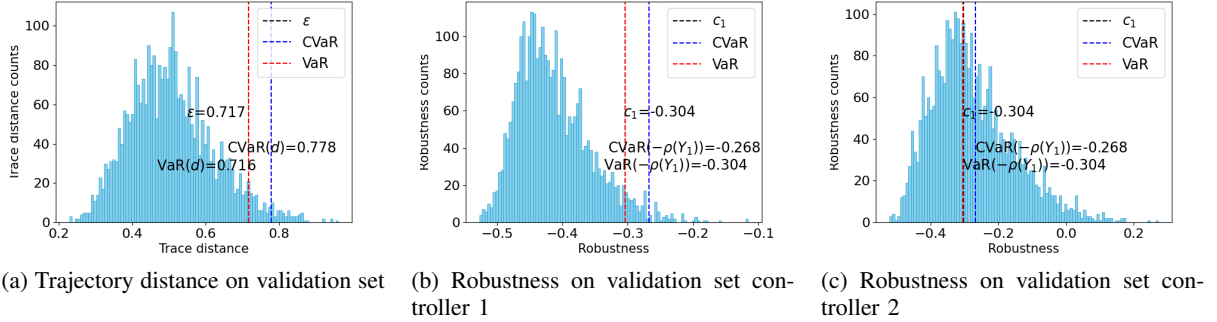


Fig. 2: Distance and robustness histogram for Dubin's car with $\delta = \bar{\delta} = 0.05$. We use $CVaR(d)$ to denote $CVaR(d(Y_1, Y_2))$. The c_1 and ϵ are the values of conformal prediction on the calibration set of $\rho^{\phi_{dubin}}(Y_1)$ and $d_{\infty}(Y_1, Y_2)$.

for various distance metrics. The results are shown in Table I. The table shows that with smaller sizes of the calibration set, we get a more conservative ϵ for d_{∞} (which translates into a higher validation score). The VaR is almost identical to the value of ϵ at larger D_{cal} sizes. We note that the $CVaR$ values change with the value of VaR . A similar trend can be observed the Skorokhod distance and the L_2 -metric.

Empirical evaluation of transference. We empirically demonstrate that Theorem 1 holds. We use $C(Y) = \rho^{\phi_{dubin}}(Y)$, i.e., the robust semantics w.r.t. the property ϕ_{dubin} , and the L_{∞} signal metric d_{∞} . The results are shown in Table II. We can see that the predicted upper bound for the robustness of realizations of Y_2 w.r.t. ϕ_{dubin} is negative ($c_1 - \epsilon$), so it is not possible to conclude that the second system satisfies ϕ_{dubin} with probability greater than $1 - \delta - \bar{\delta}$. However, we note that c_2 is indeed greater than the bound ($c_1 - \epsilon$). Similarly, we show that Theorem 3 is also empirically validated by computing the $CVaR$ values for the first system and the risk measure on $d_{\infty}(Y_1, Y_2)$. We show the empirical distributions of $d_{\infty}(Y_1, Y_2)$, and $\rho^{\phi_{dubin}}(Y_i)$ for $i = 1, 2$ in Figure 2.

Empirical evaluation of Theorem 2. We next apply Algorithms 1 and 2 to this case study. We grid the initial set of states evenly into 25 cells with a grid size of $\kappa = 0.02$. We sample 650 trajectories on each cell to obtain their calibration sets. Algorithm 2 gives $\bar{Z} = 0.7562$ and $L\kappa = 0.0687$, giving $\bar{Z} + L\kappa = 0.8249$. We then evaluate on two test sets of unseen initial conditions with $|D_{test}| = 1000, 2500$. The success rate

on the test sets are 0.9996 and 1.0, with the goal success rate being 0.9. The experiments demonstrate the effectiveness of Theorem 6.

B. F-16 aircraft.

The F-16 aircraft control system from [47] uses a 13-dimensional non-linear plant model based on a 6 d.o.f. airplane model, and its dynamics describe force equations, moments, kinematics, and engine behavior. We alter the original system S_1 from [47] to a modified version S_2 by changing the controller gains. We evaluate the performance of the two systems on the ground collision avoidance scenario with the specification $\phi_{gcas} := G_{[0,T]}(h \geq 1000)$ where T is the mission time and h is the altitude. For data collection, we perform uniform sampling of the initial states. We assume that the x-center of gravity (xcg) of the aircraft is a stochastic parameter with uniform distribution on $[0, 0.8]$. We obtain a calibration set D_{cal} of size 1000 by uniform sampling of the initial states and the xcg parameter. We separately sample 3000 signals for D_{test} . The results of transference and risk estimates are shown in Table III.

C. Autonomous Driving using the CARLA simulator.

CARLA is a high-fidelity simulator for testing of autonomous driving systems [48]. We consider two learning-based lane-keeping controllers from [14], one being an imitation learning controller (S_1) and another being a learned

Distance	$ D_{cal} $	ϵ	$d(Y_1, Y_2)$		
Metric			$VS(d(Y_1, Y_2))$	$Var(d(Y_1, Y_2))$	$CVaR(d(Y_1, Y_2))$
d_∞	50	0.7825	0.987	0.7183	0.7947
	1000	0.7163	0.956	0.7148	0.7647
	2000	0.7122	0.952	0.712	0.7814
	3000	0.7118	0.952	0.7117	0.7862
d_{sk} (Skorokhod Distance)	50	0.6723	0.953	0.6517	0.7181
	1000	0.6722	0.972	0.6711	0.7156
	2000	0.6645	0.96	0.6639	0.7106
	3000	0.6619	0.952	0.6613	0.7079
d_2	50	2.6086	0.937	2.503	2.612
	1000	2.7339	0.96	2.732	3.048
	2000	2.7071	0.944	2.706	3.044
	3000	2.7238	0.955	2.722	3.0929

TABLE I: Effect of calibration set size on the validation score and risk measures. The size of the test set, i.e., $|D_{test}|$, is 1000. We use the conformal prediction procedure from Section V to obtain ϵ as defined in Definition 1 for $\delta = 0.05$.

	$ D_{cal} $	c_1	ϵ	$VS(\rho_1)$	$VS(d_\infty)$	c_2	Thm 1	$CVaR$			Thm.3
							valid?	$-d_\infty$	$-\rho_1$	$-\rho_2$	valid?
$\delta = 0.2,$	100	0.31	0.59	0.95	0.76	0.21	Y	0.90	-0.28	0.00	Y
$\bar{\delta} = 0.05$	3K	0.30	0.60	0.95	0.79	0.20	Y	0.93	-0.27	0.03	Y
$\delta = 0.1,$	1K	0.30	0.67	0.96	0.92	0.15	Y	0.79	-0.27	0.02	Y
$\bar{\delta} = 0.05$	3K	0.30	0.66	0.95	0.91	0.15	Y	0.81	-0.27	0.03	Y
$\delta = 0.05,$	2K	0.31	0.71	0.94	0.95	0.11	Y	0.78	-0.27	0.02	Y
$\bar{\delta} = 0.05$	3K	0.30	0.71	0.95	0.95	0.11	Y	0.79	-0.27	0.03	Y

TABLE II: Empirical evaluation of transference. Let ρ_i be short-hand for $\rho^{\varphi_{dubin}}(Y_i)$ for $i = 1, 2$, and d_∞ be short-hand for $d_\infty(Y_1, Y_2)$. Using Theorem 5, we show $\text{Prob}(\rho_1 \geq c_1) > 1 - \bar{\delta}$, and $\text{Prob}(d_\infty \leq \epsilon) > 1 - \delta$. The validity scores for each guarantee on a test set D_{test} with 1000 samples are shown. The value c_2 is obtained using Theorem 5 on ρ_2 and observe that it exceeds $c_1 - \epsilon$, validating Theorem 1. Similarly, we report the $CVaR$ values for $-\rho_1$ and d_∞ , and $CVaR(-\rho_1) + CVaR(d_\infty) \geq CVaR(-\rho_2)$ for all cases, validating Theorem 3.

Case Study	Spec	$ D_{cal} $	$ D_{test} $	$VS(\rho_1)$	$VS(d_\infty)$	ϵ	$Var(d_\infty)$
F-16	ϕ_{gcas}	1K	3K	0.95	0.98	200	200
CARLA	ϕ_{cte}	700	300	0.94	0.96	1.88	1.87
Satellite	ϕ_{sat}	7K	3K	0.96	0.97	0.18	0.18

TABLE III: Transference results for various case studies. We use $\delta = 0.05$ and $\bar{\delta} = 0.05$. As before, ρ_1 is used as short-hand for $\rho^\phi(Y_1)$ for each spec, and d_∞ is used as short-hand for $d_\infty(Y_1, Y_2)$.

barrier function controller (S_2). We obtain 1000 trajectories from each controller during a 180 degree left turn, and we use 700 of them for calibration and 300 for testing. In this data, the initial states (c_e, θ_e) are drawn uniformly from $[-1, 1] \times [-0.4, 0.4]$ where c_e is the deviation from the center of the lane center (cross track error) and θ_e is the orientation error. The STL specification $\phi_{cte} := G(|c_e| \leq 2.25)$ restricts $|c_e|$ to be bounded by 2.25. The results are shown in Table III.

D. Spacecraft Rendezvous

Next, we consider a spacecraft rendezvous problem from [49]. Here, a deputy spacecraft is to rendezvous with a master spacecraft while staying within a line-of-sight cone. The system is a 4D model $s = [x, y, v_x, v_y]^\top$ where $x, y \in \mathbb{R}$ are the relative horizontal and vertical distances between the two spacecrafts and $v_x, v_y \in \mathbb{R}$ are the relative vertical and horizontal velocities. There are two different feedback controllers, using the same control algorithms we used in Dubin's car example (i.e., the controllers from [44], [45] and [46]). The STL specification is a reach-

Case Study	Spec	$ D_{cal} $	$ D_{test} $	δ	$CVaR$		
					d_∞	$-\rho_1$	$-\rho_2$
F-16	ϕ_{gcas}	1K	3K	0.01	200.3	-62.3	-62.3
CARLA	ϕ_{quad}	7K	3K	0.01	2.04	-0.31	0.88
Satellite	ϕ_{sat}	7K	3K	0.01	0.19	0.0	0.08

TABLE IV: Empirical validation of risk transference for all case studies. As before, ρ_i is short-hand for $\rho^\phi(Y_i)$, and d_∞ is short-hand for $d_\infty(Y_1, Y_2)$. Here, we set the risk level $\beta = \delta$ in each case.

avoid specification (visually depicted in Fig. 1d), which requires the system to always stay in the yellow region and eventually reach the target rectangle $\mathcal{T}$ shown: $\phi_{sat} := G_{[1,5]}(y, |y|, |v_x|, |v_y| \leq -|x|, y_{max}, v_{x,max}, v_{y,max}) \wedge F_{[1,5]}^s \in \mathcal{T}$. The set of initial states is $\mathcal{I} = [-0.1, 0.1] \times [-0.1, 0.1]$. The system is assumed to have additive Gaussian process noise with zero mean and a diagonal covariance matrix with variances $10^{-4}, 10^{-4}, 5 \times 10^{-8}, 5 \times 10^{-8}$. We uniformly sample 100 different initial states from $\mathcal{I}$ and 100 noise values sampled from the noise distribution. We divide the dataset into D_{cal} and D_{test} with sizes 7K and 3K respectively. The results are shown in Table III.

Discussion on results for Transference across case studies. We omit the column for Table III that shows the proportion of D_{test} of the realizations of Y_2 for which the bound $c_1 - \epsilon$ exceeds c_2 , where the c_i 's are the conformal bounds on ρ_i 's. For all case studies this ratio was either 1.0 or close to 1.0, establishing the empirical validity of Theorem 1. We also observe that above results show that it is feasible to use stochastic conformance in a control improvisation loop, where we want to change a system controller (perhaps for optimizing a performance objective) while allowing only some degradation on probabilistic safety guarantees.

VII. RELATED WORK

Conformance has found applications in cyber-physical system design [50], [51] as well as in drug testing and other applications [52]–[54]. Our work is inspired by existing works for **conformance of deterministic systems** by which we mean that systems are non-stochastic, see [21], [55] for surveys. The authors in [23]–[25] considered conformance testing between hybrid system. To capture distance between hybrid system trajectories that may exhibit discontinuities, signal metrics were considered that simultaneously quantify distance in space and time, resembling notions of system closeness in the hybrid systems literature [56], [57]. For instance, [23] proposes $(T, J, (\tau, \epsilon))$ -closeness where τ and ϵ capture both timing distortions and state value mismatches, respectively, and where T and J quantify limits on the total time and number of discontinuities, respectively. A stronger notion compared to $(T, J, (\tau, \epsilon))$ -closeness was proposed in [22] by using the Skoroghod metric. The benefit of [22] over the other notion is that it preserves the timing structure. All these works derive transference results with respect to timed linear temporal logic or metric interval temporal logic specifications.

Conformance of stochastic systems has been less studied. The authors in [58] propose precision and recall conformance measures based on the notion of entropy of stochastic automata. The authors in [59] use the Wasserstein distance to quantify distance between two stochastic systems, which is fundamentally different from our approach. (Bi)simulation relations for stochastic systems were studied in [60]–[62]. Such techniques can define behavioral relations for systems [63], [64], and they can be used to transfer verification results between systems [65]. The authors in [66] utilize such behavioral relations to verify RL policies between a concrete and an abstract system. We remark that bisimulations are difficult to compute, see e.g., [67], unlike our approach. Probably closest to our work is [34]. However, in this paper conformance is task specific which allows two systems to be conformant w.r.t. a system specification even when the systems produce completely different trajectories. Additionally, we consider a worst-case notion of conformance where no information about the input that excites both stochastic systems is available.

VIII. CONCLUSION

We studied conformance of stochastic dynamical systems. Particularly, we defined conformance between two stochastic systems as probabilistic bounds over the distribution of distances between model trajectories. Additionally, we proposed the non-conformance risk to reason about the risk of stochastic systems not being conformant. We showed that both notions have the transference property, meaning that conformant systems satisfy similar system specifications. Lastly, we showed how stochastic conformance and the non-conformance risk can be estimated from data using statistical tools such as conformal prediction.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their feedback. The National Science Foundation supported this work through the following grants: CAREER award (SHF-2048094), CNS-1932620, CNS-2039087, FMITF-1837131, CCF-SHF-1932620, the Airbus Institute for Engineering Research, and funding by Toyota R&D and Siemens Corporate Research through the USC Center for Autonomy and AI.

REFERENCES

- [1] E. Bartocci, J. Deshmukh, A. Donzé, G. Fainekos, O. Maler, D. Ničković, and S. Sankaranarayanan, “Specification-based monitoring of cyber-physical systems: a survey on theory, tools and applications,” *Lectures on Runtime Verification: Introductory and Advanced Topics*, pp. 135–175, 2018.

- [2] Q. Thibault, J. Anderson, A. Chandratte, G. Pedrielli, and G. Fainekos, "Psy-taliro: A python toolbox for search-based test generation for cyber-physical systems," in *Formal Methods for Industrial Critical Systems: 26th International Conference, FMICS 2021, Paris, France, August 24–26, 2021, Proceedings* 26. Springer, 2021, pp. 223–231.
- [3] T. Akazaki, S. Liu, Y. Yamagata, Y. Duan, and J. Hao, "Falsification of cyber-physical systems using deep reinforcement learning," in *Formal Methods: 22nd International Symposium, FM 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 15–17, 2018, Proceedings* 22. Springer, 2018, pp. 456–465.
- [4] J. V. Deshmukh and S. Sankaranarayanan, "Formal techniques for verification and testing of cyber-physical systems," *Design Automation of Cyber-Physical Systems*, pp. 69–105, 2019.
- [5] X. Qin, N. Aréchiga, A. Best, and J. Deshmukh, "Automatic testing with reusable adversarial agents," *arXiv preprint arXiv:1910.13645*, 2021.
- [6] C. Baier and J.-P. Katoen, *Principles of model checking*. MIT press, 2008.
- [7] Y. Wang, M. Zarei, B. Bonakdarpour, and M. Pajic, "Statistical verification of hyperproperties for cyber-physical systems," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 18, no. 5s, pp. 1–23, 2019.
- [8] A. Legay, B. Delahaye, and S. Bensalem, "Statistical model checking: An overview," in *Proceedings of the International conference on runtime verification*, St. Julians, Malta, November 2010, pp. 122–135.
- [9] A. Legay and M. Viswanathan, "Statistical model checking: challenges and perspectives," *International Journal on Software Tools for Technology Transfer*, vol. 17, pp. 369–376, 2015.
- [10] G. Agha and K. Palmisano, "A survey of statistical model checking," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 28, no. 1, pp. 1–39, 2018.
- [11] A. Abate, A. Edwards, M. Giacobbe, H. Punchihewa, and D. Roy, "Quantitative verification with neural networks for probabilistic programs and stochastic systems," 2023.
- [12] Y. Zhao and K. Y. Rozier, "Probabilistic model checking for comparative analysis of automated air traffic control systems," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2014, pp. 690–695.
- [13] X. Qin, Y. Xia, A. Zutshi, C. Fan, and J. V. Deshmukh, "Statistical verification of cyber-physical systems using surrogate models and conformal inference," in *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPs)*, 2022, pp. 116–126.
- [14] L. Lindemann, L. Jiang, N. Matni, and G. J. Pappas, "Risk of stochastic systems for temporal logic specifications," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 3, pp. 1–31, 2023.
- [15] 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," *arXiv preprint arXiv:2204.09833*, 2022.
- [16] A. D. Pimentel, "Exploring exploration: A tutorial introduction to embedded systems design space exploration," *IEEE Design & Test*, vol. 34, no. 1, pp. 77–90, 2016.
- [17] W. H. Schilders, H. A. Van der Vorst, and J. Rommes, *Model order reduction: theory, research aspects and applications*. Springer, 2008, vol. 13.
- [18] R. Alur, T. A. Henzinger, G. Lafferriere, and G. J. Pappas, "Discrete abstractions of hybrid systems," *Proceedings of the IEEE*, vol. 88, no. 7, pp. 971–984, 2000.
- [19] C. Belta, B. Yordanov, and E. A. Gol, *Formal methods for discrete-time dynamical systems*. Springer, 2017, vol. 15.
- [20] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *Journal of Intelligent & Robotic Systems*, vol. 86, no. 2, pp. 153–173, 2017.
- [21] H. Roehm, J. Oehlerking, M. Woehrle, and M. Althoff, "Model conformance for cyber-physical systems: A survey," *ACM Transactions on Cyber-Physical Systems*, vol. 3, no. 3, pp. 1–26, 2019.
- [22] J. V. Deshmukh, R. Majumdar, and V. S. Prabh, "Quantifying conformance using the skorokhod metric," in *Computer Aided Verification: 27th International Conference, CAV 2015, San Francisco, CA, USA, July 18–24, 2015, Proceedings, Part II* 27. Springer, 2015, pp. 234–250.
- [23] H. Abbas, B. Hoxha, G. Fainekos, J. V. Deshmukh, J. Kapinski, and K. Ueda, "Conformance testing as falsification for cyber-physical systems," *arXiv preprint arXiv:1401.5200*, 2014.
- [24] H. Abbas, H. Mittelmann, and G. Fainekos, "Formal property verification in a conformance testing framework," in *2014 Twelfth ACM/IEEE Conference on Formal Methods and Models for Codesign (MEMOCODE)*. IEEE, 2014, pp. 155–164.
- [25] H. Y. Abbas, *Test-based falsification and conformance testing for cyber-physical systems*. Arizona State University, 2015.
- [26] A. Majumdar and M. Pavone, "How should a robot assess risk? towards an axiomatic theory of risk in robotics," in *Robotics Research*. Springer, 2020, pp. 75–84.
- [27] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in *Proceedings of the Conference on Neural Information Processing Systems*, vol. 32, Vancouver, Canada, December 2019.
- [28] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.
- [29] O. Maler and D. Nickovic, "Monitoring temporal properties of continuous signals," in *Proceedings of the Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems*, Grenoble, France, September 2004, pp. 152–166.
- [30] X. Qin, N. Hashemi, L. Lindemann, and J. V. Deshmukh, "Conformance testing for stochastic cyber-physical systems," *arXiv preprint arXiv:2308.06474*, 2023.
- [31] A. Donzé and O. Maler, "Robust satisfaction of temporal logic over real-valued signals," in *Proceedings of the International Conference on Formal Modeling and Analysis of Timed Systems*, Klosterneuburg, Austria, September 2010, pp. 92–106.
- [32] G. E. Fainekos and G. J. Pappas, "Robustness of temporal logic specifications for continuous-time signals," *Theoretical Computer Science*, vol. 410, no. 42, pp. 4262–4291, 2009.
- [33] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.
- [34] Y. Wang, M. Zarei, B. Bonakdarpour, and M. Pajic, "Probabilistic conformance for cyber-physical systems," in *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, 2021, pp. 55–66.
- [35] M. Cleaveland, L. Lindemann, R. Ivanov, and G. J. Pappas, "Risk verification of stochastic systems with neural network controllers," *Artificial Intelligence*, vol. 313, p. 103782, 2022.
- [36] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.
- [37] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.
- [38] M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: A unified review of theory and new challenges," *Bernoulli*, vol. 29, no. 1, pp. 1 – 23, 2023.
- [39] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.
- [40] M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi, "Robust validation: Confident predictions even when distributions shift," *arXiv preprint arXiv:2008.04267*, 2020.
- [41] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.
- [42] P. Massart, "The tight constant in the dvoretzky-kiefer-wolfowitz inequality," *The annals of Probability*, pp. 1269–1283, 1990.
- [43] Y. Wang and F. Gao, "Deviation inequalities for an estimator of the conditional value-at-risk," *Operations Research Letters*, vol. 38, no. 3, pp. 236–239, 2010.
- [44] A. P. Vinod and M. M. Oishi, "Affine controller synthesis for stochastic reachability via difference of convex programming," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 7273–7280.
- [45] K. Lesser, M. Oishi, and R. S. Erwin, "Stochastic reachability for control of spacecraft relative motion," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 4705–4712.
- [46] M. P. Vitis and C. J. Tomlin, "On feedback design and risk allocation in chance constrained control," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, 2011, pp. 734–739.
- [47] P. Heidlauf, A. Collins, M. Bolender, and S. Bak, "Verification challenges in f-16 ground collision avoidance and other automated maneuvers," in *ARCH@ ADHS*, 2018, pp. 208–217.
- [48] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the Conference on robot learning*. Mountain View, California: PMLR, November 2017, pp. 1–16.
- [49] A. P. Vinod, J. D. Gleason, and M. M. Oishi, "Sreachtools: a matlab stochastic reachability toolbox," in *Proceedings of the 22nd ACM*

international conference on hybrid systems: computation and control, 2019, pp. 33–38.

- [50] X. Jin, J. V. Deshmukh, J. Kapinski, K. Ueda, and K. Butts, “Benchmarks for model transformations and conformance checking,” in *1st International Workshop on Applied Verification for Continuous and Hybrid Systems (ARCH)*, 2014.
- [51] H. Araujo, G. Carvalho, M. Mohaqeqi, M. R. Mousavi, and A. Sampaio, “Sound conformance testing for cyber-physical systems: Theory and implementation,” *Science of Computer Programming*, vol. 162, pp. 35–54, 2018.
- [52] R. Dimitrova, M. Gazda, M. R. Mousavi, S. Biewer, and H. Hermanns, “Conformance-based doping detection for cyber-physical systems,” in *Formal Techniques for Distributed Objects, Components, and Systems: 40th IFIP WG 6.1 International Conference, FORTE 2020, Held as Part of the 15th International Federated Conference on Distributed Computing Techniques, DisCoTec 2020, Valletta, Malta, June 15–19, 2020, Proceedings 40*. Springer, 2020, pp. 59–77.
- [53] S. Biewer, R. Dimitrova, M. Fries, M. Gazda, T. Heinze, H. Hermanns, and M. R. Mousavi, “Conformance relations and hyperproperties for doping detection in time and space,” *arXiv preprint arXiv:2012.03910*, 2020.
- [54] R. Dimitrova, M. Gazda, M. R. Mousavi, S. Biewer, and H. Hermanns, “Conformance-based doping detection for cyber-physical systems,” in *Formal Techniques for Distributed Objects, Components, and Systems*, A. Gotsman and A. Sokolova, Eds. Cham: Springer International Publishing, 2020, pp. 59–77.
- [55] N. Khakpour and M. R. Mousavi, “Notions of conformance testing for cyber-physical systems: Overview and roadmap,” in *26th International Conference on Concurrency Theory (CONCUR 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.
- [56] R. Goebel, R. G. Sanfelice, and A. R. Teel, “Hybrid dynamical systems,” *IEEE control systems magazine*, vol. 29, no. 2, pp. 28–93, 2009.
- [57] —, “Hybrid dynamical systems: modeling stability, and robustness,” Princeton, NJ, USA, 2012.
- [58] S. J. Leemans and A. Polyvyanyy, “Stochastic-aware conformance checking: An entropy-based approach,” in *Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, Proceedings 32*. Springer, 2020, pp. 217–233.
- [59] S. J. Leemans, A. F. Syring, and W. M. van der Aalst, “Earth movers’ stochastic conformance checking,” in *Business Process Management Forum: BPM Forum 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*. Springer, 2019, pp. 127–143.
- [60] A. A. Julius, A. Girard, and G. J. Pappas, “Approximate bisimulation for a class of stochastic hybrid systems,” in *2006 American Control Conference*. IEEE, 2006, pp. 6–pp.
- [61] A. A. Julius and G. J. Pappas, “Approximations of stochastic hybrid systems,” *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1193–1203, 2009.
- [62] S. Haesaert and S. Soudjani, “Robust dynamic programming for temporal logic control of stochastic systems,” *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2496–2511, 2020.
- [63] G. Bian and A. Abate, “On the relationship between bisimulation and trace equivalence in an approximate probabilistic context,” in *Foundations of Software Science and Computation Structures: 20th International Conference, FOSSACS 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22–29, 2017, Proceedings 20*. Springer, 2017, pp. 321–337.
- [64] A. A. Julius and A. Van Der Schaft, “Bisimulation as congruence in the behavioral setting,” in *Proceedings of the 44th IEEE Conference on Decision and Control*. IEEE, 2005, pp. 814–819.
- [65] K. Zhang and M. Zamani, “Infinite-step opacity of nondeterministic finite transition systems: A bisimulation relation approach,” in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 5615–5619.
- [66] F. Delgrange, A. Nowe, and G. A. Pérez, “Wasserstein auto-encoded mdps: Formal verification of efficiently distilled rl policies with many-sided guarantees,” *arXiv preprint arXiv:2303.12558*, 2023.
- [67] A. Girard and G. J. Pappas, “Approximate bisimulation: A bridge between computer science and control theory,” *European Journal of Control*, vol. 17, no. 5–6, pp. 568–578, 2011.