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, jdeshmuk}@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(y₁, y₂) be a metric defined over the space of the model behaviors. For deterministic models, two models S₁, S₂ are said to be δ-conformant if for all inputs u it holds that d(y₁, y₂) < δ where y₁ = S₁(u) and y₂ = S₂(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

https://doi.org/10.34727/2023/isbn.978-3-85448-060-0_38

This article is licensed under a Creative Commons Attribution 4.0 International License
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 $\Omega$ is the sample space, $\mathcal{F}$ is a $\sigma$-algebra$^1$ of $\Omega$, and $P : \mathcal{F} \to [0, 1]$ 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 $S_i : \mathcal{U} \times \Omega \to \mathcal{Y}$ where $\mathcal{U}$ and $\mathcal{Y}$ denote the sets of all input and output signals. We allow input signals$^2$ to be stochastic, and we use the notation $U : \mathbb{T} \times \Omega \to \mathbb{R}^m$ 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 $U : \mathbb{T} \times \Omega \to \mathbb{R}^m$ to systems $S_1$ and $S_2$, and let the resulting output signals be denoted by $Y_1 : \mathbb{T} \times \Omega \to \mathbb{R}^n$ and $Y_2 : \mathbb{T} \times \Omega \to \mathbb{R}^p$, 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 $D_1$ and $D_2$, 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 $D_1$ and $D_2$.

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 $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 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 $L_{\infty}$ 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 $R : \mathcal{F}(\Omega, \mathbb{R}) \to \mathbb{R}$ that maps the set of real-valued random variables $\mathcal{F}(\Omega, \mathbb{R})$ 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.

1. A $\sigma$-algebra on a set $\Omega$ is a nonempty collection of subsets of $\Omega$ closed under complement, countable unions, and countable intersections.

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

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

295
In this paper, we particularly consider the value-at-risk \(VaR_\beta\) and the conditional value-at-risk \(CVaR_\beta\) at risk level \(\beta \in (0, 1)\). The \(VaR_\beta\) of a random variable \(Z : \Omega \to \mathbb{R}\) is defined as

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

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

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

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 \(\alpha\)), it holds that \(CVaR_\beta(Z) = E(Z|Z \geq VaR_\beta(Z))\), 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 \(Z, Z' \in \mathcal{Z}(\Omega, \mathbb{R})\), the risk measure \(R\) is monotone if \(R(Z(\omega)) \leq R(Z'(\omega))\) for all \(\omega \in \Omega\) implies that \(R(Z) \leq R(Z')\).
- For a random variable \(Z \in \mathcal{Z}(\Omega, \mathbb{R})\), 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 \(Z, Z' \in \mathcal{Z}(\Omega, \mathbb{R})\), the risk measure \(R\) is subadditive if \(R(Z + Z') \leq R(Z) + R(Z')\).

We remark that the \(VaR_\beta\) and the \(CVaR_\beta\) satisfies all three properties [26].

### B. System specifications

To express specifications, we use Signal Temporal Logic (STL). Let \(y : T \to \mathbb{R}^n\) be a deterministic signal, e.g., a realization of the stochastic signal \(Y\). The atomic elements of STL are predicates that are functions \(\mu : \mathbb{R}^n \to \{\text{True}, \text{False}\}\). For convenience, the predicate \(\mu\) is often defined via a predicate function \(h : \mathbb{R}^n \to \mathbb{R}\) 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

\[
\phi := \text{True} \mid \mu \mid \neg \phi \mid \phi \land \phi' \mid \phi'U_1\phi'' \quad (1)
\]

where \(\phi'\) and \(\phi''\) are STL formulas. The Boolean operators \(\neg\) and \(\land\) encode negations (“not’s”) and conjunctions (“and”), respectively. The until operator \(\phi'U_1\phi''\) encodes that \(\phi'\) is to be true until \(\phi''\) becomes true at some future time within the time interval \(I \subseteq \mathbb{R}_{\geq 0}\). We derive the operators for disjunction \((\phi' \lor \phi'') := \neg(\neg\phi' \land \neg\phi'')\), eventually \((F_1\phi := \top U_1\phi)\), and always \((G_1\phi := \neg F_1\neg\phi)\).

To determine if a signal \(y\) satisfies an STL formula \(\phi\) that is imposed at time \(t\), we can define the semantics as a relation \(|=\) i.e., \((y, t) |= \phi\) means that \(\phi\) 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 \(\phi\) 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 \((y, t) |= \phi\) 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 upront, and provide motivation for these afterwards.

**Definition 1.** Let \(U : \mathcal{T} \times \Omega \to \mathbb{R}^m\) be a stochastic input signal, \(S_1, S_2 : \mathcal{U} \times \Omega \to \mathcal{Y}\) be stochastic systems, and \(Y_1, Y_2 : \mathcal{T} \times \Omega \to \mathbb{R}^n\) 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 \(d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}\) be a signal metric. Then, we say that the systems \(S_1\) and \(S_2\) under the input \(U\) are \((\epsilon, \delta)\)-conformant if

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

Additionally, let \(R : \mathcal{Z}(\Omega, \mathbb{R}) \to \mathbb{R}\) 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 \(T := \{t_0, \ldots, t_T\}\). Further, for all \(t \in T\) let \(y_1(t) := 0\) and \(y_2(t) := \epsilon\). Clearly, equation (2) is satisfied, e.g., for \(d_{\infty}\). 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 \(\epsilon\), respectively. The Kullback–Leibler divergence between these two distributions is \(\infty\) [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 $D_2$ 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 $\delta \in (0, 1)$ such that $\delta > \delta$. We then solve the quadratic equation $2\epsilon/a - \epsilon^2/a^2 = 1 - \delta$ subject to the constraint that $\epsilon \leq a$. Consequently, we find that $\epsilon \geq \sqrt{(1-\delta)(1+\sqrt{\delta})}$ results in $\text{Prob}(|y| \leq \epsilon) < 1 - \delta$ so that (2) is not satisfied. \hfill \Box

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 $\text{Prob}(Y_1, \tau) = \phi - \text{Prob}(Y_2, \tau) < |\phi| < c$. 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, not to 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 $VaR$, then we know that

$$VaR_\beta(d(Y_1, Y_2)) > r \iff \text{Prob}(d(Y_1, Y_2) \leq r) < \beta$$

since $VaR_\beta(d(Y_1, Y_2)) \leq r$ 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 $VaR_\beta(d(Y_1, Y_2)) > r$ implies that the systems $S_1$ and $S_2$ under $U$ are not $(\epsilon, \delta)$-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 $\epsilon, \delta \in \mathbb{R}$ 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 $(\epsilon, \delta)$-conformant if

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

Additionally, let $R : \mathcal{F}(\Omega, \mathbb{R}) \rightarrow \mathbb{R}$ 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

$$R\left(\sup_{U \in \mathcal{U}} d(Y_1, Y_2)\right) > r.$$  \hfill (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 $(\epsilon, \delta)$-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 $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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

$$|C(y_1) - C(y_2)| \leq Hd(y_1, y_2)^\gamma.$$  \hfill (6)

A specific example of the performance function $C$ is the robust semantics $\rho^\phi$ of an STL specification $\phi$. In fact, the robust semantics $\rho^\phi$ are Hölder continuous w.r.t. the sup-norm $d_{\sup}$ 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 $C(y) := \int_0^T g(t)^T y(t)dt$, 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

$$(C(y_1) \geq c \land d(y_1, y_2) \leq \epsilon) \iff C(y_2) \geq c - He^\gamma.$$  \hfill (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 $(\epsilon, \delta)$-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

$$\text{Prob}(C(Y_1) \geq c) \geq 1 - \tilde{\delta} \implies \text{Prob}(C(Y_2) \geq c - He^\gamma) \geq 1 - \delta - \tilde{\delta}.$$  \hfill (8)

**Proof.** By assumption, it holds that $\text{Prob}(d(Y_1, Y_1) \leq \epsilon) \geq 1 - \delta$ and $\text{Prob}(C(Y_1) \geq c) \geq 1 - \delta$ so that we know that
\( \text{Prob}(d(Y_1, Y_1) > \epsilon) \leq \delta \) and \( \text{Prob}(C(Y_1) < c) \leq \tilde{\delta} \). We can now apply the union bound over these two events so that
\[
\text{Prob}(d(Y_1, Y_1) > \epsilon \lor C(Y_1) < c) \leq \delta + \tilde{\delta}.
\]
From here, we can simply see that
\[
\text{Prob}(d(Y_1, Y_1) \leq \epsilon \land C(Y_1) \geq c) \geq 1 - \delta - \tilde{\delta}.
\]
Since \( C \) is Hölder continuous w.r.t. \( d \), which implies that equation (7) holds, it is easy to conclude that \( \text{Prob}(C(Y_2) \geq c - H\epsilon^\gamma) \geq 1 - \delta - \tilde{\delta} \).

Therefore, Theorem 1 tells us that (i) \((\epsilon, \delta)\)-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 \((\epsilon, \delta)\)-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 \((\epsilon, \delta)\)-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
\[
\text{Prob}(\inf_{U \in \mathcal{U}} C(Y_1) \geq c) \geq 1 - \tilde{\delta}
\]
\[
\Rightarrow
\text{Prob}(\inf_{U \in \mathcal{U}} C(Y_2) \geq c - H\epsilon^\gamma) \geq 1 - \delta - \tilde{\delta}.
\]

**Proof.** By assumption, it holds that \( \text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) \leq \epsilon) \geq 1 - \delta \) and \( \text{Prob}(\inf_{U \in \mathcal{U}} C(Y_1) \geq c) \geq 1 - \tilde{\delta} \) so that we know that \( \text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) > \epsilon) \leq \delta \) and \( \text{Prob}(\inf_{U \in \mathcal{U}} C(Y_1) < c) \leq \tilde{\delta} \). We can now apply the union bound over these two events so that
\[
\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) > \epsilon \lor \inf_{U \in \mathcal{U}} C(Y_1) < c) \leq \delta + \tilde{\delta}.
\]
From here, we can simply see that
\[
\text{Prob}(\sup_{U \in \mathcal{U}} d(Y_1, Y_1) \leq \epsilon \land \inf_{U \in \mathcal{U}} C(Y_1) \geq c) \geq 1 - \delta - \tilde{\delta}.
\]
This equation tells us that, for each \( U \in \mathcal{U} \), we have
\[
\text{Prob}(d(Y_1, Y_1) \leq \epsilon \land C(Y_1) \geq c) \geq 1 - \delta - \tilde{\delta}.
\]
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 \( \text{Prob}(\inf_{U \in \mathcal{U}} C(Y_2) \geq c - H\epsilon^\gamma) \geq 1 - \delta - \tilde{\delta} \).

**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
\[
R(-C(Y_2)) \leq R(-C(Y_1)) + H r.
\]

**Proof.** We can derive the following chain of inequalities
\[
R(-C(Y_2)) \leq R(-C(Y_1)) + H d(Y_1, Y_1)
\]
\[
\leq R(-C(Y_1)) + R(H d(Y_1, Y_1))
\]
\[
\leq R(-C(Y_1)) + H R(d(Y_1, Y_1))
\]
\[
\leq R(-C(Y_1)) + H r.
\]
where (a) follows since \( d \) 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., \( R(d(Y_1, Y_2)) \leq r \).
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 $Z, Z^{(1)}, \ldots, Z^{(k)}$ 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 $Z^{(1)}, \ldots, Z^{(k)}$, 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 $Z$) that depends on $Z^{(1)}, \ldots, Z^{(k)}$ such that $\text{Prob}(Z \leq Z) \geq 1 - \delta$.

By a surprisingly simple quantile argument, see [27, Lemma 1], one can obtain $Z$ to be the $(1 - \delta)$th quantile of the empirical distribution of the values $Z^{(1)}, \ldots, Z^{(k)}$ and $\infty$. By assuming that $Z^{(1)}, \ldots, Z^{(k)}$ are sorted in non-decreasing order, and by adding $Z^{(k+1)} := \infty$, we can equivalently obtain $\bar{Z} := Z^{(p)}$ where $p := [(k + 1)(1 - \delta)]$ 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 $D_{\text{cal}}$ that consists of realizations $y_1^{(i)}$ and $y_2^{(i)}$ from the stochastic signals $Y_1 \sim D_1$ and $Y_2 \sim D_2$, respectively.

**Theorem 5.** Let the premises of Definition 1 hold and $D_{\text{cal}}$ be a calibration dataset with datapoints $(y_1^{(i)}, y_2^{(i)})$ drawn from $D_1 \times D_2$. Further, define $Z^{(i)} := d(y_1^{(i)}, y_2^{(i)})$ for all $i \in \{1, \ldots, |D_{\text{cal}}|\}$ 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 $\bar{U}$ are $(\varepsilon, \delta)$-conformant if $\bar{Z} \leq \varepsilon$.

We see that checking stochastic conformance as defined in Definition 1 is computationally simple when we have a calibration dataset $D_{\text{cal}}$. 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 $U$ 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 presumably 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:** $Z$ such that $\text{Prob}(\sup_{U \in \bar{U}} d(Y_1, Y_2) \leq Z + L\kappa) \geq 1 - \delta$

1: Construct $\kappa$-net $\bar{U}$ of $U$

2: for $\bar{U} \in \bar{U}$ do

3: Obtain calibration set $D_{\text{cal}}^\bar{U}$ consisting of realizations $(y_1^{(1)}, y_2^{(1)})$ under $\bar{U}$

4: Compute $Z_{\bar{U}} := Z(\bar{U})$ by applying Theorem 5 but instead using dataset $D_{\text{cal}}^\bar{U}$

5: $Z := \max_{\bar{U} \in \bar{U}} Z_{\bar{U}}$

Our approach is summarized in Algorithms 1 and 2. Algorithm 1 computes $Z$ such that $\text{Prob}(\sup_{U \in \bar{U}} d(Y_1, Y_2) \leq Z + L\kappa) \geq 1 - \delta$ when $L$ is known and where $\kappa$ 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 $\kappa$-net $\bar{U}$ of $U$, i.e., we construct a finite set $\bar{U}$ so that for each $U \in \bar{U}$ there exists a $\bar{U} \in \bar{U}$ such that $d(U, \bar{U}) \leq \kappa$ where $d : U \times U \rightarrow \mathbb{R}$ is a metric. For this purpose, simple gridding strategies can be used as long as the set $\bar{U}$ has a convenient representation. Alternatively, randomized algorithms can be used that sample from $U$ [41]. In lines 2-4, we apply Theorem 5 for each element $\bar{U} \in \bar{U}$. Therefore, we obtain realizations $(y_1^{(1)}, y_2^{(1)})$ from $D_1 \times D_2$ under $\bar{U}$ (line 3). We then compute $Z_{\bar{U}}$ so that $\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot))) \leq Z_{\bar{U}}) \geq 1 - \delta$ (line 4). Finally, we set $Z := \max_{\bar{U} \in \bar{U}} Z_{\bar{U}}$ (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 $D_1 \times D_2$ under $U'$ and realizations $(y_1'', y_2'')$ from $D_1 \times D_2$ under $U'''$ (line 3), and compute the non-conformity score $\bar{L}(\cdot)$ (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 $\text{Prob}(\sup_{U \in \bar{U}} d(Y_1, Y_2) \leq \bar{Z} + L\kappa) \geq 1 - \delta$

Thus, the systems $S_1$ and $S_2$ are $(\varepsilon, \delta)$-conformant if $\bar{Z} + L\kappa \leq \varepsilon$. Otherwise, let $\delta_L \in (0, 1)$ be a failure probability, then the output $\bar{L}$ of Algorithm 2 ensures that $\text{Prob}(\sup_{U \in \bar{U}} d(Y_1, Y_2) \leq Z + L\kappa) \geq 1 - \delta - \delta_L$

where $\text{Prob}$ is defined over the randomness introduced in Algorithm 2.

**Proof.** From line 4 of Algorithm 1 we know that $\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot))) \leq Z_{\bar{U}}) \geq 1 - \delta$ for each $\bar{U} \in \bar{U}$.
Due to Lipschitz continuity, we can conclude that for each $U \in \mathcal{U}$ that is such that $d(U, U) \leq \alpha$ it holds that

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

Since $\mathcal{U}$ is a $\kappa$-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 $\text{Prob}(d(Y_1(\bar{U}, \cdot), Y_2(\bar{U}, \cdot)) \leq \bar{Z}_{\bar{U}} \land L \leq \bar{L}) \geq 1 - \delta - \delta_L$.

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

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 $D_{\text{cal}}$ be a calibration dataset with datapoints $(y_1^{(i)}, y_2^{(i)})$ drawn from $D_1 \times D_2$. 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 $i \in \{1, \ldots, |D_{\text{cal}}|\}$ and assume that $\text{Prob}(Z \leq \alpha)$ is continuous in $\alpha$. Then,

$$\text{Prob}(\max_{\alpha \in \mathbb{R}} \left\{ \inf_{\alpha} \mathbb{P} \left[ \text{Prob}(Z \leq \alpha) - \sqrt{\frac{\ln(2/\gamma)}{2|D_{\text{cal}}|}} \geq \beta \right] \right\} \geq 1 - \gamma.$$  

where we have

$$\mathbb{P} \left[ \text{Prob}(Z \leq \alpha) - \sqrt{\frac{\ln(2/\gamma)}{2|D_{\text{cal}}|}} \geq \beta \right] := \inf_{\alpha \in \mathbb{R}} \mathbb{P} \left[ \text{Prob}(Z \leq \alpha) - \sqrt{\frac{\ln(2/\gamma)}{2|D_{\text{cal}}|}} \geq \beta \right]$$

with the empirical cumulative distribution function

$$\text{Prob}(Z \leq \alpha) := \frac{1}{|D_{\text{cal}}|} \sum_{i=1}^{D_{\text{cal}}} \mathbb{I}(Z^{(i)} \leq \alpha)$$

and the indicator function $\mathbb{I}$.

For estimating $\text{CVaR}_\beta(Z)$, 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 $D_{\text{cal}}$ be a calibration dataset with datapoints $(y_1^{(i)}, y_2^{(i)})$ drawn from $D_1 \times D_2$. 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 $i \in \{1, \ldots, |D_{\text{cal}}|\}$ and assume that $\text{Prob}(d(Y_1, Y_2) \in [a, b]) = 1$. Then, it holds that

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

where $\text{CVaR}_\beta := \text{CVaR}_\beta + \sqrt{\frac{\ln(3/\gamma)}{|D_{\text{cal}}|(1-\beta)}}(b - a)$ and $\text{CVaR}_\beta := \text{CVaR}_\beta - \sqrt{\frac{\ln(3/\gamma)}{|D_{\text{cal}}|(1-\beta)}}(b - a)$ where the empirical estimate of $\text{CVaR}_\beta(Z)$ is $\hat{\text{CVaR}}_\beta := \inf_{\alpha \in \mathbb{R}} (\alpha + (|D_{\text{cal}}|(1 - \beta))^{-1} \sum_{i=1}^{D_{\text{cal}}}[Z^{(i)} - \alpha]^+)$. 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 $\text{CVaR}_\beta \geq \alpha$ or $\text{CVaR}_\beta \geq \alpha$ 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 $Z$ using a conformal prediction procedure for a nonconformity score defined by the random variable $Z$, i.e., such that $\text{Prob}(Z \leq Z) \geq 1 - \delta$. Then, given a test set $D_{\text{test}}$, the validation score is defined as

$$V_{\text{S}}(Z) := |\{z \in D_{\text{test}} \mid z \leq Z\}|/|D_{\text{test}}|.$$  

A. Dubin’s car.

Dubin’s car models the motion of a point mass vehicle. The state variables are the $x$ and $y$ position, $\theta$ denotes the steering angle and $v$ the longitudinal velocity. While both $\theta$ 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 $\theta(t) := T_s \pi + \sum_{i=1}^{s} \omega(i) T_s$ where $a \beta := 0.1$s. In this example, we assume that $\omega(i) := \frac{\pi}{\sqrt{|t|}}$ for $i \in \{1, 25\}$, and $\omega(i) := -\frac{\pi}{\sqrt{|t|}}$ 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:

$$x(t + 1) = x(t) + T_s v(t) \cos(\theta(t)) + \eta^x(t)$$  

$$y(t + 1) = y(t) + T_s v(t) \sin(\theta(t)) + \eta^y(t)$$

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 $I := [-1, 0] \times [-1, 0]$. The controller aims to ensure that the system trajectory stays within a series of sets $T_1$ through $T_{50}$. The corresponding STL specification is $\phi_{\text{dubin}} := \bigwedge_{i=1}^{50} F_{[1-i]} ([x_i, y_i] \in T_i)$. For the experiments that follow, we uniformly sampled initial states from $I$ and noise $\eta^x, \eta^y$ 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 $D_{\text{cal}}$.
Spacecraft Trajectories

Robustness on validation set con-

Fig. 2: Distance and robustness histogram for Dubin’s car with

pair of realizations is the same.

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.

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 $\epsilon$ for $d_\infty$ (which translates into a higher validation score). The $VaR$ is almost identical to the value of $\epsilon$ 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 $L_\kappa$-metric.

\textit{Empirical evaluation of transference.} We empirically demonstrate that Theorem 1 holds. We use $C(Y) = \rho^{\phi_{\text{dubin}}}(Y)$, i.e., the robust semantics w.r.t. the property $\phi_{\text{dubin}}$, and the $L_\infty$ signal metric $d_\infty$. 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. $\phi_{\text{dubin}}$ is negative ($c_1 - \epsilon$), so it is not possible to conclude that the second system satisfies $\phi_{\text{dubin}}$ with probability greater than $1 - \delta - \delta$. 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 $\rho^{\phi_{\text{dubin}}}(Y_i)$ for $i = 1, 2$ in Figure 2.

\textit{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 $Z = 0.7562$ and $L_K = 0.0687$, giving $Z + L_K = 0.8249$. 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.

\textbf{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 $\phi_{\text{geas}} := G^0_{[0, T]}(h \geq 1000)$ 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.

\textbf{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
of the lane center (cross track error) and $\theta$ 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 $\theta$ 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.

### 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 $\epsilon$ as defined in Definition 1 for $\delta = 0.05$.

| $|D_{cal}|$ | $|D_{test}|$ | $c_1$ | $c_2$ | $\epsilon$ | $V S(\rho_1)$ | $V S(\rho_2)$ | $CV aR(d_\infty)$ | $CV aR(-\rho_1)$ | $CV aR(-\rho_2)$ | Valid? |
|------|------|------|------|------|------------|------------|---------------|---------------|---------------|------|
| 50   | 100  | 0.31 | 0.30 | 0.59 | 0.95       | 0.76       | 0.21          | Y             | -0.28         | 0.00 | Y     |
| 1000 | 3K   | 0.30 | 0.30 | 0.60 | 0.95       | 0.79       | 0.20          | Y             | -0.27         | 0.03 | Y     |
| 50   | 100  | 0.30 | 0.30 | 0.67 | 0.96       | 0.92       | 0.15          | Y             | -0.27         | 0.02 | Y     |
| 1K   | 1K   | 0.30 | 0.30 | 0.66 | 0.95       | 0.91       | 0.15          | Y             | -0.27         | 0.03 | Y     |
| 1K   | 3K   | 0.30 | 0.30 | 0.71 | 0.94       | 0.95       | 0.11          | Y             | -0.27         | 0.02 | Y     |
| 1K   | 3K   | 0.30 | 0.30 | 0.71 | 0.94       | 0.95       | 0.11          | Y             | -0.27         | 0.03 | Y     |

### Table II: Empirical evaluation of transference. Let $\rho_i$ be short-hand for $\rho^\delta(Y_i)$ for $i = 1,2$, and $d_\infty$ be short-hand for $d_\infty(Y_1, Y_2)$. Using Theorem 5, we show $\text{Prob}(\rho_1 \geq c_1 > 1 - \delta)$, 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 $\rho_2$ and observe that it exceeds $c_1 - \epsilon$, validating Theorem 1. Similarly, we report the $CV aR$ values for $-\rho_1$ and $d_\infty$, and $CV aR(-\rho_1) + CV aR(d_\infty) \geq CV aR(-\rho_2)$ for all cases, validating Theorem 3.

| $|D_{cal}|$ | $c_1$ | $c_2$ | $\epsilon$ | $V S(\rho_1)$ | $V S(\rho_2)$ | $CV aR(d_\infty)$ | $CV aR(-\rho_1)$ | $CV aR(-\rho_2)$ | Valid? |
|------|------|------|------|------------|------------|---------------|---------------|---------------|------|
| 50   | 100  | 0.31 | 0.30 | 0.59       | 0.95       | 0.21          | Y             | -0.28         | 0.00 | Y     |
| 1000 | 3K   | 0.30 | 0.30 | 0.60       | 0.95       | 0.20          | Y             | -0.27         | 0.03 | Y     |
| 50   | 100  | 0.30 | 0.30 | 0.67       | 0.96       | 0.15          | Y             | -0.27         | 0.02 | Y     |
| 1K   | 1K   | 0.30 | 0.30 | 0.66       | 0.95       | 0.15          | Y             | -0.27         | 0.03 | Y     |
| 1K   | 3K   | 0.30 | 0.30 | 0.71       | 0.94       | 0.11          | Y             | -0.27         | 0.02 | Y     |
| 1K   | 3K   | 0.30 | 0.30 | 0.71       | 0.94       | 0.11          | Y             | -0.27         | 0.03 | Y     |

### Table III: Transference results for various case studies. We use $\delta = 0.05$ and $\bar{\delta} = 0.05$. As before, $\rho_1$ is used as short-hand for $\rho^\delta(Y_i)$ for each spec, and $d_\infty$ is used as short-hand for $d_\infty(Y_1, Y_2)$.

| Case Study | Spec | $|D_{cal}|$ | $|D_{test}|$ | $V S(\rho_1)$ | $V S(\rho_2)$ | $V aR(d_\infty)$ |
|------------|------|------------|-------------|---------------|---------------|-----------------|
| F-16       | $\phi_{cte}$ | 1K          | 3K          | 0.95          | 0.98          | 200             |
| CARLA      | $\phi_{cte}$ | 700         | 300         | 0.94          | 0.96          | 1.88            |
| Satellite  | $\phi_{sat}$ | 7K          | 3K          | 0.96          | 0.97          | 0.18            |

### 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-
and from the noise distribution. We divide the dataset into different initial states from $I$.

Discussion on results for Transference across case studies. We omit the column for $D_{test}$ of $I\overline{D}$ shown: $\phi_{sat} := G_{[1,0]}(y, \bar{y}; |v_x|, |v_y| \leq -|x|, y_{max}, v_{x,max}, v_{y,max}) \wedge F_{[1,5]} \in T$. The set of initial states is $I = [-0.1, 0.1] \times [-0.1, 0.1]$. The system is assumed to have additive Gaussian process noise with zero mean and a diagonal covariance matrix with variances $10^{-4}, 10^{-4}, 5 \times 10^{-6}, 5 \times 10^{-5}$. We uniformly sample 100 different initial states from $I$ 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.

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 in [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, FMiF-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


| Case Study | Spec    | $|D_{cal}|$ | $|D_{test}|$ | $\delta$ | $d_{\infty}$ | $-\rho_1$ | $-\rho_2$ |
|------------|---------|-----------|-----------|--------|-------------|----------|---------|
| F-16       | $\phi_{gas}$ | 1K        | 3K        | 0.01   | 200.3       | -62.3    | -62.3   |
| CARLA      | $\phi_{quad}$ | 7K        | 3K        | 0.01   | 2.04        | -0.31    | 0.88    |
| Satellite  | $\phi_{sat}$  | 7K        | 3K        | 0.01   | 0.19        | 0.0      | 0.08    |

TABLE IV: Empirical validation of risk transference for all case studies. As before, $\rho_1$ is short-hand for $\rho^1(Y_i)$, and $d_{\infty}$ is short-hand for $d_{\infty}(Y_1, Y_2)$. Here, we set the risk level $\beta = \delta$ in each case.

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 $\tau$ and $\epsilon$ 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 Skorohod 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.
[50] X. Jin, J. V. Deshmukh, J. Kapinski, K. Ueda, and K. Butts, “Bench-
marks 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
54, 2018.

“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,

and M. R. Mousavi, “Conformance relations and hyperproperties for
doping detection in time and space,” arXiv preprint arXiv:2012.03910,
2020.

“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

for cyber-physical systems: Overview and roadmap,” in 26th Interna-
tional Conference on Concurrency Theory (CONCUR 2015). Schloss


stochastic conformance checking.” in Business Process Management
Forum: BPM Forum 2019, Vienna, Austria, September 1–6, 2019,

[58] A. A. Julius, A. Girard, and G. J. Pappas, “Approximate bisimulation
for a class of stochastic hybrid systems,” in 2006 American Control


[60] S. Haesaert and S. Soudjani, “Robust dynamic programming for tempo-
ral logic control of stochastic systems,” IEEE Transactions on Automatic

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,
pp. 321–337.

[62] A. A. Julius and A. Van Der Schaft, “Bisimulation as congruence in
the behavioral setting.” in Proceedings of the 44th IEEE Conference on

finite transition systems: A bisimulation relation approach,” in 2017
IEEE 56th Annual Conference on Decision and Control (CDC). IEEE,
2017, pp. 5615–5619.

between computer science and control theory,” European Journal of
Control, vol. 17, no. 5-6, pp. 568–578, 2011.