Evaluation Uncertainty in Data-Driven Self-Driving Testing

Zhiyuan Huang[1], Mansur Arief[2], Henry Lam[3], and Ding Zhao[2]

***Abstract*— Safety evaluation of self-driving technologies has been extensively studied. One recent approach uses Monte Carlo based evaluation to estimate the occurrence probabilities of safety-critical events as safety measures. These Monte Carlo samples are generated from stochastic input models constructed based on real-world data. In this paper, we propose an approach to assess the impact on the probability estimates from the evaluation procedures due to the estimation error caused by data variability. Our proposed method merges the classical bootstrap method for estimating input uncertainty with a likelihood ratio based scheme to reuse experiment outputs. This approach is economical and efficient in terms of implementation costs in assessing input uncertainty for the evaluation of self-driving technology. We use an example in autonomous vehicle (AV) safety evaluation to demonstrate the proposed approach as a diagnostic tool for the quality of the fitted input model.**

I. Introduction

The competitive race toward the mass deployment of self-driving cars driving side-by-side with human-driven vehicles on public roads advocates for an accurate and highly precise safety evaluation framework to ensure safe driving. However, achieving meaningful precision is a challenging task when the safety-critical events under study are rare in naturalistic situations. A recent method has been developed which which adopts Monte Carlo method empowered by the Importance Sampling technique as a variance reduction scheme; this has produced appealing results. In [1], it is shown that the efficiency is enhanced by ten thousand times with the incorporation of large-scale driving data sets and the employed statistical models. This improved efficiency is highly appealing for autonomous vehicle (AV) researchers as the required testing effort is overly demanding, an estimate of 8.8 billion driving miles required to provide 'sufficient' evidence to compare the safety of AV driving and human driving from logged data [2].

A common framework adopted to estimate the safety measure is Monte Carlo simulation, which generates a large number of samples and simulates experiments using each sample. Then an empirical expectation with confidence interval is obtained via central limit theorem. By properly integrating the Monte Carlo approach into the modeling scheme, researchers have been able to estimate the risk from driving situations based on control and trajectory prediction [3], [4] and corner cases in various driving scenarios [1], [5].

In the studies, the AV system is viewed as a black box and a Monte Carlo simulation is used to evaluate its safety performance. Using this view, the evaluation metric is estimated from Monte Carlo samples of variables that encompasses the uncertainties in the system, e.g. traffic environment or noise in control and observation. Usually, the stochastic models that generate the Monte Carlo samples are fitted using real-world data, and the model training is implemented separately with the safety evaluation. However, since parameters in the stochastic models are estimated from data, the variability of the data has a direct effect on the test result and the reliability of the evaluation. In this paper, this effect is referred to as *input uncertainty* [6]. In order to provide a convincing test evaluation, the input uncertainty needs to be addressed and highlighted as part of the test results.

The goal of this paper is to provide a way to construct a confidence interval for the evaluation results as a quantitative measurement of the input uncertainty. In AV safety evaluation, since the experiment or simulation can be quite expensive, the accelerated evaluation approach studied in [7], [1], [5] was proposed to increase the efficiency of safety evaluation. Similarly, for input uncertainty quantification, an ideal method should "minimize" the number of experiments and computing time to be computationally economic.

In this paper, we propose an extension of the classic bootstrap technique [8] for assessing the input uncertainty in Monte Carlo based evaluation approach. Our approach use a likelihood ratio based scheme for estimations in bootstrap replications, which reuses the Monte Carlo estimators for the initial evaluation of safety measures. The proposed approach significantly reduces the computation burden in the standard bootstrap implementation and enables parallel computation. Moreover, The likelihood ratio based scheme can be perfectly adapted to the accelerated evaluation approach. It can efficiently utilize the accelerated evaluation structure and provides good importance sampling estimator for each bootstrap replication in the subsequent stages.

The remainder of the paper is structured as follows: Section II introduces the notations and sets up the problem. Section III reviews classic bootstrap schemes and presents the proposed approach. Section IV illustrates the implementation details in the proposed approach using numerical examples and demonstrates using an AV evaluation example. The conclusion is provided in Section V.

We gratefully acknowledge the support from the National Science Foundation under grants CAREER CMMI-1653339/1834710, IIS-1849280, and IIS-1849304.

[1]Zhiyuan Huang is with the Department of Industrial and Operations Engineering at University of Michigan, 1205 Beal Ave, MI, USA zhyhuang@umich.edu

[2] Mansur Arief and Ding Zhao are with the Department of Mechanical Engineering at Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA marief@andrew.cmu.edu , dingzhao@cmu.edu

[3]Henry Lam is with the Department of Industrial Engineering and Operations Research, Columbia University, 500 W. 120th Street, NY, USA henry.lam@columbia.edu

II. Problem Setting Under AV Testing

In this section, we set up the notations for the AV evaluation problem. Then we will analyze the influence of input uncertainty in the evaluation and the potential issues that may arise if the input uncertainty is ignored.

A. AV Safety Evaluation Setting

The goal of the Monte Carlo based test approach is to understand the performance of the AV system under uncertain environment. We use $\xi \in \mathbb{R}^d$ to denote d-dimensional uncertain factors to the AV system, where each element represents one attribute of the environment. We use ξ_i to denote a realization of the random vector ξ. In Accelerated Evaluation approaches [1], [5], [7], the traffic environment is considered an uncertainty for the AV system under study and is represented by ξ.

A parametric stochastic model $p(\xi|\theta)$ is used to represent the uncertainty, where $\theta \in \mathbb{R}^m$ represents the parameters in parametric stochastic model. The model can be static [7], [1], or dynamic, which represents stochastic processes [5]. The underlying assumption is that the parametric model contains the true distribution. In another word, there exists a parameter θ_0 such that $\xi \sim p(\xi|\theta_0)$.

Here, we use $f(\xi)$ to denote the performance measurement of an AV system under environment ξ, which is referred to as *performance function*. For instance, we are interested in a certain type of safety-critical event (e.g. a crash) that occurs to the AV system under the environment ξ, and we use $\varepsilon \subseteq \mathbb{R}^d$ to denote the set of safety-critical event. Then the performance function is defined as $f(\xi) = I_\varepsilon(\xi) \in \{0, 1\}$, which indicates whether a certain type of safety-critical event (e.g. a crash) is occurred to the AV system under the environment ξ, where 0 and 1 represent whether a safety-critical event occurring (1 for crash). For example, $f(\xi)$ can be the test results of a trail in computer simulation, a real-world on-track trail, or a trail in a particular zone of a public street. Therefore it can be rather expensive to run experiment trail for $f(\xi)$.

Our goal is to estimate the expectation of f given by

$$E[f(\xi)|\theta_0] = \int_\xi f(\xi) p(\xi|\theta_0) d\xi. \tag{1}$$

In Accelerated Evaluation, this expectation is the probability of the safety critical event, which is revealed by the equality

$$E[f(\xi)|\theta_0] = E[I_\varepsilon(\xi)|\theta_0] = P(\xi \in \varepsilon). \tag{2}$$

This measure is used as the criterion for the safety of a tested AV, which is denoted by γ in later discussion.

Usually the performance function is defined by complex systems, and the expectation (1) is hard to be analytically computed even if $p(\xi|\theta_0)$ is fully known. Hence the Monte Carlo approach is applied to estimate $E[f(\xi)|\theta_0]$. For crude Monte Carlo, we generate samples $\xi_1, ..., \xi_n$ from $p(\xi|\theta_0)$ and estimate $E[f(\xi)|\theta_0]$ using the sample mean

$$\hat{E}[f(\xi)|\theta_0] = \frac{\sum_{i=1}^n f(\xi_i)}{n}. \tag{3}$$

Each evaluation of the performance function $f(\xi_i)$ at a certain sample ξ_i is referred to as one *experiment trail*.

In the context of AV testing, we expect the safety-critical event to be very rare ($\gamma < 10^{-5}$), where crude Monte Carlo is inefficient. The inefficiency is reflected in the large relative error ($error/p$) of the crude Monte Carlo estimator. To intuitively explained this, we can consider that every ξ drawn from $p(\xi, \theta_0)$ returns $f(\xi) = 0$ with probability $1-p$, and therefore huge number of samples are required to obtain a safety-critical event. The computation cost is usually prohibitive for obtaining an accurate estimation (in terms of relative error) due to expensive experiment trials.

To improve the efficiency in estimating $E[f(\xi)|\theta_0]$, [1] uses importance sampling estimator to reduce the variance. Instead of drawing samples from $p(\xi|\theta_0)$, we construct an accelerating distribution $\tilde{p}(\xi)$ based on information of $p(\xi|\theta_0)$ and ε. With samples $\xi_1, ..., \xi_n$ from $\tilde{p}(\xi)$, we use the estimator

$$g(\xi_i, \theta_0) = \frac{p(\xi_i|\theta_0)}{\tilde{p}(\xi_i)} f(\xi_i), \tag{4}$$

which can be proved to be unbiased. With a good selection of $\tilde{p}$, the importance sampling estimator can be very efficient. [7] has shown that the importance sampling estimator can achieve the same accuracy as the crude Monte Carlo estimator using only 10^{-3} of the crude Monte Carlo samples. Then we use the sample mean

$$\bar{g}_0 = \frac{\sum_{i=1}^n g(\xi_i, \theta_0)}{n} \tag{5}$$

to estimate the expectation $E[f(\xi)|\theta_0]$.

Usually a confidence interval is constructed as a reference of the accuracy of the estimation. For a confidence interval with confidence level $1-\alpha$, we want to have

$$P(E[f(\xi)|\theta_0] \in [C_L, C_U]) \geq 1 - \alpha, \tag{6}$$

i.e. we want the confidence interval $[C_L, C_U]$ to cover the truth with probability greater than $1-\alpha$. The most commonly used confidence interval for sample mean is derived from central limit theorem [9], which uses

$$C_L = \bar{g}_0 - z_{\alpha/2} \sqrt{\hat{Var}_\xi(\bar{g}_0)} \tag{7}$$

and

$$C_U = \bar{g}_0 + z_{1-\alpha/2} \sqrt{\hat{Var}_\xi(\bar{g}_0)}, \tag{8}$$

where z_α denotes the α quantile of standard Gaussian distribution.

B. Input Uncertainty in Safety Evaluation

In this paper, we consider the situation where θ_0 is unknown but a finite number of data from $p(\xi|\theta_0)$ is available. We use the maximum likelihood estimation (MLE) $\hat{\theta}$ for the parameter θ_0 in the stochastic model. Note that $\hat{\theta}$ is a consistent estimator of θ_0, i.e. $\hat{\theta}$ converges to θ_0 in probability as the number of samples increases. Since $\hat{\theta}$ is estimated from data of ξ, it is random due to the variability of the samples.

In practice, if the environment modeling and the evaluation are implemented as separate tasks, the estimated parameter $\hat{\theta}$ will be used as the true parameter. That is, the estimator is given by

$$\bar{g} = \frac{\sum_{i=1}^{n} g(\xi_i, \hat{\theta})}{n}, \tag{9}$$

where ξ_i's are generated from $p(\xi|\hat{\theta})$. The confidence interval from (7) and (8) uses the variance estimated from the samples, that is

$$\frac{\sum_{i=1}^{n}(g(\xi_i, \hat{\theta}) - \bar{g})^2}{n-1}. \tag{10}$$

However, the above approach ignores the variation sourced from the estimated parameter $\hat{\theta}$, where we consider as the input uncertainty. The influence of input uncertainty can be revealed by a decomposition of the variance of $\bar{g}$:

$$Var(\bar{g}) = Var_{\hat{\theta}}\left(E_{\xi}\left[\bar{g}|\hat{\theta}\right]\right) + E_{\hat{\theta}}\left[Var_{\xi}\left(\bar{g}|\hat{\theta}\right)\right]. \tag{11}$$

In this decomposition, the first term is the input uncertainty and the second term is referred to as *simulation uncertainty*. We note that if we ignore the variation of $\hat{\theta}$, only $Var_{\xi}\left(\bar{Y}|\hat{\theta}\right)$ would be considered as the variance of the estimator. The resulted confidence interval will be likely to undercover for the truth $E[f(\xi)|\theta_0]$. Under the AV evaluation context, confidence intervals that undercover for the truth are harmful for the reliability of the evaluation.

In this paper, our goal is to construct confidence intervals that target to cover $E[f(\xi)|\theta_0]$ with confidence level $1-\alpha$. This is a way to quantify the input uncertainty and therefore provide an assessment of the reliability of the evaluation results.

III. Measurement of Input Uncertainty

In this section, we first introduce some well-studied bootstrap framework. We then propose our approach based on these techniques.

A. Classic Bootstrap Approach

The bootstrap technique dates back to [8], [10], which is studied to estimate the variability of statistical estimators by judiciously reusing the data. [11] considers a parametric version of bootstrap for assessing the input uncertainty in simulation. For further interests of input uncertainty quantification, one can refer to [12], [13], [14], [15], [16], [6] and [17], Section 7.2.

We first clarify some notations to avoid confusion. Note that the random vector ξ and its samples appears in both the input modeling part and the simulation part. We use X_i's to denote samples that we collected from the real world and used to estimate θ. We use ξ_i's to represent the samples in the simulation part, which are generated from a certain distribution $\tilde{p}$ and are used to evaluate the estimator $Y(\xi_i, \theta)$.

In general, a bootstrap scheme for quantifying input uncertainty is as follows. We first generate samples $\hat{\theta}^1, ..., \hat{\theta}^B$ that approximately follows the true distribution of $\hat{\theta}$. For each $\hat{\theta}_i$, we generate samples $\xi_1, ..., \xi_r$ from $p(\xi; \theta)$ and estimate $\bar{g}^i$ using

$$\bar{g}^i = \frac{1}{r}\sum_{j=1}^{r} g(\xi_j, \hat{\theta}^i). \tag{12}$$

After computing $\bar{g}^1, ..., \bar{g}^B$, we find the $\alpha/2$ and $1-\alpha/2$ - the empirical quantiles of $\bar{g}^1, ..., \bar{g}^B$ as lower and upper bound of the confidence interval, respectively. We denote as $C_L = \hat{q}_{\alpha/2}(\bar{g}^i)$ the lower bound and as $C_U = \hat{q}_{1-\alpha/2}(\bar{g}^i)$ the upper bound. We list three bootstrap approaches in the Appendix.

B. The Proposed Approach: A Likelihood Ratio Based Estimation for Bootstrap

To motivate the proposed approach, we first consider the computation cost for a classic bootstrap scheme. No matter what bootstrap scheme we use, after we obtain the bootstrapped parameters $\hat{\theta}^1, ..., \hat{\theta}^B$, we would need to estimate $E[g|\hat{\theta}^i]$ using $\bar{g}^i$. To obtain a good empirical quantile, we usually require B to be as large as possible (usually hundreds or more)[8]. Also, in order to reduce the simulation uncertainty to avoid obtaining an over-covered confidence interval, we want r to be as large as possible. The number of experiment trials in total will be rB, which is B times more than estimating the probability. When the experiment is expensive and time-consuming, the price for assessing the input uncertainty might not be affordable. Here, we propose an approach that can assess the input uncertainty with no additional cost for experiment trials.

Assume we have already estimated the average performance measure from samples $\xi_1, ..., \xi_n$ from $\tilde{p}(\xi)$ using (9), where $\tilde{p}(\xi)$ can be $p(\xi, \hat{\theta})$ or an appropriate accelerating distribution for $p(\xi, \hat{\theta})$. Then, we obtain bootstrap parameters $\hat{\theta}^1, ..., \hat{\theta}^B$ using any bootstrap scheme. For each $\hat{\theta}^i$, instead of generate a new sample from $p(\xi, \hat{\theta}^i)$, we use the same set of samples $\xi_1, ..., \xi_n$, and estimate $\bar{g}^i$ using

$$\bar{g}^i = \frac{1}{n}\sum_{j=1}^{n} \frac{p(\xi_j, \hat{\theta}^i)}{p(\xi_j, \hat{\theta})} g(\xi_j, \hat{\theta}), \tag{13}$$

We should note that each $\bar{g}^i$ is still an unbiased estimator, i.e. we have

$$E\left[\frac{p(\xi, \hat{\theta}^i)}{p(\xi, \hat{\theta})} g(\xi, \hat{\theta})\right] = E\left[g(\xi, \hat{\theta}^i)\right]. \tag{14}$$

Note that by estimating $\bar{g}_i$ in this way, we do not need to evaluate $f(\xi)$ (which is hidden in g) at any new realization of ξ.

This approach can fit into the accelerated evaluation framework, i.e. when $\tilde{p}$ is a good accelerated distribution for $p(\xi, \hat{\theta})$ and g is defined by (4). By using the likelihood ratio adjustment, the samples across different resampled value of $\hat{\theta}^1, ..., \hat{\theta}^B$ are now correlated. It is unclear how this would affect the reliability of our estimate, but this issue would likely go away when r is large enough (which is the case in accelerated evaluation). Secondly, the likelihood ratio adjustment can sometimes blow up the magnitude of the

output estimate, especially when the when the estimation of $\hat{\theta}$ is highly uncertain. Since this is also a sign that the stochastic model is unreliable, this issue would not affect the use of the proposed approach. Usually we can speculate the $\tilde{p}(\xi)$ would also be a good accelerating distribution for $p(\xi, \hat{\theta}^i)$. For instance, if we use exponential tilting of Exponential distribution, the optimal $\tilde{p}(\xi)$ for a certain performance function $f(x)$ is the same for any parameter values θ for the exponential distribution. By using the proposed approach, we saved $r(B-1)$ experiment trials compared to the classical bootstrap approaches.

IV. Numerical Experiments and Discussion

In this section, we present some numerical experiments to illustrate the proposed approach and discuss some implementation details. We first discuss the performance of the three bootstrap schemes under different scenarios. We then use a simple illustrative problem to demonstrate the proposed approach. Lastly, we apply the proposed approach on an AV testing example problem.

A. Comparison of Bootstrap Schemes

In Appendix, we introduce three bootstrap schemes that are applicable to our framework. Here we use some numerical studies to show the advantages of each scheme and provide a guideline of choosing suitable scheme in different conditions.

The purpose of the experiment is to investigate if $\hat{\theta}^i$'s generated using these bootstrap schemes are roughly close to the true distribution of $\hat{\theta}$ with different numbers of samples k. In the experiment, we first generate k samples from $p(\xi|\theta_0)$. For each bootstrap scheme, we use these samples to generate $\hat{\theta}^1, ..., \hat{\theta}_B$ with $B = 1000$. We use the $\alpha/2$ and $1-\alpha/2$ empirical quantile of these $\hat{\theta}^i$'s as upper and lower bound for a confidence interval and check whether θ_0 is covered. We repeat this procedure for 1000 times with an independently generated sample set. We use the coverage of the truth to test the accuracy of the confidence interval obtained from these schemes. We use $\alpha = 0.05$ in our experiments.

The experiment results with different k and different distribution models are tabulated in Tables I. In the table, "Plain" represents plain bootstrap, "Parametric" represents parametric bootstrap, "Asym Cls" stands for the asymptotic distribution scheme using closed form Fisher's information and "Asym Est" stands for the asymptotic distribution scheme using empirical Fisher's information. See Appendix for details of these methods.

We consider exponential distribution for $p(X|\theta_0)$. From Table I, we observe that when $k = 10$, the coverage rates for all schemes have an obvious gap to the target 95%. For the plain bootstrap, the relatively low performance is due to the very small sample size used to construct the empirical distribution. For the parametric bootstrap, this is caused by the error in estimating $\hat{\theta}$. The asymptotic approaches suffer from both bad estimation of $\hat{\theta}$ and small value of k (note that asymptotic analyses require $k \to \infty$). Among these approaches, the parametric bootstrap has the smallest gap. This is partly because the assumption of the correct parametric model remedies the error from the variability of the samples. As we increase the value of k, the gap between target coverage and the obtained coverage reduces. When we use $k = 100$, the coverage rates for all schemes are already close to the target.

TABLE I

The CI coverage of true parameter μ in exponential distribution using three bootstrap schemes.

Samples	Approach	Object	Coverage
k=10	Plain	μ	84.70%
	Parametric	μ	92.20%
	Asym Cls	μ	88.30%
	Asym Est	μ	90.20%
k=20	Plain	μ	91.40%
	Parametric	μ	93.10%
	Asym Cls	μ	93.30%
	Asym Est	μ	92.00%
k=100	Plain	μ	94.10%
	Parametric	μ	95.10%
	Asym Cls	μ	94.30%
	Asym Est	μ	95.20%

TABLE II

The coverage and average width of confidence intervals from various approaches

Samples	100	1000	10000
Coverage CF	0.9432	0.9451	0.9505
CI Width CF	1.33e-05	8.85e-07	2.20e-07
Coverage LR	0.9426	0.9444	0.9486
CI Width LR	1.33e-05	8.85e-07	2.20e-07
Coverage SU	0.0177	0.0630	0.1903
CI Width SU	8.28e-08	3.08e-08	2.72e-08

CF (closed form), LR (likelihood ratio),
SU (simulation uncertainty only)

In summary, the parametric bootstrap provides a better coverage of the truth. especially when the number of samples is very small. The coverage for these schemes is similar when the number of samples is large enough. In a sufficient sample size situation, the asymptotic schemes have an upper hand for the efficiency of generating the parameters.

B. Illustrative Problem

We consider a simple probability estimation problem to demonstrate the effectiveness of the proposed approach in providing a valid confidence interval. This is shown in two aspects: a) the coverage of the proposed approach is close to the target; b) the confidence interval width is relatively narrow.

We consider estimating the probability of $P(\xi > \beta)$, where ξ follows a standard Gaussian distribution and we use $\beta = 5$. The choice of the problem is because we have an analytic solution for the probability, which makes it easier to validate whether the constructed confidence interval covers the truth or not. We use different numbers of sample size k for estimating $\hat{\theta} = \{\hat{\mu}, \hat{\sigma}\}$. We use $B = 1000$ bootstrap samples for constructing the confidence interval. For the estimation of $\bar{g}^i$, we consider two approaches. The first is to use the proposed approach with 10,000 importance

sampling estimators. To show the constructed CI has a relative narrow width, we also consider using the analytic solution for $P(\xi > \beta)$ for each $\hat{\theta}^i$ as a baseline (so that there is no simulation uncertainty). We repeat for 10000 total replications and compute the coverage of the confidence interval.

The experiment results are summarized in Table II, where we show the coverage and width of the confidence intervals computed using the closed form probability (CF), likelihood ratio (LR) estimation, and Equation (7) and (8) that only consider simulation uncertainty (SU), i.e. with input uncertainty ignored.

We have two main observations from these experiment results. First, the proposed likelihood ratio scheme provides a good estimation of the probability of interest. This claim is supported by the similar coverage rates and confidence interval width for the closed form baseline approach and our proposed approach. Second, the confidence interval computed without incorporating the input uncertainty is problematic. This observation is revealed by the low coverage rates (especially when the estimator has high variability, e.g. when the sample size is small) and narrow confidence interval width.

C. Accelerated Evaluation Example

To demonstrate the proposed approach, we consider the AV evaluation problem and AV model discussed in [1]. The lane change test scenario is shown in Figure 1, where we evaluate the safety level of a test AV by estimating the probability of crash when a frontal car cut into the lane. Crash is determined by whether the minimum range of two vehicles during the lane change procedure reaches 0. The traffic environment in this scenario is represent by v, the initial velocity of the frontal vehicle, R, the initial range between the two vehicles, and TTC, the time-to-collision value defined by $TTC = R/\dot{R}$.

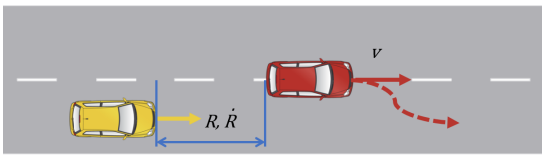

Fig. 1. An illustration of the lane change scenario in AV evaluation.

In our problem, we consider the frontal car to have an initial velocity $v = 30m/s$, which is a common speed in highway driving. We extract 12,304 lane change scenario samples identified from the SPMD dataset [18] with similar velocity. We use the samples to fit R^{-1} and TTC^{-1} with exponential distribution. We used the cross-entropy method to find an optimal accelerating distribution by exponential tilting for R^{-1} and TTC^{-1}; we generated 10^5 samples from the accelerating distribution. We then use the proposed approach to construct confidence interval that incorporates the input uncertainty.

In Figure 2, we present the probability estimation and the two types of confidence intervals we construct given different

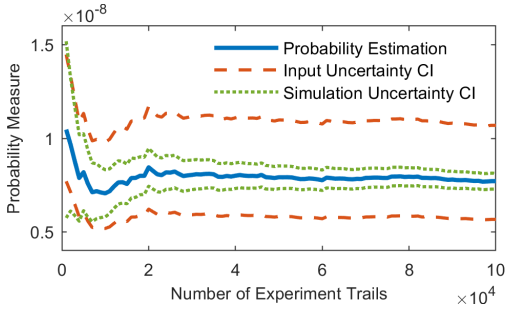

Fig. 2. The estimates of safety critical events rate (probability of crash) and their confidence intervals with different number of experiments.

number of samples. The confidence interval for simulation uncertainty is estimated using (7) and (8). We observe that the confidence interval for simulation uncertainty has a much smaller width than the input uncertainty width. This observation indicates that if the input uncertainty is ignored, the evaluation results can be misleading. For instance, if we use the confidence upper bound to interpret the safety level of a vehicle, the input uncertainty upper bound is roughly 1.5 times of the simulation uncertainty, hence using only the simulation uncertainty would underestimate the risk of crash.

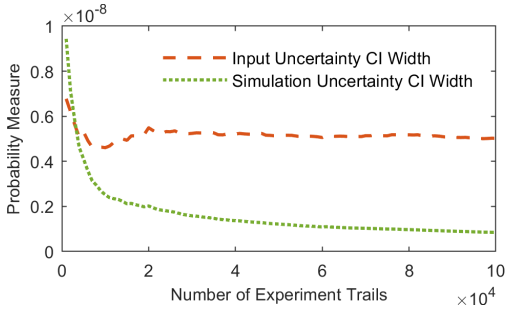

Fig. 3. The width of confidence intervals for estimates of safety critical events rate (probability of crash) with different number of experiments.

Figure 3 shows how the widths of the two intervals changes as the number of experiment trials increases. As known in literature, the width of the simulation uncertainty confidence interval shrinks in the order of $O(1/\sqrt{n})$. This trend can be easily observed from the figure. On the other hand, since we are not changing the number of samples we use to estimate $\hat{\theta}$, the input uncertainty confidence interval should not change as n increases, which is confirmed by the figure. When the number of experiment trials is sufficiently large, the simulation uncertainty decreases while the input uncertainty remains the same as the number of trials increases. When n is small, the CI width of input uncertainty is not as stable as when n is large. This is because when we do not have enough samples, the simulation uncertainty becomes large and can perturb the estimation of the input uncertainty.

This implies that input uncertainty can partially help reveal the consistency of information from the data and ignoring it could lead to misleading results and wrong conclusion. In the case where one can ensure the n data provide

good representativeness for the whole unseen data, input uncertainty analysis help describe the information richness of the collected data with regard to the model and evaluation results, which are essential for rigorous AV evaluation purposes.

V. Conclusion

In this paper, we propose an approach to assess the input uncertainty in Monte Carlo- based AV testing methods, which requires zero additional experiment trails. The proposed approach is shown to be computationally efficient and easy to implement while providing valid confidence intervals that incorporate input uncertainty. In the future, we would consider extending our study to model-free input uncertainty analysis for a wider application domains.

APPENDIX

A. Examples of Bootstrap Methods

Here, we introduce three different schemes for generate samples $\hat{\theta}^1, ..., \hat{\theta}^B$ that are straight-forward and easy to implement. For a sound empirical study on the performance of bootstrap schemes, refer to [19].

These bootstrap schemes assume that we start with a sample set $\{X_1, ..., X_k\}$ and the MLE $\hat{\theta}$ is estimated using these samples. Note that in the discussed approaches, we restrict the resampling size to be equal to the original sample size, namely k. This is not required for the bootstrap technique, but we adopt this setting for convenience and simplicity.

1) Plain Bootstrap: Plain bootstrap considers the sample set $\{X_1, ..., X_k\}$ as an empirical distribution, say $\hat{f}$, and use it as an approximation of the real distribution of X_i. We draw k samples from $\hat{f}$, i.e. resample from $\{X_1, ..., X_k\}$ with replacement, and then use these samples to estimate $\hat{\theta}^1$ (using MLE). We repeat this procedure for B times to obtain $\hat{\theta}^1, ..., \hat{\theta}^B$.

2) Parametric Bootstrap: Here we use $p(X, \hat{\theta})$ as an approximation of the real distribution of X_i. We draw k samples from $p(X, \hat{\theta})$ and use them to estimate $\hat{\theta}^1$. We repeat this for B times and collect $\hat{\theta}^i,\ i = 1, ..., B$.

3) Sample Parameters from Asymptotic Distribution: Since the $\hat{\theta}$ is estimated using MLE, we know the asymptotic behavior of $\hat{\theta}$. That is when $k \to \infty$, we have

$$\sqrt{k}(\hat{\theta} - \theta_0) \sim N(0, I^{-1}(\theta_0)), \tag{15}$$

where $I^{-1}(\theta)$ is the inverse of Fisher's information matrix of the parametric distribution $p(\cdot|\theta)$. Since θ_0 is unknown, we can use its MLE $\hat{\theta}$ to obtain an approximation of the asymptotic distribution $N(0, I^{-1}(\hat{\theta}))$.

In practice, one can use the empirical Fisher's information matrix, which is an estimation based on the samples. That is

$$\hat{I}(\theta) = -\frac{1}{k}\sum_{i=1}^{k} \frac{\partial^2}{\partial \theta^2} \log p(X_i|\theta), \tag{16}$$

where X_i's are the samples we use to fit the model. Thus, we can direct sample $\hat{\theta}^1, ..., \hat{\theta}^B$ from $N(\hat{\theta}, I^{-1}(\hat{\theta})/k)$ or $N(\hat{\theta}, \hat{I}^{-1}(\hat{\theta})/k)$, which reduces computation cost from re-sampling and estimating $\hat{\theta}^i$.

References

[1] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE transactions on intelligent transportation systems*, vol. 18, no. 3, pp. 595–607, 2017.

[2] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[3] A. Broadhurst, S. Baker, and T. Kanade, "Monte carlo road safety reasoning," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.* IEEE, 2005, pp. 319–324.

[4] A. Eidehall and L. Petersson, "Statistical threat assessment for general road scenes using monte carlo sampling," *IEEE Transactions on intelligent transportation systems*, vol. 9, no. 1, pp. 137–147, 2008.

[5] D. Zhao, X. Huang, H. Peng, H. Lam, and D. J. LeBlanc, "Accelerated evaluation of automated vehicles in car-following maneuvers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 733–744, 2018.

[6] H. Lam, "Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation," in *2016 Winter Simulation Conference (WSC)*. IEEE, 2016, pp. 178–192.

[7] Z. Huang, H. Lam, D. J. LeBlanc, and D. Zhao, "Accelerated evaluation of automated vehicles using piecewise mixture models," *IEEE Transactions on Intelligent Transportation Systems*, no. 99, pp. 1–11, 2017.

[8] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[9] S. Asmussen and P. W. Glynn, *Stochastic simulation: algorithms and analysis*. Springer Science & Business Media, 2007, vol. 57.

[10] B. Efron, *The jackknife, the bootstrap, and other resampling plans*. Siam, 1982, vol. 38.

[11] R. C. Cheng and W. Holloand, "Sensitivity of computer simulation experiments to errors in input data," *Journal of Statistical Computation and Simulation*, vol. 57, no. 1-4, pp. 219–241, 1997.

[12] R. Barton, S. Chick, R. Cheng, S. Henderson, A. Law, B. Schmeiser, L. Leemis, L. Schruben, and J. Wilson, "Panel discussion on current issues in input modeling," in *Proceedings of the 2002 Winter Simulation Conference*. IEEE, 2002, pp. 353–369.

[13] S. G. Henderson, "Input model uncertainty: Why do we care and what should we do about it?" in *Winter Simulation Conference*, vol. 1, 2003, pp. 90–100.

[14] S. E. Chick, "Bayesian ideas and discrete event simulation: why, what and how," in *Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference, 2006, pp. 96–105.

[15] R. R. Barton, "Input uncertainty in outout analysis," in *Proceedings of the Winter Simulation Conference*. Winter Simulation Conference, 2012, p. 6.

[16] E. Song, B. L. Nelson, and C. D. Pegden, "Advanced tutorial: Input uncertainty quantification," in *Proceedings of the Winter Simulation Conference 2014*. IEEE, 2014, pp. 162–176.

[17] B. Nelson, *Foundations and methods of stochastic simulation: a first course*. Springer Science & Business Media, 2013.

[18] D. Bezzina and J. R. Sayer, "Safety Pilot: Model Deployment Test Conductor Team Report," UMTRI, Tech. Rep., 2014.

[19] R. R. Barton, B. L. Nelson, and W. Xie, "A framework for input uncertainty analysis," in *Proceedings of the Winter Simulation Conference*. Winter Simulation Conference, 2010, pp. 1189–1198.