Assessing Modeling Variability in Autonomous Vehicle Accelerated Evaluation

Authors: Zhiyuan Huang, Mansur Arief, Henry Lam, and Ding Zhao

Presenter: Jesung Park
Facilitators: Parham Hamouni, Susan Chang

June 24, 2019

aisc

Introduction

- Achieving meaningful precision for safety in autonomous vehicles is difficult under real-world driving conditions
  - 8.8 billion miles of driving required
- Monte Carlo samples are drawn from empirical distributions of real-world data or stochastic models fitted from real-world data
  - Reliability depends on the correctness of the underlying models, whose parameters are estimated from the data (i.e. input uncertainty)
  - Construct a confidence interval for the evaluation results as a measurement of the input uncertainty
- Recent approach adopting Monte Carlo method empowered by importance sampling used to improve efficiency by 10,000 times

Notation

- $\xi \in \mathbb{R}^d$ denotes the uncertain factors to the AV system i.e. the environment
- $\theta \in \mathbb{R}^m$ are the parameters in a parametric stochastic model with density function $f(\xi;\theta)$, assume θ_0 is the truth i.e. $\xi \sim f(\xi;\theta_0)$
- $G(\xi)$ denotes the performance measurement of the AV system under environment ξ
    - Also called performance function
    - Goal is to measure $\mathrm{E}[G(\xi)|\theta_0] = \int_\xi G(\xi) f(\xi;\theta_0)\, d\xi$ but this is difficult even if $f(\xi;\theta_0)$ is fully known
    - Each evaluation of the performance function $G(\xi_i)$ at a certain sample ξ_i is referred to as one experiment trail

Notation Example

- You’re driving down the highway when a giant wall appears out of thin air
- ξ is the velocity of the car and distance between the car and the wall
  - $\xi = \{d, \dot{d}\}$
- Assuming a Gaussian distribution, θ is the mean and covariance of ξ
- $G(\xi)$ can be defined as 0 for not crashing, 1 for crashing

Estimating the Performance Function

- Given samples $\xi_1, \cdots, \xi_n$ generated from a certain distribution $f(\xi)$, denote estimator as $Y(\xi; \theta_0)$

$$\hat{\mathrm{E}}[G(\xi)|\theta_0] = \bar{Y} = \frac{\sum_{i=1}^{n} Y(\xi_i; \theta_0)}{n}$$

- Since θ_0 is not observable, use $\hat{\theta}$
- Decompose variance of $\bar{Y}$

$$var(\bar{Y}) = \underbrace{var_{\hat{\theta}}\left[E_{\xi}\left(\bar{Y}|\hat{\theta}\right)\right]}_{\text{Input uncertainty}} + \underbrace{E_{\hat{\theta}}\left[var_{\xi}\left(\bar{Y}|\hat{\theta}\right)\right]}_{\text{Simulation uncertainty}}$$

Accelerated Evaluation

- Performance function is defined as $G(\xi) = I_\epsilon(\xi) \in \{0,1\}$ depending on whether or not a crash occurred
    - ϵ is the set of safety-critical events
- Average performance measure is the probability of a safety-critical event happening
    - $\mathrm{E}[G(\xi)|\theta_0] = \mathrm{E}[I_\epsilon(\xi)|\theta_0] = P(\xi \in \epsilon)$ denoted by p
- Given that p is very small $(p < 10^{-5})$, it is very inefficient to simply run MC simulations

Accelerated Evaluation

- Importance sampling is used to estimate expected values under one distribution given samples from another
  - Used in off-policy reinforcement learning
- Construct an accelerating distribution $\tilde{f}(\xi)$ based on information of $f(\xi;\theta_0)$ and ϵ
- With $\xi_1, \cdots, \xi_n$ from $\tilde{f}(\xi)$, use an unbiased estimator

$$Y(\xi_i;\theta_0) = \underbrace{\frac{f(\xi_i;\theta_0)}{\tilde{f}(\xi_i)}}_{\text{Importance sampling ratio}} G(\xi_i)$$

Classic Bootstrap Approach

- X_i's denote samples collected from the real world
  - Used to estimate θ
- ξ_i's represent the samples in the simulation part
  - Generated from a certain distribution $\tilde{f}$
  - Used to evaluate the estimator $Y(\xi_i;\theta)$

Classic Bootstrap Approach

- Generate samples $\hat{\theta}^1, \cdots, \hat{\theta}^B$ that approximately follows the true distribution of $\hat{\theta}$
- For each $\hat{\theta}^i$, generate samples $\xi_1, \cdots, \xi_r$ from $f(\xi; \theta)$
- Estimate $\bar{Y}^i$ using

$$\bar{Y}^i = \frac{1}{r}\sum_{j=1}^{r} Y(\xi_j; \hat{\theta}^i)$$

- Fit a confidence interval based on $\{\bar{Y}^1, \bar{Y}^2 \ldots \bar{Y}^B\}$

Direct Bootstrap

- Consider the sample $\{X_1, \cdots, X_k\}$ as an empirical distribution, say $\hat{f}$
- Use it as an approximation for the real distribution of X_i
- Draw k samples from $\hat{f}$ with replacement
- Use samples to estimate $\hat{\theta}^1$
- Repeat B times to obtain $\hat{\theta}^1, \cdots, \hat{\theta}^B$

Parametric Bootstrap

- Use $f\left(X, \hat{\theta}\right)$ as an approximation of the real distribution of X_i
- Draw k samples from $f\left(X, \hat{\theta}\right)$ and use them to estimate $\hat{\theta}^1$
- Repeat B times to obtain $\hat{\theta}^1, \cdots, \hat{\theta}^B$

Sample Parameters from Asymptotic Distribution

- When $k \to \infty$, we have

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

  - Where $I^{-1}(\theta)$ is the inverse of Fisher's information matrix of the parametric distribution $f(\cdot\,; \theta)$

- In practice, a closed form of Fisher's information matrix might not be available. Use the empirical Fisher's information matrix instead

$$\hat{I}(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\frac{\delta^2}{\delta\theta^2}\log f(X_i; \theta)$$

- Sample from $N\left(\hat{\theta}, I^{-1}(\hat{\theta})/k\right)$ or $N\left(\hat{\theta}, \hat{I}^{-1}(\hat{\theta})/k\right)$

Direct Bootstrap Example

- Sample B times from $\hat{f} = \{30.39, 28.42, \ldots, 31.15\}$ to estimate $\hat{\theta}^i$

32.34	24.68	32.34	30.39	28.42
30.92	28.42	35.14	30.92	35.14

- $\hat{\theta}^1 = \{\hat{\mu}: 30.87, \hat{\sigma}: 3.19\}$
- Generate r samples from $N(30.87, 3.19)$ and evaluate performance

30.17	29.91	32.34	27.70	30.81
36.06	33.12	33.07	33.61	36.70

- $rB = 100$ experiment trails of $Y(\xi_j; \hat{\theta}^i)$

$$\bar{Y}^i = \frac{1}{r} \sum_{j=1}^{r} Y(\xi_j; \hat{\theta}^i)$$

Likelihood Ratio Based Estimation for Bootstrap

- In the classic bootstrap scheme, B requires > 30 and is usually ≥ 100, and we want r to be as large as possible to minimize simulation uncertainty
- Number of experiment trails in total will be rB, which is B times more than estimating the probability
- With samples $\xi_1, \cdots, \xi_n$ that were already generated from $\tilde{f}(\xi)$, obtain $\hat{\theta}^1, \cdots, \hat{\theta}^B$ from any bootstrap scheme. Estimate $\bar{Y}^i$ using

$$\bar{Y}^i = \frac{1}{n} \sum_{j=1}^{n} \underbrace{\frac{f\left(\xi_j; \hat{\theta}^i\right)}{f\left(\xi_j; \hat{\theta}\right)}}_{\text{Importance sampling ratio}} \overbrace{Y(\xi_j; \hat{\theta})}^{\text{Probability of crash given } \hat{\theta}}$$

Likelihood Ratio Example

- Obtain $\hat{\theta}^1 = \{\hat{\mu}: 30.87, \hat{\sigma}: 3.19\}$ with direct bootstrap
- Take the n samples from $\tilde{f}(\xi)$ and evaluate performance

29.93	35.71	28.86	29.42	31.86
27.27	31.21	29.16	26.74	16.06

- $\xi_1 = 29.93,\ \hat{\theta} = \{\hat{\mu}: 28.62, \hat{\sigma}: 5.09\}$
- $\frac{f(\xi_1;\hat{\theta}^1)}{f(\xi_1;\hat{\theta})} = \frac{0.120}{0.076} = 1.58$
- $n = 10$ experiment trails of $Y(\xi_j; \hat{\theta})$

$$\bar{Y}^i = \frac{1}{n}\sum_{j=1}^{n} \frac{f(\xi_j; \hat{\theta}^i)}{f(\xi_j; \hat{\theta})} Y(\xi_j; \hat{\theta})$$

Brake!

Comparison of Bootstrap Schemes

TABLE I

THE CI COVERAGE OF TRUE PARAMETER μ IN EXPONENTIAL DISTRIBUTION USING THREE BOOTSTRAP SCHEMES.

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

- Generate k samples from $f(X;\theta_0)$ then generate $\hat{\theta}^1, \cdots, \hat{\theta}^B$ with a bootstrap scheme
- Look at coverage of μ by confidence intervals generated with $\hat{\theta}^1, \cdots, \hat{\theta}^B$
  - $B = 1000,\ \alpha = 0.05$
- Repeat 1,000 times
- When $\mathrm{k} = 10$, all schemes have an obvious gap to the target 95%

Comparison of Bootstrap Schemes

TABLE II

THE CI COVERAGE OF TRUE PARAMETERS μ, σ IN GAUSSIAN DISTRIBUTION USING THREE BOOTSTRAP SCHEMES.

Samples	Approach	Object	Coverage
k=20	Direct	μ	92.10%
		σ	88.60%
	Parametric	μ	92.30%
		σ	93.00%
	Asym Cls	μ	92.90%
		σ	92.80%
	Asym Est	μ	92.60%
		σ	93.50%
k=100	Direct	μ	95.20%
		σ	91.70%
	Parametric	μ	94.80%
		σ	93.70%
	Asym Cls	μ	95.00%
		σ	93.50%
	Asym Est	μ	94.90%
		σ	93.40%

- Overall, the parametric bootstrap provides a better coverage of the truth, especially with lower sample sizes
- With large enough sample sizes, asymptotic schemes are preferable due to their efficiency in generating the parameters
- $k = 100$ is insufficient for getting a good coverage of σ

Comparison of Baseline vs. Proposed Approach

- Goal: Estimate the probability of $P(\xi > \gamma)$ with $\xi \sim N(\mu, \sigma)$ and $\gamma = 5$
  - Easy to assess accuracy of CI since analytical solution exists
- Estimate $\hat{\theta} = \{\hat{\mu}, \hat{\sigma}\}$ with $B = 1000$
- Likelihood ratio provides a good approx.
- Simulation uncertainty shouldn't be ignored
  - CF: closed form probability for each bootstrap parameter
  - LR: likelihood ratio estimation
  - SU: simulation uncertainty only

TABLE III

THE COVERAGE AND AVERAGE WIDTH OF CONFIDENCE INTERVALS CONSTRUCTED BY A BASELINE APPROACH, THE PROPOSED APPROACH AND THE APPROACH THAT ONLY CONSIDERS SIMULATION UNCERTAINTY.

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

Accelerated Evaluation Example

- Evaluate the safety level of a test AV by estimating the probability of crash when a frontal car cuts into lane
    - v, the initial velocity of the frontal vehicle
    - R, the initial range between the two vehicles
    - TTC, time-to-collision defined by $TTC = R/\dot{R}$
- Consider $v = 30m/s$ and extract 12,304 lane change samples from SPMD dataset
- Generate 10,000 samples from the accelerating distribution

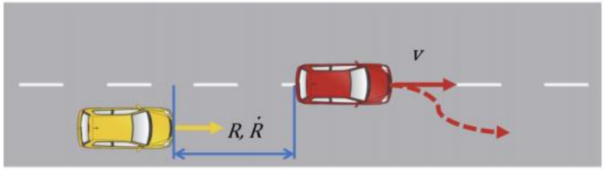

Fig. 1. An illustration of the lane change scenario in AV evaluation.

Accelerated Evaluation Results

- CI width for simulation uncertainty is much smaller than the width of input uncertainty
- Looking only at simulation uncertainty will underestimate the risk of crash

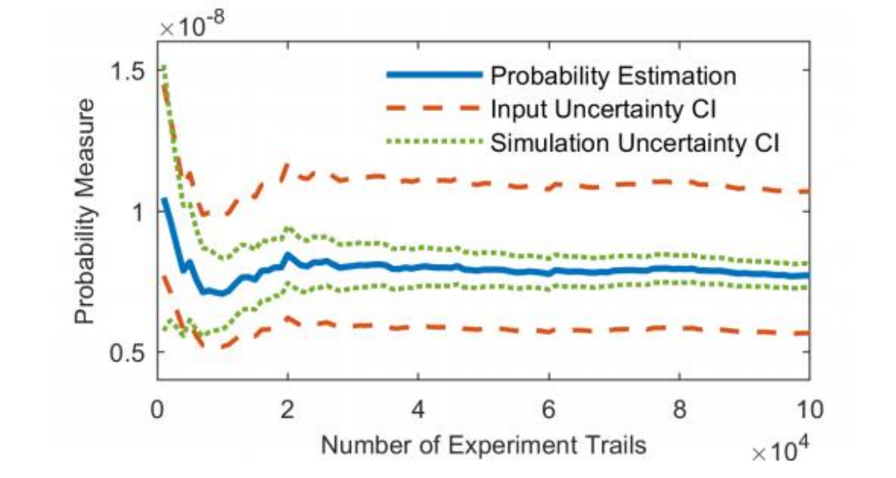

Fig. 2. The estimations of safety critical events rate and their confidence intervals with different number of experiments.

Accelerated Evaluation Results

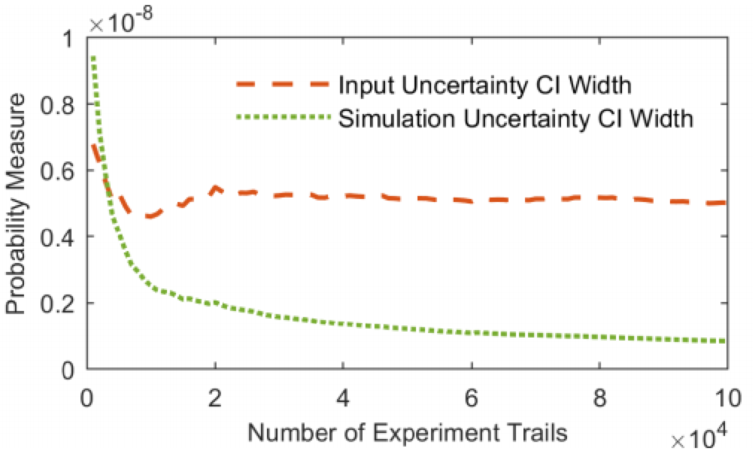

Fig. 3. The width of confidence intervals for estimations of safety critical events rate with different number of experiments.

- Width of the simulation uncertainty shrinks in the order of $O(1/\sqrt{n})$
- Since the number of samples used to estimate $\hat{\theta}$ does not change, input uncertainty width remains the same

Key Takeaways

- By using the proposed approach, we saved $r(B-1)$ experiment trails compared to the classical bootstrap approaches
- Input uncertainty should not be ignored in estimating model variability

Discussion Points

- Besides autonomous vehicles, what are other applications that can leverage this approach?
- What other performance function $G(\xi)$ can we define to evaluate AV safety?
- Would picking a different importance sampling ratio result in an improvement of the coverage?
- Potential of distribution assumption apart from Gaussian, and/or combining techniques such as maximum likelihood estimation