CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

**Report No. UMTRI-2023-5** **March 2023**

**Project Start Date: March, 2019**

**Project End Date: February, 2020**

# Accelerated Training for Connected and Automated Vehicles Based on Adaptive Evaluation Method

**by**
**Henry Liu, Professor**
**Yiheng Feng, Assistant Research Scientist**
**University of Michigan**

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3551747105 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Liu, H.X. & Feng, Y. (2020). Accelerated Training for Connected and Automated Vehicles Based on Adaptive Evaluation Method. Final Report. UMTRI-2023-5.
DOI: 10.7302/7017

## Contacts

For more information:

Dr. Henry X. Liu
University of Michigan
2350 Hayward, Ann Arbor, MI, 48109
Phone: (734) 647-4796
Email: henryliu@umich.edu

Dr. Yiheng Feng
University of Michigan
2901 Baxter Rd, Ann Arbor, MI, 48109
Phone: (734) 936-1052
Email: yhfeng@umich.edu

**Center for Connected and Automated Transportation**
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48152
umtri-ccat@umich.edu
ccat.umtri.umich.edu
(734) 763-2498

**Technical Report Documentation Page**

| **1. Report No.** UMTRI-2023-5 | **2. Government Accession No.** | **3. Recipient's Catalog No.** |
|---|---|---|
| **4. Title and Subtitle** Accelerated Training for Connected and Automated Vehicles Based on Adaptive | | **5. Report Date** March 2020 |

| | |
|---|---|
| Evaluation Method<br>DOI: 10.7302/7017 | **6. Performing Organization Code** |
| **7. Author(s)**<br>Liu, Henry, Ph.D., https://orcid.org/0000-0002-3685-9920<br>Feng, Yiheng, Ph.D., https://orcid.org/0000-0001-5656-3222 | **8. Performing Organization Report No.** |
| **9. Performing Organization Name and Address**<br>UMTRI<br>2901 Baxter Road<br>Ann Arbor, MI 48109 | **10. Work Unit No.**<br><br>**11. Contract or Grant No.**<br>Contract No. 69A3551747105 |
| **12. Sponsoring Agency Name and Address**<br>Center for Connected and Automated Transportation<br>University of Michigan Transportation Research Institute<br>2901 Baxter Road<br>Ann Arbor, MI 48109 | **13. Type of Report and Period Covered**<br>Final Report<br>March 2019 – February 2020<br>**14. Sponsoring Agency Code** |

**15. Supplementary Notes**

Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.

**16. Abstract**

How to generate testing scenario libraries for connected and automated vehicles (CAVs) is a major challenge faced by the industry. In previous studies, to evaluate maneuver challenge of a scenario, surrogate models (SMs) are often used without explicit knowledge of the CAV under test. However, performance dissimilarities between the SM and the CAV under test usually exist, and it can lead to the generation of suboptimal scenario libraries. In this project, an adaptive testing scenario library generation (ATSLG) method is proposed to solve this problem. A customized testing scenario library for a specific CAV model is generated through an adaptive process. To compensate for the performance dissimilarities and leverage each test of the CAV, Bayesian optimization techniques are applied with classification-based Gaussian Process Regression and a newly designed acquisition function. Comparing with a pre-determined library, a CAV can be tested and evaluated in a more efficient manner with the customized library. To validate the proposed method, a cut-in case study is investigated and the results demonstrate that the proposed method can further accelerate the evaluation process by a few orders of magnitude.

| | | | |
|---|---|---|---|
| **17. Key Words**<br>Connected and automated vehicles, Testing scenario library, Adaptive testing and evaluation, Bayesian optimization | **18. Distribution Statement**<br>No restrictions. | | |
| **19. Security Classif. (of this report)**<br>Unclassified | **20. Security Classif. (of this page)**<br>Unclassified | **21. No. of Pages**<br>32 | **22. Price** |

## Table of Contents

List of Figures ..... 2

Project Summary ..... 3

1. Introduction ..... 4

2. Adaptive Testing Scenario Library Generation Method ..... 6
   - 2.1 Revisit the TSLG Method ..... 6
   - 2.2 Problem Formulation ..... 8
     - 2.2.1 ATSLG Problem ..... 8
     - 2.2.2 Bayesian Optimization Scheme ..... 9
   - 2.3 Adaptive Testing Scenario Library Generation ..... 10
     - 2.3.1 Initial Testing Scenarios ..... 11
     - 2.3.2 Classification-based Gaussian Process Regression ..... 11
     - 2.3.3 Surrogate Model Update and Library Generation ..... 14
     - 2.3.4 Acquisition Function Design ..... 14
     - 2.3.5 Overall Algorithm ..... 16

3. Cut-in Case Study ..... 16
   - 3.1 Case Description ..... 16
   - 3.2 Offline Library Generation ..... 17
   - 3.3 Adaptive Library Generation ..... 20
   - 3.4 CAV Evaluation ..... 23

4. Findings and Recommendations ..... 25

5. Outputs ..... 25

6. Impacts ..... 26

References ..... 27

## List of Figures

Figure 1: Illustration of suboptimal scenarios for a test CAV. ..... 4
Figure 2: Illustration of the adaptive testing scenario library generation process. ..... 5
Figure 3 Illustration of the cut-in case. ..... 17
Figure 4 An illustration of the cut-in events distribution in Michigan area (Gong, et al., 2018). ..... 17
Figure 5 The probability of the cut-in range and range rate in NDD, i.e., $P(x)$. ..... 18
Figure 6 Safety performance of the SM. ..... 19
Figure 7 The offline generated library of the cut-in case for safety evaluation based on the FVDM. ..... 20
Figure 8. Results of 50 initial testing scenarios, where the black dots denote the dissimilar scenarios, and the orange dots denote the similar scenarios. ..... 21
Figure 9. The results of the adaptive library generation for the cut-in case. ..... 22
Figure 10. The improved library of the cut-in case for safety evaluation. ..... 23
Figure 11. The evaluation results of the cut-in case with NDD evaluation (a, b), offline library evaluation (c-e, blue line), and adaptive library evaluation (c-e, red line). ..... 25

## Project Summary

How to generate testing scenario libraries for connected and automated vehicles (CAVs) is a major challenge faced by the industry. In previous studies, to evaluate maneuver challenge of a scenario, surrogate models (SMs) are often used without explicit knowledge of the CAV under test. However, performance dissimilarities between the SM and the CAV under test usually exist, and it can lead to the generation of suboptimal scenario libraries. In this project, an adaptive testing scenario library generation (ATSLG) method is proposed to solve this problem. A customized testing scenario library for a specific CAV model is generated through an adaptive process. To compensate for the performance dissimilarities and leverage each test of the CAV, Bayesian optimization techniques are applied with classification-based Gaussian Process Regression and a newly designed acquisition function. Compared with a pre-determined library, a CAV can be tested and evaluated in a more efficient manner with the customized library. To validate the proposed method, a cut-in case study is investigated, and the results demonstrate that the proposed method can further accelerate the evaluation process by a few orders of magnitude.

## 1. Introduction

Testing scenario library generation (TSLG) is a major challenge in evaluating connected and automated vehicles (CAVs). A scenario describes the temporal development in a sequence of scenes, where a scene is a snapshot of the environment including stationary elements (e.g., road geometry) and dynamic elements (e.g., background vehicles) [1]. Given an operational design domain (ODD) [2], there could exist millions of scenarios with different parameters, e.g., different maneuvers of background vehicles. A testing scenario library is defined as a critical subset of scenarios that can be used for the evaluation of certain performance metrics (e.g., safety). In the past few years, increasing research efforts have been made to solve the TSLG problem [3-13] (see [14] and references therein). However, most existing methods have limitations in either scenario types that can be handled, CAV models that can be applied, or performance metrics that can be evaluated.

To overcome these limitations, a systematic framework was proposed in our previous studies [14-16]. Each testing scenario was evaluated by a newly proposed measure, scenario criticality, which can be computed as a combination of exposure frequency and maneuver challenge. The exposure frequency can be obtained by using naturalistic driving data (NDD). To evaluate the maneuver challenge, a surrogate model (SM) is utilized as the exact CAV model is not available. Performance dissimilarities between the SM and the specific CAV under evaluation, however, usually exist and can lead to the generation of suboptimal scenario library. The suboptimal library may increase the number of tests in order to reach a required evaluation precision, therefore may become the major source of evaluation inefficiency.

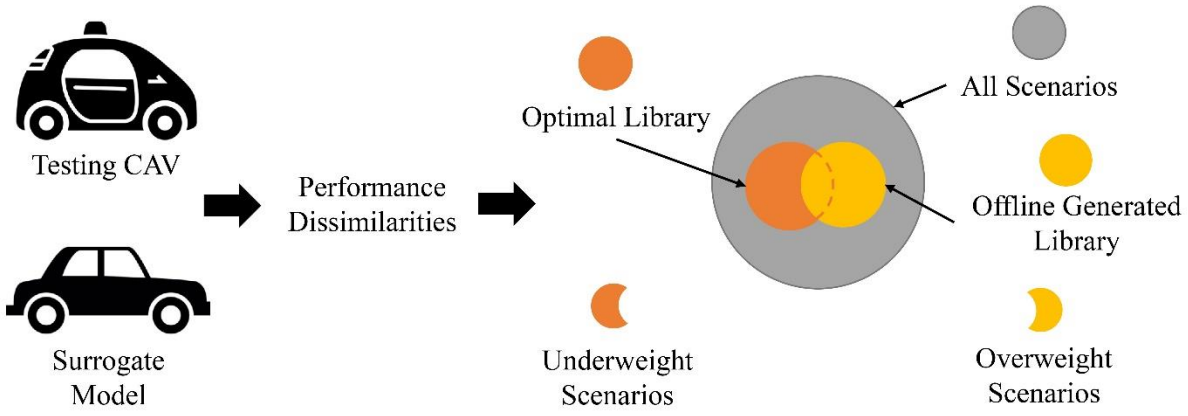



Figure 1: Illustration of suboptimal scenarios for a test CAV.

Two types of suboptimal scenarios exist, as shown in Figure *1*. Underweight scenarios represent the critical scenarios that are ignored by the library, and overweight scenarios represent the uncritical scenarios that are included in the library. If we denote the scenario library generated

by using the SM as “offline generated library”, and a customized library that includes all critical scenarios specifically designed for a CAV as “optimal library”, the differences between these two libraries include both underweight and overweight scenarios.

The goal of this project is to generate the customized optimal library by reducing the number of suboptimal scenarios through an adaptive testing process. An illustration of this process is shown in Figure *2*. The customization process starts with the test of CAV using a small set of scenarios sampled from the off-line generated library. After the initial testing, at each iteration, the most informative scenario is selected and tested, following that the SM is dynamically updated and the customized library is progressively improved, until the threshold for the dissimilarity compensation is reached. With the customized library, the CAV can be tested and evaluated in a more efficient manner, comparing with the evaluation method utilizing the offline generated library.

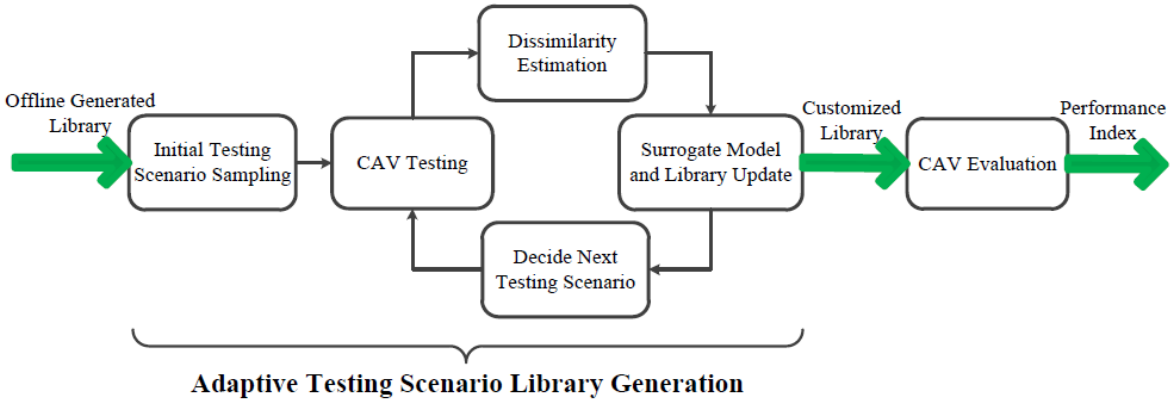



Figure 2: Illustration of the adaptive testing scenario library generation process.

In the adaptive testing process, to leverage each CAV test, Bayesian optimization techniques [17, 18] are applied. The classification-based Gaussian Process Regression (GPR) [19] is used to estimate the nonstationary performance dissimilarities, and a new acquisition function is designed to determine the most informative testing scenario in each iteration. Both the prior knowledge (e.g., SM and offline generated library) and observations (e.g., results from the adaptive testing process) are utilized to customize the library. To validate the proposed framework, a cut-in case is studied in similar settings to those in [15]. Comparing with the TSLG framework in [14], the new adaptive framework can further accelerate the evaluation process by a few orders of magnitude, e.g., 10-100.

## 2. Adaptive Testing Scenario Library Generation Method

In this section, we will introduce the proposed adaptive testing scenario library generation (ATSLG) method. For the convenience of the readers, Section 2.1 briefly revisits the offline library generation method discussed in [14-16]. In Section 2.2, the problem of the adaptive testing process is formulated. The ATSLG method is elaborated in Section 2.3.

### 2.1 Revisit the TSLG Method

The goal of the TSLG is to is to generate a set of critical scenarios, which can be used to evaluate CAVs for certain performance indices more efficiently. If an event of interest during a testing scenario is denoted as $\mathrm{A}$, e.g., an accident event, the performance index can be defined as its occurrence probability, i.e.,

$$P(A|\theta) = \sum_{x \in X} P(A|x,\theta)P(x|\theta), \tag{2-1}$$

where $x$ denotes the decision variables of testing scenarios (e.g., maneuvers of background vehicles), $X$ denotes the feasible set of $x$, and $\theta$ denotes the pre-determined parameters in the operational design domain (ODD) of CAVs, e.g., road type, weather condition, number of lanes, speed limit, *etc*. Since $\theta$ keeps constant in the library generation process, it will be omitted from now on to simplify the notations. So, the Eq. (2-1) is rewritten as

$$P(A) = \sum_{x \in X} P(A|x)P(x). \tag{2-2}$$

Essentially the on-road test is to evaluate the performance index in a naturalistic driving environment. Taking the cut-in case as an example, if a testing CAV drives on public roads, experiences $n$ cut-in scenarios, and has $m$ accident events, the accident rate of the CAV in the cut-in scenarios is estimated as

$$\begin{aligned} P(A) &= \sum_{x \in X} P(A|x)P(x), \\ &\approx \frac{1}{n}\sum_{i=1}^{n} P(A|x_i), x_i \sim P(x), \\ &\approx \frac{m}{n}, \end{aligned} \tag{2-3}$$

where the last two equations are derived by Monte Carlo theory [20]. Here the cut-in scenarios on public roads follow the naturalistic distribution, i.e., $x_i \sim P(x)$. Because the accident event $\mathrm{A}$ under the naturalistic driving environment is very rare, the required number of tests is intolerably large for reasonable estimation precision [21].

To mitigate this issue, importance sampling techniques were applied by [6] as

$$\begin{aligned} P(A) &= \sum_{x \in X} P(A|x)P(x), \\ &= \sum_{x \in X} \frac{P(A|x)P(x)}{q(x)} q(x), \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \frac{P(A|x_i)P(x_i)}{q(x_i)}, x_i \sim q(x), \end{aligned} \tag{2-4}$$

where $q(x)$ denotes an importance function, and $P(A|x_i)$ is obtained by the testing results. Compared with Eq. (2-3), testing scenarios are sampled via the importance function $q(x)$ instead of $P(x)$. Intuitively, if $q(x)$ can increase the testing frequency of critical scenarios where the accident events happen more frequently, the evaluation efficiency can be improved.

For a certain estimation precision, the minimal number of tests (i.e., evaluation efficiency) is determined by the importance function. The required estimation precision can be measured by relative half-width for a given confidence level [22]. With the confidence level at $100(1-\alpha)\%$, the relative half-width is defined as

$$l_r = \frac{\Phi^{-1}(1-\alpha/2)}{\mu_A} \sqrt{Var(\mu_A)} = \frac{\Phi^{-1}(1-\alpha/2)}{\mu_A} \frac{\sigma}{\sqrt{n}}, \tag{2-5}$$

where $\mu_A = P(A)$, $\Phi^{-1}$ denotes the inverse cumulative distribution function of standard normal distribution $\mathcal{N}(0,1)$, and $Var(\mu_A) = \sigma^2/n$ denotes the estimation variance. For a pre-determined half-width $\beta$, the minimal number of tests is derived as

$$n \geq \left( \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{\mu_A \beta} \right)^2 \sigma^2. \tag{2-6}$$

Therefore, the evaluation process has higher efficiency with a smaller $\sigma$. By importance sampling theory [23], the estimation variance can be derived as

$$\sigma^2(q) = \sum_{x \in X} \frac{\left(P(A|x)P(x)\right)^2}{q(x)} - \mu_A^2, \tag{2-7}$$

which is determined by the importance function. To obtain an importance function with small variance, a heuristic searching method was proposed in [6], which performs well in simple cases for safety evaluation (e.g., cut-in). For complex cases and other metrics (e.g., functionality), construction of a proper importance function becomes a big challenge.

To solve this problem, the criticality of scenarios was defined in [14], and critical scenarios were obtained to construct a library by efficient searching methods. The critical scenarios as well as their criticality values essentially construct the important function. Specifically, the criticality is defined as the combination of maneuver challenge and exposure frequency as

$$V(x) \stackrel{\text{def}}{=} P(S|x)P(x), \tag{2-8}$$

where $S$ denotes the event of interest with the SM of CAVs. The SM denotes what we know about common features of CAVs. Integrated with a $\epsilon$-greedy sampling policy, the importance function is constructed as

$$q(x) = \begin{cases} (1-\epsilon)V(x)/W, & x \in \Phi \\ \dfrac{\epsilon}{N(X)-N(\Phi)}, & x \notin \Phi \end{cases}, \tag{2-9}$$

where $\Phi$ denotes the critical scenarios (i.e., the library), $N(X)$ and $N(\Phi)$ denote the total number of feasible scenarios and critical scenarios respectively, and W is a normalization factor as

$$\mathrm{W} = \sum_{x \in \Phi} V(x). \tag{2-10}$$

The importance function was justified by theoretical analysis and case studies regarding evaluation accuracy and efficiency.

As discussed above, the maneuver challenge ($P(S|x)$) is evaluated by using an SM of CAV. However, performance dissimilarities between the SM and CAV models usually exist and can lead to the generation of suboptimal scenario library. The suboptimal library may increase the variance $\sigma^2$ and therefore decrease the evaluation efficiency. To further improve the evaluation efficiency, the problem of ATSLG is formulated and addressed in this project.

**2.2 Problem Formulation**

In this section, the problem of ATSLG is formulated as a Bayesian optimization problem. Specifically, the ATSLG problem is analyzed in Subsection 2.2.1. In Subsection 2.2.2, the Bayesian optimization scheme is presented, and major challenges are analyzed.

*2.2.1 ATSLG Problem*

The goal of the ATSLG is to minimize the estimation variance $\sigma^2$ by as few numbers of tests as possible. As discussed above, the key is to compensate for the performance dissimilarities between the SM and the CAV under test. The dissimilarity function can be defined as

$$f(x) = P(A|x) - P(S|x), x \in \mathrm{X}.$$

Each test of the CAV will provide one observation of $f(x)$. Denote $\tilde{f}(x)$as an estimation of $f(x)$, and then the SM can be updated with the compensation as

$$P(S'|x) = P(S|x) + \tilde{f}(x), x \in \mathrm{X}.$$

where $S'$ denotes the event of interest with the updated SM. Therefore, with the compensation $\tilde{f}(x)$, the estimation variance should be reduced. If the mapping relation is denoted as a function $\sigma^2(\tilde{f}(x))$, the ATSLG problem can be formulated as

$$\min_{\tilde{f} \in F} \sigma^2(\tilde{f})$$

where $F$ denotes the function space of $\tilde{f}$.

As indicated in Theorem 2 in [14], the optimal solution is obtained if the dissimilarities are exactly compensated, i.e., $\tilde{f} = f$. Generally, more observations of $f$ can lead to better compensation. However, each observation of $f$ required one real vehicle testing, which is time-consuming and cost expensive. Therefore, the objective function should be optimized with as few observations as possible.

To solve the problem, there are two critical subproblems. The first is how to select each test scenario $x$ for the new observation of $f(x)$. The informativeness of each scenario should be evaluated in the sense that how much information the new observation can provide for reducing the estimation variance. At each iteration, the most informative scenario should be selected for the next observation. The second is how to update the compensation function $\tilde{f}(x)$ for smaller $\sigma^2$ by leveraging all the existing observations and prior knowledge.

*2.2.2 Bayesian Optimization Scheme*

Bayesian optimization tries to optimize an unknown function $f(x)$ by as few observations as possible [17]. It has been widely applied in various fields including intelligent transportation systems [24-29] (see [18], [30] and references therein). It provides a powerful and flexible scheme especially for the optimization problems with expensive and black-box objective functions. The basic idea is to assume a prior probabilistic model for $f(x)$ and then exploit this model to decide where to observe $f(x)$ next, while integrating out uncertainty. Prior knowledge can be well utilized in the construction of the prior probabilistic model. To decide the next point for observation, various acquisition functions have been proposed for the measurement of the informativeness [18], e.g., expected improvement, knowledge gradient, entropy search, and predictive entropy search. With a properly designed acquisition function, the most informative scenario can be selected. Posterior knowledge can be obtained by integrating prior knowledge and observations.

Table 1. Algorithm scheme of the ATSLG process

| |
|---|
| **Input**: SM and offline generated library;<br>**Output**: Evaluation results of the CAV |
| **Step 1**: Observe $f$ by testing the CAV with initial testing scenarios. |

**Step 2**: While the stop criteria (e.g., budget or precision) is not satisfied

**Step 2.1**: Obtain the estimation $\tilde{f}$;

**Step 2.2**: Update SM and library;

**Step 2.3**: Decide next iteration of testing scenarios;

**Step 2.4**: Observe $f$ by testing the CAV with new testing scenarios;

**End**

**Step 3**: Test and evaluate the CAV with the customized library.

In this project, we propose to apply the Bayesian optimization scheme for the ATSLG problem. Specifically, the scheme of the ATSLG problem is described in Table *1*. The SM and the offline generated library can be utilized as prior knowledge. The informativeness of each scenario can be evaluated by the acquisition function, and $\tilde{f}(x)$ can be estimated as the posterior knowledge. Then, the SM as well as the library can be improved accordingly.

When applying the Bayesian optimization scheme to the ATSLG problem, there are three major challenges as follows:

First, the ATSLG problem optimizes in the function space, $\tilde{f} \in F$, instead of the parameter space, $x \in X$. Essentially, the function space is infinite-dimensional, and optimization in the function space belongs to the domain of infinite dimensional analysis [31]. For the common Bayesian optimization problems, however, the decision variable $x \in X$ is finite-dimensional, which is less complex and challenging. Although the function space can be simplified as a finite-dimensional space after the discretization, its dimension is still much higher than the decision variable. In the cut-in case of this project, for example, the dimension of $\tilde{f}(x)$ is 3,420 after discretization, while the dimension of $x$ is only 2.

Second, performances of a CAV may change more drastically in certain scenario neighborhoods than others, and therefore the covariance of the dissimilarity function can be highly non-stationary and nonlinear.

Third, the objective function $\sigma^2$ is unavailable for the ATSLG problem. $\sigma^2$cannot be calculated unless $\mu_A^2$ is known, which is exactly what needs to be evaluated. However, most existing acquisition functions of Bayesian optimization methods are calculated based on the availability of objective functions. Consequently, a new acquisition function needs to be designed.

We aim to address the above challenges in the following section.

### 2.3 Adaptive Testing Scenario Library Generation

In Subsection 2.3.1, to “prime the pump” with initial testing scenarios, a sampling mechanism that balances the exploitation of the offline generated library and exploration outside the library is designed. Such a sampling mechanism will provide a sketch of the dissimilarity function. In Subsection 2.3.2, different from most Bayesian optimization methods where explicit objective functions are estimated, the dissimilarity function is estimated by the Gaussian process regression (GPR) method. To handle the non-stationary challenge, scenarios are classified into two groups before applying the GPR method, resulting in the classification based GPR method. In Subsection 2.3.3, the SM is compensated with the estimated dissimilarity function, and the new library is generated accordingly. Furthermore, in Subsection 2.3.4, the informativeness of each scenario is measured by the estimated improvement of the evaluation efficiency, and then a new acquisition function is designed. Finally, the overall algorithm is summarized in Subsection 2.3.5.

*2.3.1 Initial Testing Scenarios*

To provide a sketch of the dissimilarity function, we should balance the exploitation of the offline generated library and exploration outside the library. To this end, a simple yet effective policy is proposed as follows. Since scenarios of the library have higher testing priority, they are more likely to be overweighted. To find overweight scenarios, the library is sampled according to scenario criticality values. Similarly, scenarios outside the library are more likely to be underweighted. To find underweight scenarios, scenarios outside the library are randomly sampled with a probability $\gamma$. Comparing with the $\epsilon$, the value of $\gamma$ is much larger, e.g., 0.5. Similar to the ``No Free Lunch Theorem'' [32], if there is no additional information about locations of the underweight scenarios, any searching scheme is no better than random sampling. Incorporating all these considerations, the initial testing scenarios are sampled as

$$P(x_0) = \begin{cases} \dfrac{(1-\gamma)V(x_0)}{W}, & x_0 \in \Phi \\ \dfrac{\gamma}{N(X) - N(\Phi)}, & x_0 \notin \Phi \end{cases},$$

where $x_0$ denotes an initial testing scenario. To better explore the underweight scenarios, the value of $\gamma$ (e.g., 0.5) is usually greater than $\epsilon$ (e.g., 0.1) in the $\epsilon$-sampling policy .

*2.3.2 Classification-based Gaussian Process Regression*

The dissimilarity function is estimated by the GPR method [19], because of the following advantages. As a non-parametric method, it is not limited by a functional form and thus is flexible and powerful for estimating highly nonlinear functions. Moreover, it is also convenient to add prior knowledge of the specific problem by selecting different covariance functions. In this project, a non-stationary covariance function is designed by the classification-based GPR method. Furthermore, besides the function estimation, it can also provide a probability distribution over

the function estimation, which captures the estimation uncertainty. The informativeness of each scenario can be evaluated based on the estimation uncertainty.

The basic idea is to use a Gaussian process (GP) to describe a probability distribution over the functions. Specifically, the value of $f(x)$ at each scenario $x$ is viewed as a Gaussian random variable, and values of $f(x)$ at all scenarios follow a joint Gaussian distribution. As a result, $f(x)$ can be represented by the GP as

$$f(x) \sim GP\big(m(x), k(x,x')\big),$$

where both $x$ and $x'$ denote scenarios, $m(x)$ denotes the mean function, and $k(x,x')$ denotes the covariance function as

$$m(x) = E\big(f(x)\big),$$
$$k(x,x') = E\big[\big(f(x) - m(x)\big)\big(f(x') - m(x')\big)\big].$$

Let $X_N = \{x_n \in X\}_{n=1}^{N}$ denotes $N$ points of sampled scenarios with observations. An observation of $f(x)$ is equivalent to one test run of the CAV. The observations are denoted as $f(X_N)$. Let $X_{N^*} = \{x_{n^*} \in X\}_{n^*=1}^{N^*}$ denote the $N^*$ points of scenarios without observations, i.e., $f(X_{N^*})$. By properties of GP [19], $f(X_{N^*})$ can be estimated by the posterior probability distribution as

$$f(X_{N^*})|f(X_N) \sim GP\big(\tilde{f}_{X_N}(X_{N^*}), \sigma^2_{\tilde{P},X_N}(X_{N^*})\big),$$

where

$$\tilde{f}_{X_N}(X_{N^*}) = K(X_{N^*}, X_N)K(X_N, X_N)^{-1}f(X_N),$$
$$\sigma^2_{\tilde{P},X_N}(X_{N^*}) = K(X_{N^*}, X_{N^*}) - K(X_{N^*}, X_N)K(X_N, X_N)^{-1}K(X_N, X_{N^*}),$$

and $K(X_{N^*}, X_N)$ denotes the $N \times N_*$ matrix of the covariance functions evaluated at all pairs of $X_N$ and $X_{N^*}$. The mean function $\tilde{f}_{X_N}(X_{N^*})$ is naturally the estimation of $f(x)$ at scenarios $X_{N^*}$. The variance function $\sigma^2_{\tilde{P},X_N}(X_{N^*})$ denotes a type of confidence of the estimation. The overall process is denoted as the Gaussian process regression (GPR) method [19].

Design of the mean function $m(x)$ and covariance function $k(x,x')$ is critical to the GPR. Zero mean function is applied. Covariance functions typically have the property that points closer in the variable space are more strongly correlated as

$$\|x - x'\| < \|x - x''\| \Rightarrow k(x,x') > k(x,x''),$$

where $\|\cdot\|$ denotes a distance measurement. The commonly used covariance functions include squared exponential covariance function, Matern class of covariance functions, and piecewise polynomial covariance functions [19]. Hyperparameters of covariance functions can be determined by the maximum likelihood estimate method based on the observations [19]. All these covariance functions have a stationary assumption, i.e., covariances are only determined by the distances between scenarios. For the dissimilarity function $f(x)$, however, the covariance

could be heterogenous, i.e., performances of a CAV may change more drastically in certain scenario neighborhoods than others. To solve this issue, non-stationary covariance functions should be applied, e.g., dot product covariance function [19].

In this project, the Gaussian process classification (GPC) is integrated into the GPR to solve the heterogenous issue. The idea is similar to the treed Gaussian process models [30], which divides the variable space by a decision tree and applies the GPR method in each region respectively. In this project, the input space is divided into two classes, i.e., dissimilar scenarios and similar scenarios, by the values of $f(x)$ as

$$y(x) = \begin{cases} +1, & f(x) \neq 0, \\ -1, & f(x) = 0, \end{cases} \tag{2-11}$$

where $y(x)$ denotes the class label, i.e., $+1$ for dissimilar scenarios and $-1$ for similar scenarios. The class labels of the scenarios $X_N$, i.e., $y(X_N)$, are calculated based on the observations $f(X_N)$ by Eq. (2-11). Let $X_{N1}$ denote the dissimilar scenarios in $X_N$ and $X_{N2}$ denote the similar scenarios, i.e., $N_1 + N_2 = N$. To classify the scenarios without observations, i.e., $y(X_{N^*})$, the GPC places a prior probabilistic model over a latent function $g(x)$, computes posterior distribution of $g(x)$, and "squashes" the function through a logistic function to provide a probabilistic prediction of the class labels of scenarios, i.e., $p\big(y(X_{N^*})\big)$ [19]. Note that the purpose of the latent function $g(x)$ is solely to allow a convenient formulation of the model, and we are not particularly interested in the values of $g(x)$. Then, the probabilistic prediction of the class labels is

$$p(y(x) = +1 | y(X_N)) = \int logistic(g(x)) \times p(g(x)|y(X_N)) dg(x), x \in X_{N^*}$$

where the logistic function is defined as

$$logistic(z) = \frac{1}{1 + e^{-z}}.$$

With a pre-determined threshold $P_{th}$, scenarios without observations (i.e., $X_{N^*}$) can be classified into two classes including dissimilar scenarios (i.e., $X_{N_1^*}$) and similar scenarios (i.e., $X_{N_2^*}$). After the classification process, the GPR method is applied to two classes $\{X_{N1}, X_{N_1^*}\}$ and $\{X_{N2}, X_{N_2^*}\}$ respectively. As the covariance functions are calibrated respectively, the classification-based GPR is non-stationary and can capture the heterogeneous characteristics of the dissimilarity function.

Finally, based on the observations of scenarios $X_N$, the values of $f(x)$ can be estimated as

$$f_{X_N}(x) \sim \begin{cases} \mathcal{N}\big(\tilde{f}_{X_{N_1}}(x), \sigma^2_{P,X_{N_1}}(x)\big), & p\big(y(x) = +1 | y(X_N)\big) > P_{th}, \\ \mathcal{N}\big(\tilde{f}_{X_{N_2}}(x), \sigma^2_{P,X_{N_2}}(x)\big), & p\big(y(x) = +1 | y(X_N)\big) \leq P_{th}, \end{cases}$$

where $\mathcal{N}\big(\tilde{f}_{X_{N1}}(x), \sigma^2_{P,X_{N1}}(x)\big)$ denotes the GPR results in the dissimilar scenarios, and $\mathcal{N}\big(\tilde{f}_{X_{N2}}(x), \sigma^2_{P,X_{N2}}(x)\big)$ denotes the results in the similar scenarios.

*2.3.3 Surrogate Model Update and Library Generation*

Based on the estimated dissimilarity function $f_{X_N}(x)$, the SM can be improved. A straightforward method is to add the estimated dissimilarity function to the original SM as

$$P(S_{X_N}|x) = P(S_0|x) + f_{X_N}(x), \tag{2-12}$$

where $S_0$ denotes the event of interest with the original SM, and $S_{X_N}$ denotes the event with the updated SM based on the observations $X_N$. The updated SM can be applied in the calculation of acquisition function to decide next iteration of testing scenarios.

A new library can be constructed based on the new SM. One problem is the errors brought by the pre-determined threshold $P_{th}$, which could be amplified in the library generation process. To mitigate these errors, an expectation-based SM is constructed for library generation. To keep the rareness property of the SM, the values of $P(S_{X_N}|x)$ keep zero for scenarios $x \in U$ as

$$U = \{P(S_0|x) = 0, P(S_{X_N}|x) = 0\}, \tag{2-13}$$

because neither prior knowledge ($P(S_0|x)$) nor posterior knowledge ($P(S_{X_N}|x)$) (see Eq. (4C-1)) indicates the scenarios $x \in U$ to be critical. For other scenarios $x \in X / U$, the SM values are recalculated by the expectation of classification probability to mitigate the influence of $P_{th}$ as

$$\begin{aligned} P_E(S_{X_N}|x) = P(S_0|x) + p(y(x) = +1|y(X_N)) \cdot \tilde{f}_{X_{N_1}}(x) \\ + p(y(x) = -1|y(X_N)) \cdot \tilde{f}_{X_{N_2}}(x), x \in X/U, \end{aligned} \tag{2-14}$$

where $\tilde{f}_{X_{N1}}(x)$ and $\tilde{f}_{X_{N2}}(x)$ denotes the regression result (i.e., mean value) at scenario $x$ based on the observations at scenarios $X_{N1}$ and $X_{N2}$ respectively.

*2.3.4 Acquisition Function Design*

The goal of the acquisition function is to decide next iteration of observations, i.e., next round of testing scenarios for the CAV. The objective function is unobservable unless $\mu_A$ is known, which is exactly what needs to be evaluated. Therefore, traditional acquisition functions based on objective functions cannot be directly applied. To solve this issue, a new acquisition function is designed. Both the classification uncertainty and regression variance are considered, and the exploitation and exploration are balanced.

The expected improvement is the most commonly used acquisition function for Bayesian optimization methods as

$$EI_{X_N}(x) \triangleq E_{X_N}\left[(f(x) - f^*_{X_N})^+\right], \tag{2-15}$$

where $E_{X_N}[\cdot] = E[\cdot\,|f(X_N)]$ denotes the expectation taken from the posterior distribution given observations, $f^*_{X_N} = \max_{n\le N} f(x_n)$, and

$$a^+ = \begin{cases} a, & a > 0 \\ 0 & a \le 0 \end{cases}.$$

For the ATSLG problem, the objective function is the estimation variance, so the ideal expected improvement of observing scenario $x_{N+1}$ is

$$IEI_{X_N}(x_{N+1}) \triangleq E_{X_N}\left[\left(\sigma^2(q_{X_N}) - \sigma^2(q_{X_{N+1}})\right)^+\right], \tag{2-16}$$

where $X_{N+1} = \{X_N, x_{N+1}\}$, and $q_{X_N}$ denotes the generated library by the observations in scenarios $X_N$.

As discussed before, the calculation of Eq. (2-16) is infeasible based on observations. To solve this issue, a pointwise contribution to the estimation variance is defined to replace $\sigma^2$ as

$$PI_{X_N}(x) \triangleq \frac{\left(P(S_{X_N}|x)P(x)\right)^2}{q_{X_N}(x)}. \tag{2-17}$$

Compared with Eq. (2-7), the maximal improvement of $\sigma^2$ by testing the CAV at scenario $x$ is bounded by $PI_{X_N}(x)$. Then the expected value of $PI_{X_N}(x)$ is derived as

$$EPI_{X_N}(x) \triangleq E\left[\frac{\left(P(S_{X_N}|x)P(x)\right)^2}{q_{X_N}(x)}\right]. \tag{2-18}$$

Applying the integration by parts, the analytic form of Eq. (2-18) is as follow.

**Theorem 1**. The analytic form of $EPI_{X_N}(x)$ can be derived as

$$EPI_{X_N}(x) = \frac{P^2(x)}{q_{X_N}(x)}\left(\left(P(S_0|x) + \tilde{f}_{X_{N_i}}(x)\right)^2 + \sigma^2_{\tilde{P},X_{N_i}}(x)\right), \tag{2-19}$$

where $i = 1$ for $P\big(y(x) = +1\big|y(X_N)\big) > P_{th}$, i.e., dissimilar scenarios, and $i = 2$ for $p\big(y(x) = +1\big|y(X_N)\big) \le P_{th}$, i.e., similar scenarios.

The $EPI_{X_N}(x)$ does not include the classification uncertainty. To explore the boundaries of the classification, the classification variance, i.e., $\sigma^2_{C,X_N}(x)$, is integrated into the acquisition function as

$$I_{X_N}(x) = w\frac{EPI_{X_N}(x)}{U_E} + \frac{\sigma^2_{C,X_N}(x)}{U_C}, x \in X/U, \tag{2-20}$$

where $U$ denotes the set defined in Eq. (2-13), $U_E = \max_x EPI_{X_N}(x)$ and $U_C = \max_x \sigma^2_{C,X_N}(x)$ are normalization factors to make the metrics comparable, $w$ is a weight to balance the two terms, and $\sigma^2_{C,X_N}(x)$ can be calculated similarly. Recall neither prior knowledge nor posterior knowledge

indicates that scenarios in the set $U$ are critical. Therefore, we cannot exploit an acquisition function to search potential critical scenarios in the set $U$. Instead, a small probability ($\beta$) of random sampling is applied to explore these scenarios. Finally, the next iteration of testing scenario is decided by

$$x_{N+1} = \begin{cases} \max_{x \in X_N} I_{X_N}(x), x \in X/U, & \text{with a probability } 1\text{-}\beta \\ \text{random sampling for } x \in U, & \text{with a probability } \beta \end{cases}. \quad (2\text{-}21)$$

*2.3.5 Overall Algorithm*

As shown in Algorithm 1, the test of a CAV includes three steps, described in the following:

The first step is to test the CAV with initial scenarios generated. The testing results provide a sketch of the dissimilarity function.

Based on the sketch, the second step is to test the CAV with the most informative scenario iteratively. At each iteration, the dissimilarity function is estimated, the SM as well as the library is updated, and the acquisition function is calculated to determine the next test scenario. The iterative process will stop if the number of tests is larger than the pre-determined budget or the estimation precision is satisfied.

With the updated library, the third step is to test and evaluate the CAV with the epsilon-greedy sampling policy. The minimal number of tests can be determined, and the CAV performance can be evaluated.

## 3. Cut-in Case Study

this section, the proposed method is demonstrated in a cut-in case for safety evaluation.

### 3.1 Case Description

Fig. 3 illustrates the cut-in case, where a background vehicle (BV) makes a lane change in front of the testing CAV. Similar to the previous work [6] [15], the decision variables in this case are determined as

$$x = (R, \dot{R}),$$

where $R$ and $\dot{R}$ denote the range (i.e., vehicle distance) and range rate (i.e., speed difference) at the cut-in moment. The safety performance is evaluated by the accident rate of the CAV in public road test. The accident event can be defined by reaching a threshold of minimal distance between two vehicles, i.e., $d_{min}$.

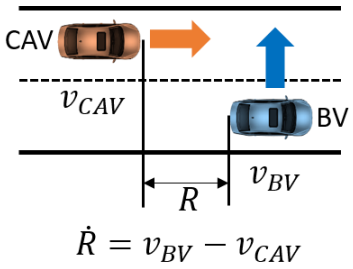



Figure 3 Illustration of the cut-in case.

### 3.2 Offline Library Generation

The TSLG method in [14] is conducted to generate the offline library, including naturalistic driving data (NDD) analysis, SM construction, and library generation.

The NDD from the Safety Pilot Model Deployment (SPMD) program at University of Michigan [34] is utilized to estimate the exposure frequency of the cut-in scenarios. The SPMD database is one of the largest databases in the world that records naturalistic driving behaviors over 34.9 million miles from 2,842 equipped vehicles in Ann Arbor, Michigan. The following query criteria are designed to extract all cut-in events [6] [15]: (a) the vehicles' speeds at the cut-in moment belong to $(2m/s, 40m/s)$; (b) the range at the cut-in moment belongs to $(0.1m, 90m)$. A total number of 414,770 qualified cut-in events are successfully obtained. The location distribution of the events is shown in Figure *4*. The joint probability distribution of the cut-in range and range rate (i.e., $P(x)$) is shown in Figure *5*.

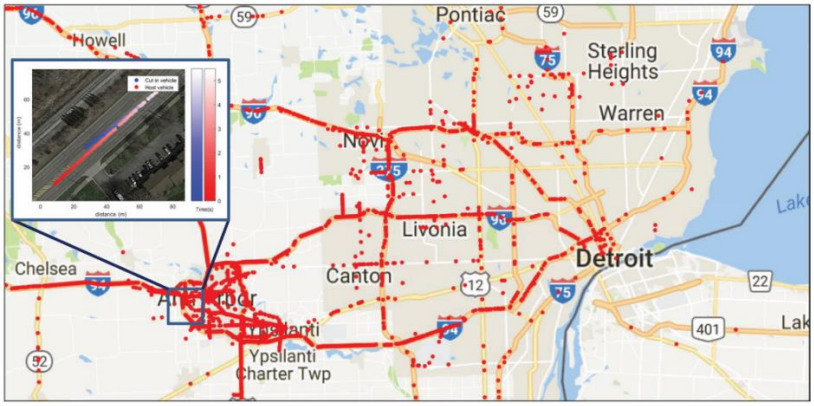



Figure 4 An illustration of the cut-in events distribution in Michigan area [6].

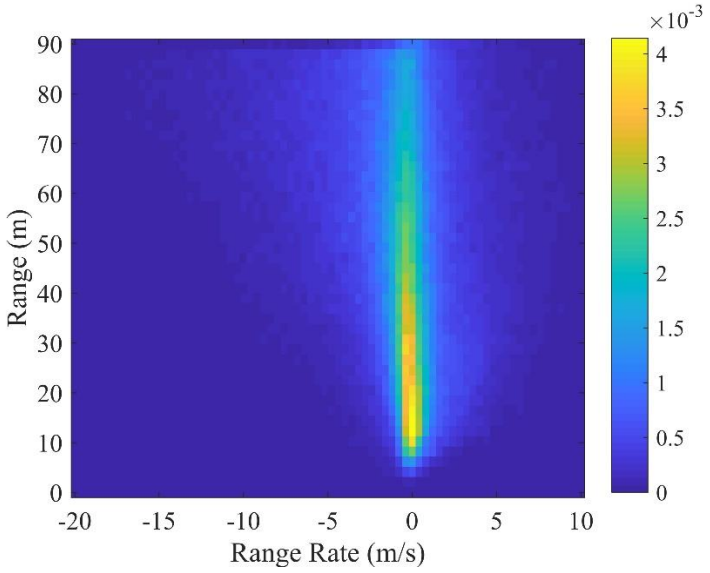



Figure 5 The probability of the cut-in range and range rate in NDD, i.e., $P(x)$.

SM denotes what we know about the common features of CAVs. In this proejct, one commonly used model, i.e., Full Velocity Difference Model (FVDM) [35], is adopted as the SM as the car-following behaviors of CAVs after the cut-in event as

$$u(k+1) = C_0\left[V_1 + V_2 \tanh(C_1(R(k) - L) - C_2) - \dot{R}(k)\right],$$

where $u(k+1)$ denotes the acceleration of the CAV at time step $k+1$, $C_0$, $V_1$, $V_2$, $C_1$, $L$, and $C_2$ are constant parameters. Similar to [15], the constraints of acceleration and velocity are added to make the model more practical, i.e., model accident-prone behaviors, as

$$v_{min} \le v \le v_{max}, a_{min} \le u \le a_{max}.$$

All parameters in [35] are adopted and are calibrated by SPMD data as listed in Table *2*. Figure *6* shows the safety performance of the constructed SM, where the SM has accidents in the yellow region.

Table 2 The values of the parameters for the cut-in case.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $C_0$ | 0.85 | $V_1$ | 6.75 |
| $V_2$ | 7.91 | $C_1$ | 0.13 |
| $L$ | 5 | $C_2$ | 1.57 |
| $v_{min}$ | 2 | $v_{max}$ | 40 |

| | | | |
|---|---|---|---|
| $a_{min}$ | -4 | $a_{max}$ | 2 |
| $P_{th}$ | 0.7 | $w$ | 0.5 |
| $\gamma$ | 0.5 | $\beta$ | 0.1 |

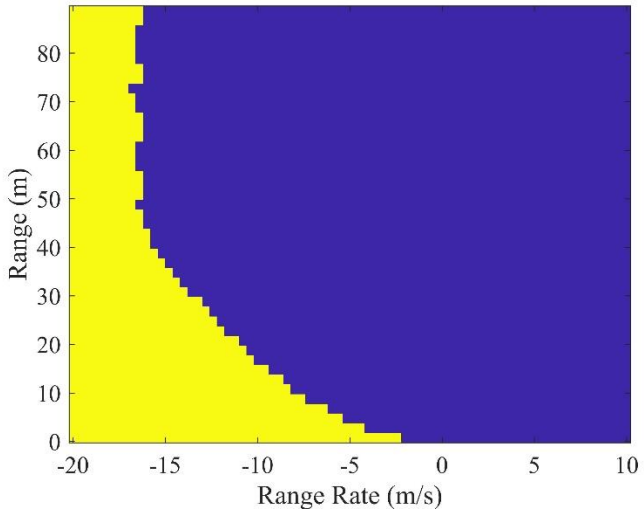



Figure 6 Safety performance of the SM.

To obtain critical scenarios and construct the library, the threshold for critical scenarios is determined as

$$P(x|S) > \frac{1}{N(X)} = 2.9 \times 10^{-4},$$

where $N(X)$ denotes the total number of scenarios as $N(X) = 47 \times 76 = 3{,}420$. The range and range rate are discretized by $2m$ and $0.4m/s$ respectively, and their boundaries are $(0,90]$ and $[-20,10]$. Figure 7 shows the obtained probability distribution combining both exposure frequency Figure 5 and maneuver challenge Figure 6. The colors denote the sampling probabilities of the scenarios, i.e., $q(x)$ in Eq. (2-9). In this case, the generated library contains a total number of 342 critical scenarios, which is about $10\%$ of all scenarios.

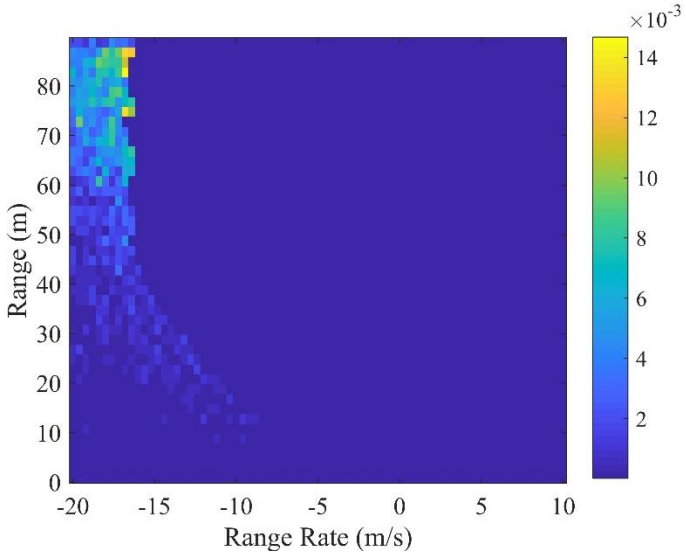



Figure 7 The offline generated library of the cut-in case for safety evaluation based on the FVDM.

### 3.3 Adaptive Library Generation

After the offline scenario library is generated, 50 scenarios are sampled as initial testing scenarios (Step 1 in Algorithm 1). Then 50 iterations of adaptive testing are conducted (Step 2 in Algorithm 1). The MATLAB toolbox in [37] is utilized to execute the GPR/GPC. The squared exponential with automatic relevance determination covariance function is applied for the regression and classification as

$$k(x, x') = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_d - x'_d}{\lambda_d}\right)^2\right],$$

where $D$ denotes the dimensions of $x$. $\sigma_f$ and $\lambda_d$ are hyperparameters, which are determined by optimizing the marginal likelihood [19]. Since $\lambda_d$ determines the relevancy of input features to the regression and classification, the covariance function is called "automatic relevance determination".

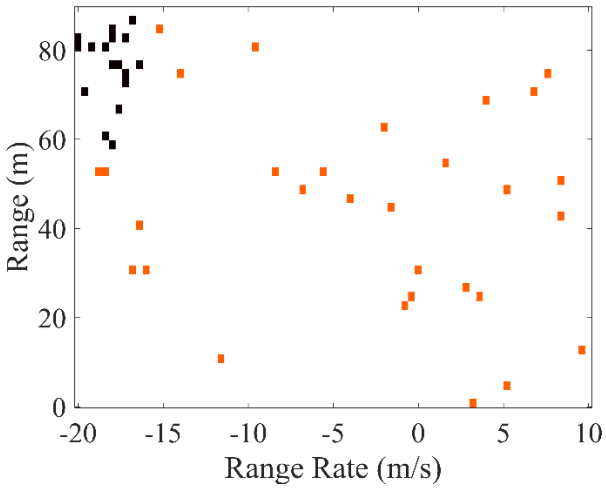
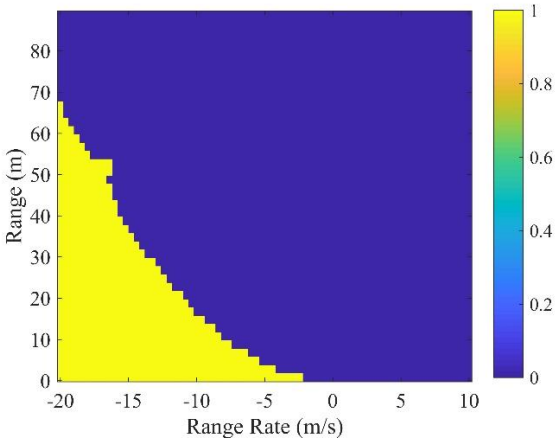
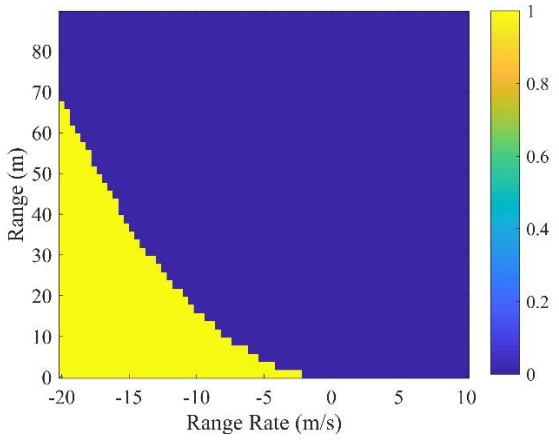



Figure 8. Results of 50 initial testing scenarios, where the black dots denote the dissimilar scenarios, and the orange dots denote the similar scenarios.

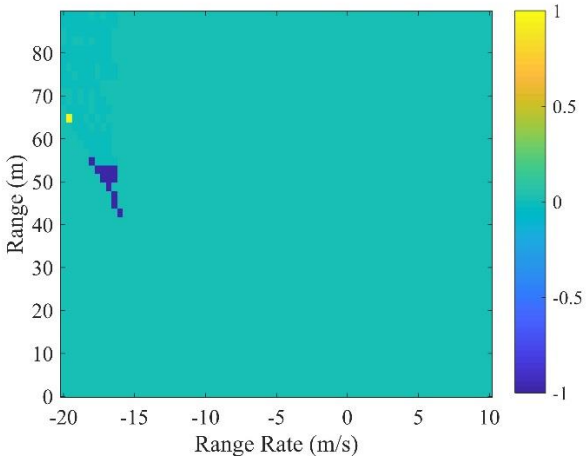
Figure 8-10 show the results of the adaptive library generation process. The initial testing results are shown in Figure 8, where the black dots denote the dissimilar scenarios, and the orange dots denote the similar scenarios. A sketch of the dissimilarity function is obtained. As shown in Figure 9 (a), after 5 iterations of adaptive testing and library generation, dissimilarities between the SM and the CAV are much decreased. Fig. 9 (e) shows that the acquisition function can capture both the classification uncertainty and the regression variances. After 50 iterations, the SM has been well developed and the dissimilarities are almost eliminated, as shown in Figure 9 (b) and (d). Compared with the offline generated library in Figure 7, the improved library in Figure 10 has been changed significantly. If more adaptive test budget is allocated, the acquisition function can further improve the SM.



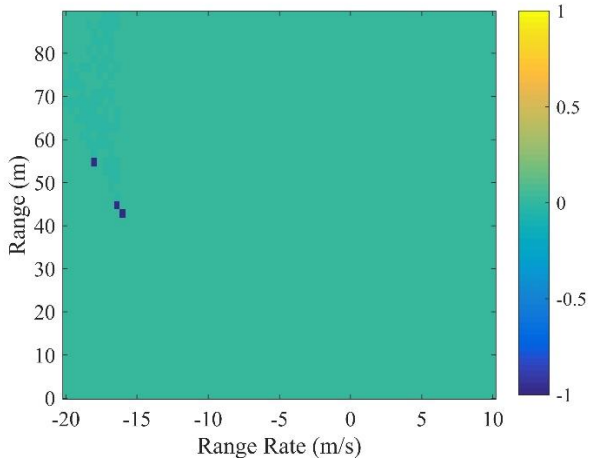

(a) Iteration 5: SM



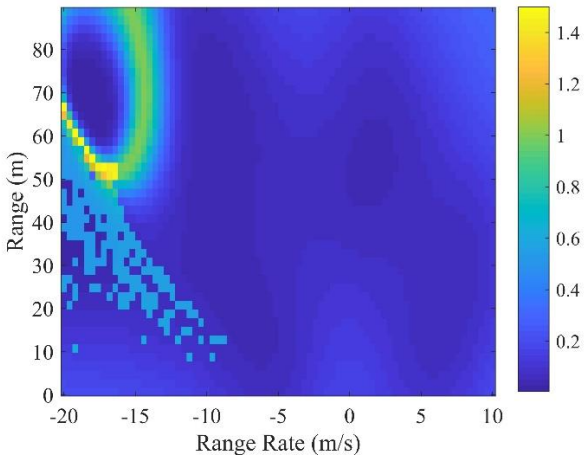

(b) Iteration 50: SM



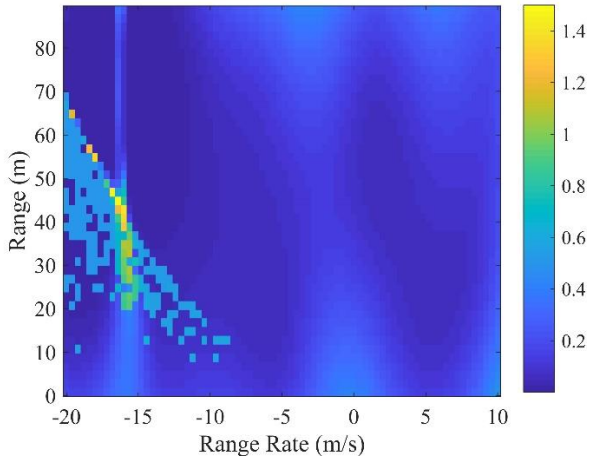

(c) Iteration 5: Remaining Dissimilarities




(d) Iteration 50: Remaining Dissimilarities




(e) Iteration 5: Acquicision Function




(f) Iteration 50: Acquicision Function

Figure 9. The results of the adaptive library generation for the cut-in case.

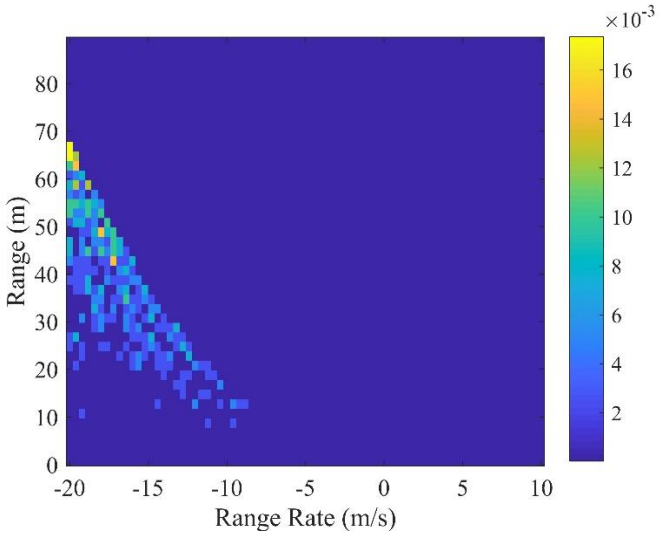
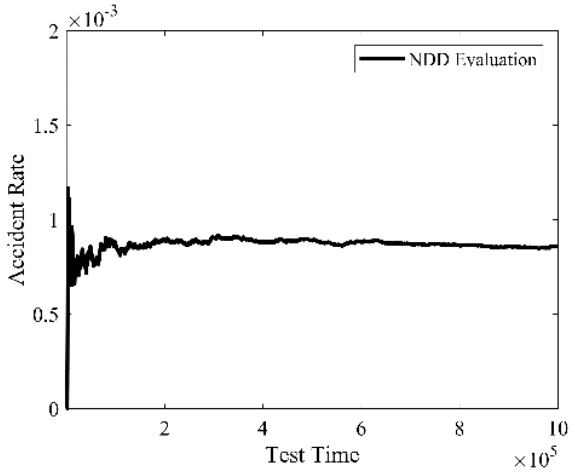


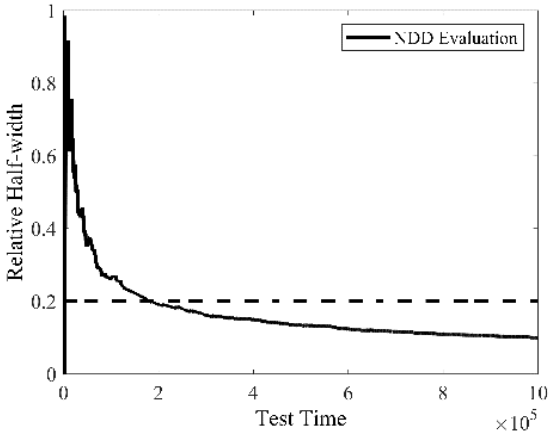

Figure 10. The improved library of the cut-in case for safety evaluation.

### 3.4 CAV Evaluation

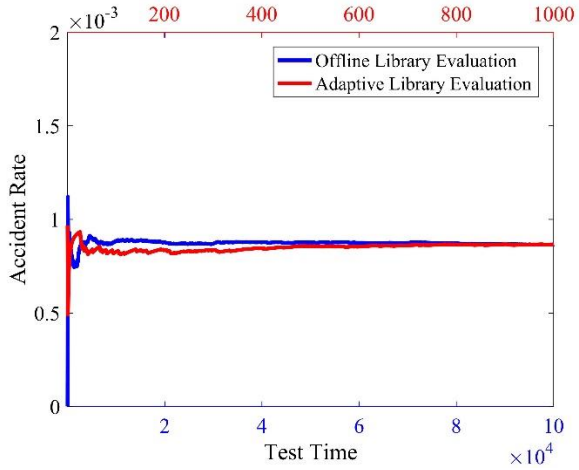
Based on the improved library, the CAV evaluation process is then performed. The accident rate of the CAV is estimated by public road test method as in Eq. (2-3), offline library evaluation method as in Eq. (2-4), and the adaptive library evaluation method. Figure 11 shows the evaluation results of the cut-in case. Results show that all three methods can converge to the same accident rate after sufficient number of tests (Figure 11 (a) and (c)). To compare the convergence speed, the relative half-width is estimated by Eq. (2-5) with the three methods in Figure 11 (b), Figure 11 (d, blue line), and Figure 11 (d, red line) respectively. To reach the 0.2 relative half-width, the total required number of tests are $1.9 \times 10^5$, 2,090, and 121 respectively. Note that the 121 tests with the adaptive library evaluation method already include 100 tests during the adaptive library generation process. Therefore, the proposed method in this project accelerates the evaluation process by 1570 times and 17 times respectively. Figure 11 (e) shows the numbers of required tests with different required relative half-widths (i.e., precision). With higher precision requirements (i.e., decreasing of the relative half-width), the original TSLG method becomes very inefficient, because the dissimilarities can't be eliminated.


×10^-3
NDD Evaluation
Accident Rate
Test Time
×10^5

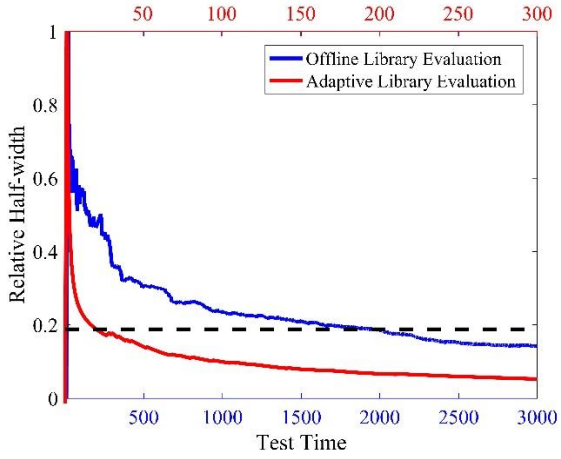
(a)


NDD Evaluation
Relative Half-width
Test Time
×10^5

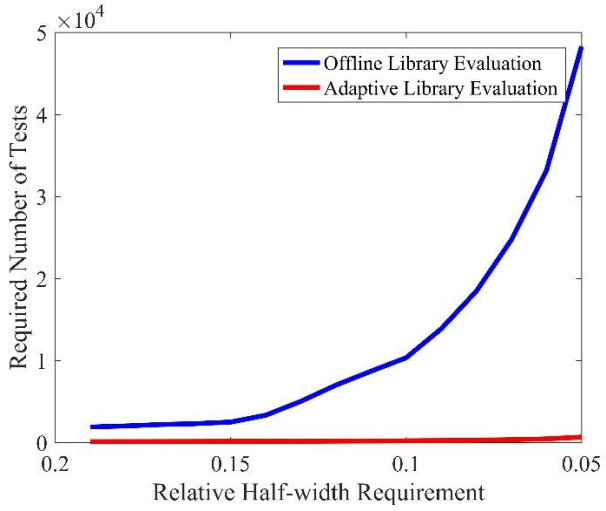
(b)


×10^-3
Offline Library Evaluation
Adaptive Library Evaluation
Accident Rate
Test Time
×10^4

(c)


Offline Library Evaluation
Adaptive Library Evaluation
Relative Half-width
Test Time

(d)




(e)

Figure 11. The evaluation results of the cut-in case with NDD evaluation (a, b), offline library evaluation (c-e, blue line), and adaptive library evaluation (c-e, red line).

## 4. Findings and Recommendations

In this project, we proposed an adaptive testing scenario library generation method for CAV evaluation. The major idea is to generate the customized library by compensating the dissimilarities between SM and CAV through an adaptive testing process. To leverage each test of CAV, the Bayesian optimization scheme is applied. A classification-based Gaussian process regression is adopted to estimate the non-stationary dissimilarity function, and a new acquisition function is designed to determine new testing scenarios in each iteration. A cut-in case is investigated for safety evaluation. Comparing with the TSLG method, the total number of required tests is further decreased by a few orders of magnitude (e.g., 10-100 times). More importantly, the acceleration of the evaluation process is more prominent if higher evaluation precision is required. To the best of our knowledge, this is the first study that identifies the adaptive TSLG problem and solves it systematically. It provides guidelines in generating testing scenario libraries for closed testing facilities to enable accurate and efficient CAV evaluation.

## 5. Outputs

The following outputs were generated during the performance of this project:

- Conference Presentations: 2020 TRB Annual Meeting and 2020 Automated Vehicle Symposium

- Journal Paper: Feng, S., Feng, Y., Sun, H., Zhang, Y., & Liu, H. X. (2020). Testing scenario library generation for connected and automated vehicles: an adaptive framework. IEEE Transactions on Intelligent Transportation Systems. DOI: 10.1109/TITS.2020.3023668.


## 6. Impacts

The impacts from the development of an adaptive testing scenario generation framework are significant. This has the potential to save automobile manufacturers and their suppliers millions of dollars in testing by improving the testing process with a few magnitudes. With the proposed framework, the automobile manufacturers don't need to deploy real vehicles on the road to perform NDD evaluation for billions of miles to collect statistic significant result. This cost savings can be cascaded to consumers, making the cost of a CAV more affordable. In turn, this may increase the penetration of CAVs faster.

## References

[1] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, “Defining and substantiating the terms scene, situation, and scenario for automated driving,” in 2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE, 2015, pp. 982–988.

[2] Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, jun 2018. [Online]. Available: https://doi.org/10.4271/J3016-201806/

[3] D. Jung, D. Jung, C. Jeong, Y. Kou, and H. Peng, “Worst case scenarios generation and its application on driving,” SAE Technical Paper, Tech. Rep., 2007.

[4] H. Hunger, “Test specifications for highly automated driving functions: Highway pilot,” Tech. Rep., 2017. [Online]. Available: https://www.pegasusprojekt.de

[5] L. Li, W.-L. Huang, Y. Liu, N.-N. Zheng, and F.-Y.Wang, “Intelligence testing for autonomous vehicles: A new approach,” IEEE Transactions on Intelligent Vehicles, vol. 1, no. 2, pp. 158–166, 2016.

[6] 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.

[7] D. Zhao, X. Huang, H. Peng, H. Lam, and D. J. LeBlanc, “Accelerated evaluation of automated vehicles in carfollowing maneuvers,” IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 733–744, 2018.

[8] L. Li, Y.-L. Lin, N.-N. Zheng, F.-Y. Wang, Y. Liu, D. Cao, K. Wang, and W.-L. Huang, “Artificial intelligence test: a case study of intelligent vehicles,” Artificial Intelligence Review, vol. 50, no. 3, pp. 441–465, 2018.

[9] G. E. Mullins, P. G. Stankiewicz, R. C. Hawthorne, and S. K. Gupta, “Adaptive generation of challenging scenarios for testing and evaluation of autonomous vehicles,” Journal of Systems and Software, vol. 137, pp. 197–215, 2018.

[10] S. Zhang, H. Peng, D. Zhao, and H. E. Tseng, “Accelerated evaluation of autonomous vehicles in the lane change scenario based on subset simulation technique,” in 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 3935–3940.

[11] M. Koren, S. Alsaif, R. Lee, and M. J. Kochenderfer, “Adaptive stress testing for autonomous vehicles,” in 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 1–7.

[12] L. Li, X. Wang, K. Wang, Y. Lin, J. Xin, L. Chen, L. Xu, B. Tian, Y. Ai, J. Wang et al., “Parallel

testing of vehicle intelligence via virtual-real interaction," Sci. Robot, vol. 4, 2019.

[13] L. Li, N. Zheng, and F.-Y. Wang, "A theoretical foundation of intelligence testing and its application for intelligent vehicles," IEEE Transactions on Intelligent Transportation Systems, 2020, doi:10.1109/TITS.2020.2991039.

[14] S. Feng, Y. Feng, C. Yu, Y. Zhang, and H. X. Liu, "Testing scenario library generation for connected and automated vehicles, Part I: Methodology," IEEE Transactions on Intelligent Transportation Systems, 2020, doi:10.1109/TITS.2020.2972211.

[15] S. Feng, Y. Feng, H. Sun, S. Bao, Y. Zhang, and H. X. Liu, "Testing scenario library generation for connected and automated vehicles, Part II: Case Studies," IEEE Transactions on Intelligent Transportation Systems, 2020, doi:10.1109/TITS.2020.2988309.

[16] S. Feng, Y. Feng, X. Yan, S. Shen, S. Xu, and H. X. Liu, "Safety assessment of highly automated driving systems in test tracks: A new framework," Accident Analysis &

Prevention, vol. 144, p. 105664, 2020.

[17] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in Advances in neural information processing systems, 2012, pp. 2951–2959.

[18] P. I. Frazier, "A tutorial on bayesian optimization," arXiv preprint arXiv:1807.02811, 2018.

[19] C. K. Williams and C. E. Rasmussen, Gaussian processes for machine learning. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[20] R. Y. Rubinstein and D. P. Kroese, Simulation and the Monte Carlo method. John Wiley & Sons, 2016, vol. 10.

[21] 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.

[22] S. M. Ross, Introductory statistics. Academic Press, 2017.

[23] A. B. Owen, Monte Carlo theory, methods and examples, 2013. [Online]. Available: https://statweb.stanford.edu/owen/mc/

[24] J. Deshmukh, M. Horvat, X. Jin, R. Majumdar, and V. S. Prabhu, "Testing cyber-physical systems through bayesian optimization," ACM Transactions on Embedded Computing Systems (TECS), vol. 16, no. 5s, pp. 1–18, 2017.

[25] L. Schultz and V. Sokolov, "Bayesian optimization for transportation simulators," Procedia computer science, vol. 130, pp. 973–978, 2018.

[26] T. Otsuka, H. Shimizu, T. Iwata, F. Naya, H. Sawada, and N. Ueda, "Bayesian optimization

for crowd traffic control using multi-agent simulation," in 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 1981–1988.

[27] X. Chen, Z. He, and L. Sun, "A bayesian tensor decomposition approach for spatiotemporal traffic data imputation," Transportation research part C: emerging technologies, vol. 98, pp. 73–84, 2019.

[28] J. Duan, F. Gao, and Y. He, "Test scenario generation and optimization technology for intelligent driving systems," IEEE Intelligent Transportation Systems Magazine, 2020.

[29] T. Liu, Y. Liu, J. Liu, L. Wang, L. Xu, G. Qiu, and H. Gao, "A bayesian learning based scheme for online dynamic security assessment and preventive control," IEEE Transactions on Power Systems, vol. 35, no. 5, 2020.

[30] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of bayesian optimization," Proceedings of the IEEE, vol. 104, no. 1, pp. 148–175, 2015.

[31] A. H. Guide, Infinite dimensional analysis. Springer, 2006.

[32] D. H. Wolpert, W. G. Macready et al., "No free lunch theorems for optimization," IEEE transactions on evolutionary computation, vol. 1, no. 1, pp. 67–82, 1997.

[33] R. B. Gramacy, Bayesian treed Gaussian process models. Citeseer, 2005.

[34] D. Bezzina and J. Sayer, "Safety pilot model deployment: Test conductor team report," Report No. DOT HS, vol. 812, p. 171, 2014.

[35] J. W. Ro, P. S. Roop, A. Malik, and P. Ranjitkar, "A formal approach for modeling and simulation of human car-following behavior," IEEE transactions on intelligent transportation systems, vol. 19, no. 2, pp. 639–648, 2017.

[36] S. Hamdar and H. Mahmassani, "From existing accident free car-following models to colliding vehicles: exploration and assessment," Transportation Research Record: Journal of the Transportation Research Board, no. 2088, pp. 45–56, 2008.

[37] C. E. Rasmussen, "Gaussian processes in machine learning," in Summer School on Machine Learning. Springer, 2003, pp. 63–71.