Adaptive Testing Scenario Library Generation for Connected and Automated Vehicles

Shuo Feng, Yiheng Feng, Haowei Sun, Yi Zhang, Member, IEEE, and Henry X. Liu, Member, IEEE

Abstract—How to generate testing scenario library 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 library. In this paper, an adaptive testing scenario library generation (ATSLG) method is proposed to solve this problem. A customized testing scenario library for a specific CAV model will be generated as the result of the adaptive process. To estimate the performance dissimilarities and leverage each test of the CAV, Bayesian optimization techniques are applied with classification-based Gaussian Process Regression and a new-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 and a highway exit case are studied for safety and functionality evaluation respectively. For both two cases, the proposed method can further accelerate the evaluation process by a few orders of magnitudes.

Index Terms—Connected and Automated Vehicles, Testing Scenario Library, Adaptive Testing and Evaluation, Bayesian Optimization

I. 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 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][4][5][6][7][8] (see [10] and references therein). However, most existing methods have limitations in either scenario types that can be handled (e.g., low-dimensional scenarios only), CAV models that can be applied (e.g., a specific CAV only), or performance metrics that can be evaluated (e.g., safety evaluation only).

To overcome these limitations, a systematic framework was proposed in our previous studies [10][11]. 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 we assume the exact model of CAV 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 precision of CAV evaluation, therefore may become the major source of evaluation inefficiency.

Two types of suboptimal scenarios can be identified, as shown in Fig. 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 difference between these two libraries include all underweight and overweight scenarios.

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

The goal of this paper 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 Fig. 2. The customization process starts with the test of CAV using a small set 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
Offline Generated Library

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

threshold for the dissimilarity is reached. With the customized library, the CAV can be tested and evaluated in a more efficient manner, if compared with the evaluation method utilizing the offline generated library.

For the adaptive testing process, to leverage each test of CAV, Bayesian optimization techniques [12][13] are applied. The classification-based Gaussian Process Regression (GPR) [14] is used to estimate the nonstationary performance dissimilarities, and a new acquisition function is designed to determine the most informative testing scenario at 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 and a highway-exit case are studied in similar settings to those in [11]. If compared with the framework in [10], the new framework can further accelerate the evaluation process by a few orders of magnitudes, e.g., 10-100.

The rest of this paper is organized as follows. For the convenience of the readers, Section II briefly revisits the offline library generation method discussed in [10][11]. In Section III, the problem of adaptive testing process is formulated. The Bayesian optimization based method is elaborated in Section IV. In Section V and Section VI, the cut-in case and highway-exit case are presented to demonstrate the performance of the proposed method. Finally, Section VII concludes the paper.

II. REVISIT THE TSLG METHOD

The goal of the TSLG method [10] is to generate a set of critical scenarios, which can be used to evaluate CAVs for certain performance indices. If an event of interest is denoted as $A$, e.g., an accident event, the performance index can be defined as its occurrence probability:

$$P(A|\theta) = \sum_{x \in \mathbb{X}} P(A|x, \theta)P(x|\theta),$$

where $x$ denotes the decision variables of testing scenarios (e.g., maneuvers of background vehicles), $\mathbb{X}$ denotes the feasible set of $x$, and $\theta$ denotes the pre-determined parameters by the ODD. Since $\theta$ keeps constant for a certain ODD, it will be omitted from now on to simplify the notations. So, the Eq. (1) is rewritten as

$$P(A) = \sum_{x \in \mathbb{X}} P(A|x)P(x).$$

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 test 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

$$P(A) = \frac{1}{n} \sum_{i=1}^{n} P(A|x_i), x_i \sim P(x),$$

where the last two equations are derived by Monte Carlo theory [15]. Here the cut-in scenarios on public roads follow the naturalistic distribution, i.e., $x_i \sim P(x)$. Because the accident event $A$ under the naturalistic driving environment is very rare, the required number of tests is intolerably large for reasonable estimation precision [16].

To mitigate this issue, importance sampling techniques were applied by [5] as

$$P(A) = \sum_{x \in \mathbb{X}} P(A|x)P(x),$$

$$= \sum_{x \in \mathbb{X}} P(A|x) \frac{q(x)}{q(x)},$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \frac{P(A|x_i)P(x_i)}{q(x_i)}, x_i \sim q(x),$$

where $q(x)$ denotes an importance function satisfying

$$q(x) \in [0, 1], \sum_{x \in \mathbb{X}} q(x) = 1, P(x) > 0 \Rightarrow q(x) > 0.$$
increase the testing priority of critical scenarios, the evaluation efficiency can be improved.

For a certain estimation precision, the minimal number of tests is determined by the importance function, and the required estimation precision can be measured by relative half-width for a given confidence level [17]. With the confidence level at 100(1 − α)%%, the relative half-width is defined as

\[
l_r = \Phi^{-1}(1-\alpha/2) \frac{\text{Var}(\mu_A)}{\mu_A},
\]

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

\[
n \geq \left( \frac{\Phi^{-1}(1-\alpha/2)}{\mu_A \beta} \right)^2 \sigma^2.
\]

Therefore, the evaluation process has higher efficiency with a smaller \( \sigma^2 \). By importance sampling theory [18], the estimation variance can be derived as

\[
\sigma^2 = \sum_{x \in \mathcal{X}} \frac{(P(A|x)P(x))^2}{q(x)} - \mu_A^2,
\]

which is determined by the importance function. To obtain an importance function with small variance, a heuristic searching method was proposed in [5], 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 remains a huge challenge.

To solve this problem, the scenario criticality was newly defined in [10] as a combination of maneuver challenge \( (P(S|x)) \) and exposure frequency \( (P(x)) \) as

\[
V(x) \overset{\text{def}}{=} P(S|x)P(x),
\]

where \( S \) denotes the event of interest with the SM of CAVs. Integrated with a \( \varepsilon \)-greedy sampling policy, the importance function is essentially constructed as

\[
q(x) = \begin{cases} 
(1-\varepsilon)V(x)/W, & x \in \Phi \\
\varepsilon/N(\mathcal{X}) - N(\Phi), & x \notin \Phi
\end{cases}
\]

where \( \Phi \) denotes the set of critical scenarios (i.e., the library), \( N(\mathcal{X}) \) and \( N(\Phi) \) denote the scenario numbers of the sets, and \( W \) is a normalization factor as

\[
W = \sum_{x \in \Phi} V(x).
\]

The constructed importance function was justified by theoretical analysis and case studies regarding evaluation accuracy and efficiency were provided in [10][11].

As discussed above, maneuver challenge \( (P(S|x)) \) is evaluated by using a 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 adaptive testing scenario library generation (ATSLG) is formulated and addressed in this paper.

### III. Problem Formulation

In this Section, the problem of ATSLG is formulated as a Bayesian optimization problem. The ATSLG problem is analyzed in Subsection III-A, and the Bayesian optimization scheme is presented in Subsection III-B.

#### A. ATSLG Problem

The goal of the ATSLG is to minimize the estimation variance \( \sigma^2 \) by a minimized number of tests. As discussed above, the key to minimizing \( \sigma^2 \) is to reduce the performance dissimilarities between the SM and the test CAV. The dissimilarity function is defined as

\[
f(x) \overset{\text{def}}{=} P(A|x) - P(S|x), x \in \mathcal{X}.
\]

Every test of the CAV will provide one observation of \( f(x) \). Here we denote \( \tilde{f}(x) \) as an estimation of \( f(x) \), then the SM can be updated as

\[
P(S'|x) = P(S|x) + \tilde{f}(x), x \in \mathcal{X},
\]

where \( S' \) denotes the event of interest with the updated SM. Substituting it into Eq. (9) and Eq. (10), the constructed importance function can be derived as

\[
q = h(\tilde{f}),
\]

where \( h \) denotes a mapping from \( \tilde{f} \) to \( q \). Substituting Eq. (14) into Eq. (6), the estimation variance can be rewritten as a function of \( \tilde{f} \). Finally, the problem of the ATSLG can be formulated as

\[
\min_{\tilde{f} \in \mathcal{F}} \sigma^2(\tilde{f}),
\]

where \( \mathcal{F} \) denotes the feasible function space of \( \tilde{f} \).

As indicated in Theorem 2 in [10], the optimal solution of Eq. (15) is that the dissimilarity function is exactly known, i.e., \( \tilde{f}^* = f \). In general, more observations of \( f \) can lead to more accurate estimation of \( \tilde{f} \). However, each observation of \( f \) is time-consuming and cost-expensive. Therefore, the objective function should be optimized by as few observations as possible. To this end, the Bayesian optimization scheme is applied.

#### B. Bayesian Optimization Scheme

Bayesian optimization is an approach to optimize an unknown function \( f(x) \) by as few observations as possible [12][13]. 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. The prior probability model is usually constructed based on prior knowledge of the problem. For the ATSLG problem, the prior knowledge comes from the SM and the offline generated library. To decide next point for observation, various acquisition functions have been proposed [13], e.g., expected improvement, knowledge gradient, entropy search, and predictive.
entropy search. With a properly designed acquisition function, each observation of \( f(x) \) can be leveraged. Integrating the prior knowledge and observations, the estimation of \( f(x) \), \( \hat{f}(x) \), can be obtained by regression methods. Then the SM as well as the library can be improved. The overall scheme is described in Algorithm 1.

Algorithm 1: The overall scheme of the ATSLG process.

Input: SM and offline generated library
Output: Customized library
1 Step 1: Observe \( f \) with initial testing scenarios. (Sec 4.A)
2 Step 2: while the stop criteria (e.g., budget or estimation precision) is not satisfied do
3  \hspace{1cm} Step 2.1: Obtain the estimation \( \hat{f} \) (Sec 4.B);
4  \hspace{1cm} Step 2.2: Update SM and library (Sec 4.C);
5  \hspace{1cm} Step 2.3: Decide next iteration of testing scenarios (Sec 4.D);
6  \hspace{1cm} Step 2.4: Observe \( f \) by testing the CAV with new scenarios;
7 end

IV. ADAPTIVE TESTING SCENARIO LIBRARY GENERATION

In this section, all steps of algorithm 1 are elaborated: at Step 1, a sampling mechanism of initial testing scenarios is designed to provide a sketch of the dissimilarity function; at Step 2.1, a classification-based GPR method is applied to estimate the dissimilarity function; at Step 2.2, the SM and library are improved based on the estimated dissimilarity function; and at Step 2.3, an acquisition function is designed to decide next iteration of testing scenarios.

A. Initial Testing Scenarios

The goal of the initial testing scenarios is to provide a sketch of the dissimilarity function. The major difficulty lies in the trade-off between exploitation of the offline generated library and exploration outside the library. 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 \( \gamma \) (e.g., 0.5). Similar to the “No Free Lunch Theorem” [19], 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

\[
P(x_0) = \begin{cases} 
(1 - \gamma)V(x_0)/W, & x_0 \in \Phi, \\
\gamma/(N(\mathcal{X}) - N(\Phi)), & x_0 \in \mathcal{X}\setminus\Phi,
\end{cases}
\]

where \( x_0 \) denotes an initial testing scenario.

B. Classification-based Gaussian Process Regression

In this paper, the Gaussian process (GP) is applied to provide a prior probabilistic model for the dissimilarity function. 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 as

\[
f(x) \sim \mathcal{G}\mathcal{P}(m(x), k(x, x'))
\]

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

\[
m(x) = E(f(x)),
\]

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

Based on the GP, Gaussian process regression (GPR) can be applied to estimate the values of \( f(x) \) for unobserved scenarios. Denote \( N \) points of scenarios with observations as \( \mathcal{X}_N = \{x_n \in \mathcal{X}\}^{N}_{n=1} \), and \( N^* \) points of scenarios without observations as \( \mathcal{X}_N^* = \{x_n^* \in \mathcal{X}\}^{N^*}_{n^*=1} \).

An observation of \( f(x) \) is equivalent to one test of the CAV, and the observation results are denoted as \( f(\mathcal{X}_N) \). By properties of GP, \( f(\mathcal{X}_N^*) \) can be estimated by the posterior probability distribution as

\[
f(\mathcal{X}_N^*)|f(\mathcal{X}_N) \sim \mathcal{G}\mathcal{P}\left(\hat{f}_{\mathcal{X}_N}(\mathcal{X}_N^*), \sigma_{\mathcal{P}_{\mathcal{X}_N}}^2(\mathcal{X}_N^*)\right),
\]

where the mean \( \hat{f}_{\mathcal{X}_N}(\mathcal{X}_N^*) \) and variance \( \sigma_{\mathcal{P}_{\mathcal{X}_N}}^2(\mathcal{X}_N^*) \) are determined by the observations, the mean function \( m(x) \), and the covariance function \( k(x, x') \) [14]. In this paper, the zero mean function is applied. It is worth noting that the covariance function could be heterogeneous, e.g., performances of a CAV may change more drastically in certain scenario neighborhoods than others.

To handle the heterogeneous issue, the Gaussian process classification (GPC) is incorporated. The idea is based on the treed GPR [20], which divides the variable space into several regions by a decision tree and applies GPR in each region respectively. Different from the deterministic classification method, GPC provides a probability distribution of different classes for each variable. As a result, a variable could belong to multiple classes with different probabilities and, therefore, be estimated by the GPR in each class respectively. The final estimation of the variable is the expectation of all these estimation results. In this paper, scenarios are divided into two classes, suboptimal scenarios and optimal scenarios, by the values of \( f(x) \) as

\[
y(x) = \begin{cases} 
+1, & f(x) \neq 0, \\
-1, & f(x) = 0,
\end{cases}
\]

where \( y(x) \) denotes the class label, i.e., +1 for suboptimal scenarios and -1 for optimal scenarios. The class labels of the scenarios \( \mathcal{X}_N \), i.e., \( y(\mathcal{X}_N) \), are calculated based on the observations. Let \( \mathcal{X}_{N_1} \) denote the observed suboptimal scenarios and \( \mathcal{X}_{N_2} \) denote the observed optimal scenarios. To
classify the unobserved scenarios, \( y(\mathcal{X}_{N^*}) \) can be estimated by the posterior probability as
\[
P(y(x) = +1|y(\mathcal{X}_N)), \ x \in \mathcal{X}_{N^*},
\]
where the analytic equations can be found in [20]. For notation simplification, Eq. (23) is denoted as \( P_{1,\mathcal{X}_N}(x) \), and
\[
P_{2,\mathcal{X}_N}(x) = 1 - P_{1,\mathcal{X}_N}(x).
\]

Finally, the GPC-based GPR results of \( f(x) \) can be represented as
\[
f_{\mathcal{X}_N}(x) \sim \begin{cases} \mathcal{N}(\hat{f}_{\mathcal{X}_N_1}(x), \sigma^2_{\mathcal{X}_N_1}(x)), \ & \text{with } P_{1,\mathcal{X}_N}(x), \\ \mathcal{N}(\hat{f}_{\mathcal{X}_N_2}(x), \sigma^2_{\mathcal{X}_N_2}(x)), \ & \text{with } P_{2,\mathcal{X}_N}(x), \end{cases}
\]
where \( \mathcal{N}(\hat{f}_{\mathcal{X}_N_1}(x), \sigma^2_{\mathcal{X}_N_1}(x)) \) denotes the GPR results in suboptimal scenarios, and \( \mathcal{N}(\hat{f}_{\mathcal{X}_N_2}(x), \sigma^2_{\mathcal{X}_N_2}(x)) \) denotes the results in optimal scenarios. The estimation of \( f(x) \) can be obtained by the expectation as
\[
\hat{f}_{\mathcal{X}_N}(x) = P_{1,\mathcal{X}_N}(x)\hat{f}_{\mathcal{X}_N_1}(x) + P_{2,\mathcal{X}_N}(x)\hat{f}_{\mathcal{X}_N_2}(x).
\]

\[\text{C. Surrogate Model Update and Library Generation}\]

Based on the estimation of the dissimilarity function, the SM can be updated as
\[
P(S_{\mathcal{X}_N}|x) = P(S_0|x) + \hat{f}_{\mathcal{X}_N}(x), \ x \in \mathcal{X},
\]
where \( S_0 \) denotes the offline used SM. As the rareness property of CAVs, the values of \( P(S_{\mathcal{X}_N}|x) \) are more likely to be zero for most scenarios. The Gaussian assumption, however, would produce huge number of small yet non-zero estimation values. To keep the rareness property, a set \( \mathcal{U} \) is defined as
\[
\mathcal{U} = \{P(S_0|x) = 0, P_{1,\mathcal{X}_N}(x) \leq P_{th}\},
\]
where \( P_{th} \) is a pre-determined probability threshold for classification, e.g., 0.5. Scenarios \( x \in \mathcal{U} \) are indicated unclassified by both the prior knowledge \( P(S_0|x) = 0 \) and the posterior knowledge \( P_{1,\mathcal{X}_N}(x) \leq P_{th} \). Therefore, Eq. (27) is modified as
\[
P_E(S_{\mathcal{X}_N}|x) = \begin{cases} P(S_0|x) + \hat{f}_{\mathcal{X}_N}(x), \ & x \in \mathcal{X}/\mathcal{U} \\ 0, \ & x \in \mathcal{U}. \end{cases}
\]

Based on the updated SM, a new importance function \( q_{\mathcal{X}_N}(x) \) as well as a library can be constructed by Eq. (10).

\[\text{D. Acquisition Function Design}\]

The acquisition function is designed to determine next iteration of testing scenario for the CAV. To leverage each test of the CAV, the evaluation variance \( \sigma^2 \) in Eq. (15) should be reduced as much as possible by testing each new scenario. As shown in Eq. (8), however, \( \sigma^2 \) cannot be calculated unless \( \mu_A \) is known, which is exactly what needs to be evaluated. To mitigate this issue, a point-wise index is defined as
\[
PI_{\mathcal{X}_N}(x) \overset{\text{def}}{=} \frac{[P(A_{\mathcal{X}_N}|x)P(x)]^2}{q_{\mathcal{X}_N}(x)},
\]
where \( P(A_{\mathcal{X}_N}|x) \) denotes the estimation results of \( P(A|x) \) as
\[
P(A_{\mathcal{X}_N}|x) \sim P(S_0|x) + f_{\mathcal{X}_N}(x),
\]
where \( f_{\mathcal{X}_N}(x) \) is given in Eq. (25). If compared with Eq. (8), \( PI_{\mathcal{X}_N}(x) \) measures the maximal reduction of \( \sigma^2 \) by the testing scenario \( x \).

With the new index, an acquisition function is built by the expected improvement method, which is commonly used for Bayesian optimization [13]. The expected value of \( PI_{\mathcal{X}_N}(x) \) is defined as
\[
EPI_{\mathcal{X}_N}(x) \overset{\text{def}}{=} E\left[\frac{(P(A_{\mathcal{X}_N}|x)P(x))^2}{q_{\mathcal{X}_N}(x)}\right].
\]

Applying the integration by parts and Eq. (31), the analytical form of Eq. (32) is derived as
\[
EPI_{\mathcal{X}_N}(x) = P_{1,\mathcal{X}_N}(x) \cdot EPI_{1,\mathcal{X}_N}(x) + P_{2,\mathcal{X}_N}(x) \cdot EPI_{2,\mathcal{X}_N}(x),
\]
where
\[
EPI_{1,\mathcal{X}_N}(x) = \frac{P^2(x)}{q_{\mathcal{X}_N}(x)} \left(\frac{P(S_0|x) + \hat{f}_{\mathcal{X}_N}(x)}{\sigma_{\mathcal{X}_N}(x)}\right)^2 + \sigma_{\mathcal{X}_N}(x),
\]
where \( i = 1 \) for suboptimal scenarios, and \( i = 2 \) for optimal scenarios.

To better explore the boundaries of the classification, the classification variance \( \sigma^2_{\mathcal{X}_N}(x) \) is incorporated into the acquisition function as
\[
I_{\mathcal{X}_N}(x) = w \frac{EPI_{\mathcal{X}_N}(x)}{U_E} + \frac{\sigma^2_{\mathcal{X}_N}(x)}{U_C},
\]
where \( w \) is a weight to balance the two terms, and \( U_E, U_C \) are normalization factors to make the metrics comparable. The classification variance can be calculated by the GPC method [20]. Recall that the scenarios \( x \in \mathcal{U} \) are indicated unclassified. Therefore, the acquisition function, which exploits existing information, is unlikely to explore these scenarios. To search possible “unexpected” suboptimal scenarios, a small probability (\( \beta \)) of random sampling is applied. Finally, the next iteration of testing scenario is decided by
\[
x_{N+1} = \begin{cases} \max_x I_{\mathcal{X}_N}(x), x \in \mathcal{X}/\mathcal{U}, \ & \text{with } 1 - \beta \\ \text{random sampling for } x \in \mathcal{U}, \ & \text{with } \beta \end{cases}.
\]

\[\text{V. Cut-in Case Study}\]

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

\[\text{A. Case Description}\]

The cut-in case is illustrated in Fig. 3 (a), where a background vehicle (BV) makes a lane change in front of the test CAV. Similar to previous work [5][11], the decision variables are constructed as
\[
x = (R, \dot{R}),
\]
where \( R \) and \( \dot{R} \) denote the range and range rate of two vehicles at the cut-in moment. The accident event is defined as reaching
a threshold of minimal distance between the two vehicles, i.e., $d_{\text{min}} = 1\, \text{m}$. The safety performance is evaluated by the accident rate of the CAV on public roads. A CAV car-following model used in \cite{5} \cite{11}, which combines adaptive cruise control and autonomous emergency braking functions, is evaluated.

![Fig. 3: Illustrations of the cut-in case and highway exist case.](image)

### B. Offline Library Generation

The TSLG method in \cite{10} is conducted to generate the offline library. To estimate the exposure frequency of the cut-in scenarios, NDD from the Safety Pilot Model Deployment program at the University of Michigan \cite{21} is utilized. A total number of 414,770 qualified cut-in events are successfully obtained. The location distribution of the events is shown in Fig. 4. The joint probability distribution of the cut-in range and range rate (i.e., $P(x)$) is shown in Fig. 5.

![Fig. 4: An illustration of the cut-in events distribution in Michigan area.](image)

To determine the maneuver challenge, the Full Velocity Difference Model (FVDM) \cite{23} is adopted as the SM 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 \cite{24}, the constraints of acceleration and velocity are added to make the model more practical, i.e., model accident-prone behaviors, as

$$v_{\text{min}} \leq v \leq v_{\text{max}}, a_{\text{min}} \leq a \leq a_{\text{max}}.$$  \hfill (38)

All calibrated parameters in \cite{23} are adopted as listed in Table I. Fig. 6 shows the safety performance of the constructed SM, where the SM has accidents in the yellow region. It is worth noting that, to make the dissimilarity obvious, the selected SM in this case is different from the Intelligent Driving Model adopted in \cite{11}.

| 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  |
| $v_{\text{min}}$ | 2 | $v_{\text{max}}$ | 40 |
| $a_{\text{max}}$ | 2 | $a_{\text{min}}$ | -4 |
| $P_{\text{th}}$ | 0.7 | $w$      | 0.5   |
| $\gamma$  | 0.5   | $\epsilon$ | 0.1   |

TABLE I: The parameter values of the cut-in case.

Fig. 5: The exposure frequency of the cut-in range and range rate.

Fig. 6: The maneuver challenge of the SM regarding safety.

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

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

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

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

C. Adaptive Library Generation

After the offline scenario library is generated, 50 scenarios are sampled as initial testing scenarios, and then 50 iterations of adaptive testing are conducted. The MATLAB toolbox in [14] is utilized to execute the GPR and GPC. The squared exponential with automatic relevance determination covariance function is applied for the regression and classification as

$$k(x, x') = \sigma_f^2 \exp \left[ -\frac{1}{2} \sum_{d=1}^{D} \left( \frac{x_d - x'_d}{\lambda_d} \right)^2 \right],$$  \hspace{1cm} (40)

where $D$ denotes the dimensions of $x$. $\sigma_f$ and $\lambda_d$ are hyper-parameters, which are determined by optimizing the marginal likelihood [14].

Fig. 8-10 show the results of the adaptive library generation process. The initial testing results are shown in Fig. 8 where the black dots denote the observed suboptimal scenarios, and the orange dots denote the observed optimal scenarios. A sketch of the dissimilarity function is obtained. As shown in Fig. 9 (a), after 5 iterations of adaptive testing process, performance 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 Fig. 9 (b) and (d). If compared with the offline generated library in Fig. 7 the customized library has been improved significantly, as shown in Fig. 10.

D. CAV Evaluation

With the customized library, the CAV is further tested and evaluated. The accident rate of the CAV is estimated by on-road test method (i.e., NDD evaluation) and the evaluation method with the offline generated library (i.e., offline library evaluation) as two baselines. Results are shown in Fig. 11. The blue line denotes the results of the offline library evaluation method, and the bottom $x$-axis denotes its number of tests. The red line denotes the results of the adaptive library evaluation method, and the top $x$-axis denotes its number of tests. Results show that all three methods can converge to the same accident
rate after sufficient number of tests (Fig. 11 (a) and (c)). To compare the convergence speed, the relative half-width is estimated by Eq. (6) with the three methods in Fig. 11 (b) and Fig. 11 (d). 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 of the adaptive library evaluation method already include 100 tests at the adaptive testing process. Therefore, the proposed ATSLG method accelerates the evaluation process by 1570 times and 17 times respectively, if compared with the on-road test method and the evaluation method with the offline generated library. Fig. 12 shows the numbers of required tests with different required relative half-widths. By decreasing of the relative half-width, the evaluation precision is increasing, and the advantage of the proposed method becomes more obvious.

VI. HIGHWAY EXIT CASE STUDY

In this section, the proposed method is further demonstrated in the highway exit case for functionality evaluation. Functionality is another important performance metric, which is defined by whether a CAV can complete a given task in a specific scenario. Consider a scenario that a CAV needs to make a lane change to the right and exit the highway within a certain distance, with several BVs driving on the right lane. If the CAV is very conservative and keeps a long safety distance with surrounding vehicles, it may fail to complete the lane-change task before the freeway exit. In such case, the vehicle may pass the safety evaluation but fail in the functionality evaluation. Similar to safety evaluation, the functionality of a CAV can be evaluated by the failure rates of the CAV in completing certain driving tasks with different environment settings and BVs’ trajectories.

A. Case Description

The highway exit case is illustrated in Fig. 3 (b), where the test CAV needs to make a lane change to the right and exit the highway within a certain distance. Similar to previous work [11], for simplification of the problem, the initial position and velocity of the CAV are pre-determined as $p_0$ and $v_0$, and only two BVs are considered. The two BVs will keep their initial velocity unless the distance between them is less than a threshold $d_{cf}$. As a result, the decision variables are formulated as

$$x = (p_{0,1}, v_{0,1}, p_{0,2}, v_{0,2}),$$  \hfill (41)

where $p_{0,i}, v_{0,i}$ denote the initial position and speed of the $i$-th BV. The discrete time and position intervals are denoted as $\Delta t$ and $\Delta p$ respectively. The parameter values used in this study are summarized in Table II. The functionality performance is evaluated by the task failure rate of the CAV on public roads. The task failure event is defined as the CAV fails to make a lane change to the right and exit the highway before the off-ramp.
from about 0.12% of all scenarios. The total number of critical scenarios in the library is 1,895, testing scenario library of the highway exit case is generated.

where \( v \) and \( n \) denote the acceleration of the new follower and old follower respectively. As it is desirable to complete the lane change, the politeness factor is set close to zero, e.g., \( \rho \) is determined as
\[
\rho = 21 - \Delta p, \tag{42}
\]
and complexity of the highway exit case, the testing budget is increased as 300 initial testing scenarios and 300 iterations of adaptive testing. Similar to the cut-in case, the same covariance function and MATLAB toolbox are utilized.

### D. CAV Evaluation

Results of the highway exit case are shown in Fig. 13 from the same perspectives as the cut-in case. It is obvious that all three methods converge to the same task failure rate after sufficient number of tests. For a 0.2 relative half-width precision, the required number of tests are \( 9.35 \times 10^5 \), \( 1.58 \times 10^4 \), and \( 1,617 \), respectively. Therefore, the adaptive library evaluation method accelerates the evaluation process by 578 and 9.7 times, if compared with the on-road test method and the evaluation method with the offline generated library. Fig. 14 shows that the acceleration effectiveness becomes more significant with higher precision requirement.

\[
P(x) = P(p_{0,1})P(v_{0,1}, R, v_{0,2}), \tag{42}
\]

where \( R = p_{0,1} - p_{0,2}, P(p_{0,1}) \) denotes the probability of the initial position of the leading vehicle, which is assumed following uniform distribution, and \( P(v_{0,1}, R, v_{0,2}) \) is obtained from the distribution of car-following trajectories in the NDD.

To compute the maneuver challenge, the MOBIL (‘minimizing overall braking induces by lane changes’) model [26] is applied as the SM. It provides the utility measurement method for deciding which gap has a desirable lane change position as
\[
U_{LG} = \bar{u} - u + \rho_{LG} (\bar{u}_{new} - u_{new} + \bar{u}_{old} - u_{old}), \tag{43}
\]

where \( \bar{u} \) denotes the new acceleration of the CAV after the lane change, \( \rho_{LG} \) is the politeness factor, and \( u_{new}, u_{old} \) denote the acceleration of the new follower and old follower respectively. As it is desirable to complete the lane change, the politeness factor is set close to zero, e.g., \( \rho_{LG} = 0.1 \). To predict the CAV’s trajectories before the lane-change, the Model Predictive Control (MPC) [27] is applied, and the trajectory with higher predictive utility of lane change, i.e., \( U_{LG} \), will be chosen as the solution to the task.

Similar to the cut-in case, the threshold of critical scenarios is determined as
\[
V(x) > \frac{1}{N(X)} = 6.1 \times 10^{-7}, \tag{44}
\]

where \( N(X) = n_p^2 \times n_v^2 = 1.64 \times 10^6 \). Here \( n_p = 61 \) and \( n_v = 21 \) denote the number of feasible values of \( p_{0,i} \) and \( v_{0,j} \). After applying the critical scenario searching method, the testing scenario library of the highway exit case is generated. The total number of critical scenarios in the library is 1,895, which is about 0.12% of all scenarios.

### B. Offline Library Generation

To compute exposure frequency, NDD from the Integrated Vehicle-Based Safety System project is used [25]. A total of \( 5 \times 10^4 \) car-following events and \( 1.47 \times 10^6 \) points of car-following trajectories are obtained. The exposure frequency of a scenario can be estimated as

\[
P(\Delta t, v, m) = \left( \frac{\text{Number of car-following events}}{\text{Total number of points}} \right) \times \left( \frac{\text{Distance covered}}{\text{Scenario length}} \right),
\]

where \( \Delta t \) is the time difference between two points, \( v \) is the speed of the following vehicle, and \( m \) is the mass of the following vehicle.

The CA V lane-change model developed in [28] is used in this paper. It changes the task failure rate from about \( 10^{-3} \) to about \( 10^{-4} \). Considering the dimensions of a scenario, the threshold of critical scenarios is defined as

\[
\Delta \gamma = \frac{\text{Dimension of car-following trajectory}}{\text{Scenario length}} \times \text{Minimum safety gap}.
\]

### C. Adaptive Library Generation

The CAV lane-change model developed in [28] is used for evaluation in this case study. To make the dissimilarity obvious, the minimum safety gap \( d_s \) in [28] is changed from 1 m to 0.1 m in this paper. It changes the task failure rate from about \( 10^{-3} \) to about \( 10^{-4} \). Considering the dimensions of a scenario, the threshold of critical scenarios is defined as

\[
\Delta \gamma = \frac{\text{Dimension of car-following trajectory}}{\text{Scenario length}} \times \text{Minimum safety gap}.
\]
VII. Conclusions

In this paper, the adaptive testing scenario library generation (ATSLG) method is proposed to generate customized library for CAV testing and evaluation. Compared with the TSLG method discussed in [10][11], the proposed method is more efficient and robust.

The major idea to generate the customized library is by reducing 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 heterogeneous dissimilarity function, and a new acquisition function is designed to determine each iteration of testing scenario. Two cases studies, cut-in and highway-exit, are investigated for safety and functionality evaluation respectively. If compared with the TSLG method, the total number of required tests is further decreased by a few orders of magnitudes (e.g., 10-100 times). More importantly, the acceleration of the evaluation process is more prominent if higher precision is required.

There are still many interesting topics that can be further investigated. For example, the ATSLG problem for high-dimensional scenarios becomes more complex, and how to address the high-dimensional issue in adaptive process remains as a problem. Moreover, it is interesting to apply the proposed method in more realistic CAV testing platforms with pre-established scenario libraries.

References

[1] S. Ulbrich, T. Menzel, A. Reschka, F. Schultd, 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. [Online]. Available: https://doi.org/10.12711/3016-20180880

[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.pegassusprojekt.de]

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. Feng, Y. Feng, C. Yu, Y. Zhang, and H. X. Liu, “Testing scenario library generation for connected and automated vehicles, part i: Methodology,” arXiv preprint arXiv:1905.03419, 2019.

[11] S. Feng, Y. Feng, H. Sun, S. Bao, A. Misra, Y. Zhang, and H. X. Liu, “Testing scenario library generation for connected and automated vehicles, part ii: Case studies,” arXiv preprint arXiv:1905.03428, 2019.

[12] 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.

[13] P. I. Frazier, “A tutorial on bayesian optimization,” arXiv preprint arXiv:1807.02811, 2018.

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

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

[16] 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.

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

[18] A. B. Owen, Monte Carlo theory, methods and examples, 2013.

[19] 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.

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

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

[22] X. Gong, Y. Guo, Y. Feng, J. Sun, and D. Zhao, “Evaluation of the energy efficiency in a mixed traffic with automated vehicles and human controlled vehicles,” arXiv preprint arXiv:1806.00377, 2018.

[23] 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.

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

[25] J. R. Sayer, S. E. Bogard, M. L. Buonarosa, D. J. LeBlanc, D. S. Funkhouser, S. Bao, A. D. Blankespoor, and C. B. Winkler, “Integrated vehicle-based safety systems light-vehicle field operational test key findings report,” 2011.

[26] A. Kesting, M. Treiber, and D. Helbing, “General lane-changing model mobil for car-following models,” Transportation Research Record, vol. 1999, no. 1, pp. 86–94, 2007.

[27] J. B. Rawlings and D. Q. Mayne, Model predictive control: Theory and design. Nob Hill Pub. Madison, Wisconsin, 2009.

[28] J. Nilsson, J. Silvin, M. Brannstrom, E. Coelingh, and J. Fredriksson, “If, when, and how to perform lane change maneuvers on highways,” IEEE Intelligent Transportation Systems Magazine, vol. 8, no. 4, pp. 68–78, 2016.

Shuo Feng is currently a postdoctoral researcher at Department of Civil and Environmental Engineering at University of Michigan. He received the Ph.D. degree in Control Science and Engineering from Tsinghua University, China, in 2019 and the B.S. degree in Department of Automation from Tsinghua University, China, in 2014. He was also a visiting Ph.D. student at Department of Civil and Environmental Engineering at University of Michigan, Ann Arbor, in 2017-2019. His current research interests include connected and automated vehicle evaluation, optimal platoon control, and transportation data analysis.
Yiheng Feng is currently an Assistant Research Scientist at University of Michigan Transportation Research Institute. He graduated from the University of Arizona with a Ph.D degree in Systems and Industrial Engineering in 2015. He has a Master degree from the Civil Engineering Department, University of Minnesota, Twin Cities in 2011. He also earned the B.S. and M.E. degree from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China in 2005 and 2007 respectively. His research interests include traffic signal systems control and security, and connected and automated vehicles testing and evaluation.

Haowei Sun is currently a graduate at Department of Civil and Environmental Engineering at University of Michigan. He received the bachelors degree in Department of Automation from Tsinghua University, China, in 2019, and he visited University of Michigan for a summer research internship in 2018. His research interests include intelligent transportation, optimization method and deep reinforcement learning.

Yi Zhang received the BS degree in 1986 and MS degree in 1988 from Tsinghua University in China, and earned the Ph.D. degree in 1995 from the University of Strathclyde in UK. He is a professor in the control science and engineering at Tsinghua University with his current research interests focusing on intelligent transportation systems. His active research areas include intelligent vehicle-infrastructure cooperative systems, analysis of urban transportation systems, urban road network management, traffic data fusion and dissemination, and urban traffic control and management. His research fields also cover the advanced control theory and applications, advanced detection and measurement, systems engineering, etc.

Henry X. Liu is a Professor of Civil and Environmental Engineering at the University of Michigan, Ann Arbor and a Research Professor of the University of Michigan Transportation Research Institute. He also directs the USDOT Region 5 Center for Connected and Automated Transportation. Dr. Liu received his Ph.D. degree in Civil and Environmental Engineering from the University of Wisconsin at Madison in 2000 and his Bachelor degree in Automotive Engineering from Tsinghua University in 1993. Dr. Liu’s research interests focus on transportation network monitoring, modeling, and control, as well as mobility and safety applications with connected and automated vehicles. On these topics, he has published more than 100 refereed journal articles. Dr. Liu is the managing editor of Journal of Intelligent Transportation Systems and an associate editor of Transportation Research Part C.