Certifiable Evaluation for Autonomous Vehicle Perception Systems using Deep Importance Sampling (Deep IS)

Mansur Arief[1], Zhepeng Cen[1], Zhenyuan Liu[2], Zhiyuan Huang[3], Bo Li[4], Henry Lam[2], and Ding Zhao[1]

***Abstract*— Evaluating the performance of autonomous vehicles (AV) and their complex AI-driven functionalities to high precision under naturalistic conditions remains a challenge, especially when the failure or dangerous cases are rare. Rarity does not only require an enormous sample size for a naive method to achieve high confidence residual risk estimation, but it can also cause serious risk underestimation issues that is hard to detect. Meanwhile, the state-of-the-art rare safety-critical event evaluation approach that comes with a correctness guarantee can compute an upper bound for the true risk under certain conditions, which limits its practical uses. In this work, we propose Deep Importance Sampling (Deep IS) framework that utilizes a deep neural network to obtain an efficient less biased risk estimate, with an efficiency that is on par with that of the state-of-the-art method. In the numerical experiment evaluating the misclassification rate of a traffic sign classifier, Deep IS only needs 1/40-th of the samples required by a naive sampling method to achieve 10% relative error. Furthermore, the estimate produced by Deep IS is 10 times less conservative compared to the risk upper bound and only off by at most 10% difference to the true target. This efficient deep-learning-based IS procedure promises a highly efficient method to deal with often high-dimensional functional safety problems with rare naturalistic failure cases that are prevalent in AV domains.**

I. Introduction

Synthesis of artificial intelligence (AI), transportation sciences, and robotics have shown enormous potentials in the development of safe and efficient intelligent transportation systems. Autonomous vehicles (AVs), for instance, are now driving alongside human drivers on public roads [1], relying on various sensors, actuators, and complex algorithms to handle various tasks including perception, decision making, and controls. AI enables efficient processing of large streams of input data, supplying its predictions to downstream tasks that actualize driving maneuvers [2]. While the results have been promising and appealing under normal and common scenarios, it remains a challenge to get a precise and accurate estimate of the residual risk of AVs under whole naturalistic situations, particularly because failures and safety-critical scenarios are uncommon and rare [3]–[5]. Safety assurance thus emerges as the forefront issue to solve in order to reap the full potentials of AVs, such as improved road safety and reduced traffic congestion at scale [6], [7].

One important element of safety assurance is a rigorous evaluation framework prior to deployment. The rigor in evaluation methods plays a huge role in ensuring that not only the evaluated system passed the minimum safety bar (e.g. safer than human driver), but also the precision and uncertainties of the evaluation results are properly taken into consideration. Addressing evaluation uncertainty is particularly important for AV applications since the failure rate is tiny (e.g. 1.26 fatal cases per 100 million miles of driving [8]), but comes with serious consequences [9], thus should not be ignored.

In this regard, traditional evaluation methods are insufficient, even for functional evaluation of AVs. Formal verification [10], scenario-based approaches [11], [12], or function-based approaches [13] are limiting the test scope and could not capture the full complexity and stochasticity of the interactions between AV and its environments to maintain tractability. Even the more recent evaluation approaches, such as shadow testing, staged introduction, field tests, and simulations, are also subject to their limitations. Shadow testing approach [14] could only compare the similarity between AI decisions and human decisions, which are not always the safest. Staged introduction (e.g. geofencing [15]) only suits testing up to SAE Level 4 autonomy [13], [16]. Finally, naturalistic field operational test, despite its highest testing fidelity and popularity, is well-known to be very expensive and inefficient, requiring billions of miles to achieve meaningful statistical significance to estimate rare dangerous or fatal cases [17]. Its computer-simulated counterparts are also subject to the same problem, requiring a huge number of replications that blow up very quickly as the failures become rarer.

Earlier works have been proposed to deal with AV testing inefficiencies due to the rarity of the failure cases. Accelerated Monte Carlo methods based on Importance Sampling (IS) have been proposed in [18]–[21] with orders of magnitude of efficiency improvement. However, these methods can only deal with relatively low-dimensional problems and requires some level of information of the underlying systems. Adversarial environment generator for IS design constructed using reinforcement learning is proposed in [22], but lacks efficiency guarantee. A framework called Deep-PrAE with efficiency guarantee and suitable for black-box systems is proposed in [23], but can only compute a risk upper-bound, hence less useful for evaluations that aim to select the safest design or estimate the residual risk of an AV prototype.

[1]Mansur Arief, Zhepeng Cen, and Ding Zhao are with the Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2]Zhenyuan Liu and Henry Lam are with the Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA

[3]Zhiyuan Huang is with the Department of Management Science and Engineering, Tongji University, Shanghai, China

[4]Bo Li is with the Department of Computer Science, University of Illinois Urbana Champaign, Urbana Champaign, IL, USA

In this work, we adopt the use of a deep learning classifier to design IS from [23], but aiming at outputting an estimator that is closer to the true target. This is particularly important for evaluating AV perception systems since object detection and classification algorithms are customarily complex and ranked based on their prediction accuracy (or correspondingly their misclassification rate), often really close to one another. In this situation, the upper bound of the misclassification rate is far less meaningful compared to the true misclassification rate as a basis for selecting the best algorithms to implement on AVs.

To achieve this, we propose Deep Importance Sampling (Deep IS) framework. We combine the versatility and strength of deep learning models to approximate high-dimensional complex sets and the rigor and efficiency of IS for rare-event simulation to match the level of complexity of the evaluated system. Numerical experiments on the evaluation of a traffic sign classifier (that deals with 32 by 32 image, i.e. 1,024 input dimensions) show the potential of the method, requiring 50 times smaller sample size compared to Monte Carlo to achieve 10% relative error on a target of an order 10^{-6}.

In sum, our contribution is twofold. First, we propose a handy and practical framework to design an efficient importance sampler for large-dimensional evaluation problems. Second, we achieve similar level of efficiency performance with the state-of-the-art method and even extend it to compute a less biased estimator for the target risk. These contributions allow us to devise a robust method that can evaluate the performance of AV functionalities to a high degree of precision even under rare failure cases. Furthermore, the method can bring us closer to evaluating the safety of the intended functionality (SOTIF) of AV as a system.

II. Problem Formulation

Our evaluation task deals with estimating the failure probability of AV perception systems (usually deals with high-dimensional input). The notion of failure here means the perception system fails to provide a correct interpretation of its surrounding environment under naturalistic conditions. A similar notion of evaluation tasks that focuses on AV decision-making modules (usually deals with inputs that are low or moderately high dimensional, depending on the fidelity of the evaluated AV model) are available in [18], [20], [22]. By being "naturalistic", we mean the environment behavior, represented by a multivariate random variable $X \in \mathbb{R}^d$, follows certain natural randomness. We model it using a probability distribution p. The set of failure events given an AV model, say $\mathcal{S} \subset \mathbb{R}^d$, is a subset of all realizations that leads to AV making incorrect predictions. In terms of AV traffic sign classifier, such set may contain traffic sign classes with low representations in the training set or traffic sign images that have been perturbed heavily by noise (see illustration in Fig. 1). Note that even a high-performing classifier might still make incorrect predictions on such examples, due to its low occurrence during training.

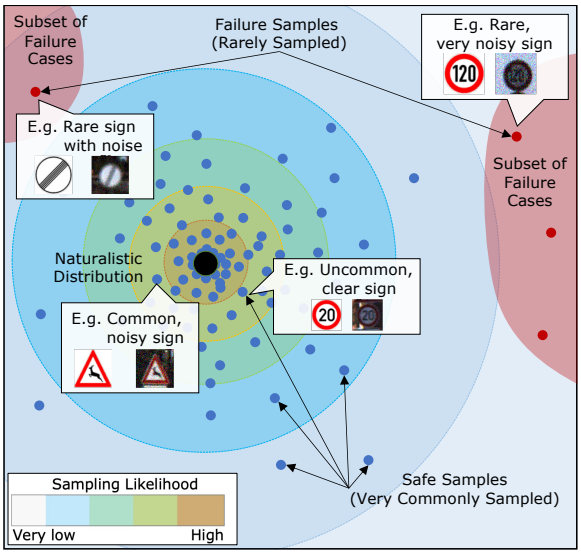

Fig. 1. Illustration of rare-event failure cases in AV traffic sign classifier. In this case, the rare failure subsets may contain traffic sign images that are likely misclassified by a trained deep learning classifiers, e.g. traffic signs with low representation in the training set or common signs subjected to very noisy perturbations. Despite the low likelihood of these dangerous cases, it could cause safety issue if occurred during deployment.

Extra efforts such as data augmentation and robust training methods might help reduce such failures to a certain extent, but the problem could persist. To account for the fact that the set might shrink as the classifier becomes more robust (e.g. through data augmentation [24] or robust training [25]), we parameterize the failure set with some parameter rarity γ, (hence written as $\mathcal{S}_\gamma$), emphasizing how the shrinkage might change the rarity of observing failure cases.

With this setting, the above goal of functional evaluation can be formulated as estimating the failure rate

$$\mu = \mathbb{P}(X \in \mathcal{S}_\gamma) = \mathbb{E}_{X\sim p}[1_{X\in\mathcal{S}_\gamma}], \tag{1}$$

where 1_ξ is the indicator function with value 1 if the statement ξ is true and 0 otherwise, and the expectation is taken with respect to naturalistic distribution p. Thus, μ represents the probability that the AV encounters failure cases under the naturalistic randomness of its operational design domains. This notion of safety is helpful because it allows ranking the alternative designs during the design process and comparing the residual risk of the selected classifier with that of human perception.

Evaluation Goal and Sample Size Requirement

The sample size requirement is related to the target confidence level of our estimator $\hat{\mu}_n$ that uses n samples to estimate μ. Since μ can be tiny, the error between $\hat{\mu}$ and μ should be measured relatively (i.e. using relative error instead of absolute error). This is important since $\hat{\mu}$ is only meaningful to make decisions about the performance if the error is small enough relative to the target μ.

Suppose that we aim to achieve ϵ relative error w.r.t. μ with high confidence and use n samples $Z_1, \cdots, Z_n$ where

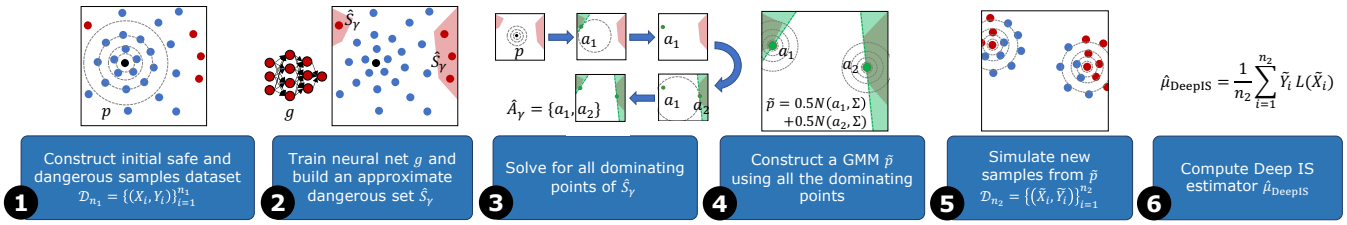

Fig. 2. Illustration of Deep IS Framework.

$Z_i = 1_{X_i \in \mathcal{S}_\gamma}$. Then, we need the sample size n to ensure

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \epsilon\mu) \leq \delta, \tag{2}$$

where δ is our target confidence level (e.g., $\delta = 5\%$). By applying Markov inequality, we obtain that the sample size n must satisfy

$$n \geq \frac{\text{Var}(Z_i)}{\mu^2 \delta \epsilon^2}. \tag{3}$$

This value of n is prohibitively large when μ is tiny, raising the well-known rare-event challenge in AV safety evaluations [19].

Importance Sampling (IS) Approaches

IS has been successfully applied to improve sampling efficiency for rare-event problems [26]. The main idea is to bias the probability measure toward the regions of failure events. This is realized by generating more failure samples from another distribution $\tilde{p}$ (instead of p), and outputs the sample average weighted by the likelihood ratio $L = \frac{dp}{d\tilde{p}}$. The critical part of this approach is to carefully select the proposal distribution $\tilde{p}$ so that the resulting estimate attains a small relative error (RE), computed as

$$RE(\hat{\mu}_n) = \frac{\text{Std}(\hat{\mu}_n)}{\mathbb{E}[\hat{\mu}_n]} \times 100\%, \tag{4}$$

i.e. the ratio between the standard deviation of an estimator to its mean. Given that an estimator $\hat{\mu}_n$ is unbiased, RE quantifies its uncertainty in estimating the target with sample size n. A certifiably efficient estimator is one that requires minimal n to attain a reasonably small RE (e.g. 10% or 20%), regardless how small the target μ is.

One known algorithm to realize this is to use a mixture model that accounts for the most likely scenarios that cover all the local modes of dangerous sets. Such most likely scenarios are called dominating point [27], [28]. Dominating-point-based IS has been adopted by earlier work in this direction under various settings. In [29], the authors derived an efficient IS distribution for the evaluation of simple lane-change AV scenarios, an improvement to the earlier work [18] that uses heuristic approaches. In [19], the authors extended the approach to deal with a dynamic system with unimodal parametric class and evaluated car-following scenarios. The relaxation is proposed using a simple parametric class to include piecewise models in [30], [31] and Gaussian mixture models in [20] with theoretical guarantees provided in [32]. [23] applies the idea for black-box systems, solving for the dominating points of a neural network approximation, and computes an upper bound for the failure probability of a simple AV model system with an efficiency guarantee. The limitation of existing studies is that they all focus on developing methods for relatively low-dimensional decision-making algorithms usually based on control theories and decision trees which have been lagged by the increasingly complex AI algorithms, e.g. basic longitudinal and lateral control [33], [34]. Larger systems are addressed in [22] at the cost of the level of rigor. Here, we combine these approaches and propose Deep IS framework that achieves a high-efficiency boost and suits to evaluate AV perception functionality.

III. Deep IS Framework

The key ingredient of Deep IS is the versatile use of a deep learning classifier to approximate the failure set $\mathcal{S}_\gamma$. We illustrate the framework in Fig. 2 as a series of six-step procedure.

We assume to have collected a dataset of size n_1 that is representative enough and can be used for learning the overall shape of failure set (**Step 1**). Methods that could generate dangerous or safety-critical scenarios are helpful for this step (e.g. [35], [36]), though they are often computationally expensive. As a cheaper alternative, we use learning methods, leveraging the versatility of deep neural networks to learn complex decision boundary. The learning procedure trains a deep neural net g for classification/set learning task (**Step 2**), extracting prior information from dataset $\mathcal{D}_{n_1} = \{(X_i, Y_i)\}_{i=1}^{n_1}$ that consists of normal cases ($Y_i = 0$) and failure cases ($Y_i = 1$), where $Y_i = 1_{X_i \in \mathcal{S}_\gamma}, i = 1, 2, \cdots, n_1$. Our goal in this step is to train the parameter of g so that the positive prediction region

$$\hat{\mathcal{S}}_\gamma = \{x : g(x) \geq 0\} \tag{5}$$

provides an approximation for the failure set $\mathcal{S}_\gamma$. Here, we assume $\mathcal{D}_{n_1}$ is representative enough and the training procedure is proper, so that $\hat{\mathcal{S}}_\gamma \approx \mathcal{S}_\gamma$.

Next, we solve for the set of dominating points $\hat{A}_\gamma$ of $\hat{\mathcal{S}}_\gamma$ (**Step 3**). To find all the dominating points, we iteratively solve an optimization problem that minimizes the rate function I of the naturalistic distribution p over the portion of $\hat{\mathcal{S}}_\gamma$ not covered by the local region of any previously found dominating point. For Gaussian $p = N(\lambda, \Sigma)$ (i.e. a Gaussian distribution with mean λ and covariance Σ), the objective to

minimize is

$$I(x) = (x-\lambda)^T\Sigma^{-1}(x-\lambda). \tag{6}$$

The constraints in this case are

$$g(x) \geq 0 \tag{7}$$

$$(x^\dagger-\lambda)^T\Sigma^{-1}(x-x^\dagger) < 0, \forall x^\dagger \in \hat{A}_\gamma. \tag{8}$$

Constraint (7) ensures that the solution lies in the positive prediction region of g while (8) ensures that the current iteration search only on the space not covered by the dominating points found in earlier iterations. See Alg. 1 for a reference on how this is implemented iteratively. A more general formulation that goes beyond Gaussian p is available in the literature, for instance, the formulation in [23].

Next, we construct our proposal Gaussian mixture model (**Step 4**)

$$\tilde{p} = \frac{1}{|\hat{A}_\gamma|}\sum_{a\in\hat{A}_\gamma} N(a,\Sigma). \tag{9}$$

This serves a weighted sampling model with larger likelihood surrounding the dominating points. Finally, we perform final sampling, collecting n_2 samples and their label $\mathcal{D}_{n_2} = \{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^{n_2}$ (**Step 5**) and compute Deep IS estimator (**Step 6**)

$$\hat{\mu}_{\text{DeepIS}} = \frac{1}{n_2}\sum_{i=1}^{n_2} L(\tilde{X}_i)\tilde{Y}_i, \tag{10}$$

where

$$L(\tilde{X}_i) = \frac{\phi(\tilde{X}_i;\lambda,\Sigma)}{\frac{1}{|\hat{A}_\gamma|}\sum_{a\in\hat{A}_\gamma}\phi(\tilde{X}_i;a,\Sigma)}, \tag{11}$$

$\tilde{Y}_i = 1_{\tilde{X}_i\in\mathcal{S}_\gamma}$, and $\phi(\cdot,\lambda,\Sigma)$ is the density of Gaussian distribution with mean λ and covariance Σ. The full algorithm for Deep IS framework is summarized in Alg. 1.

Algorithm 1 Deep IS

1: SET LEARNING WITH n_1 SAMPLES:
2: Train classifier with + prediction $\hat{\mathcal{S}}_\gamma = \{x : g(x) \geq 0\}$
3: DOMINATING SET W.R.T. $p = N(\lambda,\Sigma)$ AND $\hat{\mathcal{S}}_\gamma$:
4: Initiate $\hat{A}_\gamma \leftarrow \emptyset, \mathcal{X} \leftarrow \hat{\mathcal{S}}_\gamma$
5: While $\mathcal{X} \neq \emptyset$:
6: Solve the optimization problem

$$\begin{aligned} x^* = \arg\min_x \ & (x-\lambda)^T\Sigma^{-1}(x-\lambda) \\ \text{s.t.} \ & g(x) \geq 0, \\ & (x^\dagger-\lambda)^T\Sigma^{-1}(x-x^\dagger) < 0, \ \forall x^\dagger \in \hat{A}_\gamma \end{aligned} \tag{12}$$

7: Update $\hat{A}_\gamma \leftarrow \hat{A}_\gamma \cup \{x^*\}$
8: Update $\mathcal{X} \leftarrow \mathcal{X} \cap_{x^\dagger\in\hat{A}_\gamma}\{x : (x^\dagger-\lambda)^T\Sigma^{-1}(x-x^\dagger) < 0\}$
9: COMPUTING ESTIMATOR WITH n_2 SAMPLES:
10: Sample $\tilde{X}_1, \cdots, \tilde{X}_{n_2}$ from $\tilde{p} = \frac{1}{|\hat{A}_\gamma|}\sum_{a\in\hat{A}_\gamma} N(a,\Sigma)$
11: Compute $L(\tilde{X}_i) = \frac{\phi(\tilde{X}_i;\lambda,\Sigma)}{\frac{1}{|\hat{A}_\gamma|}\sum_{a\in\hat{A}_\gamma}\phi(\tilde{X}_i;a,\Sigma)}$
12: Compute $\hat{\mu}_{\text{DeepIS}} = \frac{1}{n_2}\sum_{i=1}^{n_2} L(\tilde{X}_i)1_{\tilde{X}_i\in\mathcal{S}_\gamma}$

Practical Choices for a Tractable Optimization Problem

We note here that the architecture of g has a practical consequence in iteratively solving (12). While it is desirable to get the most accurate model g, more complex model classes lead to more complex approximation $\hat{\mathcal{S}}_\gamma$, thus a much harder optimization problem in Step 3. Since an efficient IS proposal requires a mixture model that accounts for all of the dominating points of $\hat{\mathcal{S}}_\gamma$, we need to ensure that the resulting problem is tractable.

To ensure this, we choose g to be feed-forward ReLU-activated deep neural nets. The main reason is the availability of an efficient mixed integer program formulation for a ReLU neural network in the literature [37], [38]. For this choice of g, the resulting approximation $\hat{\mathcal{S}}_\gamma$ is a union of polytopes and its boundary is piecewise linear.

In particular, suppose we use L-layer deep neural net with ReLU activation functions, and the i-th layer outputs of ReLU are denoted as

$$s_i = \max\{W_i^T s_{i-1} + b_i, 0\}, \tag{13}$$

where W_i's and b_i's are the weight and bias parameters of g. Then, we can represent the constraint $g(x) \geq 0$ in (12) can be replaced by

$$s_L \geq 0, \tag{14}$$

$$s_i \leq W_i^T s_{i-1} + b_i + M(1-z_i), \quad i = 1,\cdots,L \tag{15}$$

$$s_i \geq W_i^T s_{i-1} + b_i, \quad i = 1,\cdots,L \tag{16}$$

$$s_i \leq M z_i, \quad i = 1,\cdots,L \tag{17}$$

$$s_i \geq 0, \quad i = 1,\cdots,L \tag{18}$$

$$z_i \in \{0,1\}^{m_i}, \quad i = 1,\cdots,L \tag{19}$$

$$s_0 = x. \tag{20}$$

Constraint (14) ensures positive prediction region. Constraint (15)-(18) represent the ReLu operator. Finally, constraint (19) and (20) handle binary requirement and original data input x on layer 0, respectively. Furthermore, m_i is the number of nodes in i-th layer, and M is some practical upper-bound value. The problem size with this formulation grows linearly to the network size (the number of integer variables is equal to the number of neurons and the number of constraints is roughly four times the number of neurons). We seek to extend the model to include CNN architecture in the future works.

IV. NUMERICAL EXPERIMENTS

For our experiment, we consider the traffic sign classification problem on the German Traffic Sign Benchmark (GTSRB) dataset [39] which contains about 50,000 images of traffic signs and labels. We evaluate the performance of MicronNet [40], one of the state-of-the-art traffic sign classifiers, that achieves 98.7% accuracy on GTSRB benchmarks. We use traffic sign images sized 32 by 32 pixels.

Our Deep IS framework uses a 4-layer feed-forward ReLU-activated neural network with 32, 16, 8, and 16 neurons in each layer as g. We construct a training set with 40,000 training data, which correctly classifies correctly- or

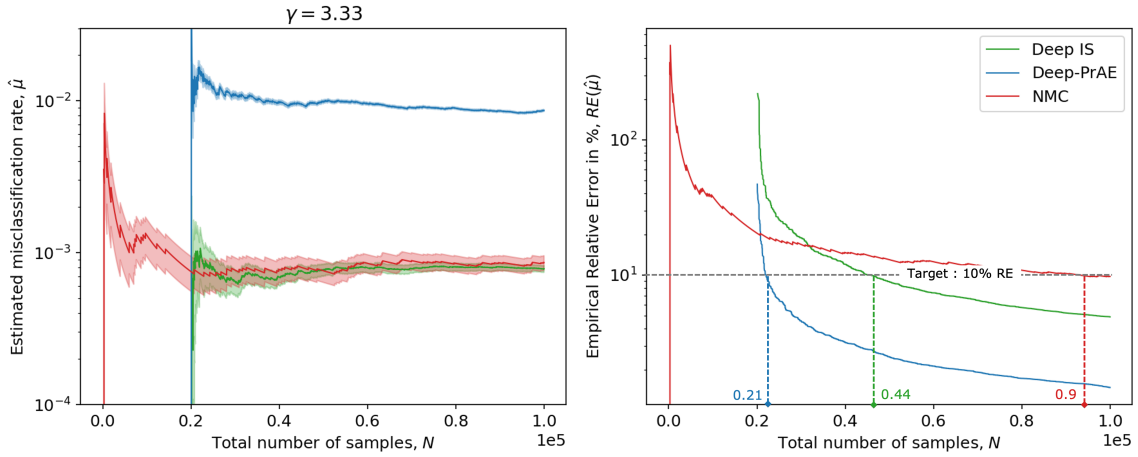

Fig. 3. Estimation of traffic sign classifier misclassification rate ((rarity level $\gamma = 3.33$ - moderately rare failure cases))

TABLE I

SAMPLE SIZE, DEGREE OF CONSERVATIVENESS, AND ACCELERATION RATE OF NMC, DEEP IS, AND DEEP-PRAE TO ACHIEVE 10% RE

Estimator	Rarity level $\gamma = 3.33$			Rarity level $\gamma = 5.0$		
	Sample Size	Degree of Conservativeness	Acceleration Rate	Sample Size	Degree of Conservativeness	Acceleration Rate
NMC	90,239	1.0 (baseline)	1.0 (baseline)	18,181,798	1.0 (baseline)	1.0 (baseline)
Deep IS	44,204	0.91	2.0	418,983	1.02	43.4
Deep-PrAE	21,203	10.02	4.3	118,370	14.26	153.6

incorrectly-classified unperturbed samples with 96.5% accuracy. We then perturb a fixed correctly classified input from each class with a Gaussian noise sampled from $N(0, \sigma^2)$ on each pixel. The parameter σ corresponds to the rarity level γ, where $\sigma = 1/\gamma$. Similar setting can be found in earlier work, e.g. [41]. We test two different rarity levels: moderate ($\sigma = 0.3$) and extremely rare ($\sigma = 0.2$), corresponding to $\gamma = 1/0.3 = 3.33$ and $\gamma = 1/0.2 = 5$, respectively. These settings represent different real-world conditions regarding how noisy traffic signs are being perceived by AV sensors.

We then run Alg. 1. We implement the sequential searching algorithm with mixed integer program formulation and use the first 100 dominating points that we obtain (which takes 72 hours). Due to the high-dimensionality of the input space and the choice of g, the number of dominating points in this problem is extremely large (much larger than 100). As benchmark, we also run Deep-PrAE (the IS-based upper-bound estimator from [42]) and NMC (naive Monte Carlo). We summarize the estimated misclassification rates and REs in Fig. 3 and Fig. 4.

In addition, we report the sample size required by each algorithm to achieve 10% RE in Table I. From this, we calculate the acceleration rate of Deep IS and Deep-PrAE, which is the ratio between NMC sample size (baseline) to Deep IS or Deep-PrAE sample size. We also report in Table I the degree of conservativeness (the ratio between the estimate of Deep IS or Deep-PrAE to that of NMC). Acceleration rate highlights the efficiency boost obtained by each method, while degree of conservativeness highlights how close the output estimator to the true (unbiased) target.

V. DISCUSSIONS

In this section, we discuss our findings in terms of estimator's conservativeness and sampling efficiency and highlight the limitation of the current work.

Estimator Conservativeness

We first observe from Fig. 3 (left) that Deep-IS estimator (solid green line) and NMC (solid red line) converge to approximately the same value ($\hat{\mu} \approx 8.6 \times 10^{-4}$), showing that Deep IS provides a close estimate of the target risk. Meanwhile the Deep-PrAE outputs 8.62×10^{-3} with $n = 100,000$, validating the upper bounds it produces.

This observation repeats in Fig. 4 (left) that deals with an even smaller misclassification rate (higher rarity parameter value, $\gamma = 5.0$ compared to the previous one with $\gamma = 3.33$). Deep IS computes a really close estimate ($\hat{\mu} = 2.04 \times 10^{-6}$) to the target ($\approx 2.0 \times 10^{-6}$), which is computed using NMC that we uses an extremely larget number of samples (more than 18 million samples!). Due to the extreme rarity of the misclassification rate, NMC only encounters 1 misclassification case in the first 100,000 samples, resulting in overly high estimation variance that is barely informative (Fig. 4 right). This is due to overall smaller noise variance added to the traffic signs, hence the classifier can identify most of the traffic signs correctly. Again, Deep-PrAE provide a valid upper bound ($\hat{\mu} = 2.88 \times 10^{-5}$), that is more than 14 times the value of the target. This shows the ability of Deep IS estimator to produce a less conservative estimate even under an extreme rarity problem.

The degrees of conservativeness summarized in Table I shows Deep IS outputs a much less conservative estimate

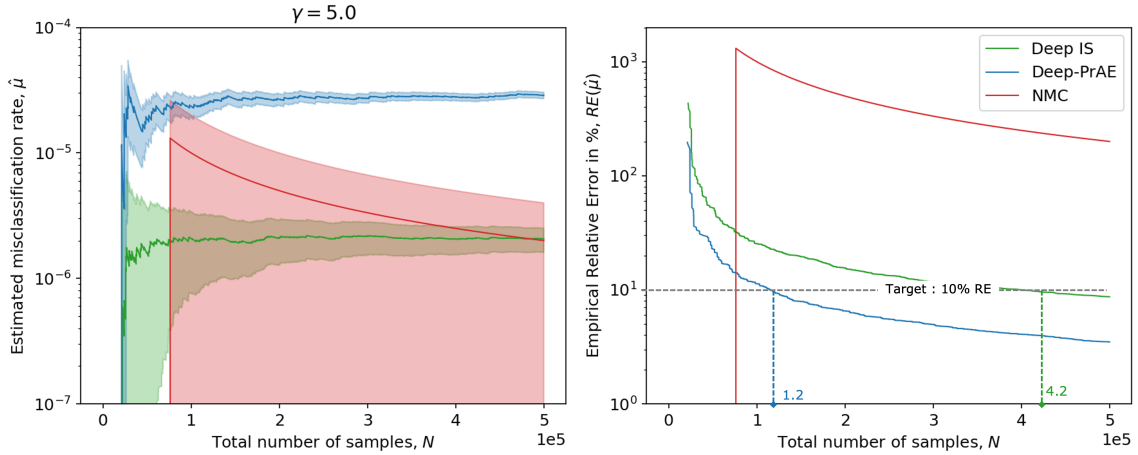

Fig. 4. Estimation of traffic sign classifier misclassification rate (rarity level $\gamma = 5.0$ - extremely rare failure cases)

than Deep-PrAE (about 1/10-th for $\gamma = 3.33$ and 1/14-th for $\gamma = 5.0$). This is a desirable especially for tasks aiming to estimate the true risk of the system. However, we note that while Deep IS performs well in these examples, the correctness certificate of this approach is much weaker compared to Deep-PrAE with provable efficiency certificate for an upper-bound. In fact, the degree of conservativeness of Deep IS estimator flips from being less conservative (0.91) for $\gamma = 3.33$ to slightly more conservative (1.02). Further analysis of Deep IS efficiency guarantee is subject to our future research.

Sampling Efficiency

We now focus on the right figures in Fig. 3, and Fig. 4, all highlighting the efficiency of each sampling algorithm. In Fig. 3 (right), we observe that NMC's RE shrinks much slower compared to the other methods, followed by Deep IS and Deep-PrAE. While it is clear that NMC slow shrinkage is due to its sampling efficiency, the case for Deep IS is slightly different. RE computes the uncertainty of the estimator as a percentage of its estimated value, which means larger uncertainty (in magnitude) is allowed for larger estimator. The large relative error of Deep IS in 4 (right) is thus mainly due to its smaller estimated value compared to the upper bound that Deep-PrAE estimates.

Regardless, suppose that our goal is to produce an estimate with 10% maximum RE. Then, for the case $\gamma = 3.33$, NMC will terminate after using 90,239 samples, 44,204 samples for Deep IS, 21,203 samples for Deep-PrAE. For the case $\gamma = 5.0$, NMC uses more than 18 million samples, 418,983 for Deep IS, and 118,370 for Deep-PrAE. NMC clearly requires much larger sample size, followed by Deep IS, and Deep-PrAE. From the acceleration rate in Table I, we immediately found that only Deep IS can both obtain relatively unbiased estimate and accelerate the procedure significantly, to up to 43.4 times (for the traffic sign classification with $\sigma = 5$). Deep-PrAE can achieve up to 153.6 acceleration rate, but at the cost of 14 times more conservative estimate. If the task at hand requires only an upper-bound for the target, then using Deep-PrAE appears to be more efficient. However, if the task demands for a closer-to-target estimate, Deep IS emerges as the best alternative.

VI. Conclusion and Future Outlook

In this paper, we proposed Deep IS, a practical framework to evaluate AV perception systems under rare-event failure cases. The proposed approach designs efficient IS distribution by combining the dominating point machinery with deep-learning-based rare-event set learning. The key property that distinguishes our approach with other accelerated safety simulation methods is our practicality and versatility. Leveraging on deep learning classifier and recent advances in guaranteed upper bound, our approach applies to generic structure and attains high efficiency improvement for evaluation tasks dealing with AV perception systems that handles large-dimensional inputs. The on-par efficiency with state-of-the-art method and estimates that is closer to the target promises a highly efficient method to deal with AV functional safety evaluation problems with rare naturalistic failure cases. Applications regarding evaluate the safety of intended functionality of AVs can be carried out using the proposed framework, which are our goals in the subsequent work.

ACKNOWLEDGMENT

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710, IIS-1849280, IIS-1849304, and CAREER CNS-2047454.

References

[1] "How Can We Speed Up Autonomous Vehicle Deployment?" [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2021/06/25/how-can-we-speed-up-autonomous-vehicle-deployment/

[2] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[3] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE International Journal of Transportation Safety*, vol. 4, no. 1, pp. 15–24, 2016.

[4] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[5] N. H. T. S. Administration, "Automated driving systems 2.0: A vision for safety," *Washington, DC: US Department of Transportation, DOT HS*, vol. 812, p. 442, 2017.

[6] "Connected and Autonomous Vehicles – The UK Economic Opportunity." [Online]. Available: https://assets.kpmg/content/dam/kpmg/images/2015/05/connected-and-autonomous-vehicles.pdf

[7] "Automated Vehicles for Safety." [Online]. Available: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

[8] "NHTSA Releases Q1 2021 Fatality Estimates, New Edition of 'Countermeasures That Work'." [Online]. Available: https://www.nhtsa.gov/press-releases/q1-2021-fatality-estimates-10th-countermeasures-that-work

[9] "Tesla Autopilot system was on during fatal California crash, adding to self-driving safety concerns." [Online]. Available: https://www.washingtonpost.com/technology/2021/05/14/tesla-california-autopilot-crash/

[10] A. Rizaldi, F. Immler, and M. Althoff, "A formally verified checker of the safe distance traffic rules for autonomous vehicles," in *NASA Formal Methods Symposium*. Springer, 2016, pp. 175–190.

[11] "PEGASUS-Project." [Online]. Available: https://www.pegasusprojekt.de/en/home

[12] A.-M. Jacobo, U. Nobuyuki, Y. Kunio, O. Koichiro, K. Eiichi, and T. Satoshi, "Development of a safety assurance process for autonomous vehicles in japan," in *Proceedings of ESV Conference*, 2019.

[13] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE access*, vol. 8, pp. 87 456–87 477, 2020.

[14] "Tesla's "Shadow" Testing Offers A Useful Advantage On The Biggest Problem In Robocars." [Online]. Available: https://www.forbes.com/sites/bradtempleton/2019/04/29/teslas-shadow-testing-offers-a-useful-advantage-on-the-biggest-problem-in-robocars/

[15] "Geofences: The Invisible Walls Surrounding Autonomous Cars." [Online]. Available: https://www.apex.one/articles/geofences-the-invisible-walls-surrounding-autonomous-cars/

[16] "SAE J3016: Levels of Driving Automation." [Online]. Available: https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart_5.3.21.pdf

[17] N. Kalra and S. M. Paddock, "Driving to Safety - How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?" 2014. [Online]. Available: http://www.rand.org/pubs/research_reports/RR1478.html

[18] D. Zhao, H. Peng, H. Lam, S. Bao, K. Nobukawa, D. J. LeBlanc, and C. S. Pan, "Accelerated evaluation of automated vehicles in lane change scenarios," in *ASME 2015 Dynamic Systems and Control Conference*. American Society of Mechanical Engineers, 2015, pp. V001T17A002—-V001T17A002.

[19] 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.

[20] Z. Huang, Y. Guo, M. Arief, H. Lam, and D. Zhao, "A versatile approach to evaluating and testing automated vehicles based on kernel methods," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 4796–4802.

[21] M. O'Kelly, A. Sinha, H. Namkoong, R. Tedrake, and J. C. Duchi, "Scalable end-to-end autonomous vehicle testing via rare-event simulation," *Advances in neural information processing systems*, vol. 31, 2018.

[22] S. Feng, X. Yan, H. Sun, Y. Feng, and H. X. Liu, "Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment," *Nature communications*, vol. 12, no. 1, pp. 1–14, 2021.

[23] M. Arief, Y. Bai, W. Ding, S. He, Z. Huang, H. Lam, and D. Zhao, "Certifiable deep importance sampling for rare-event simulation of black-box systems," *arXiv preprint arXiv:2111.02204*, 2021.

[24] V. Shakhuro, B. Faizov, and A. Konushin, "Rare traffic sign recognition using synthetic training data," in *Proceedings of the 3rd International Conference on Video and Image Processing*, 2019, pp. 23–26.

[25] Y. Han, K. Virupakshappa, and E. Oruklu, "Robust traffic sign recognition with feature extraction and k-nn classification methods," in *2015 IEEE International Conference on Electro/Information Technology (EIT)*. IEEE, 2015, pp. 484–488.

[26] J. Bucklew, *Introduction to rare event simulation*. Springer Science & Business Media, 2013.

[27] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient monte carlo estimation," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 579–588, 1990.

[28] A. Dieker and M. Mandjes, "On asymptotically efficient simulation of large deviation probabilities," *Advances in applied probability*, vol. 37, no. 2, pp. 539–552, 2005.

[29] 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*.

[30] Z. Huang, D. Zhao, H. Lam, D. J. LeBlanc, and H. Peng, "Evaluation of automated vehicles in the frontal cut-in scenario — an enhanced approach using piecewise mixture models," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 197–202.

[31] Z. Huang, H. Lam, D. J. LeBlanc, and D. Zhao, "Accelerated evaluation of automated vehicles using piecewise mixture models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 2845–2855, Sep. 2018.

[32] Z. Huang, H. Lam, and D. Zhao, "Rare-event simulation without structural information: a learning-based approach," in *2018 Winter Simulation Conference (WSC)*. IEEE, 2018, pp. 1826–1837.

[33] A. G. Ulsoy, H. Peng, and M. Çakmakci, *Automotive control systems*. Cambridge University Press, 2012.

[34] H. Peng, "Conducting the Mcity ABC Test: A Testing Method for Highly Automated Vehicles," Mcity, Tech. Rep., 2020.

[35] 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.

[36] W. Ding, C. Xu, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical scenario generation from methodological perspective," *arXiv preprint arXiv:2202.02215*, 2022.

[37] V. Tjeng, K. Xiao, and R. Tedrake, "Evaluating robustness of neural networks with mixed integer programming," *arXiv preprint arXiv:1711.07356*, 2017.

[38] Z. Huang, H. Lam, and D. Zhao, "Designing importance samplers to simulate machine learning predictors via optimization," in *2018 Winter Simulation Conference (WSC)*. IEEE, 2018, pp. 1730–1741.

[39] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.

[40] A. Wong, M. J. Shafiee, and M. S. Jules, "Micronnet: a highly compact deep convolutional neural network architecture for real-time embedded traffic sign classification," *IEEE Access*, vol. 6, pp. 59 803–59 810, 2018.

[41] Y. Bai, Z. Huang, H. Lam, and D. Zhao, "Over-conservativeness of variance-based efficiency criteria and probabilistic efficiency in rare-event simulation," *arXiv preprint arXiv:2110.12573*, 2021.

[42] M. Arief, Z. Huang, G. K. S. Kumar, Y. Bai, S. He, W. Ding, H. Lam, and D. Zhao, "Deep probabilistic accelerated evaluation: A certifiable rare-event simulation methodology for black-box autonomy," *arXiv preprint arXiv:2006.15722*, 2020.