Deep Probabilistic Accelerated Evaluation: A Robust Certifiable Rare-Event Simulation Methodology for Black-Box Safety-Critical Systems

Mansur Arief Carnegie Mellon University
Zhiyuan Huang Carnegie Mellon University
Guru K.S. Kumar Carnegie Mellon University
Yuanlu Bai Columbia University
Shengyi He Columbia University
Wenhao Ding Carnegie Mellon University
Henry Lam Columbia University
Ding Zhao Carnegie Mellon University

Abstract

Evaluating the reliability of intelligent physical systems against rare safety-critical events poses a huge testing burden for real-world applications. Simulation provides a useful platform to evaluate the extremal risks of these systems before their deployments. Importance Sampling (IS), while proven to be powerful for rare-event simulation, faces challenges in handling these learning-based systems due to their black-box nature that fundamentally undermines its efficiency guarantee, which can lead to under-estimation without diagnostically detected. We propose a framework called Deep Probabilistic Accelerated Evaluation (Deep-PrAE) to design statistically guaranteed IS, by converting black-box samplers that are versatile but could lack guarantees, into one with what we call a relaxed efficiency certificate that allows accurate estimation of bounds on the safety-critical event probability. We present the theory of Deep-PrAE that combines the dominating point concept with rare-event set learning via deep neural network classifiers, and demonstrate its effectiveness in numerical examples including the safety-testing of an intelligent driving algorithm.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

The unprecedented deployment of intelligent physical systems on many real-world applications comes with the need for safety validation and certification (Kalra and Paddock, 2016; Koopman and Wagner, 2017; Uesato et al., 2018). For systems that interact with humans and are potentially safety-critical - which can range from medical systems to self-driving cars and personal assistive robots - it is imperative to rigorously assess their risks before their full-scale deployments. The challenge, however, is that these risks are often associated precisely to how AI reacts in rare and catastrophic scenarios which, by their own nature, are not sufficiently observed.

The challenge of validating the safety of intelligent systems described above is, unfortunately, insusceptible to traditional test methods. In the self-driving context, for instance, the goal of validation is to ensure the AI-enabled system reduces human-level accident rate (in the order of 1.5 per 10^8 miles of driving), thus delivering enhanced safety promise to the public (Evan, 2016; Kalra and Paddock, 2016; NTSB, 2016). Formal verification, which mathematically analyzes and verifies autonomous design, faces challenges when applied to black-box or complex models due to the lack of analytic tractability to formulate failure cases or consider all execution trajectories (Clarke et al., 2018). Automated scenario selection approaches generate test cases based on domain knowledge (Wegener and Bühler, 2004) or adaptive searching algorithms (such as adaptive stress testing; Koren et al. 2018), which is more implementable but falls short of rigor. Test matrix approaches, such as Euro NCAP (NHTSA, 2007), use prepopulated test cases extracted from crash databases, but they only contain historical human-driver information. The closest analog to the latter for self-driving ve-

hicles is "naturalistic tests", which means placing them in real-world environments and gathering observations. This method, however, is economically prohibitive because of the rarity of the target conflict events (Zhao et al., 2017; Arief et al., 2018; Claybrook and Kildare, 2018; O'Kelly et al., 2018).

Because of all these limitations, simulation-based tests surface as a powerful approach to validate complex black-box designs (Corso et al., 2020). This approach operates by integrating the target intelligent algorithm into an interacting virtual simulation platform that models the surrounding environment. By running enough Monte Carlo sampling of this (stochastic) environment, one hopes to observe catastrophic conflict events and subsequently conduct statistical analyses. This approach is flexible and scalable, as it hinges on building a virtual environment instead of physical systems, and provides a probabilistic assessment on the occurrences and behaviors of safety-critical events (Koopman and Wagner, 2018).

Nonetheless, similar to the challenge encountered by naturalistic tests, because of their rarity, safety-critical events are seldom observed in the simulation experiments. In other words, it could take an enormous amount of Monte Carlo simulation runs to observe one "hit", and this in turn manifests statistically as a large estimation variance per simulation run relative to the target probability of interest (i.e., the so-called *relative error*; L'ecuyer et al. 2010). This problem, which is called rare-event simulation (Bucklew, 2013), is addressed conventionally under the umbrella of variance reduction, which includes a range of techniques from importance sampling (IS) (Juneja and Shahabuddin, 2006; Blanchet and Lam, 2012) to multi-level splitting (Glasserman et al., 1999; Villén-Altamirano and Villén-Altamirano, 1994). Typically, to ensure the relative error is dramatically reduced, one has to analyze the underlying model structures to gain understanding of the rare-event behaviors, and leverage this knowledge to design good Monte Carlo schemes (Juneja and Shahabuddin, 2006; Dean and Dupuis, 2009). For convenience, we call such relative error reduction guarantee an *efficiency certificate*.

Our main focus of this paper is on rare-event problems with the underlying model unknown or too complicated to support analytical tractability. In this case, traditional variance reduction approaches may fail to provide an efficiency certificate. Moreover, we will explain how some existing "black-box" variance reduction techniques, while versatile and powerful, could lead to dangerous *under-estimation* of a rare-event probability without detected diagnostically due to a lack of efficiency certificate. This motivates us to study a framework to convert these black-box methods into one that has rigorous certificate. More precisely, our framework consists of three ingredients:

Relaxed efficiency certificate: We shift the estimation of target rare-event probability to an upper (and lower) bound, in a way that supports the integration of learning errors into variance reduction without giving up estimation correctness.

Set-learning with one-sided error: We design learning algorithms based on deep neural network classifier to create outer (or inner) approximations of rare-event sets. This classifier has a special property that, under a geometric property called orthogonal monotonicity, it exhibits zero false negative rates.

Deep-learning-based IS: With the deep-learning based rare-event set approximation, we search the so-called *dominating points* in rare-event analysis to create IS that achieves the relaxed efficiency certificate.

We call our framework consisting of the three ingredients above *Deep Probabilistic Accelerated Evaluation (Deep-PrAE)*, where "Accelerated Evaluation" follows terminologies in recent approaches for the safety-testing of autonomous vehicles (Zhao et al., 2016; Huang et al., 2018). In the set-learning step in Deep-PrAE, the samples fed into our deep classifier can be generated by any black-box algorithms including the cross-entropy (CE) method (De Boer et al., 2005; Rubinstein and Kroese, 2013) and particle approaches such as adaptive multi-level splitting (AMS) (Au and Beck, 2001; Cérou and Guyader, 2007; Webb et al., 2018). Deep-PrAE turns these samples into an IS with an efficiency certificate against undetected under-estimation. Our approach is robust in the sense that it provides a tight bound for the target rare-event probability if the underlying classifier is expressive enough, while it still provides a correct, though conservative, bound if the classifier is weak. To our best knowledge, such type of guarantees and robustness features is the first of its kind in the rare-event simulation literature, and we envision our work to lay the foundation for further improvements to design certified methods for evaluating more sophisticated intelligent designs.

2 Statistical Challenges in Black-Box Rare-Event Simulation

Our evaluation goal is the probabilistic assessment of a complex physical system invoking rare but catastrophic events in a stochastic environment. For concreteness, we write this rare-event probability $\mu = P(X \in \mathcal{S}_\gamma)$. Here X is a random vector in $\mathbb{R}^d$ that denotes the environment, and is distributed according to p. $\mathcal{S}_\gamma$ denotes a safety-critical set on the interaction between the physical system and the environment. The "rarity"

parameter $\gamma \in \mathbb{R}$ is considered a large number, with the property that as $\gamma \to \infty$, $\mu \to 0$ (Think of, e.g., $\mathcal{S}_\gamma = \{x : f(x) \geq \gamma\}$ for some risk function f and exceedance threshold γ). We will work with Gaussian p for the ease of analysis, but our framework is more general (i.e., applies to Gaussian mixtures and other light-tailed distributions). Here, we explain intuitively the main concepts and challenges in black-box rare-event simulation, leaving the details to Appendix A.

Monte Carlo Efficiency. Suppose we use a Monte Carlo estimator $\hat{\mu}_n$ to estimate μ, by running n simulation runs in total. Since μ is tiny, the error of a meaningful estimation must be measured in relative term, i.e., we would like

$$P(|\hat{\mu}_n - \mu| > \epsilon\mu) \leq \delta \tag{1}$$

where δ is some confidence level (e.g., $\delta = 5\%$) and $0 < \epsilon < 1$.

Suppose that $\hat{\mu}_n$ is unbiased and is an average of n i.i.d. simulation runs, i.e., $\hat{\mu}_n = (1/n)\sum_{i=1}^n Z_i$ for some random unbiased output Z_i. We define the *relative error* $RE = Var(Z_i)/\mu^2$ as the ratio of variance (per-run) and squared mean. Importantly, to attain (1), a sufficient condition is $n \geq RE/(\delta\epsilon^2)$. So, when RE is large, the required Monte Carlo size is also large.

Challenges in Naive Monte Carlo. Let $Z_i = I(X_i \in \mathcal{S}_\gamma)$ where $I(\cdot)$ denotes the indicator function, and X_i is an i.i.d. copy of X. Since Z_i follows a Bernoulli distribution, $RE = (1-\mu)/\mu$. Thus, the required n scales linearly in $1/\mu$ (when μ is tiny). This demanding condition is a manifestation of the difficulty in hitting $\mathcal{S}_\gamma$. In the standard large deviations regime (Amir Dembo, 2010; Dupuis and Ellis, 2011) where μ is exponentially small in γ, the required Monte Carlo size n would grow *exponentially* in γ.

Variance Reduction. The severe burden when using naive Monte Carlo motivates techniques to drive down RE. First we introduce the following notion:

Definition 1. *We say an estimator $\hat{\mu}_n$ satisfies an efficiency certificate to estimate μ if it achieves (1) with $n = \tilde{O}(\log(1/\mu))$, for given $0 < \epsilon, \delta < 1$.*

In the above, $\tilde{O}(\cdot)$ denotes a polynomial growth in $\cdot$. If $\hat{\mu}_n$ is constructed from n i.i.d. samples, then the efficiency certificate can be attained with $RE = \tilde{O}(\log(1/\mu))$. Note that in the large deviations regime, the sample size n used in a certifiable estimator is reduced from exponential in γ in naive Monte Carlo to *polynomial* in γ.

Importance sampling (IS) is a prominent technique to achieve efficiency certificate (Glynn and Iglehart, 1989). IS generates X from another distribution $\tilde{p}$ (called IS distribution), and outputs $\hat{\mu}_n = (1/n)\sum_{i=1}^n L(X_i)I(X_i \in \mathcal{S}_\gamma)$ where $L = dp/d\tilde{p}$ is the likelihood ratio, or the Radon-Nikodym derivative, between p and $\tilde{p}$. Via a change of measure, it is easy to see that $\hat{\mu}_n$ is unbiased for μ. The key is to control its RE by selecting a good $\tilde{p}$. This requires analyzing the behavior of the likelihood ratio L under the rare event, and in turn understanding the rare-event sample path dynamics (Juneja and Shahabuddin, 2006).

Perils of Black-Box Variance Reduction Algorithms. Unfortunately, in black-box settings where complete model knowledge and analytical tractability are unavailable, the classical IS methodology faces severe challenges. To explain this, we first need to understand how efficiency certificate can be obtained based on the concept of *dominating points*. From now on, we consider input $X \in \mathbb{R}^d$ from a Gaussian distribution $N(\lambda, \Sigma)$ where Σ is positive definite.

Definition 2. *A set $A_\gamma \subset \mathbb{R}^d$ is a* dominating set *for the set $\mathcal{S}_\gamma \subset \mathbb{R}^d$ associated with the distribution $N(\lambda, \Sigma)$ if for any $x \in \mathcal{S}_\gamma$, there exists at least one $a \in A_\gamma$ such that $(a-\lambda)^T\Sigma^{-1}(x-a) \geq 0$. Moreover, this set is minimal in the sense that if any point in A_γ is removed, then the remaining set no longer satisfies the above condition. We call any point in A_γ a dominating point.*

The dominating set comprises the "corner" cases where the rare event occurs (Sadowsky and Bucklew, 1990). In other words, each dominating point a encodes, in a local region, the most likely scenario should the rare event happen, and this typically corresponds to the highest-density point in this region. Locality here refers to the portion of the rare-event set that is on one side of the hyperplane cutting through a (see Figure 1(a)).

Intuitively, to increase the frequency of hitting the rare-event set (and subsequently to reduce variance), an IS would translate the distributional mean from λ to the global highest-density point in the rare-event set. The delicacy, however, is that this is *insufficient* to control the variance, due to the "overshoots" arising from sampling randomness. In order to properly control the overall variance, one needs to divide the rare-event set into local regions governed by dominating points, and using a mixture IS distribution that accounts for *all* of them. This approach gives a certifiable IS, described as follows:

Suppose $\mathcal{S}_\gamma = \bigcup_j \mathcal{S}_\gamma^j$, where each $\mathcal{S}_\gamma^j$ is a "local" region corresponding to a dominating point $a_j \in A_\gamma$ associated with the distribution $N(\lambda, \Sigma)$, with conditions stated precisely in Theorem 4 in the Appendix. Then the IS estimator constructed by n i.i.d. outputs drawn from the IS distribution $\sum_j \alpha_j N(a_j, \Sigma)$ achieves an efficiency certificate in estimating $\mu = P(X \in \mathcal{S}_\gamma)$.

On the contrary, if the Gaussian (mixture) IS distri-

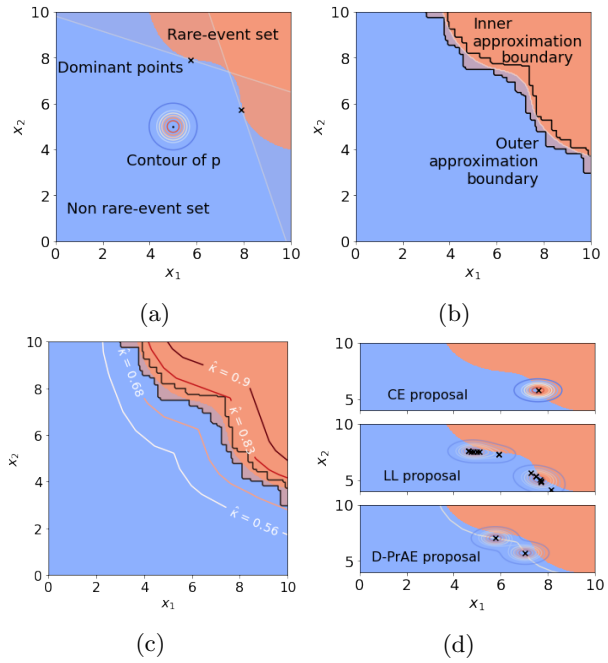


Figure 1: (a) An example of $\mathcal{S}_\gamma$ with two dominating points (b) Outer- and inner- approximations of $\mathcal{S}_\gamma$ (c) $\hat{\kappa}$ tuning for Stage 1 Alg. 1 (d) IS proposals: too few dominating points for CE with too simple parametric class, too many for LL, and a balance for Deep-PrAE.

bution misses any of the dominating points, then the resulting estimate may be utterly unreliable for two reasons. First, not only that efficiency certificate may fail to hold, but its RE can be arbitrarily large. Second, even more dangerously, this poor performance can be empirically hidden and leads to a systematic *under-estimation* of the rare-event probability without being detected. In other words, in a given experiment, we may observe a reasonable empirical relative error (i.e., sample variance over squared sample mean), yet the estimate is much lower than the correct value. These are revealed in the following example:

Theorem 1 (Perils of under-estimation)**.** *Suppose we estimate* $\mu = P(X \geq \gamma \text{ or } X \leq -k\gamma)$ *where* $X \sim p = N(0,1)$ *and* $0 < k < 3$*. We choose* $\tilde{p} = N(\gamma, 1)$ *as the IS distribution to obtain* $\hat{\mu}_n$*. Then 1) The relative error of* $\hat{\mu}_n$ *grows exponentially in* γ*. 2) If* n *is polynomial in* γ*, we have* $P\left(\left|\hat{\mu}_n - \bar{\Phi}(\gamma)\right| > \varepsilon\bar{\Phi}(\gamma)\right) = O\left(\frac{\gamma}{n\varepsilon^2}\right)$ *for any* $\varepsilon > 0$ *where* $\bar{\Phi}(\gamma) = P(X \geq \gamma) < \mu$*, and the empirical relative error* $= O(n^2)$ *with probability higher than* $1 - 1/2^n$*.*

The second conclusion in Theorem 1 implies that the estimator $\hat{\mu}_n$, built from an IS with a missed dominating point, systematically under-estimates the target μ, yet with high probability its empirical relative error grows only polynomially in γ, thus wrongly fooling the user that the estimator is efficient.

With this, we now explain why using black-box variance reduction algorithms can be dangerous - in the sense of not having an efficiency certificate and, more importantly, the risk of an unnoticed systematic under-estimation. In the literature, there are two lines of techniques that apply to black-box problems. The first line is the CE method, which uses optimization to search for a good parametrization over a parametric class of IS. The objective criteria include the cross-entropy (with an oracle-best zero-variance IS distribution; Rubinstein and Kroese 2013; De Boer et al. 2005) and estimation variance (Arouna, 2004). Without closed-form expressions, and also to combat the rare-event issue, one typically solves a sequence of empirical optimization problems, starting from a "less rare" problem (i.e., smaller γ) and gradually increasing the rarity with updated empirical objectives using better IS samples. Achieving efficiency requires both a sufficiently expressive parametric IS class and parameter convergence (so that at the end all the dominating points are accounted for). The second line of methods is the multi-level splitting or subsimulation (Au and Beck, 2001; Cérou and Guyader, 2007), a particle method in lieu of IS, which relies on enough mixing of descendant particles. Full analyses on these methods to reach efficiency certificate appear challenging, and without one the estimators could be under-estimated, and without detected, as illustrated in Theorem 1. We discuss more details of CE and AMS in Appendix E.

Note that there are other variants of CE and AMS. The former include enhanced CE such as Markov chain IS (Botev et al., 2013, 2016; Grace et al., 2014), neural network IS (Müller et al., 2019) and nonparametric CE (Rubinstein, 2005).

The latter include RESTART which works similarly as subset simulation and splitting but performs a number of simulation retrials after entering regions with a higher importance function value (Villén-Altamirano, 2010). Similar to standard CE and AMS, these methods also face challenges in satisfying an efficiency certificate.

Lastly, we briefly review several other methods with guarantees similar to our efficiency certificate, but relies heavily on structual knowledge. The first one is large-deviations-based IS including sequential exponential tilting (Bucklew, 2004; Asmussen and Glynn, 2007; Siegmund, 1976) and mixture-based proposals (Chen et al., 2019). Another method, which is especially powerful for heavy tailed problems, is conditional Monte Carlo which reduces the variance by sampling conditional on some auxiliary random variables (Asmussen and Kroese, 2006).

Compared to existing methods as reviewed above, our

novelty is to tackle black-box problems while sustaining a correctness guarantee, via a new certificate and a careful integration of set-learning with the dominating point machinery.

3 The Deep Probabilistic Accelerated Evaluation Framework

We propose the Deep-PrAE framework to overcome the challenges faced by black-box variance reduction algorithms. This framework comprises two stages: First is to learn the rare-event set from a first-stage sample batch, by viewing set learning as a classification task. These first-stage samples can be drawn from any rare-event sampling methods including CE and AMS. Second is to apply an efficiency-certified IS on the rare-event probability over the learned set. Algorithm 1 shows our main procedure. The key to achieving an ultimate efficiency certificate lies in how we learn the rare-event set in Stage 1, which requires two properties:

Small one-sided generalization error: "One-sided" generalization error here means the learned set is either an outer or an inner approximation of the unknown true rare-event set, with probability 1. Converting this into a classification, this means the false negative (or positive) rate is exactly 0. "Small" here then refers to the other type of error being controlled.

Decomposability: The learned set is decomposable according to dominating points in the form of Theorem 2, so that an efficient mixture IS can apply.

The first property ensures that, even though the learned set can contain errors, the learned rare-event probability is either an upper or lower bound of the truth. This requirement is important as it is difficult to translate the impact of generalization errors into rare-event estimation errors. By Theorem 1, we know that any non-zero error implies the risk of missing out on important regions of the rare-event set, undetectably. The one-sided generalization error allows a shift of our target to valid upper and lower bounds that can be correctly estimated, which is the core novelty of Deep-PrAE.

To this end, we introduce a new efficiency notion:

Definition 3. *We say an estimator $\hat{\mu}_n$ satisfies an upper-bound relaxed efficiency certificate to estimate μ if $P(\hat{\mu}_n - \mu < -\epsilon\mu) \leq \delta$ with $n \geq \tilde{O}(\log(1/\mu))$, for given $0 < \epsilon, \delta < 1$.*

Compared with the efficiency certificate in (1), Definition 3 is relaxed to only requiring $\hat{\mu}_n$ to be an upper bound of μ, up to an error of $\epsilon\mu$. An analogous lower-bound relaxed efficiency certificate can be seen in Appendix D. From a risk quantification viewpoint, the upper bound for μ is more crucial, and the lower bound serves to assess an estimation gap. The following provides a handy certification:

Algorithm 1: Deep-PrAE to estimate $\mu = P(X \in \mathcal{S}_\gamma)$.

Input: Black-box evaluator $I(\cdot \in \mathcal{S}_\gamma)$, initial Stage 1 samples $\{(\tilde{X}_i, Y_i)\}_{i=1,\ldots,n_1}$, Stage 2 sampling budget n_2, input distribution $N(\lambda, \Sigma)$.

Output: IS estimate $\hat{\mu}_n$.

Stage 1 (Set Learning):

Train classifier with positive decision region $\overline{\mathcal{S}}_\gamma^\kappa = \{x : \hat{g}(x) \geq \kappa\}$ using $\{(\tilde{X}_i, Y_i)\}_{i=1,\ldots,n_1}$;

Replace κ by $\hat{\kappa} = \max\{\kappa \in \mathbb{R} : (\overline{\mathcal{S}}_\gamma^\kappa)^c \subset \mathcal{H}(T_0)\}$;

Stage 2 (Mixture IS based on Searched dominating points):

Start with $\hat{A}_\gamma = \emptyset$;

While $\{x : \hat{g}(x) \geq \hat{\kappa}, (x_j^* - \lambda)^T\Sigma^{-1}(x - x_j^*) < 0, \ \forall x_j^* \in \hat{A}_\gamma\} \neq \emptyset$ **do**

Find a dominating point x^* by solving the optimization problem

$$x^* = \arg\min_x \ (x-\lambda)^T\Sigma^{-1}(x-\lambda)$$
$$\text{s.t.} \quad \hat{g}(x) \geq \hat{\kappa},$$
$$(x_j^* - \lambda)^T\Sigma^{-1}(x - x_j^*) < 0, \ \forall x_j^* \in \hat{A}_\gamma$$

and update $\hat{A}_\gamma \leftarrow \hat{A}_\gamma \cup \{x^*\}$;

End

Sample $X_1, \ldots, X_{n_2}$ from the mixture distribution $\sum_{a \in \hat{A}_\gamma} (1/|\hat{A}_\gamma|) N(a, \Sigma)$.

Compute the IS estimator $\hat{\mu}_n = (1/n_2)\sum_{i=1}^{n_2} L(X_i) I(X_i \in \bar{\mathcal{S}}_\gamma^{\hat{\kappa}})$, where the likelihood ratio $L(X_i) = \phi(X_i; \lambda, \Sigma)/(\sum_{a \in \hat{A}_\gamma}(1/|\hat{A}_\gamma|)\phi(X_i; a, \Sigma))$ and $\phi(\cdot; \alpha, \Sigma)$ denotes the density of $N(\alpha, \Sigma)$.

Proposition 1 (Achieving relaxed efficiency certificate)**.** *Suppose $\hat{\mu}_n$ is upward biased, i.e., $\overline{\mu} := E[\hat{\mu}_n] \geq \mu$. Moreover, suppose $\hat{\mu}_n$ takes the form of an average of n i.i.d. simulation runs Z_i, with $RE = Var(Z_i)/\overline{\mu}^2 = \tilde{O}(\log(1/\overline{\mu}))$. Then $\hat{\mu}_n$ possesses the upper-bound relaxed efficiency certificate.*

Proposition 1 stipulates that a relaxed efficiency certificate can be attained by an upward biased estimator that has a logarithmic relative error with respect to the biased mean. Appendix C shows an extension of Proposition 1 to two-stage procedures, where the first stage determines the upward biased mean. This upward biased mean, in turn, can be obtained by learning an outer approximation for the rare-event set, giving:

Corollary 1 (Set-learning + IS)**.** *Consider estimat-*

ing $\mu = P(X \in \mathcal{S}_\gamma)$. Suppose we can learn a set $\overline{\mathcal{S}}_\gamma$ with any number n_1 of i.i.d. samples D_{n_1} (drawn from some distribution) such that $\overline{\mathcal{S}}_\gamma \supset \mathcal{S}_\gamma$ with probability 1. Also suppose that there is an efficiency certificate for an IS estimator for $\overline{\mu}(D_{n_1}) := P(X \in \overline{\mathcal{S}}_\gamma)$. Then a two-stage estimator where a constant n_1 number of samples D_{n_1} are first used to construct $\overline{\mathcal{S}}_\gamma$, and $n_2 = \tilde{O}(\log(1/\overline{\mu}(D_{n_1}))$ samples are used for the IS in the second stage, achieves the upper-bound relaxed efficiency certificate.

To execute the procedure in Corollary 1, we need to learn an outer approximation of the rare-event set. To this end, consider set learning as a classification problem. Suppose we have collected n_1 Stage 1 samples $\{(\tilde{X}_i, Y_i)\}_{i=1,\ldots,n_1}$, where $Y_i = 1$ if $\tilde{X}_i$ is in the rare-event set $\mathcal{S}_\gamma$, and 0 otherwise. Here, it is beneficial to use Stage 1 samples that have sufficient presence in $\mathcal{S}_\gamma$, which can be achieved via any black-box variance reduction methods. We then consider the pairs $\{(\tilde{X}_i, Y_i)\}$ where $\tilde{X}_i$ is regarded as the feature and Y_i as the binary label, and construct a classifier, say $\hat{g}(x) : \mathbb{R}^d \to [0, 1]$, from some hypothesis class $\mathcal{G}$ that (nominally) signifies $P(Y = 1|X = x)$. The learned rare-event set $\overline{\mathcal{S}}_\gamma$ is taken to be $\{x : \hat{g}(x) \geq \kappa\}$ for some threshold $\kappa \in \mathbb{R}$.

The outer approximation requirement $\overline{\mathcal{S}}_\gamma \supset \mathcal{S}_\gamma$ means that all true positive (i.e., 1) labels must be correctly classified, or in other words, the false negative (i.e., 0) rate is zero, i.e.,

$$P(X \in \overline{\mathcal{S}}_\gamma^c, Y = 1) = 0 \tag{2}$$

Typically, achieving such a zero "Type I" misclassification rate is impossible for any finite sample except in degenerate cases. However, this is achievable under a geometric premise on the rare-event set $\mathcal{S}_\gamma$ that we call *orthogonal monotonicity*. To facilitate discussion, suppose from now on that the rare-event set is known to lie entirely in the positive quadrant $\mathbb{R}^d_+$, so in learning the set, we only consider sampling points in $\mathbb{R}^d_+$ (analogous development can be extended to the entire space).

Definition 4. *We call a set $\mathcal{S} \subset \mathbb{R}^d_+$* orthogonally monotone *if for any two points $x, x' \in \mathbb{R}^d_+$, we have $x \leq x'$ (where the inequality is defined coordinate-wise) and $x \in \mathcal{S}$ implies $x' \in \mathcal{S}$ too.*

Definition 4 means that any point that is more "extreme" than a point in the rare-event set must also lie inside the same set. This is an intuitive assumption that appears to hold in some safety-critical rare-event settings (see Section 4). Note that, even with such a monotonicity property, the boundary of the rare-event set can still be very complex. The key is that, with orthogonal monotonicity, we can now produce a classification procedure that satisfies (2). In fact, the simplest approach is to use what we call an orthogonally monotone hull:

Definition 5. *For a set of points $D = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d_+$, we define the* orthogonally monotone hull *of D (with respect to the origin) as $\mathcal{H}(D) = \cup_i \mathcal{R}(x_i)$, where $\mathcal{R}(x_i)$ is the rectangle that contains both x_i and the origin as two of its corners.*

In other words, the orthogonally monotone hull consists of the union of all the rectangles each wrapping each point x_i and the origin 0. Now, denote $T_0 = \{\tilde{X}_i : Y_i = 0\}$ as the non-rare-event sampled points. Evidently, if $\mathcal{S}_\gamma$ is orthogonally monotone, then $\mathcal{H}(T_0) \subset \mathcal{S}_\gamma^c$ (where complement is with respect to $\mathbb{R}^d_+$), or equivalently, $\mathcal{H}(T_0)^c \supset \mathcal{S}_\gamma$, i.e., $\mathcal{H}(T_0)^c$ is an outer approximation of the rare-event set $\mathcal{S}_\gamma$. Figure 1(b) shows this outer approximation (and also the inner counterpart). Moreover, $\mathcal{H}(T_0)^c$ is the smallest region (in terms of set volume) such that (2) holds, because any smaller region could exclude a point that has label 1 with positive probability.

Lazy-Learner IS. We now consider an estimator for μ where, given the n_1 samples in Stage 1, we build the mixture IS depicted in Theorem 1 to estimate $P(X \in \mathcal{H}(T_0)^c)$ in Stage 2. Since $\mathcal{H}(T_0)^c$ takes the form $(\cup_{i:Y_i=0} \mathcal{R}(\tilde{X}_i))^c$, it has a finite number of dominating points, which can be found by a sequential algorithm (similar to the one that we will discuss momentarily). We call this the "lazy-learner" approach. Its problem, however, is that $\mathcal{H}(T_0)^c$ tends to have a very rough boundary. This generates a large number of dominating points, many of which are unnecessary in that they do not correspond to any "true" dominating points in the original rare-event set $\mathcal{S}_\gamma$ (see the middle of Figure 1(d)). This in turn leads to a large number of mixture components that degrades the IS efficiency, as the RE bound in Theorem 1 scales linearly with the number of mixture components.

Deep-Learning-Based IS. Our main approach is a deep-learning alternative that resolves the statistical degradation of the lazy learner. We train a neural network classifier, say $\hat{g}$, using all the Stage 1 samples $\{(\tilde{X}_i, Y_i)\}$, and obtain an approximate non-rare-event region $(\overline{\mathcal{S}}_\gamma^\kappa)^c = \{x : \hat{g}(x) < \kappa\}$, where κ is say 1/2. Then we adjust κ minimally away from 1/2, say to $\hat{\kappa}$, so that $(\overline{\mathcal{S}}_\gamma^{\hat{\kappa}})^c \subset \mathcal{H}(T_0)$, i.e., $\hat{\kappa} = \max\{\kappa \in \mathbb{R} : (\overline{\mathcal{S}}_\gamma^\kappa)^c \subset \mathcal{H}(T_0)\}$. Then $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}} \supset \mathcal{H}(T_0)^c \supset \mathcal{S}_\gamma$, so that $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ is an outer approximation for $\mathcal{S}_\gamma$ (see Figure 1(c), where $\hat{\kappa} = 0.68$). Stage 1 in Algorithm 1 shows this procedure. With this, we can run mixture IS to estimate $P(X \in \overline{\mathcal{S}}_\gamma^{\hat{\kappa}})$ in Stage 2.

The execution of this algorithm requires the set $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}} = \{x : \hat{g}(x) \geq \hat{\kappa}\}$ to be in a form susceptible to The-

orem 2 and the search of all its dominating points. When $\hat{g}$ is a ReLU-activated neural net, the boundary of $\hat{g}(x) \geq \hat{\kappa}$ is piecewise linear and $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ is a union of polytopes, and Theorem 2 applies. Finding all dominating points is done by a sequential "cutting-plane" method that iteratively locates the next dominating point by minimizing $(x-\mu)^T\Sigma^{-1}(x-\mu)$ over the remaining portion of $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ not covered by the local region of any previously found points x_j^*. These optimization sequences can be solved via mixed integer program (MIP) formulations for ReLU networks (Tjeng et al. 2017; Huang et al. 2018; see Appendix B). Note that a user can control the size of these MIPs via the neural net architecture. Regardless of the expressiveness of these networks, Algorithm 1 enjoys the following guarantee:

Theorem 2 (Relaxed efficiency certificate for deep-learning-based mixture IS)**.** *Suppose $\mathcal{S}_\gamma$ is orthogonally monotone, and $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ satisfies the same conditions for $\mathcal{S}_\gamma$ in Theorem 2. Then Algorithm 1 attains the upper-bound relaxed efficiency certificate by using a constant number of Stage 1 samples.*

Figure 1(d) shows how our deep-learning-based IS achieves superior efficiency compared to other alternatives. The cross-entropy method can miss a dominating point (1) and result in systematic under-estimation. The lazy-learner IS, on the other hand, generates too many dominating points (64) and degrades efficiency. Algorithm 1 finds the right number (2) and approximate locations of the dominating points.

Moreover, whereas the upper-bound certificate is guaranteed in our design, in practice, the deep-learning-based IS also appears to work well in controlling the conservativeness of the bound, as dictated by the false positive rate $P(X \in \overline{\mathcal{S}}_\gamma^{\hat{\kappa}}, Y = 0)$ (see our experiments next). We close this section with a finite-sample bound on the false positive rate. Here, in deriving our bound, we assume the use of a sampling distribution q in generating independent Stage 1 samples, and we use empirical risk minimization (ERM) to train $\hat{g}$, i.e., $\hat{g} := \text{argmin}_{g\in\mathcal{G}}\{R_{n_1}(g) := \frac{1}{n_1}\sum_{i=1}^{n_1}\ell(g(\tilde{X}_i), Y_i)\}$ where ℓ is a loss function and $\mathcal{G}$ is the considered hypothesis class. Correspondingly, let $R(g) := E_{X\sim q}\ell(g(X), I(X \in \mathcal{S}_\gamma))$ be the true risk function and $g^* := \arg\min_{g\in\mathcal{G}} R(g)$ its minimizer. Also let $\kappa^* := \min_{x\in\mathcal{S}_\gamma} g^*(x)$ be the true threshold associated with g^* in obtaining the smallest outer rare-event set approximation.

Theorem 3 (Conservativeness)**.** *Consider $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ obtained in Algorithm 1 where $\hat{g}$ is trained from an ERM. Suppose the density q has bounded support $K \subset [0, M]^d$ and $0 < q_l \leq q(x) \leq q_u$ for any $x \in K$. Also suppose there exists a function h such that for any $g \in \mathcal{G}$, $g(x) \geq \kappa$ implies $\ell(g(x), 0) \geq h(\kappa) > 0$. (e.g., if ℓ is the squared loss, then $h(\kappa)$ could be chosen as $h(\kappa) = \kappa^2$). Then, with probability at least $1-\delta$,*

$$P_{X\sim q}\left(X \in \bar{\mathcal{S}}_\gamma^{\hat{\kappa}} \setminus \mathcal{S}_\gamma\right) \leq \frac{R(g^*) + 2\sup_{g\in\mathcal{G}}|R_{n_1}(g) - R(g)|}{h(\kappa^* - t(\delta, n_1)\sqrt{d}Lip(g^*) - \|\hat{g} - g^*\|_\infty)}.$$

Here, $Lip(g^*)$ is the Lipschitz parameter of g^*, and $t(\delta, n_1) = 3\left(\frac{\log(n_1 q_l) + d\log M + \log\frac{1}{\delta}}{n_1 q_l}\right)^{\frac{1}{d}}$.

Theorem 3 reveals a tradeoff between overfitting (measured by $\sup_{g\in\mathcal{G}}|R_{n_1}(g) - R(g)|$ and $\|\hat{g} - g^*\|_\infty$) and underfitting (measured by $R(g^*) = \inf_{g\in\mathcal{G}} R(g)$). Appendix C discusses related results on the sharp estimates of these quantities for deep neural networks, a more sophisticated version of Theorem 3 that applies to the cross-entropy loss, a corresponding bound for the lazy learner, as well as results to interpret Theorem 3 under the original distribution p.

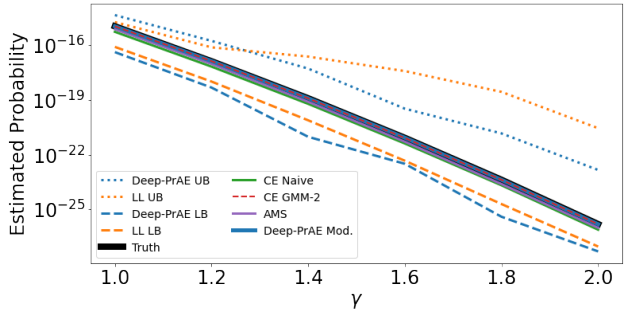
Finally, we point out some works in the literature that approximate the Pareto frontier of a monotone function (Wu et al., 2018; Legriel et al., 2010). While the boundary of an orthogonally monotone set looks similar to the Pareto frontier, our recipe (outer/inner approximation using piecewise-linear-activation NN) is designed to minimize the number of dominating points while simultaneously achieve the relaxed efficiency certificate for rare-event estimation. Such a guarantee is novelly beyond these previous works.

4 Numerical Experiments

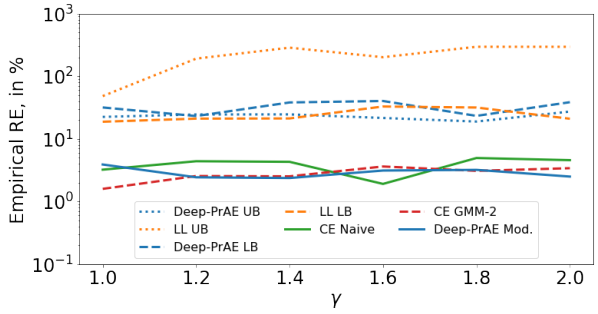
We implement and compare the estimated probabilities and the REs of deep-learning-based IS for the upper bound (Deep-PrAE UB) and lazy-learner IS (LL UB). We also show the corresponding lower-bound estimator (Deep-PrAE LB and LL LB) and benchmark with the cross entropy method (CE), adaptive multilevel splitting (AMS), and naive Monte Carlo (NMC). For CE, we run a few variations testing different parametric classes and report two in the following figures: CE that uses a single Gaussian distribution (CE Naive), representing an overly-simplified CE implementation, and CE with Gaussian Mixture Model with k components (CE GMM-k), representing a more sophisticated CE implementation. For Deep-PrAE, we use the samples from CE Naive as the Stage 1 samples. We also run a modification of Deep-PrAE (Deep-PrAE Mod) that replaces $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ by $\mathcal{S}_\gamma$ in the last step of Algorithm 1 as an additional comparison. We use a 2-dimensional numerical example and a safety-testing of an intelligent driving algorithm for a car-following scenario. These two experiments are representative as the former is low-dimensional (visualizable) yet with *extremely* rare

events while the latter is moderately high-dimensional, challenging for most of the existing methods.

2D Example. We estimate $\mu = P(X \in \mathcal{S}_\gamma)$ where $X \sim N([5,5]^T, 0.25I_{2\times 2})$, and γ ranging from 1.0 to 2.0. We use $n = 30,000$ (10,000 for Stage 1 and 20,000 for Stage 2), and use 10,000 of the CE samples as our Stage 1 samples. Figure 1 illustrates the shape of $\mathcal{S}_\gamma$, which has two dominating points. This probability is microscopically small (e.g., $\gamma = 1.8$ gives $\mu = 4.1 \times 10^{-24}$) and serves to investigate our performance in ultra-extreme situations.



(a) Estimated rare-event probability



(b) Estimator's empirical relative error

Figure 2: 2-dimensional example. Naive Monte Carlo failed in all cases and hence not shown.

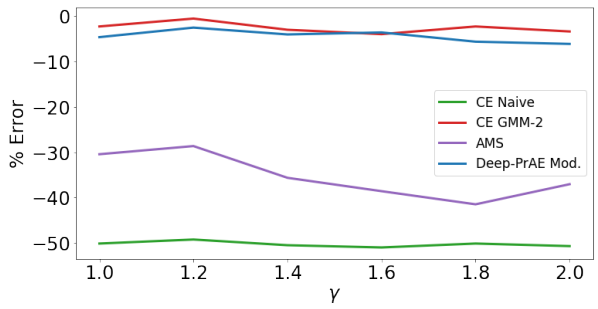
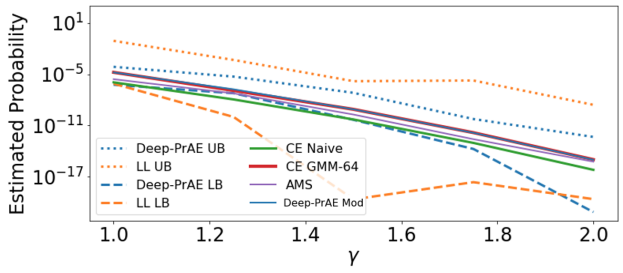


Figure 3: Percentage error of CE, AMS, and modified Deep-PrAE (minus % error means under-estimation)

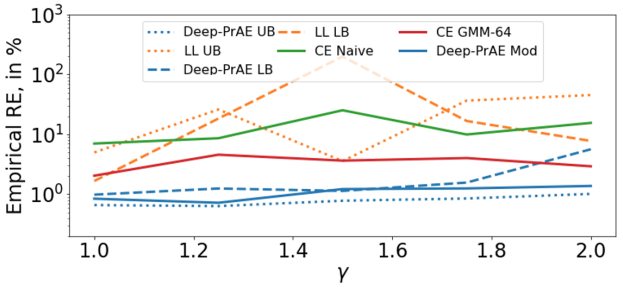
Figure 2 compares all approaches to the true value, which we compute via a proper mixture IS with 50,000 samples assuming the full knowledge of $\mathcal{S}_\gamma$. It shows several observations. First, Deep-PrAE and LL (both UB and LB) always provide valid bounds that contain the truth. Second, the UB for LL is more conservative than Deep-PrAE in up to two orders of magnitudes, which is attributed to the overly many (redundant) dominating points. Correspondingly, the RE of LL UB is tremendously high, reaching over 500% when $\gamma = 2.0$, compared to around 40% for Deep-PrAE UB, Deep-PrAE LB, and LL LB. Third, CE Naive, which finds only one dominating point, consistently under-estimates the truth by about 50%, yet it gives an over-confident RE, e.g., $< 5\%$ when $\gamma < 2$. This shows a systematic undetected under-estimation issue when CE is implemented overly-naively. AMS also underestimates the true value by 30%-40%, while CE GMM-2 and Deep-PrAE Mod perform empirically well. Figure 3 summarizes the zoomed-in performances of CE Naive, CE GMM-2, AMS, and Deep-PrAE Mod in terms of percentage error, which is the difference between the estimated and true probability as a percentage of the true value. It shows that while CE performs well when the IS parametric class is well-chosen (CE GMM-2), a poor CE parametric class (CE Naive) as well as AMS could under-estimate. Yet our Deep-PrAE, despite using samples from a poor CE class in Stage 1, can recover valid results: Deep-PrAE provides a valid UB, and Deep-PrAE Mod gives an estimate as good as the good CE class.

Intelligent Driving Example. In this example, we evaluate the crash probability of car-following scenarios involving a human-driven lead vehicle (LV) followed by an autonomous vehicle (AV). The AV is controlled by the Intelligent Driver Model (IDM) to maintain safety distance while ensuring smooth ride and maximum efficiency. IDM model is widely used for autonomy evaluation and microscopic transportation simulations (Treiber et al., 2000; Wang et al., 2018; Orzechowski et al., 2019). The state at time t is given by 6 states consisting of the position, velocity, and acceleration of both LV and AV. The dynamic system has a stochastic input u_t related to the acceleration of the LV and subject to uncertain human behavior. We consider an evaluation horizon $T = 60$ seconds and draw a sequence of 15 Gaussian random actions at a 4-second epoch, leading to a 15-dimensional LV action space. A (rare-event) crash occurs at time $t \le T$ if the longitudinal distance r_t between the two vehicles is negative, with γ parameterizing the AV maximum throttle and brake pedals. This rare-event set is analytically challenging (see Zhao et al. 2017 for a similar setting). More details are in Appendix F.

Figure 4 shows the performances of competing approaches, using $n = 10,000$. For CE, we use a single



(a) Estimated rare-event probability



(b) Estimator's empirical relative error

Figure 4: Intelligent driving example. Naive Monte Carlo failed in all cases and hence not shown.

Gaussian (CE Naive) and a large number of mixtures (CE GMM-64). Deep-PrAE and LL (UB and LB) appear consistent in giving upper and lower bounds for the target probability, and Deep-PrAE produces tighter bounds than LL (10^{-2} vs 10^{-6} in general). LL UB has $5,644$ dominating points when $\gamma = 1$ vs 42 in Deep-PrAE, and needs 4 times more time to search for them than Deep-PrAE. Moreover, the RE of Deep-PrAE is 3 times lower than LL across the range (in both UB and LB). Thus, Deep-PrAE outperforms LL in both tightness, computation speed, and RE. CE Naive and AMS seem to give a stable estimation, but evidence shows they are under-estimated: Deep-PrAE Mod and CE GMM-64 lack efficiency certificates and thus could under-estimate, and the fact that their estimates are higher than CE Naive and AMS suggests both CE Naive and AMS are under-estimated. Lastly, NMC fails to give a single hit in all cases (thus not shown on the graphs). Thus, among all approaches, Deep-PrAE stands out as giving reliable and tight bounds with low RE.

Summary of Practical Benefits. Our investigation shows strong evidence of the practical benefits of using Deep-PrAE for rare-event estimation. It generates valid bounds for the target probability with low RE and improved efficiency. The use of classifier prediction helps reduce the computational effort from running more simulations. For example, to assess whether the AV crash rate is below 10^{-8} for $\gamma = 1.0$, only 1000 simulation runs would be needed by Deep-PrAE UB or LB to get around 1% RE, which takes about 400 seconds in total. This is in contrast to 3.7 months for naive Monte Carlo.

5 Discussion and Future Work

In this paper, we proposed a robust certifiable approach to estimate rare-event probabilities in safety-critical applications. The proposed approach designs efficient IS distribution by combining the dominating point machinery with deep-learning-based rare-event set learning. We study the theoretical guarantees and present numerical examples. The key property that distinguishes our approach with existing black-box rare-event simulation methods is our correctness guarantee. Leveraging on a new notion of relaxed efficiency certificate and the orthogonal monotonicity assumption, our approach avoids the perils of undetected under-estimation as potentially encountered by other methods.

We discuss some key assumptions in our approach and related prospective follow-up works. First, the orthogonal monotonicity assumption appears an important first step to give new theories on black-box rare-event estimation beyond the existing literature. Indeed, we show that even with this assumption, black-box approaches such as CE and splitting can suffer from the dangerous pitfall of undiagnosed under-estimation, and our approach corrects for it. The real-world values of our approach are: (1) We rigorously show why our method has better performances in the orthogonally monotone cases; (2) For tasks close to being orthogonal monotonic (e.g., the IDM example), our method is empirically more robust; (3) For non-orthogonally-monotone tasks, though directly using our approach does not provide guarantees, we could potentially train mappings to latent spaces that are orthogonally monotone. We believe such type of geometric assumptions comprises a key ingredient towards a rigorous theory for black-box rare-event estimation that warrants much further developments.

Second, the dominating point search algorithm in our approach assumes Gaussian randomness. To this end we can relax it in two directions: by fitting a GMM with a sufficiently large number of components, or using light-tailed distributions (i.e., with finite exponential moments), since the dominating point machinery applies. These extensions will be left for future work.

Finally, the tightness of the upper bound depends on the sample quality. An ideal method in Stage 1 would generate samples close to the rare-event boundary to produce good approximations. Cutting the Stage 1 effort by, e.g., designing iterative schemes between Stages 1 and 2, will also be a topic for future investigation.

Acknowledgments

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710, IIS-1849280 and IIS-1849304. Mansur Arief and Wenhao Ding are supported in part by Bosch.

References

Amir Dembo, O. Z. (2010). *Large Deviations Techniques and Applications*. Springer-Verlag.

Anil, C., Lucas, J., and Grosse, R. (2019). Sorting out Lipschitz function approximation. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301, Long Beach, California, USA. PMLR.

Arief, M., Glynn, P., and Zhao, D. (2018). An accelerated approach to safely and efficiently test preproduction autonomous vehicles on public streets. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2006–2011. IEEE.

Arouna, B. (2004). Adaptative monte carlo method, a variance reduction technique. *Monte Carlo Methods and Applications*, 10(1):1–24.

Asmussen, S. and Glynn, P. W. (2007). *Rare-Event Simulation*, pages 158–205. Springer New York, New York, NY.

Asmussen, S. and Kroese, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability*, 38(2):545–558.

Au, S.-K. and Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277.

Blanchet, J. and Lam, H. (2012). State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1):38–59.

Botev, Z. I., L'Ecuyer, P., and Tuffin, B. (2013). Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing*, 23(2):271–285.

Botev, Z. I., Ridder, A., and Rojas-Nandayapa, L. (2016). Semiparametric cross entropy for rare-event simulation. *Journal of Applied Probability*, 53(3):633–649.

Bucklew, J. (2013). *Introduction to Rare Event Simulation*. Springer Science & Business Media.

Bucklew, J. A. (2004). *Rare Event Simulation for Level Crossing and Queueing Models*, pages 195–206. Springer New York, New York, NY.

Cao, Y. and Gu, Q. (2019). Tight sample complexity of learning one-hidden-layer convolutional neural networks. In *Advances in Neural Information Processing Systems 32*, pages 10612–10622. Curran Associates, Inc.

Cérou, F. and Guyader, A. (2007). Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443.

Cérou, F., Guyader, A., and Rousset, M. (2019). Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4):043108.

Chen, B., Blanchet, J., Rhee, C.-H., and Zwart, B. (2019). Efficient rare-event simulation for multiple jump events in regularly varying random walks and compound poisson processes. *Mathematics of Operations Research*, 44(3):919–942.

Clarke, E. M., Henzinger, T. A., Veith, H., and Bloem, R. (2018). *Handbook of Model Checking*, volume 10. Springer.

Claybrook, J. and Kildare, S. (2018). Autonomous vehicles: No driver... no regulation? *Science*, 361(6397):36–37.

Corso, A., Moss, R. J., Koren, M., Lee, R., and Kochenderfer, M. J. (2020). A survey of algorithms for black-box safety validation. *arXiv preprint arXiv:2005.02979*.

Cérou, F. and Guyader, A. (2016). Fluctuation analysis of adaptive multilevel splitting. *The Annals of Applied Probability*, 26(6):3319 – 3380.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67.

Dean, T. and Dupuis, P. (2009). Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic Processes and Their Applications*, 119(2):562–587.

Dieker, A. B. and Mandjes, M. (2006). Fast simulation of overflow probabilities in a queue with gaussian input. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 16(2):119–151.

Dupuis, P. and Ellis, R. S. (2011). *A Weak Convergence Approach to the Theory of Large Deviations*, volume 902. John Wiley & Sons.

Ellis, R. S. (1984). Large deviations for a general class of random vectors. *Ann. Probab.*, 12(1):1–12.

Evan, A. (2016). Fatal Tesla Self-Driving Car Crash Reminds Us That Robots Aren't Perfect. *IEEE Spectrum*.

Glasserman, P., Heidelberger, P., Shahabuddin, P., and Zajic, T. (1999). Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600.

Glynn, P. W. and Iglehart, D. L. (1989). Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392.

Grace, A. W., Kroese, D. P., and Sandmann, W. (2014). Automated state-dependent importance sampling for markov jump processes via sampling from the zero-variance distribution. *Journal of Applied Probability*, 51(3):741–755.

Guyader, A., Hengartner, N., and Matzner-Løber, E. (2011). Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics & Optimization*, 64(2):171–196.

Gärtner, J. (1977). On large deviations from the invariant measure. *Theory of Probability & Its Applications*, 22(1):24–39.

Harvey, N., Liaw, C., and Mehrabian, A. (2017). Nearly-tight VC-dimension bounds for piecewise linear neural networks. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1064–1068, Amsterdam, Netherlands. PMLR.

Huang, Z., Lam, H., LeBlanc, D. J., and Zhao, D. (2018). Accelerated evaluation of automated vehicles using piecewise mixture models. *IEEE Transactions on Intelligent Transportation Systems*, 19(9):2845–2855.

Huang, Z., Lam, H., and Zhao, D. (2018). Designing importance samplers to simulate machine learning predictors via optimization. In *2018 Winter Simulation Conference (WSC)*, pages 1730–1741. IEEE.

Juneja, S. and Shahabuddin, P. (2006). Rare-event simulation techniques: An introduction and recent advances. *Handbooks in Operations Research and Management Science*, 13:291–350.

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

Koopman, P. and Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1):90–96.

Koopman, P. and Wagner, M. (2018). Toward a framework for highly automated vehicle safety validation. Technical report, SAE Technical Paper.

Koren, M., Alsaif, S., Lee, R., and Kochenderfer, M. J. (2018). Adaptive stress testing for autonomous vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE.

L'ecuyer, P., Blanchet, J. H., Tuffin, B., and Glynn, P. W. (2010). Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 20(1):1–41.

Legriel, J., Le Guernic, C., Cotton, S., and Maler, O. (2010). Approximating the pareto front of multi-criteria optimization problems. In Esparza, J. and Majumdar, R., editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 69–83, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6231–6239. Curran Associates, Inc.

Müller, T., Mcwilliams, B., Rousselle, F., Gross, M., and Novák, J. (2019). Neural importance sampling. *ACM Trans. Graph.*, 38(5).

NHTSA (2007). The new car assessment program suggested approaches for future program enhancements. *DOT HS*, 810:698.

NTSB (2016). Preliminary Report, Highway HWY16FH018.

O'Kelly, M., Sinha, A., Namkoong, H., Tedrake, R., and Duchi, J. C. (2018). Scalable end-to-end autonomous vehicle testing via rare-event simulation. In *Advances in Neural Information Processing Systems*, pages 9827–9838.

Orzechowski, P. F., Li, K., and Lauer, M. (2019). Towards responsibility-sensitive safety of automated vehicles with reachable set analysis. In *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*, pages 1–6.

Rubinstein, R. Y. (2005). A stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation. *Methodology and Computing in Applied Probability*, 7(1):5–50.

Rubinstein, R. Y. and Kroese, D. P. (2013). *The Cross-entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer Science & Business Media.

Sadowsky, J. S. and Bucklew, J. A. (1990). On large deviations theory and asymptotically efficient monte carlo estimation. *IEEE Transactions on Information Theory*, 36(3):579–588.

Siegmund, D. (1976). Importance Sampling in the Monte Carlo Study of Sequential Tests. *The Annals of Statistics*, 4(4):673 – 684.

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

Treiber, M., Hennecke, A., and Helbing, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805–1824.

Tuffin, B. and Ridder, A. (2012). Probabilistic bounded relative error for rare event simulation learning techniques. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12. IEEE.

Uesato, J., Kumar, A., Szepesvari, C., Erez, T., Ruderman, A., Anderson, K., Heess, N., and Kohli, P. (2018). Rigorous agent evaluation: An adversarial approach to uncover catastrophic failures. *arXiv preprint arXiv:1812.01647*.

van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes. *Springer Series in Statistics*.

Villén-Altamirano, M. and Villén-Altamirano, J. (1994). Restart: a straightforward method for fast simulation of rare events. In *Proceedings of Winter Simulation Conference*, pages 282–289. IEEE.

Villén-Altamirano, J. (2010). Importance functions for restart simulation of general jackson networks. *European Journal of Operational Research*, 203(1):156–165.

Wang, X., Jiang, R., Li, L., Lin, Y., Zheng, X., and Wang, F. (2018). Capturing car-following behaviors by deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):910–920.

Webb, S., Rainforth, T., Teh, Y. W., and Kumar, M. P. (2018). A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*.

Wegener, J. and Bühler, O. (2004). Evaluation of different fitness functions for the evolutionary testing of an autonomous parking system. In *Genetic and Evolutionary Computation Conference*, pages 1400–1412. Springer.

Wu, X., Gomes-Selman, J., Shi, Q., Xue, Y., García-Villacorta, R., Anderson, E., Sethi, S., Steinschneider, S., Flecker, A., and Gomes, C. P. (2018). Efficiently approximating the pareto frontier: Hydropower dam placement in the amazon basin. In *AAAI*.

Zhao, D., Huang, X., Peng, H., Lam, H., and LeBlanc, D. J. (2017). Accelerated evaluation of automated vehicles in car-following maneuvers. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):733–744.

Zhao, D., Lam, H., Peng, H., Bao, S., LeBlanc, D. J., Nobukawa, K., and Pan, C. S. (2016). Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques. *IEEE Transactions on Intelligent Transportation Systems*, 18(3):595–607.