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.

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](https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12388) [Koopman and Wagner, 2017](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6288908/) [Uesato et al., 2018](https://arxiv.org/abs/1806.08432)). 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](https://www.blastfromtheblog.com/2016/11/24/advancing-safety-in-automotive-intelligence/) [Kalra and Paddock, 2016](https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12388) [NTSB, 2016](https://www.ntsb.gov/Public_Assets/Reports/2016/NTSB_A2016-015.pdf)). 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](https://ieeexplore.ieee.org/document/8557501)). Automated scenario selection approaches generate test cases based on domain knowledge ([Wegener and Buhler, 2004](https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9707.00163?journalCode=tiec) or adaptive searching algorithms (such as adaptive stress testing, [Koren et al., 2018](https://arxiv.org/abs/1806.08432)), which is more implementable but falls short of rigor. Test matrix approaches, such as Euro NCAP ([NTSBA, 2007](https://www.ntsba.org/content/ntsb-policy-manual5)), 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-
Deep Probabilistic Accelerated Evaluation (Deep-PrAE)

Deep Probabilistic Accelerated Evaluation (Deep-PrAE) is a framework that addresses the challenges of rare-event simulation, particularly in the context of safety-critical systems. It combines the power of deep learning with traditional variance reduction techniques to create an efficient and scalable solution for estimating the probability of rare events. The framework consists of three main ingredients:

1. **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.

2. **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.

3. **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.

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.

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.

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.

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.

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.

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.

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.

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.

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.
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., \( S_\gamma = \{ x : f(x) \geq \gamma \} \) for some risk function \( f \) and exceedance threshold \( \gamma \)). 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 \( \mu \), by running \( n \) simulation runs in total. Since \( \mu \) is tiny, the error of a meaningful estimation must be measured in relative term, i.e., we would like

\[
P(|\hat{\mu}_n - \mu| > \epsilon \mu) \leq \delta
\]

(1)

where \( \delta \) 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 \( RE = \text{Var}(Z_i)/\mu^2 \) as the ratio of variance (per-run) and squared mean. Importantly, to attain (1), a sufficient condition is \( n \geq RE/(\delta^2) \). So, when \( RE \) is large, the required Monte Carlo size is also large.

Challenges in Naive Monte Carlo. Let \( Z_i = I(X_i \in S_\gamma) \) 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 \( \mu \) is tiny). This demanding condition is a manifestation of the difficulty in hitting \( S_\gamma \). In the standard large deviations regime (Amir Dembo 2010 Dupuis and Ellis 2011) where \( \mu \) is exponentially small in \( \gamma \), the required Monte Carlo size \( n \) would grow exponentially in \( \gamma \).

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 \( \mu \) if it achieves \( (1) \) with \( n = O(\log(1/\epsilon)) \), for given \( 0 < \epsilon, \delta < 1 \).

In the above, \( O(\cdot) \) 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 \( RE = \tilde{O}(\log(1/\epsilon)) \). Note that in the large deviations regime, the sample size \( n \) used in a certifiable estimator is reduced from exponential in \( \gamma \) naive Monte Carlo to polynomial in \( \gamma \).

Importance sampling (IS) is a prominent technique to achieve efficiency certificate (Glynn and Iglehart 1989). IS generates \( X \) from another distribution \( \hat{p} \) (called IS distribution), and outputs \( \hat{\mu}_n = (1/n) \sum_{i=1}^{n} L(X_i)I(X_i \in S_\gamma) \) where \( L = dp/d\hat{p} \) is the likelihood ratio, or the Radon-Nikodym derivative, between \( p \) and \( \hat{p} \). Via a change of measure, it is easy to see that \( \hat{\mu}_n \) is unbiased for \( \mu \). The key is to control its \( RE \) by selecting a good \( \hat{p} \). 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 2000).

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 \( \Sigma \) is positive definite.

Definition 2. A set \( A_\gamma \subset \mathbb{R}^d \) is a dominating set for the set \( S_\gamma \subset \mathbb{R}^d \) associated with the distribution \( N(\lambda, \Sigma) \) if for any \( x \in S_\gamma \), 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 \( A_\gamma \) is removed, then the remaining set no longer satisfies the above condition. We call any point in \( A_\gamma \) a dominating point.

The dominating set comprises the “corner” cases where the rare event occurs (Sadowsky and Buckley 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 \( \lambda \) 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 \( S_\gamma = \bigcup_j \mathcal{S}^j_\gamma \), where each \( \mathcal{S}^j_\gamma \) 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 (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 \( \mu = P(X \in S_\gamma) \).

On the contrary, if the Gaussian (mixture) IS distri-
Theorem 1 (Perils of under-estimation). Suppose we estimate \( \mu = P(X \geq \gamma) \) where \( X \sim p = N(0,1) \) and \( 0 < k < 3 \). We choose \( \bar{p} = N(\gamma, 1) \) as the IS distribution to obtain \( \hat{\mu}_n \). Then 1) The relative error of \( \hat{\mu}_n \) grows exponentially in \( \gamma \). 2) If \( n \) is polynomial in \( \gamma \), we have \( P\left( |\hat{\mu}_n - \Phi(\gamma)| > \varepsilon \Phi(\gamma) \right) = O\left( \varepsilon^{-2} \right) \) for any \( \varepsilon > 0 \) where \( \Phi(\gamma) = P(X \geq \gamma) < \mu \), and the empirical relative error grows only polynomially in \( \gamma \), 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) 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 \( \gamma \)) 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 (Atanassov 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.

Figure 1: (a) An example of \( \mathcal{S}_r \) with two dominating points (b) Outer- and inner- approximations of \( \mathcal{S}_r \) (c) \( \bar{\kappa} \) tuning for Stage 1 Alg. (d) IS proposals: too few dominating points for CE with no simple IS class, too many for LL, and a balance for Deep-PrAE.
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 2, 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 \( \mu \) if \( P(\hat{\mu}_n - \mu < -\epsilon \mu) \leq \delta \) with \( n \geq O(\log(1/\mu)) \), for given \( 0 < \epsilon, \delta < 1 \).

Compared with the efficiency certificate in [14]. Definition 3 is relaxed to only requiring \( \hat{\mu}_n \) to be an upper bound of \( \mu \), 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 \( \mu \) is more crucial, and the lower bound serves to assess an estimation gap. The following provides a handy certification:

**Proposition 1** (Achieving relaxed efficiency certificate). Suppose \( \hat{\mu}_n \) is upward biased, i.e., \( \mathbb{E}[\hat{\mu}_n] \geq \mu \). Moreover, suppose \( \hat{\mu}_n \) takes the form of an average of \( n \) i.i.d. simulation runs \( Z_i \), with \( RE = Var(Z_i)/\mathbb{E}[Z_i]^2 = O(\log(1/\mathbb{E}[Z_i])) \). 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 D 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 \( \mu = P(X \in S_\gamma) \). Suppose we can learn a set \( \mathcal{S}_\gamma \) with any number \( n_1 \) of i.i.d. samples \( D_{n_1} \) (drawn from some distribution) such that \( \mathcal{S}_\gamma \supset S_\gamma \) with probability 1. Also suppose that there is an efficiency certificate for an IS estimator for \( \Pi(D_{n_1}) := P(X \in \mathcal{S}_\gamma) \). Then a two-stage estimator where a constant \( n_2 \) number of samples \( D_{n_2} \) are first used to construct \( \mathcal{S}_\gamma \), and \( n_2 = \tilde{O}(\log(1/\Pi(D_{n_1}))) \) 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 \( \{(\bar{X}_i, Y_i)\}_{i=1,...,n_1} \), where \( Y_i = 1 \) if \( \bar{X}_i \) is in the rare-event set \( S_\gamma \), and 0 otherwise. Here, it is beneficial to use Stage 1 samples that have sufficient presence in \( S_\gamma \), which can be achieved via any black-box variance reduction methods. We then consider the pairs \( \{(\bar{X}_i, Y_i)\} \) where \( \bar{X}_i \) 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 \( \mathcal{S}_\gamma \) is taken to be \( \{x : \hat{g}(x) \geq \kappa\} \) for some threshold \( \kappa \in \mathbb{R} \).

The outer approximation requirement \( \mathcal{S}_\gamma \supset S_\gamma \) 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.,

\[
P(X \in \mathcal{S}_\gamma, Y = 1) = 0 \tag{2}
\]

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 \( S_\gamma \), 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 \( S \subset \mathbb{R}^d_+ \) orthogonally monotone if for any two points \( x, x' \in \mathbb{R}^d_+ \), we have \( x \preceq x' \) (where the inequality is defined coordinate-wise) and \( x \in S \) implies \( x' \in S \) 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 \( \mathcal{H}(D) = \cup_i R(x_i) \), where \( R(x_i) \) 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 \( T_0 = \{\bar{X}_i : Y_i = 0\} \) as the non-rare-event sampled points. Evidently, if \( S_\gamma \) is orthogonally monotone, then \( \mathcal{H}(T_0) \subset \mathcal{S}_\gamma \) (where complement is with respect to \( \mathbb{R}^d_+ \)), or equivalently, \( \mathcal{H}(T_0)^c \supset S_\gamma \), i.e., \( \mathcal{H}(T_0)^c \) is an outer approximation of the rare-event set \( S_\gamma \). 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 \( \mu \) 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 \( \cup_i (R(x_i))^c \), 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 \( S_\gamma \) (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 \( \{(\bar{X}_i, Y_i)\} \), and obtain an approximate non-rare-event region \( \mathcal{S}_\gamma^c = \{x : \hat{g}(x) < \kappa\} \), where \( \kappa \) is say 1/2. Then we adjust \( \kappa \) minimally away from 1/2, say to \( \tilde{\kappa} \), so that \( \mathcal{S}_\gamma^c \subset \mathcal{H}(T_0) \), i.e., \( \tilde{\kappa} = \max\{\kappa \in \mathbb{R} : \mathcal{S}_\gamma^c \subset \mathcal{H}(T_0)\} \). Then \( \mathcal{S}_\gamma^c \subset \mathcal{H}(T_0)^c \supset S_\gamma \), so that \( \mathcal{S}_\gamma^c \) is an outer approximation for \( S_\gamma \) (see Figure 1(c), where \( \tilde{\kappa} = 0.68 \)). Stage 1 in Algorithm 4 shows this procedure. With this, we can run mixture IS to estimate \( P(X \in \mathcal{S}_\gamma^c) \) in Stage 2.

The execution of this algorithm requires the set \( \mathcal{S}_\gamma^c = \{x : \hat{g}(x) \geq \tilde{\kappa}\} \) to be in a form susceptible to The-
orem and the search of all its dominating points. When \( \hat{g} \) is a ReLU-activated neural net, the boundary of \( \hat{g}(x) \geq \hat{k} \) is piecewise linear and \( \mathcal{S}_k^\gamma \) 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 \( \mathcal{S}_k^\gamma \) 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 \( \mathcal{S}_k \) is orthogonally monotone, and \( \mathcal{S}_k^\infty \) satisfies the same conditions for \( \mathcal{S}_k \) in Theorem 3. Then Algorithm 2 attains the upper-bound relaxed efficiency certificate by using a constant number of Stage 1 samples.

Figure 4 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 4 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 \( P(X \in \mathcal{S}_k^\infty, Y = 0) \) (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 \( g \) and \( \mathcal{S}_k^\infty \) is piecewise linear and \( \mathcal{S}_k^\gamma \) is orthogonally monotone. The cross-entropy method can miss a dominating point (1) and result in systematic under-estimation.

**Theorem 3** (Conservativeness). Consider \( \mathcal{S}_k^\infty \) obtained in Algorithm 4 where \( \hat{g} \) is trained from an ERM. Suppose the density \( g \) has bounded support \( K \subset [0, M]^d \) and \( 0 < q_i \leq g(x) \leq q_k \) 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 \( \ell \) is the squared loss, then \( h(\kappa) \) could be chosen as \( h(\kappa) = \kappa^2 \). Then, with probability at least \( 1 - \delta \),

\[
P_{X \sim q}(X \in \mathcal{S}_k^\infty \setminus \mathcal{S}_k) \leq \frac{d_{\text{Lip}}(g^*,\hat{g})^2}{\ell(q_M) + \log(M+\log \frac{1}{\delta}) \log \frac{q_{\kappa}}{q_i}}.
\]

Here, \( d_{\text{Lip}}(g^*,\hat{g}) \) is the Lipschitz parameter of \( g^* \), and \( \ell(q_M) = 3 \left( \log(q_{\kappa}/q_i) + \log M + \log \frac{1}{\delta} \right)^{\frac{1}{2}} \).

Theorem 3 reveals a tradeoff between overfitting (measured by \( \sup_{g \in \mathcal{G}} |R_{n_1}(g) - R(g)| \)) and \( \ell(\kappa) \) 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 novel 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 \( \mathcal{S}_k^\infty \) by \( \mathcal{S}_k \) in the last step of Algorithm 4 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
Deep Probabilistic Accelerated Evaluation (Deep-PrAE)

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

2D Example. We estimate $\mu = P(X \in S)$ where $X \sim N([5, 5]^T, 0.25I_{2 \times 2})$, and $\gamma$ 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 $S_\gamma$, 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.

![Figure 1: 2-dimensional example. Naive Monte Carlo failed in all cases and hence not shown.](image)

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)](image)

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 $t \leq T$ if the longitudinal distance $r_t$ between the two vehicles is negative, with $\gamma$ 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](image)
Arief, Huang, Kumar, Bai, He, Ding, Lam, Zhao

(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 monotone (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


