# Appendices

We present supplemental results and discussions. Appendix A expands Section 2 regarding Monte Carlo efficiency and variance reduction. Appendix B provides further details on Algorithm 1, in particular the mixed integer formulation used to solve the underlying optimization problems. Appendix C expands the efficiency and conservativeness results in Section 3. Appendix D presents the lower-bound relaxed efficiency certificate and estimators in parallel to the upper-bound results in Section 3. Appendix E provides an overview of the cross-entropy method and multi-level splitting (or subset simulation) and discusses their perils for black-box problems. Appendix F illustrates further experimental results. Finally, Appendix G shows all technical proofs.

## A Further Details for Section 2

This section expands the discussions in Section 2, by explaining in more detail the notion of relative error, challenges in naive Monte Carlo, the concept of dominating points, and the perils of black-box variance reduction algorithms.

### A.1 Explanation of the Role of Relative Error

As described in Section 2, to estimate a tiny $\mu$ using $\hat{\mu}_n$, we want to ensure a high accuracy in relative term, namely, (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 $Z_i$. The Markov inequality gives that

$$P(|\hat{\mu}_n - \mu| > \epsilon\mu) \leq \frac{Var(\hat{\mu}_n)}{\epsilon^2\mu^2} = \frac{Var(Z_i)}{n\epsilon^2\mu^2}$$

so that

$$\frac{Var(Z_i)}{n\epsilon^2\mu^2} \leq \delta$$

ensures (1). Equivalently,

$$n \geq \frac{Var(Z_i)}{\delta\epsilon^2\mu^2} = \frac{RE}{\delta\epsilon^2}$$

is a sufficient condition to achieve (1), where $RE = Var(Z_i)/\mu^2$ is the relative error defined as the ratio of variance (per-run) and squared mean.

Note that replacing the second $\mu$ with $\hat{\mu}_n$ in the left hand side of (1) does not change the condition fundamentally, as either is equivalent to saying the ratio $\hat{\mu}_n/\mu$ should be close to 1. Also, note that we focus on the nontrivial case that the target probability $\mu$ is non-zero but tiny; if $\mu = 0$, then no non-zero Monte Carlo estimator can achieve a good relative error.

### A.2 Further Explanation on the Challenges in Naive Monte Carlo

We have seen in Section 2 that for the naive Monte Carlo estimator, where $Z_i = I(X_i \in \mathcal{S}_\gamma)$, the relative error is $RE = \mu(1-\mu)/\mu^2 = (1-\mu)/\mu$. Thus, when $\mu$ is tiny, the sufficient condition for $n$ to attain (1) scales at least linearly in $1/\mu$. In fact, this result can be seen to be tight by analyzing $n\hat{\mu}_n$ as a binomial variable. To be more specific, we know that $P(|\hat{\mu}_n - \mu| > \varepsilon\mu) = P(|n\hat{\mu}_n - n\mu| > \varepsilon n\mu)$ and that $n\hat{\mu}_n$ takes values in $\{0, 1, \ldots, n\}$. Therefore, if $n\mu \to 0$, then $P(|\hat{\mu}_n - \mu| > \varepsilon\mu) \to 1$, and hence (1) does not hold.

Moreover, the following provides a concrete general statement that an $n$ that grows only polynomially in $\gamma$ would fail to estimate $\mu$ that decays exponentially in $\gamma$ with enough relative accuracy, of which (1) fails to hold is an implication.

**Proposition 2.** *Suppose that $\mu = P(X \in \mathcal{S}_\gamma)$ is exponentially decaying in $\gamma$ and $n$ is polynomially growing in $\gamma$. Define $\hat{\mu}_n = (1/n)\sum_{i=1}^n I(X_i \in \mathcal{S}_\gamma)$. Then for any $0 < \varepsilon < 1$,*

$$\lim_{\gamma\to\infty} P(|\hat{\mu}_n - \mu| > \varepsilon\mu) = 1.$$

We have used the term efficiency certificate to denote an estimator that achieves (1) with $n = \tilde{O}(\log(1/\mu))$. In the rare-event literature, such an estimator is known as "logarithmically efficient" or "weakly efficient" (Juneja and Shahabuddin, 2006; Blanchet and Lam, 2012).

### A.3 Further Explanations of Dominating Points

We have mentioned that a certifiable IS should account for all dominating points, defined in Definition 2. We provide more detailed explanations here. Roughly speaking, for $X \sim N(\lambda, \Sigma)$ and a rare-event set $\mathcal{S}_\gamma$, the Laplace approximation gives $P(X \in \mathcal{S}_\gamma) \approx e^{-\inf_{a\in\mathcal{S}_\gamma} \frac{1}{2}(a-\lambda)^T\Sigma^{-1}(a-\lambda)}$ (see the proof of Theorem 4). Thus, to obtain an efficiency certificate, IS estimator given by $Z = L(X)I(X \in \mathcal{S}_\gamma)$, where $X \sim \tilde{p}$ and $L = dp/d\tilde{p}$, needs to have $\widetilde{Var}(Z) \leq \tilde{E}[Z^2] \approx e^{-\inf_{a\in\mathcal{S}_\gamma}(a-\lambda)^T\Sigma^{-1}(a-\lambda)}$ (where $\widetilde{Var}(\cdot)$ and $\tilde{E}[\cdot]$ denote the variance and expectation under $\tilde{p}$, and $\approx$ is up to some factor polynomial in $\inf_{a\in\mathcal{S}_\gamma}(a-\lambda)^T\Sigma^{-1}(a-\lambda)$; note that the last equality relation cannot be improved, as otherwise it would imply that $\widetilde{Var}(Z) = \tilde{E}[Z^2] - (\tilde{E}[Z])^2 < 0$).

Now consider an IS that translates the mean of the distribution from $\mu$ to $a^* = \text{argmin}_{a\in\mathcal{S}_\gamma}(a-\lambda)^T\Sigma^{-1}(a-\lambda)$, an intuitive choice since $a^*$ contributes the highest density among all points in $\mathcal{S}_\gamma$ (this mean translation also bears the natural interpretation as an exponential change of measure; (Bucklew, 2013)). The likelihood ratio is $L(x) = e^{(\mu-a^*)^T\Sigma^{-1}(x-\lambda)+\frac{1}{2}(\lambda-a^*)^T\Sigma^{-1}(\lambda-a^*)}$, giving

$$\tilde{E}[Z^2] = \tilde{E}[L(X)^2 I(X \in \mathcal{S}_\gamma)] = e^{-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)}\tilde{E}[e^{-2(a^*-\lambda)^T\Sigma^{-1}(x-a^*)}I(X \in \mathcal{S}_\gamma)] \qquad (3)$$

If the "overshoot" $(a^*-\lambda)^T\Sigma^{-1}(x-a^*)$, i.e., the remaining term in the exponent of $L(x)$ after moving out $-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)$, satisfies $(a^*-\lambda)^T\Sigma^{-1}(x-a^*) \geq 0$ for all $x \in \mathcal{S}_\gamma$, then the expectation in the right hand side of (3) is bounded by 1, and an efficiency certificate is achieved. This, however, is not true for all set $\mathcal{S}_\gamma$, which motivates the following definition of the dominant set and points in Definition 2.

For instance, if $\mathcal{S}_\gamma$ is convex, then, noting that $(x-\lambda)^T\Sigma^{-1}$ is precisely the gradient of the function $(1/2)(x-\lambda)^T\Sigma^{-1}(x-\lambda)$, we get that $a^*$ gives a singleton dominant set since $(a^*-\lambda)^T\Sigma^{-1}(x-a^*) \geq 0$ for all $x \in \mathcal{S}_\gamma$ is precisely the first order optimality condition of the involved quadratic optimization. In general, if we can decompose $\mathcal{S}_\gamma = \bigcup_j \mathcal{S}_\gamma^j$ where $\mathcal{S}_\gamma^j = \{x : (a_j-\lambda)^T\Sigma^{-1}(x-a_j) \geq 0\}$ for a dominating point $a_j \in A_\gamma$, then each $\mathcal{S}_\gamma^j$ can be viewed as a "local" region where the dominating point $a_j$ is the highest-density, or the most likely point such that the rare event occurs.

The following is the detailed version of Theorem 2:

**Theorem 4** (Certifiable IS)**.** *Suppose that $A_\gamma$ is the dominant set for $\mathcal{S}_\gamma$ associated with the distribution $N(\lambda, \Sigma)$. Then we can decompose $\mathcal{S}_\gamma = \bigcup_j \mathcal{S}_\gamma^j$ where $\mathcal{S}_\gamma^j$'s are disjoint, $a_j \in \mathcal{S}_\gamma^j$ and $\mathcal{S}_\gamma^j \subset \{x : (a_j-\lambda)^T\Sigma^{-1}(x-a_j) \geq 0\}$ for $a_j \in A_\gamma$. Denote $a^* = \arg\min\{(a_j-\lambda)^T\Sigma^{-1}(a_j-\lambda) : a_j \in A_\gamma\}$. Assume that each component of $a^*$ is of polynomial growth in $\gamma$. Moreover, assume that there exist invertible matrix $B$ and positive constant $\varepsilon$ such that $\{x : B(x-a^*) \geq 0, (x-a^*)^T\Sigma^{-1}(x-a^*) \leq \varepsilon^2\} \subset \mathcal{S}_\gamma$. Then the IS distribution $\sum_j \alpha_j N(a_j, \Sigma)$ achieves an efficiency certificate in estimating $\mu = P(X \in S_\gamma)$, i.e., if we let $Z = I(X \in \mathcal{S}_\gamma)L(X)$ where $L$ is the corresponding likelihood ratio, then $\tilde{E}[Z^2]/\tilde{E}[Z]^2$ is at most polynomially growing in $\gamma$. This applies in particular to $\mathcal{S}_\gamma = \{x : f(x) \geq \gamma\}$ where $f(x)$ is a piecewise linear function.*

We contrast Theorem 4 with existing works on dominating points. The latter machinery has been studied in (Sadowsky and Bucklew, 1990; Dieker and Mandjes, 2006). These papers, however, consider regimes where the Gärtner-Ellis Theorem (Gärtner, 1977; Ellis, 1984) can be applied, which requires the considered rare-event set to scale proportionately with the rarity parameter. This is in contrast to the general conditions on the dominating points used in Theorem 4.

### A.4 Further Explanation of the Example in Theorem 1

In the theorem, there are two dominating points $\gamma$ and $-k\gamma$ but the IS design only considers the first one. As a result, there could exist "unlucky" scenario where the sample falls into the rare-event set, so that $I(X \in \mathcal{S}_\gamma) = 1$, while the likelihood ratio $L(X)$ explodes, which leads to a tremendous estimation variance. Part 2 of the theorem further shows how this issue is undetected empirically, as the empirical RE appears small (polynomially in $n$ and hence $\gamma$ by our choice of $n$) while the estimation concentrates at a value that can be severely under the correct one (especially when $k < 1$). This is because the samples all land on the neighborhood of the solely considered dominating point. If the missed dominating point is a significant contributor to the rare-event probability, then the empirical performance would look as if the rare-event set is *smaller*, leading to a systematic under-estimation. Note that this phenomenon occurs even if the estimator is unbiased, which is guaranteed by IS by default.

## B Further Details on Implementing Algorithm 1

We provide further details on implementing Algorithm 1. In particular, we present how to solve the optimization problem

$$x^* = \arg\min_x \ (x-\lambda)^T\Sigma^{-1}(x-\lambda) \quad \text{s.t.} \quad \hat{g}(x) \geq \hat{\kappa}, \quad (x_j^* - \lambda)^T\Sigma^{-1}(x - x_j^*) < 0 \ \forall x_j^* \in \hat{A}_\gamma \tag{4}$$

to obtain the next dominating point in the sequential cutting-plane approach in Stage 2. Moreover, we also present how to tune

$$\hat{\kappa} = \max\{\kappa \in \mathbb{R} : (\overline{\mathcal{S}}_\gamma^\kappa)^c \subset \mathcal{H}(T_0)\} \tag{5}$$

in Stage 1.

**MIP formulations for ReLU-activated neural net classifier.** The problem (4) can be reformulated into a mixed integer program (MIP), in the case where $\hat{g}(x)$ is trained via a ReLU-activated neural net classifier, which is used in our deep-learning-based IS. Since the objective is convex quadratic and second set of constraints is linear in (4), we focus on the first constraint $\hat{g}(x) \geq \gamma$. The neural net structure $\hat{g}(x)$ in our approach (say with $n_g$ layers) can be represented as $\hat{g}(x) = (\hat{g}_{n_g} \circ ... \circ \hat{g}_1)(x)$, where each $\hat{g}_i(\cdot)$ denotes a ReLU-activated layer with linear transformation, i.e. $\hat{g}_i(\cdot) = \max\{LT(\cdot), 0\}$, where $LT(\cdot)$ denotes a certain linear transformation in the input. In order to convert $\hat{g}(\cdot)$ into an MIP constraint, we introduce $M$ as a practical upper bound for $x_1, ..., x_n$ such that $|x_i| < M$. The key step is to reformulate the ReLU function $y = \max\{x, 0\}$ into

$$\begin{aligned} &y \leq x + M(1-z) \\ &y \geq x \\ &y \leq Mz \\ &y \geq 0 \\ &z \in \{0,1\}. \end{aligned}$$

For simple ReLU networks, the size of the resulting MIP formulation depends linearly on the number of neurons in the neural network. In particular, the number of binary decision variables is linearly dependent on the number of ReLU neurons, and the number of constraints is linearly dependent the total number of all neurons (here we consider the linear transformations as independent neurons).

The MIP reformulation we discussed can be generalized to many other popular piecewise linear structures in deep learning. For instance, linear operation layers, such as normalization and convolutional layers, can be directly used as constraints; some non-linear layers, such as ReLU and max-pooling layers, introduce non-linearity by the "max" functions. A general reformulation for the max functions can be used to convert these non-linear layers to mixed integer constraints.

Consider the following equality defined by a max operation $y = \max\{x_1, x_2, ..., x_n\}$. Then the equality is equivalent to

$$\begin{aligned} &y \leq x_i + 2M(1-z_i), i = 1, ..., n \\ &y \geq x_i, i = 1, ..., n \\ &\sum_{i=1,...,n} z_i = 1 \\ &z_i \in \{0,1\}. \end{aligned}$$

**Tuning $\hat{\kappa}$.** We illustrate how to tune $\hat{\kappa}$ to achieve (5). This requires checking, for a given $\kappa$, whether $(\overline{\mathcal{S}}_\gamma^\kappa)^c \subset \mathcal{H}(T_0)$. Then, by discretizing the range of $\kappa$ or using a bisection algorithm, we can leverage this check to obtain (5).

We use an MIP to check $(\overline{\mathcal{S}}_\gamma^\kappa)^c \subset \mathcal{H}(T_0)$. Recall that $\mathcal{H}(T_0) = \bigcup_{i:Y_i=0}\{x \in \mathbb{R}_+^d : x \leq \tilde{X}_i\}$. We want to check if $\{x \in \mathbb{R}_+^d : \hat{g}(x) \leq \kappa\}$ for a given $\kappa$ lies completely inside the hull, where $\hat{g}(x)$ is trained with a ReLU-activated neural net. This can be done by solving an optimization problem as follows. First, we rewrite $\mathcal{H}(T_0)$ as $\{x \in \mathbb{R}_+^d : \min_{i=1,\dots,n}\max_{j=1,\dots,d}\{x^j - \tilde{X}_i^j\} \leq 0\}$, where $x^j$ and $x_i^j$ refer to the $j$-th components of $x$ and $\tilde{X}_i$ respectively. Then we solve

$$\begin{array}{ll} \max_{x\in\mathbb{R}^d} & \min_{i=1,\dots,n}\max_{j=1,\dots,d}\{x^j - \tilde{X}_i^j\} \\ \text{subject to} & \hat{g}(x) \leq \kappa \\ & x \geq 0 \end{array} \tag{6}$$

If the optimal value is greater than 0, this means $\{x \in \mathbb{R}_+^d : \hat{g}(x) \leq \kappa\}$ is not completely inside $\mathcal{H}(T_0)$, and vice versa. Now, we rewrite (6) as

$$\begin{array}{ll} \max_{x\in\mathbb{R}^d,\beta\in\mathbb{R}} & \beta \\ \text{subject to} & \max_{j=1,\dots,d}\{x^j - \tilde{X}_i^j\} \geq \beta \ \forall i = 1,\dots,n \\ & \hat{g}(x) \leq \kappa \\ & x \geq 0 \end{array} \tag{7}$$

We then rewrite (7) as an MIP by introducing a large real number $M$ as a practical upper bound for all coordinates of $x$:

$$\begin{array}{ll} \max_{x\in\mathbb{R}^d,\beta\in\mathbb{R}} & \beta \\ \text{subject to} & x^j - \tilde{X}_i^j + 4M(1 - z_{ij}) \geq \beta \quad \forall i = 1,\dots,n, j = 1,\dots,d \\ & \sum_{j=1,\dots,d} z_{ij} \geq 1 \quad \forall i = 1,\dots,n \\ & z_{ij} \in \{0,1\} \quad \forall i = 1,\dots,n, j = 1,\dots,d \\ & \hat{g}(x) \leq \kappa \\ & x \geq 0 \end{array} \tag{8}$$

Note that the set of points $T_0$ to be considered in constructing $\mathcal{H}(T_0)$ can be reduced to its "extreme points". More concretely, we call a point $x \in T_0$ an extreme point if there does not exist any other point $x' \in T_0$ such that $x \leq x'$. We can eliminate all points $x \in T_0$ such that $x \leq x'$ for another $x' \in T_0$, and the resulting orthogonal monotone hull would remain the same. If we carry out this elimination, then in (7) we need only consider $\tilde{X}_i$ that are extreme points in $\mathcal{H}(T_0)$, which can reduce the number of integer variables needed to add. In practice, we can also randomly remove points in $T_0$ to further reduce the number of integer variables. This would not affect the correctness of our approach, but would increase the conservativeness of the final estimate.

# C Further Results for Section 3

Here we present and discuss several additional results for Section 3 regarding estimation efficiency and conservativeness. The latter includes further theorems on the lazy-learner classifier and classifiers constructed using the difference of two functions, translation of the false positive rate under the Stage 1 sampling distribution to under the original distribution, and interpretations and refinements of the conservativeness results.

## C.1 Extending Upper-Bound Relaxed Efficiency Certificate to Two-Stage Procedures

We present an extension of Proposition 1 to two-stage procedures, which is needed to set up Corollary 1.

**Proposition 3** (Extended relaxed efficiency certificate). *Suppose constructing $\hat{\mu}_n = \hat{\mu}_{n_2}(D_{n_1})$ consists of two stages, with $n = n_1 + n_2$: First we sample $D_{n_1} = \{\tilde{X}_1,\dots,\tilde{X}_{n_1}\}$, where $\tilde{X}_i$ are i.i.d. (following some sampling distribution), and given $D_{n_1}$, we construct $\hat{\mu}_{n_2}(D_{n_1}) = (1/n_2)\sum_{i=1}^{n_2} Z_i$ where $Z_i$ are i.i.d. conditional on $D_{n_1}$ (following some distribution). Suppose $\hat{\mu}_n$ is conditionally upward biased almost surely, i.e., $\overline{\mu}(D_{n_1}) := E[\hat{\mu}_n|D_{n_1}] \geq \mu$, and the conditional relative error given $D_{n_1}$ in the second stage satisfies $RE(D_{n_1}) := Var(Z_i|D_{n_1})/\overline{\mu}(D_{n_1})^2 = \tilde{O}(\log(1/\overline{\mu}(D_{n_1})))$. If $n_1 = \tilde{O}(\log(1/\mu))$ (such as a constant number), then $\hat{\mu}_n$ possesses the upper-bound relaxed efficiency certificate.*

### C.2 Conservativeness of Lazy Learner

We provide a result to quantify the conservativeness of the lazy-learner IS in terms of the false positive rate. Recall that the lazy learner constructs the outer approximation of the rare-event set using $\mathcal{H}(T_0)^c$, which is the complement of the orthogonal monotone hull of the set of all non-rare-event samples. The conservativeness is measured concretely by the set difference between $\mathcal{H}(T_0)^c$ and $\mathcal{S}_\gamma$, for which we have the following result:

**Theorem 5** (Conservativeness of lazy learner)**.** *Suppose that the density $q$ has bounded support $K \subset [0, M]^d$, and $0 < q_l \le q(x) \le q_u$ for any $x \in K$. Then, with probability at least $1 - \delta$,*

$$P_{X\sim q}(X \in \mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma) \le M^{d-1} q_u \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1)$$

$$= \sqrt{\frac{e}{\pi(d-1)}} \left(\frac{1}{2}\pi e\right)^{\frac{d-1}{2}} q_u t(\delta, n_1)(1 + O(d^{-1})).$$

*Here $t(\delta, n_1) = 3\left(\frac{\log(n_1 q_l) + d\log M + \log\frac{1}{\delta}}{n_1 q_l}\right)^{\frac{1}{d}}$, $w_d$ is the volume of a $d-$dimensional Euclidean ball of radius 1, and the last $O(\cdot)$ is as $d$ increases.*

### C.3 Translating the False Positive Rate to under the Original distribution

Theorems 3 and 5 are stated with respect to $q$, the sampling distribution used in the first stage. We explain how to translate the false positive rate results to under the original distribution $p$. In the discussion below, we will consider Theorem 3 (and Theorem 5 can be handled similarly). In this case, our target is to give an upper bound to $P_{X\sim p}(X \in \mathcal{S}_\gamma^{\hat\kappa} \backslash \mathcal{S}_\gamma)$ based on the result of Theorem 3.

If the true input distribution $p$ does not have a bounded support, we can first choose $M$ to be large to make sure that $P_{X\sim p}(X \notin [0, M]^d)$ is small compared to the probability of $\mathcal{S}_\gamma$. We argue that we do not need $M$ to be too large here. Indeed, if $p$ is light tail (e.g., a distribution with tail probability exponential in $M$), then the required $M$ grows at most polynomially in $\gamma$.

Having selected $M$, and with the freedom in selecting $q$ in Stage 1, we could make sure that in $[0, M]^d$, $q(x)$ is bounded away from 0 (e.g., we can choose $q$ to be the uniform distribution over $[0, M]^d$). Then, by Theorem 3 and a change of measure argument, we can give a bound for $P_{X\sim p}(X \in [0, M]^d, X \in \mathcal{S}_\gamma^{\hat\kappa} \backslash \mathcal{S}_\gamma)$. Finally, we bound the false positive rate with respect to $p$ by $P_{X\sim p}(X \in \mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma) \le P_{X\sim p}(X \notin [0, M]^d) + P_{X\sim p}(X \in [0, M]^d, X \in \mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma)$.

### C.4 Conservativeness Results for Classifiers Constructed Using Differences of Two Trained Functions

Theorem 3 presents a conservativeness result when $\hat g$ is trained with an empirical risk minimization (ERM). In this subsection, we will show a more sophisticated version of Theorem 3, which corresponds more closely to the $\hat g$ that we implemented in our experiments. Suppose that the Stage 1 samples are generated in the same way as in Algorithm 1. We let $\mathcal{F} := \{f_\theta\}$ denote the function class induced by the model. Here a main difference with previously is that we allow functions in $\mathcal{F}$ to be 2-dimensional, and both the loss function and the classification boundary will be constructed from these 2-dimensional functions.

Suppose that $f_\theta$ is the output a neural network with 2 neurons in the output layer, and denote them as $f_{\theta,0}, f_{\theta,1}$. Let the loss function evaluated at the $i$-th sample be $\ell(f_\theta(\tilde X_i), Y_i)$. For example, the cross-entropy loss is given by $-\left[I(Y_i = 0)\log\frac{e^{f_{\theta,0}(\tilde X_i)}}{e^{f_{\theta,0}(\tilde X_i)} + e^{f_{\theta,1}(\tilde X_i)}} + I(Y_i = 1)\log\frac{e^{f_{\theta,1}(\tilde X_i)}}{e^{f_{\theta,0}(\tilde X_i)} + e^{f_{\theta,1}(\tilde X_i)}}\right]$. Like in the ERM approach in Theorem 3, we compute $\hat f = f_{\hat\theta} \in \mathcal{F}$ which is the minimizer of the empirical risk, i.e., $\hat f = \arg\min_{f_\theta \in \mathcal{F}} R_{n_1}(f_\theta)$. For each function $f_\theta \in \mathcal{F}$, define function $g_\theta$ as $g_\theta := f_{\theta,1} - f_{\theta,0}$. In this modified approach, the learned rare-event set would be given by $\tilde{\mathcal{S}}_\gamma^\kappa := \{x : g_{\hat\theta}(x) \ge \kappa\}$, and to make sure that $\mathcal{S}_\gamma \subset \tilde{\mathcal{S}}_\gamma^\kappa$, we would replace $\kappa$ by $\hat\kappa = \min\{g_{\hat\theta}(x) : x \notin \mathcal{H}(T_0)\}$ as in Step 1 of Algorithm 1.

We give a theorem similar to Theorem 3 for this more sophisticated procedure. To this end, we begin by giving some definitions similar to the set up of Theorem 3. Let $R(f_\theta) := E_{X\sim q}\ell(f_\theta(X), I(X \in \mathcal{S}_\gamma))$ denote the true risk function. Let $f^* = \arg\min_{f\in\mathcal{F}} R(f)$ denote the true risk minimizer within function class $\mathcal{F}$. Define $g^* = f_1^* - f_0^*$ accordingly and let $\kappa^* := \min_{x\notin\mathcal{S}_\gamma} g^*(x)$ denote the true threshold associated with $f^*$ in obtaining the smallest outer rare-event approximation.

**Theorem 6.** *Suppose that the density $q$ has bounded support $K \subset [0,M]^d$ and $0 < q_l \le q(x) \le q_u$ for any $x \in K$. Also suppose that there exists a function $h$ such that for any $f_\theta \in \mathcal{F}$, if $g_\theta(x) \ge \kappa$, we have $\ell(f_\theta(x), 0) \ge h(\kappa) > 0$ (for the cross entropy loss, this happens if we know that $f_\theta$ has a bounded range). Then, for the set $\tilde{\mathcal{S}}_\gamma^{\hat\kappa}$, with probability at least $1-\delta$,*

$$
\begin{aligned}
&P_{X\sim q}\left(X \in \tilde{\mathcal{S}}_\gamma^{\hat\kappa}, X \in \mathcal{S}_\gamma^c\right)\\
&\le \left(h(\kappa^* - t(\delta, n_1)\sqrt{d}Lip(g^*) - \|\hat g - g^*\|_\infty)\right)^{-1}\left(R(f^*) + 2\sup_{f_\theta\in\mathcal{F}} |R_{n_1}(f_\theta) - R(f_\theta)|\right)
\end{aligned}
$$

*Here $Lip(g^*)$ is the Lipschitz parameter of $g^*$, and $t(\delta, n_1)$ is defined as in Theorem 3.*

### C.5 Implications of Theorem 3 and Related Results in the Literature

First, we explain the trade-off between overfitting and underfitting. If the function class $\mathcal{G}$ is not rich, then $R(g^*) = \inf_{g\in\mathcal{G}} R(g)$ may be big because of the lack of expressive power. On the other hand, if the function class is too rich, then the generalization error will be huge. Here, the generalization error is represented by $\sup_{g_\theta\in\mathcal{G}} |R_{n_1}(g_\theta) - R(g_\theta)|$ as well as $t(\delta, n_1)\sqrt{d}\text{lip}(g^*) + \|\hat g - g^*\|_\infty$, which characterize the difference between the right hand side of the bound in the theorem and its limit as $n_1 \to \infty$.

Another question is how to give a more refined bound for the false positive rate based on Theorem 3 that depends on explicit constants of the classification model or training process. This would involve theoretical results for deep neural networks that are under active research. Let us examine the terms appearing in Theorem 3 and give some related results. In machine learning theory, the term $\sup_{g_\theta\in\mathcal{G}} |R_{n_1}(g_\theta) - R(g_\theta)|$ is often bounded by the Rademacher complexity of the function class (some results about the Rademacher complexity for neural networks are in Harvey et al. 2017; Cao and Gu 2019). The convergence of $\|\hat g - g^*\|_\infty$ to 0 as $n_1 \to \infty$ is implied by the convergence of the parameters, which is in turn justified by the empirical process theory (van der Vaart and Wellner, 1996). A bound for $\text{Lip}(g^*)$ could be potentially derived by adding norm constraints to the parameters in the neural network (Anil et al., 2019). On the other hand, if we let the network size grow to infinity, the class of neural networks can approximate any continuous function (Lu et al., 2017), and hence $R(g^*)$ can be arbitrarily small when the neural network is complex enough. However, if we restrict the choices of networks, for instance by the Lipschitz constant, then no results regarding the sufficiency of its expressive power for arbitrary functions are available in the literature to our knowledge, and thus it appears open how to simultaneously give bounds for $\text{Lip}(g^*)$ and $R(g^*)$. Future investigations on the expressive power of restricted classes of neural networks would help refining our conservativeness results further.

## D Lower-Bound Efficiency Certificate and Estimators

In Section 3, we described an approach that gives an estimator for the rare-event probability with an upper-bound relaxed efficiency certificate. Here we present analogous definitions and results on the lower-bound relaxed efficiency certificate. This lower-bound estimator gives an estimation gap for the upper-bound estimator. Moreover, by combining both of them, we can obtain an interval for the target rare-event probability.

The lower-bound relaxed efficiency certificate is defined as follows (compare with Definition 3):

**Definition 6.** *We say an estimator $\hat\mu_n$ satisfies an lower-bound* relaxed efficiency certificate *to estimate $\mu$ if $P(\hat\mu_n - \mu > \epsilon\mu) \le \delta$ with $n \ge \tilde{O}(\log(1/\mu))$, for given $0 < \epsilon, \delta < 1$.*

This definition requires that, with high probability, $\hat\mu_n$ is a lower bound of $\mu$ up to an error of $\epsilon\mu$. We have the following analog to Proposition 1:

**Corollary 2.** *Suppose $\hat\mu_n$ is downward biased, i.e., $\overline{\mu} := E[\hat\mu_n] \le \mu$. 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 lower-bound relaxed efficiency certificate.*

This motivates us to learn an inner approximation of the rare-event set in Stage 1 and then in Stage 2, we use IS as in Theorem 2 to estimate the probability of this inner approximation set. For the inner set approximation, like the outer approximation case, we use our Stage 1 samples $\{(\tilde{X}_i, Y_i)\}_{i=1,\dots,n_1}$ to construct an approximation set $\overline{\mathcal{S}}_\gamma$ that has zero false positive rate, i.e.,

$$P(X \in \overline{\mathcal{S}}_\gamma, Y = 0) = 0. \tag{9}$$

To make sure of (9), we again exploit the knowledge that the rare event set $\mathcal{S}_\gamma$ is orthogonally monotone. Indeed, denote $T_1 := \{\tilde{X}_i : Y_i = 1\}$ as the rare-event sampled points and for each point $x \in \mathbb{R}^d_+$, let $\mathcal{Q}(x) := \{x' : x' \geq x\}$. We construct $\mathcal{J}(T_1) := \cup_{x \in T_1} \mathcal{Q}(x)$ which serves as the "upper orthogonal monotone hull" of $T_1$. The orthogonal monotonicity property of $\mathcal{S}_\gamma$ implies that $\mathcal{J}(T_1) \subset \mathcal{S}_\gamma$. Moreover, $\mathcal{J}(T_1)$ is the largest choice of $\overline{\mathcal{S}}_\gamma$ such that (9) is guaranteed. Based on this observation, in parallel to Section 3, depending on how we construct the inner approximation to the rare-event set, we propose the following two approaches.

**Lazy-Learner IS (Lower Bound).** We now consider an estimator for $\mu$ where in Stage 1, we sample a constant $n_1$ i.i.d. random points from some density, say $q$. Then, we use the mixture IS depicted in Theorem 1 to estimate $P(X \in \mathcal{J}(T_1))$ in Stage 2. Since $\mathcal{J}(T_1)$ takes the form $\cup_{x \in T_1} \mathcal{Q}(x)$, it has a finite number of dominating points, which can be found by a sequential algorithm. But as explained in Section 3, this leads to a large number of mixture components that degrades the IS efficiency.

**Deep-Learning-Based IS (Lower Bound).** We train a neural network classifier, say $\hat{g}$, using all the Stage 1 samples $\{(\tilde{X}_i, Y_i)\}$, and obtain an approximate rare-event region $\overline{\mathcal{S}}^\kappa_\gamma = \{x : \hat{g}(x) \geq \kappa\}$, where $\kappa$ is say 1/2. Then we adjust $\kappa$ minimally away from 1/2, say to $\hat{\kappa}$, so that $\overline{\mathcal{S}}^{\hat{\kappa}}_\gamma \subset \mathcal{J}(T_1)$, i.e., $\hat{\kappa} = \min\{\kappa \in \mathbb{R} : \overline{\mathcal{S}}^{\hat{\kappa}}_\gamma \subset \mathcal{J}(T_1)\}$. Then $\overline{\mathcal{S}}^{\hat{\kappa}}_\gamma$ is an inner approximation for $\mathcal{S}_\gamma$ (see Figure 1(c), where $\hat{\kappa} = 0.83$). Stage 1 in Algorithm 2 shows this procedure. With this, we can run mixture IS to estimate $P(X \in \overline{\mathcal{S}}^{\hat{\kappa}}_\gamma)$ in Stage 2.

**Algorithm 2: Deep-PrAE to estimate $\mu = P(X \in \mathcal{S}_\gamma)$ (lower bound).**

```
Input: Black-box evaluator $I(\cdot \in \mathcal{S}_\gamma)$, initial Stage 1 samples $\{(\tilde{X}_i, Y_i)\}_{i=1,\dots,n_1}$, 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 $\overline{\mathcal{S}}^\kappa_\gamma = \{x : \hat{g}(x) \geq \kappa\}$ using $\{(\tilde{X}_i, Y_i)\}_{i=1,\dots,n_1}$;
Replace $\kappa$ by $\hat{\kappa} = \min\{\kappa \in \mathbb{R} : \overline{\mathcal{S}}^{\hat{\kappa}}_\gamma \subset \mathcal{J}(T_1)\}$;
Stage 2 (Mixture IS based on Searched dominating points):
The same as Stage 2 of Algorithm 1.
```

As we can see, compared with Algorithm 1, the only difference is how we adjust $\kappa$ in Stage 1. And similar to Theorem 2, we also have that Algorithm 2 attains the lower-bound relaxed efficiency certificate:

**Theorem 7** (Lower-bound relaxed efficiency certificate for deep-learning-based mixture IS)**.** *Suppose $\mathcal{S}_\gamma$ is orthogonally monotone, and $\overline{\mathcal{S}}^{\hat{\kappa}}_\gamma$ satisfies the same conditions for $\mathcal{S}_\gamma$ in Theorem 2. Then Algorithm 2 attains the lower-bound relaxed efficiency certificate by using a constant number of Stage 1 samples.*

Finally, we investigate the conservativeness of this bound, which is measured by the false negative rate $P(X \notin \overline{\mathcal{S}}^{\hat{\kappa}}_\gamma, Y = 1)$. Like in Section 3, we use 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 $\ell$ is a loss function and $\mathcal{G}$ is the considered hypothesis class. Let $g^*$ be the true risk minimizer as described in Section 3. For inner approximation, we let $\kappa^* := \max_{x \in \mathcal{S}^c_\gamma} g^*(x)$ be the true threshold associated with $g^*$ in obtaining the largest inner rare-event set approximation. Then we have the following result analogous to Theorem 3.

**Theorem 8** (Lower-bound estimation conservativeness)**.** *Consider $\overline{\mathcal{S}}^{\hat{\kappa}}_\gamma$ obtained in Algorithm 2 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 $g \in \mathcal{G}$, $g(x) \leq \kappa$ implies $\ell(g(x), 1) \geq h(\kappa) > 0$. (e.g., if $\ell$ is the squared loss, then $h(\kappa)$ could be chosen as $h(\kappa) = (1-\kappa)^2$). Then, with probability at least $1 - \delta$,*

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

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

# E Cross Entropy and Adaptive Multilevel Splitting

We provide some details on the cross-entropy method and adaptive multilevel splitting (or subset simulation), and also discuss their challenges in black-box problems.

**Cross Entropy.** The cross-entropy (CE) method (De Boer et al., 2005; Rubinstein and Kroese, 2013) uses a sequential optimization approach to iteratively solve for the optimal parameter in a parametric class of IS distributions. The objective in this optimization sequence is to minimize the Kullback–Leibler divergence between the IS distribution and the zero-variance IS distribution (the latter is theoretically known to be the conditional distribution given the occurrence of the rare event, but is unimplementable as it requires knowing the rare-event probability itself). Specifically, assume we are interested in estimating $P(g(X) > \gamma)$ and a parametric class $p_\theta$ is considered. The cross-entropy method adaptively chooses $\gamma_1 < \gamma_2 < ... < \gamma$. At each intermediate level $k$, we use the updated IS distribution $p_{\theta_k^*}$, designed for simulating $P(g(X) > \gamma_k)$, as the sampling distribution to draw samples of $X$ that sets up an empirical optimization, from which the next $\theta_{k+1}^*$ is obtained.

While flexible and easy to use, the efficiency of CE depends crucially on the expressiveness of the parametric class $p_\theta$ and the parameter convergence induced by the empirical optimization sequence. There are good approaches to determine the parametric classes (e.g., Botev et al. 2016), and also studies on the efficiency of IS distributions parametrized by empirical optimization (Tuffin and Ridder, 2012). However, it is challenging to obtain an efficiency certificate for CE that requires iterative empirical optimization in the common form depicted above. Insufficiency on either the choice of the parametric class or the parameter convergence may lead to the undetectable under-estimation issue (e.g., as in Theorem 1).

**Adaptive Multilevel Splitting.** Adaptive multilevel splitting (AMS) (or subset simulation) (Cérou and Guyader, 2007; Au and Beck, 2001) decomposes the rare-event estimation problem into estimating a sequence of conditional probabilities. We adaptively choose a threshold sequence $\gamma_1 < \gamma_2 < ... < \gamma_K = \gamma$. Then $P(g(x) > \gamma)$ can be rewritten as $P(g(x) > \gamma) = P(g(x) > \gamma_1)\prod_{k=2}^{K} P(g(x) > \gamma_k | g(x) > \gamma_{k-1})$. AMS then aims to estimate $P(g(x) > \gamma_1)$ and $P(g(x) > \gamma_k | g(x) > \gamma_{k-1})$ for each intermediate level $k = 2, ..., K$. In standard implementation, these conditional probabilities are estimated using samples from $p(g(x) > \gamma_k | g(x) > \gamma_{k-1})$ through variants of the Metropolis-Hasting (MH) algorithms.

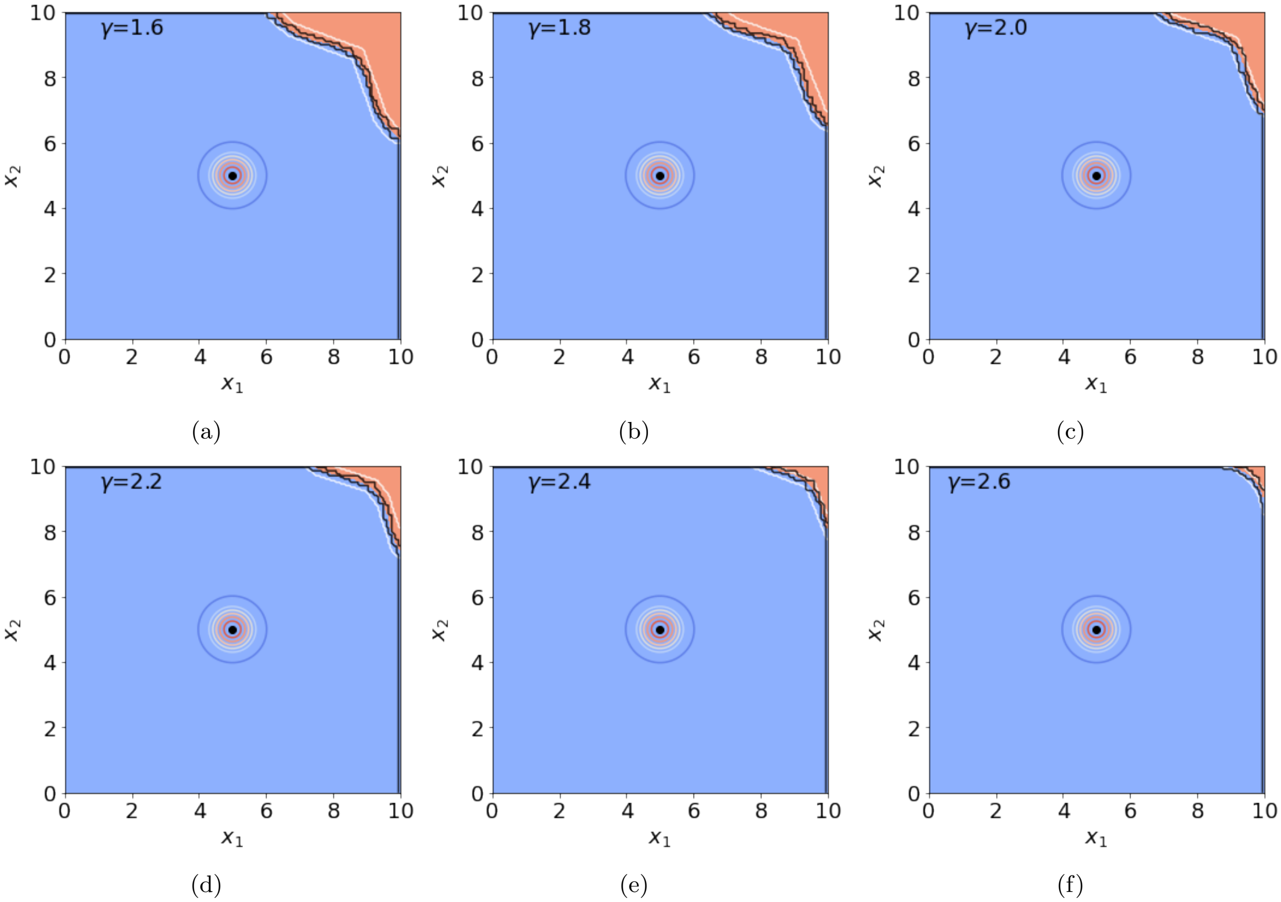
Theoretical studies have shown some nice properties of AMS, including unbiasedness and asymptotic normality (e.g., see Cérou et al. 2019). However, the variance of the estimator depends on the mixing property of the proposal distribution in the MH steps (Cérou and Guyader, 2016). Under ideal settings when direct sampling from $P(g(x) > \gamma_k | g(x) > \gamma_{k-1})$ is possible, it is shown that AMS is "almost" asymptotically optimal (Guyader et al., 2011). However, to our best knowledge, there is yet any study on provable efficiency of rare-event estimators with consideration of both AMS and MH sampling. In practice, to achieve a good performance, AMS requires a proposal distribution in the MH algorithm that can efficiently generate samples with low correlations.

# F Further Details for Numerical Experiments

This section provides more details on the two experimental examples in Section 4.

## F.1 2D Example

In the 2D example, the rarity parameter $\gamma$ governs the shape of the rare-event set $\mathcal{S}_\gamma = \{x : g(x) \geq \gamma\}$. We consider a linear combination of sigmoid functions $g(x) = \|\theta_1\psi(x-c_1-\gamma)+\theta_2\psi(x-c_2-\gamma)+\theta_3\psi(x-c_3-\gamma)+\theta_4\psi(x-c_4-\gamma)\|$ where $\theta, c$ are some constant vectors and $\psi(x) = \frac{\exp(x)}{1+\exp(x)}$. A point $x$ is a rare-event if $g(x) > \gamma$, where we take $\gamma = 1.8$ in Section 4. We use $p = N([5,5]^T, 0.25I_{2\times2})$. Figure 5 shows the rare-event set and its approximations for various $\gamma$'s. The Deep-PrAE boundaries seem tight in most cases, attributed to both the sufficiently trained NN classifier and the bisection algorithm implemented for tuning $\hat{\kappa}$ after the NN training.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: The contour of $p$, rare-event set $\mathcal{S}_\gamma$ (reddish region), outer- and inner- approximation boundaries (black lines) and Deep-PrAE UB and LB decision boundaries (white lines) for some $\gamma$ values in the 2D example.

## F.2 Intelligent Driving Safety Testing Example

We provide more details about the self-driving example, which simulates the interaction of an autonomous vehicle (AV) model that follows a human-driven lead vehicle (LV). The AV is controlled by the Intelligent Driver Model (IDM), widely used for autonomy evaluation and microscopic transportation simulation, that maintains a safety distance while ensuring smooth ride and maximum efficiency. The states of the AV are $s_t = [x_{\text{follow}}, x_{\text{lead}}, v_{\text{follow}}, v_{\text{lead}}, a_{\text{follow}}, a_{\text{lead}}]_t$ which are the position, velocity and acceleration of the AV and LV respectively. The throttle input to the AV is defined as $u_t$ which has an affine relationship with the acceleration of the vehicle. Similarly, the randomized throttle of the LV is represented by $w_t$. With a car length of $L$, the distance between the LV and AV at time $t$ is given by $r_t = x_{\text{lead},t} - x_{\text{follow},t} - L$, which has to remain below the crash threshold for safety.We describe the dynamics in more detail below. Figure 6 gives a pictorial overview of the interaction.

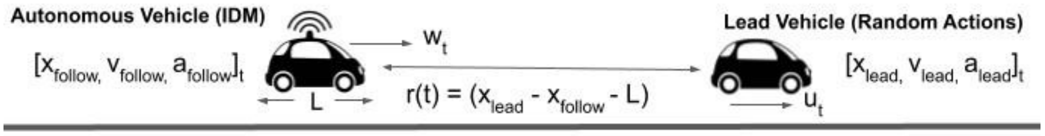

Autonomous Vehicle (IDM)
$[x_{\text{follow}}, v_{\text{follow}}, a_{\text{follow}}]_t$
$w_t$
L
$r(t) = (x_{\text{lead}} - x_{\text{follow}} - L)$
Lead Vehicle (Random Actions)
$[x_{\text{lead}}, v_{\text{lead}}, a_{\text{lead}}]_t$
$u_t$

Figure 6: The states $s_t$ and input $u_t$ of the self-driving safety-testing simulation. $w_t$ denotes the throttle input of the AV from the IDM.

Table 1: Parameters of the Intelligent Drivers Model (IDM)

| Parameters | Value |
|---|---|
| Safety distance, $s_0$ | 2 m |
| Speed of AV in free traffic, $v_0$ | 30 m/s |
| Maximum acceleration of AV, $a$ | $2\gamma$ m/s$^2$ |
| Comfortable deceleration of AV, $b$ | 1.67 m/s$^2$ |
| Maximum deceleration of AV, $d$ | $2\gamma$ m/s$^2$ |
| Safe time headway, $\bar{T}$ | 1.5 s |
| Acceleration exponent parameter, $\delta$ | 4 |
| Car length, $L$ | 4 m |

**LV actions.** The LV action contains human-driving uncertainty in decision-making modeled as Gaussian increments. For every $\Delta t$ time-steps, a Gaussian random variable is generated with the mean centered at the previous action $u_{t-\Delta t}$. We initialize $u_0 = 10$ (unitless) and $\Delta t = 4$ sec, which corresponds to zero initial acceleration and an acceleration change in the LV once every 4 seconds.

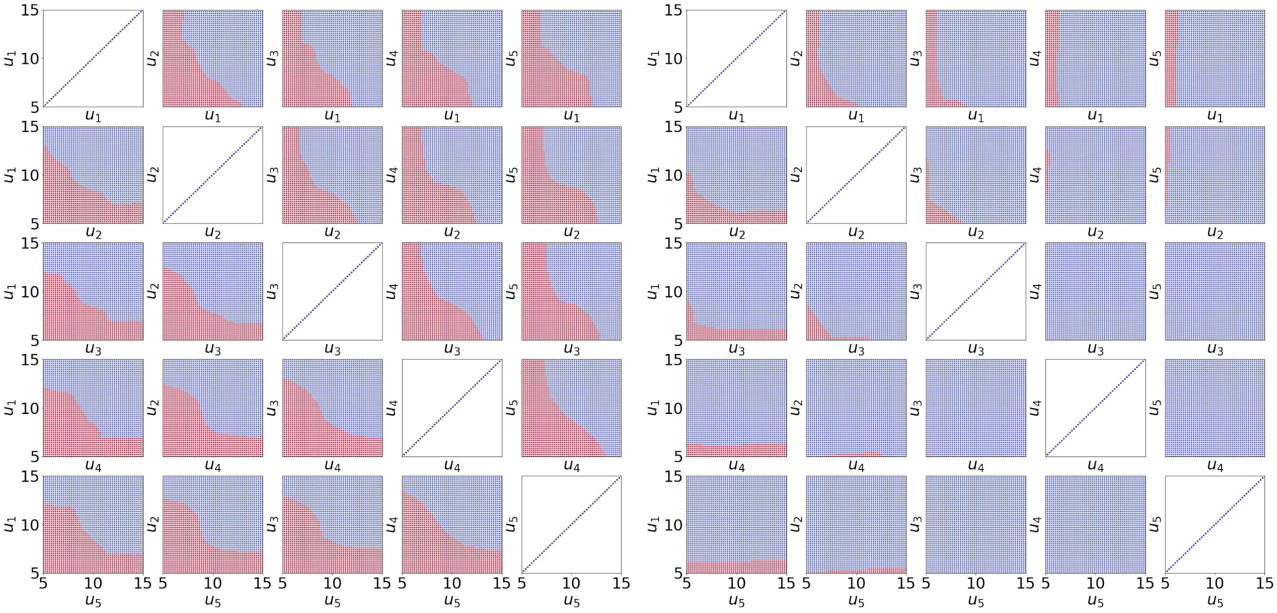
**Intelligent Driver Model (IDM) for AV.** The IDM is governed by the following equations (the subscripts "follow" and "lead" defined in Figure 6 is abbreviated to "f" and "l" for conciseness):

$$\begin{aligned}
\dot{x}_f &= v_f \\
\dot{x}_l &= v_l \\
\dot{v}_f &= \max\left(a(1-\left(\frac{v_f}{v_0}\right)^{\delta} - \left(\frac{s^*(v_f,\Delta v_f)}{s_f}\right)^2), -d\right) \\
s^*(v_f,\Delta v_f) &= s_0 + v_f\bar{T} + \frac{v_f\Delta v_f}{2\sqrt{ab}} \\
s_f &= x_l - x_f - L \\
\Delta v_f &= v_f - v_l,
\end{aligned}$$

The parameters are presented in Table 1, and $v_l \propto u_t$ and $v_f \propto w_t$, . The randomness of LV actions $u_t$'s propagates into the system and affects all the simulation states $s_t$. The IDM is governed by simple first-order kinematic equations for the position and velocity of the vehicles. The acceleration of the AV is the decision variable where it is defined by a sum of non-linear terms which dictate the "free-road" and "interaction" behaviors of the AV and LV. The acceleration of the AV is constructed in such a way that certain terms of the equations dominate when the LV is far away from the AV to influence its actions and other terms dominate when the LV is in close proximity to the AV.

**Rarity parameter $\gamma$.** Parameter $\gamma$ signifies the range invoked by the AV acceleration and deceleration pedals. Increasing $\gamma$ implies that the AV can have sudden high deceleration and hence avoid crash scenarios better and making crashes rarer. In contrast, decreasing $\gamma$ reduces the braking capability of the AV and more easily leads to crashes. For instance, $\gamma = 1.0$ corresponds to AV actions in the range $[5, 15]$ or correspondingly $a_{\text{follow},t} \in [-2, 2]$, and $\gamma = 2.0$ corresponds to $a_{\text{follow},t} \in [-4, 4]$. Figure 7 shows the approximate rare-event set by randomly sampling points and evaluating the inclusion in the set, for the two cases of $\gamma = 1.0$ and $\gamma = 2.0$. In particular, we slice the 15-dimensional space onto pairs from five of the dimensions. In all plots, we see that the crash set (red) are monotone, thus supporting the use of our Deep-PrAE framework. Although the crash set is not located in the "upper-right corner", we can implement Deep-PrAE framework for such problems by simple re-orientation.

**Sample trajectories.** Figure 8 shows two examples of sample trajectories, one successfully maintaining a safe distance, and the other leading to a crash. In Figure 8(e)-(h) where we show the crash case, the AV maintains a safe distance behind the LV until the latter starts rapidly decelerating (Figure 8(h)). Here the action corresponds to the throttle input that has an affine relationship with the acceleration of the vehicle. The LV ultimately decelerates at a rate that the AV cannot attain and its deceleration saturates after a point which leads to the crash.



(a) Case $\gamma = 1.0$

(b) Case $\gamma = 2.0$

Figure 7: Slice of pairs of the first 5 dimensions of LV action space. For any $(u_i, u_{i'})$ shown, $u_j, j \notin \{i, i'\}$ is fixed at a constant value. Blue dots = non-crash cases, red dots= crash cases.

### F.3 Code

The code and environment settings for the experiments are available at https://github.com/safeai-lab/Deep-PrAE/.

## G Proofs

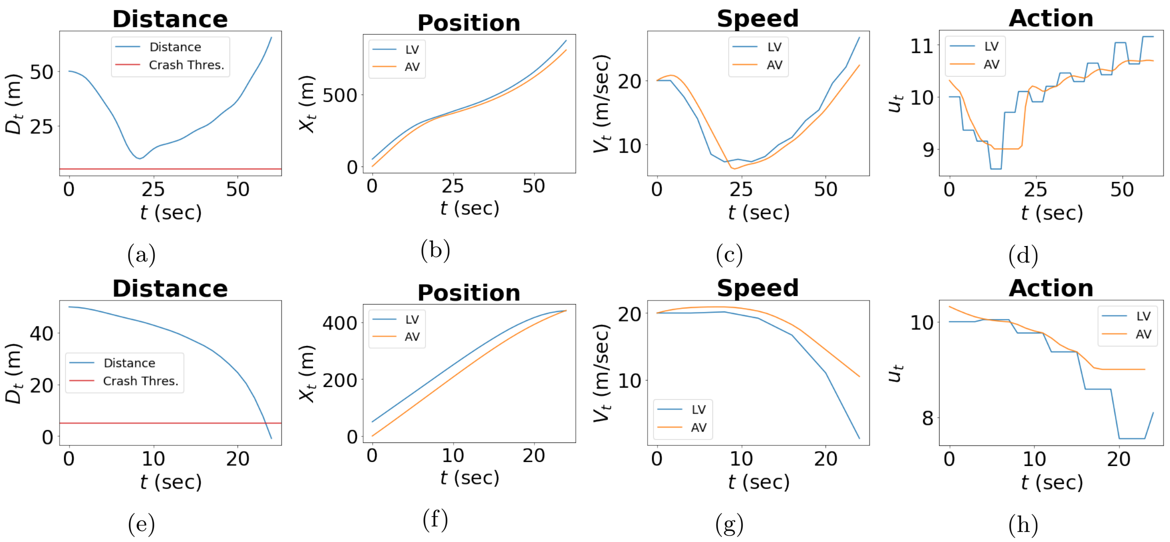
### G.1 Proofs for the Dominating Point Methodologies

*Proof of Proposition 2.* Since $\mu$ is exponentially decaying in $\gamma$ while $n$ is polynomially growing in $\gamma$, we know that $\lim_{\gamma\to\infty} n\mu = 0$. Since $n\hat{\mu}_n$ takes values in $\{0, 1, \ldots, n\}$, we get that $P(|\hat{\mu}_n - \mu| > \varepsilon\mu) = P(|n\hat{\mu}_n - n\mu| > \varepsilon n\mu) \to 1$ as $\gamma \to \infty$. □

*Proof of Theorem 4.* Throughout this proof, we write $f(\gamma) \sim g(\gamma)$ if $f(\gamma)/g(\gamma)$ changes at most polynomially in $\gamma$. We know that

$$\tilde{E}[Z^2] = \sum_j \tilde{E}[I(X \in \mathcal{S}_\gamma^j)L^2(X)] \leq \sum_j e^{-(a_j-\lambda)^T\Sigma^{-1}(a_j-\lambda)}/\alpha_j \sim e^{-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)}.$$

Denote $Y = B(X - \lambda) \sim N(0, B\Sigma B^T)$ and $s = B(a^* - \lambda)$. Define $\tilde{\varepsilon} = \varepsilon \min_{u:u^T(B\Sigma B^T)^{-1}u=1} \|u\|_\infty$. Then we also



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8: Autonomous Car Following Experiment Trajectories. Figures (a) - (d) represent a simulation episode without a crash occurring where the AV follows the LV successfully at a safe distance. Figures (e) - (h) represents a simulation episode where crash occurs at $t = 23$ seconds due to the repeated deceleration of the LV.

know that

$$\begin{aligned}
&\tilde{E}[I(X\in\mathcal{S}_\gamma)L(X)]\\
\ge& P(B(X-a^*)\ge 0,(X-a^*)^T\Sigma^{-1}(X-a^*)\le\varepsilon^2)\\
=&P(Y\ge s,(Y-s)^T(B\Sigma B^T)^{-1}(Y-s)\le\varepsilon^2)\\
=&\int_{y\ge s,(y-s)^T(B\Sigma B^T)^{-1}(y-s)\le\varepsilon^2}(2\pi)^{-d/2}|B\Sigma B^T|^{-1/2}e^{-y^T(B\Sigma B^T)^{-1}y/2}\mathrm{d}y\\
\ge&(2\pi)^{-d/2}|B\Sigma B^T|^{-1/2}e^{-\varepsilon^2/2}e^{-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)/2}\\
&\int_{y\ge s,(y-s)^T(B\Sigma B^T)^{-1}(y-s)\le\varepsilon^2}e^{-s^T(B\Sigma B^T)^{-1}(y-s)}\mathrm{d}y\\
\ge&(2\pi)^{-d/2}|B\Sigma B^T|^{-1/2}e^{-\varepsilon^2/2}e^{-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)/2}\prod_{i=1}^{d}\int_0^{\tilde\varepsilon}e^{-s^T(B\Sigma B^T)^{-1}e_iu_i}\mathrm{d}u_i\\
=&(2\pi)^{-d/2}|B\Sigma B^T|^{-1/2}e^{-\varepsilon^2/2}e^{-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)/2}\prod_{i=1}^{d}\frac{1-e^{-s^T(B\Sigma B^T)^{-1}e_i\tilde\varepsilon}}{s^T(B\Sigma B^T)^{-1}e_i}.
\end{aligned}$$

Note that it is easy to verify that $s^T(B\Sigma B^T)^{-1}e_i\ge 0$. If $s^T(B\Sigma B^T)^{-1}e_i=0$, then we naturally use $\tilde\varepsilon$ to substitute $\frac{1-e^{-s^T(B\Sigma B^T)^{-1}e_i\tilde\varepsilon}}{s^T(B\Sigma B^T)^{-1}e_i}$. Since we have assumed that the components of $a^*$ are at most polynomially growing in $\gamma$, finally we get that

$$\tilde{E}[I(X\in\mathcal{S}_\gamma)L(X)]\sim e^{-(a^*-\lambda)^T\Sigma^{-1}(a^*-\lambda)/2}$$

and hence $\tilde{E}[Z^2]/\tilde{E}[Z]^2$ is at most polynomially growing in $\gamma$.

□

*Proof of Theorem 1.* We know that $\tilde{E}[Z]=\bar\Phi(\gamma)+\bar\Phi(k\gamma)$. Moreover,

$$\tilde{E}[Z^2]=e^{\gamma^2}(\bar\Phi(2\gamma)+\bar\Phi((k-1)\gamma)).$$

If $0<k\le 1$, then $\tilde{E}[Z]=O\left(e^{-k^2\gamma^2/2}/\gamma\right)$ and $\tilde{E}[Z^2]=O\left(e^{\gamma^2}\right)$ as $\gamma\to\infty$. If $1<k<3$, then $\tilde{E}[Z]=O\left(e^{-\gamma^2/2}/\gamma\right)$ and $\tilde{E}[Z^2]=O\left(e^{(1-(k-1)^2/2)\gamma^2}/\gamma\right)$ as $\gamma\to\infty$. In both cases, we get that $\tilde{E}[Z^2]/\tilde{E}[Z]^2$ grows exponentially in $\gamma$. On the other hand, we know that

$$\begin{aligned}&\tilde{P}\left(\left|\frac{1}{n}\sum_i Z_i-\bar{\Phi}(\gamma)\right|>\varepsilon\bar{\Phi}(\gamma)\right)\\ \leq&\tilde{P}(\exists i: X_i\leq -k\gamma)+\tilde{P}\left(\left|\frac{1}{n}\sum_i I(X_i\geq\gamma)e^{\gamma^2/2-\gamma X_i}-\bar{\Phi}(\gamma)\right|>\varepsilon\bar{\Phi}(\gamma)\right).\end{aligned}$$

Clearly $\tilde{P}(\exists i: X_i\leq -k\gamma)=1-(1-\bar{\Phi}((k+1)\gamma))^n=O\left(n\bar{\Phi}((k+1)\gamma)\right)$, which is exponentially decreasing in $\gamma$ as $n$ is polynomial in $\gamma$. Moreover, by Chebyshev's inequality,

$$\begin{aligned}&\tilde{P}\left(\left|\frac{1}{n}\sum_i I(X_i\geq\gamma)e^{\gamma^2/2-\gamma X_i}-\bar{\Phi}(\gamma)\right|>\varepsilon\bar{\Phi}(\gamma)\right)\\ \leq&\frac{\tilde{E}[I(X_i\geq\gamma)e^{\gamma^2-2\gamma X_i}]}{n\varepsilon^2\bar{\Phi}^2(\gamma)}=\frac{e^{\gamma^2}\bar{\Phi}(2\gamma)}{n\varepsilon^2\bar{\Phi}^2(\gamma)}=O\left(\frac{\gamma}{n\varepsilon^2}\right).\end{aligned}$$

Thus $P(|\hat{\mu}_n-\bar{\Phi}(\gamma)|>\varepsilon\bar{\Phi}(\gamma))=O\left(\frac{\gamma}{n\varepsilon^2}\right)$. Moreover, we know that $P(\exists i: Z_i>0)\geq 1-1/2^n$ and if $Z_i>0$ for some $i$, then we have that

$$\frac{\sum_i Z_i^2/n}{(\sum_i Z_i/n)^2}\leq n^2.$$

□

### G.2 Proofs for the Relaxed Efficiency Certificate

*Proof of Proposition 1.* We have

$$P(\hat{\mu}_n-\mu<-\epsilon\mu)\leq P(\hat{\mu}_n-\overline{\mu}<-\epsilon\overline{\mu})$$

since $\overline{\mu}\geq\mu$ and $1-\epsilon>0$. Note that the Markov inequality gives

$$P(\hat{\mu}_n-\overline{\mu}<-\epsilon\overline{\mu})\leq\frac{\widetilde{Var}(Z_i)}{n\epsilon^2\overline{\mu}^2}$$

so that

$$n\geq\frac{\widetilde{Var}(Z_i)}{\delta\epsilon^2\overline{\mu}^2}=\frac{RE}{\delta\epsilon^2}=\tilde{O}\left(\log\frac{1}{\overline{\mu}}\right)=\tilde{O}\left(\log\frac{1}{\mu}\right)$$

achieves the relaxed efficiency certificate. □

*Proof of Proposition 3.* The proof follows from that of Proposition 1 with a conditioning on $D_{n_1}$. We have

$$P(\hat{\mu}_n-\mu<-\epsilon\mu|D_{n_1})\leq P(\hat{\mu}_n-\overline{\mu}(D_{n_1})<-\epsilon\overline{\mu}(D_{n_1})|D_{n_1})$$

since $\overline{\mu}(D_{n_1})\geq\mu$ almost surely and $1-\epsilon>0$. Note that the Markov inequality gives

$$P(\hat{\mu}_n-\overline{\mu}(D_{n_1})<-\epsilon\overline{\mu}(D_{n_1})|D_{n_1})\leq\frac{Var(Z_i|D_{n_1})}{n_2\epsilon^2\overline{\mu}(D_{n_1})^2}$$

so that

$$n_2\geq\frac{Var(Z_i|D_{n_1})}{\delta\epsilon^2\overline{\mu}(D_{n_1})^2}=\frac{RE(D_{n_1})}{\delta\epsilon^2}=\tilde{O}\left(\log\left(\frac{1}{\overline{\mu}(D_{n_1})}\right)\right)=\tilde{O}\left(\log\frac{1}{\mu}\right)$$

almost surely. Thus,

$$n=n_1+n_2\geq\tilde{O}\left(\log\frac{1}{\mu}\right)$$

achieves the relaxed efficiency certificate.

□

*Proof of Corollary 1.* Follows directly from Proposition 3, since $\overline{\mathcal{S}}_\gamma \supset \mathcal{S}_\gamma$ implies $\overline{\mu}(D_{n_1}) \geq \mu$ almost surely. □

*Proof of Theorem 2.* We have assumed that $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ satisfies the assumptions for $\mathcal{S}_\gamma$ in Theorem 4. Then following the proof of Theorem 4, we obtain the efficiency certificate for the IS estimator in estimating its mean. Theorem 2 is then proved by directly applying Corollary 1. □

### G.3 Proofs for Conservativeness

Recall that $T_0 = \{\tilde{X}_i : Y_i = 0\}$ where the samples are generated as in Algorithm 1. By some combinitorial argument, we can prove the following lemma which says that with high probability, each point in $\mathcal{S}_\gamma^c$ that has sufficient distance to its boundary could be covered by $\mathcal{H}(T_0)$.

**Lemma 1.** *Suppose that the density $q$ has bounded support $K \subset [0, M]^d$, and for any $x \in K$, suppose that $0 < q_l \leq q(x) \leq q_u$. Define $B_t := \{x \in \mathcal{S}_\gamma^c : x + t\mathbf{1}_{d\times 1} \in \mathcal{S}_\gamma^c\}$. Then with probability at least $1 - \delta$, we have that $B_{t(\delta,n_1)} \subset \mathcal{H}(T_0)$. Here $t(\delta, n_1) = 3\left(\frac{\log(n_1 q_l) + d\log M + \log\frac{1}{\delta}}{n_1 q_l}\right)^{\frac{1}{d}}$.*

*Proof.* The basic idea is to construct a finite number of regions, such that when there is at least one sample point in each of these regions, we would have that $B_t \subset \mathcal{H}(T_0)$. Then we could give a lower bound to the probability of $B_t \subset \mathcal{H}(T_0)$ in terms of the number of regions and the volume of each of these regions.

By dividing the first $d-1$ coordinates into $\frac{M}{\delta}$ equal parts, we partition the region $[0, M]^d$ into rectangles, each with side length $\delta$, except for the $d-$th dimension (the $\delta$ here is not exactly the $\delta$ in the statement of the lemma, since we will do a change of variable in the last step). To be more precise, the rectangles are given by

$$Z_j = \left(\prod_{i=1}^{d-1}[(j_i - 1)\delta, j_i\delta]\right) \times [0, M].$$

Here $j \in J$ and $J$ is defined by

$$J := \{j = (j_1, \cdots, j_{d-1}), j_i = 1, 2, \cdots, \frac{M}{\delta}\}.$$

Denote by $J_0$ the set which consists of $j \in J$ such that there exist a point in $B_{2\delta}$ whose first $d-1$ coordinates are $j_1\delta, j_2\delta, \cdots, j_{d-1}\delta$ respectively, i.e., $J_0 = \left\{j \in J : B_{2\delta} \cap \left(\left(\prod_{i=1}^{d-1}\{j_i\delta\}\right) \times [0, M]\right) \neq \emptyset\right\}$. For all $j \in J_0$, let $p_j$ be the point such that

i) $p_j \in B_\delta$

ii) The first $d-1$ coordinates of $p_j$ are $j_1\delta, j_2\delta, \cdots, j_{d-1}\delta$ respectively

iii) $p_j$ has $d-$th coordinate larger than $-\delta + \sup_{p \text{ satisfies i),ii)}}$ ($d$-th coordinate of $p$).

From the definition of $J_0$ and the fact that $B_\delta \supset B_{2\delta}$, $p_j$ is guaranteed to exist. We claim that $B_{2\delta} \cap Z_j \subset \mathcal{R}(p_j)$, where $\mathcal{R}(p_j)$ is the rectangle that contains 0 and $p_j$ as two of its corners. Clearly, from the definition of $Z_j$, for any point $x \in B_{2\delta} \cap Z_j$, its first $d-1$ coordinates are smaller than $j_1\delta, j_2\delta, \cdots, j_{d-1}\delta$ respectively. For the $d-$th coordinate, suppose on the contrary that there exists $x \in B_{2\delta} \cap Z_j$ with $d-$th coordinate greater than the $d-$th coordinate of $p_j$. Since $x \in Z_j$, the first $d-1$ coordinates of $x$ are at least $(j_1 - 1)\delta, (j_2 - 1)\delta, \cdots, (j_{d-1} - 1)\delta$, so we have that $x + \delta\mathbf{1}_{d\times 1} \geq p_j + \delta e_d$. Since $x \in B_{2\delta}$, we know that $x + 2\delta\mathbf{1}_{d\times 1} \in \mathcal{S}_\gamma^c$. Hence by the previous inequality and the orthogonal monotonicity of $\mathcal{S}_\gamma$, $p_j + \delta e_d + \delta\mathbf{1}_{d\times 1} \in \mathcal{S}_\gamma^c$. By definition of $B_\delta$, this implies $p_j + \delta e_d \in B_\delta$. This contradicts iii) in the definition of $p_j$. By contradiction, we have shown that each point in $B_{2\delta} \cap Z_j$ has $d-$th coordinate smaller than the $d-$th coordinate of $p_j$. So the claim that $B_{2\delta} \cap Z_j \subset \mathcal{R}(p_j)$ for any $j \in J_0$ is proved.

Then we consider those $j$ such that $j \in J - J_0$. For any point $x \in Z_j$, the first $d-1$ coordinates of $x + \delta\mathbf{1}_{d\times 1}$ are at least $j_1\delta, j_2\delta, \cdots, j_{d-1}\delta$ respectively. Since $j \notin J_0$, we have that $x + \delta\mathbf{1}_{d\times 1} \notin B_{2\delta}$. This implies $x + 3\delta\mathbf{1}_{d\times 1} \notin \mathcal{S}_\gamma^c$, or $x \notin B_{3\delta}$. So we have shown that for any $j \notin J_0$, $B_{3\delta} \cap Z_j = \emptyset$. This implies $B_{3\delta}$ has a partition given by $B_{3\delta} = \cup_{j\in J}(B_{3\delta} \cap Z_j) = \cup_{j\in J_0}(B_{3\delta} \cap Z_j)$. Notice that $B_{3\delta} \subset B_{2\delta}$, from the result in the preceding paragraph, we conclude that $B_{3\delta} \subset \cup_{j\in J_0}\mathcal{R}(p_j)$.

For each $j \in J_0$ and the constructed $p_j$, consider the region

$$G_j := \{x \in S_\gamma^c : x \geq p_j\}.$$

Observe that, if there exists a sample point in $T_0$ that lies in $G_j$, then we have $p_j \subset \mathcal{H}(T_0)$ which implies $\mathcal{R}(p_j) \subset \mathcal{H}(T_0)$. Since $p_j \in B_\delta$ and $\mathcal{S}_\gamma$ is orthogonally monotone, we have that $G_j$ contains the rectangle which contains $p_j$ and $p_j + \delta \mathbf{1}_{d\times 1}$ as two of its corners, so $\mathrm{Vol}(G_j) \geq \delta^d$. Hence the probability that $\mathcal{R}(p_j) \subset \mathcal{H}(T_0)$ has a lower bound given by

$$P(\mathcal{R}(p_j) \subset \mathcal{H}(T_0)) \geq P(T_0 \cap G_j \neq \emptyset) \geq 1 - \left(1 - \delta^d q_l\right)^{n_1} \geq 1 - e^{-n_1 q_l \delta^d}.$$

Notice that $|J_0| \leq \left(\frac{M}{\delta}\right)^{d-1}$, by union bound we have that

$$P(\cup_{j\in J_0} \mathcal{R}(p_j) \subset \mathcal{H}(T_0)) \geq 1 - \frac{M^{d-1}}{\delta^{d-1}} e^{-n_1 q_l \delta^d}.$$

Since we have shown that $B_{3\delta} \subset \cup_{j\in J_0} \mathcal{R}(p_j)$, this implies

$$P(B_{3\delta} \subset \mathcal{H}(T_0)) \geq 1 - \frac{M^{d-1}}{\delta^{d-1}} e^{-n_1 q_l \delta^d}.$$

Based on this inequality, it is not hard to check that for $t(\delta, n_1) = 3\left(\frac{\log(n_1 q_l) + d\log M + \log\frac{1}{\delta}}{n_1 q_l}\right)^{\frac{1}{d}}$, we have that $P(B_{t(\delta)} \subset \mathcal{H}(T_0)) \geq 1 - \delta$. □

*Proof of Theorem 5.* First, we show the inequality in the theorem, i.e., $P_{X\sim q}(X \in \mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma) \leq M^{d-1} q_u \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1)$. It suffices to show that with probability at least $1-\delta$, $\mathrm{Vol}\left(\mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma\right) \leq M^{d-1} \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1)$, or equivalently $\mathrm{Vol}(\mathcal{S}_\gamma^c \backslash \mathcal{H}(T_0)) \leq M^{d-1} \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1)$. Since by lemma 1 we have that $B_{t(\delta,n_1)} \subset \mathcal{H}(T_0)$ with probability at least $1-\delta$, it suffices to show that $\mathrm{Vol}(\mathcal{S}_\gamma^c \backslash B_{t(\delta,n_1)}) \leq M^{d-1} \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1)$. This latter inequality actually follows from the definition of $B_{t(\delta,n_1)}$ and some geometric argument. Indeed, by definition of $B_{t(\delta,n_1)}$, for each $x \in \mathcal{S}_\gamma^c \backslash B_{t(\delta,n_1)}$, $x$ belongs to the area which is obtained by moving the boundary of $\mathcal{S}_\gamma$ in direction $-\frac{\mathbf{1}_{d\times 1}}{\sqrt{d}}$ for a distance of $t(\delta, n_1)\sqrt{d}$. So the volume of $\mathcal{S}_\gamma^c \backslash B_{t(\delta,n_1)}$ is bounded by

$$\begin{aligned} &t(\delta, n_1)\sqrt{d} \times \mathrm{Vol}_{d-1}(\text{projection of the boundary of } S_0 \text{ in direction } \mathbf{1}_{d\times 1}) \\ \leq &t(\delta, n_1)\sqrt{d} \times \mathrm{Vol}_{d-1}(\text{projection of } [0, M]^d \text{ in direction } \mathbf{1}_{d\times 1}) \end{aligned}$$

Here $\mathrm{Vol}_{d-1}$ means computing volume in the $d-1$ dimensional space. Notice that $[0, M]^d$ is contained in a ball with radius $\frac{M\sqrt{d}}{2}$, we have that

$$\mathrm{Vol}_{d-1}(\text{projection of } [0, M]^d \text{ in direction } \mathbf{1}_{d\times 1}) \leq M^{d-1} \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1}.$$

Combining the preceding two inequalities, we have proved the inequality in the theorem. Next we show the equality in the theorem. Indeed, when $d$ is large, we have asymptotic formula $w_d = \frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d}\right)^{\frac{d}{2}} (1 + O(d^{-1}))$. Plugging this into the RHS above, we will obtain the asymptotic bound as stated in the theorem. □

*Proof of Theorem 3.* By Markov inequality and the definition of $h$,$\bar{\mathcal{S}}_\gamma^{\hat{\kappa}}$, we know that

$$P_{X\sim q}\left(X \in \bar{\mathcal{S}}_\gamma^{\hat{\kappa}}, X \in \mathcal{S}_\gamma^c\right) = P_{X\sim q}(\hat{g}(X) \geq \hat{\kappa}, X \in \mathcal{S}_\gamma^c) \leq \frac{R(\hat{g})}{h(\hat{\kappa})}. \tag{10}$$

We will compare the numerator and denominator of the RHS of (10) with their counterparts for the true minimizer $g^*$. For the numerator, since $\hat{g}$ is the empirical risk minimizer, we have that

$$\begin{aligned} R(\hat{g}) &\leq R_{n_1}(g) + \sup_{g_\theta \in \mathcal{G}} \left| R_{n_1}(g_\theta) - R(g_\theta) \right| \leq R_{n_1}(g^*) + \sup_{g_\theta \in \mathcal{G}} \left| R_{n_1}(g_\theta) - R(g_\theta) \right| \\ &\leq R(g^*) + 2 \sup_{g_\theta \in \mathcal{G}} \left| R_{n_1}(g_\theta) - R(g_\theta) \right| . \end{aligned}$$

For the denominator, from the definition of $\bar{\mathcal{S}}_\gamma^{\hat{\kappa}}$, it is not hard to verify that, in Algorithm 1, our choice of $\hat{\kappa}$ is given by $\hat{\kappa} = \min\{\hat{g}(x) : x \in \mathcal{H}(T_0)^c\}$. By lemma 1, we have that with probability at least $1-\delta$, $B_{t(\delta,n_1)} \subset \mathcal{H}(T_0)$, which implies that with probability at least $1-\delta$,

$$\begin{aligned} \hat{\kappa} &\geq \min\{\hat{g}(x) : x \in B^c_{t(\delta,n_1)}\} \geq \min\{g^*(x) : x \in B^c_{t(\delta,n_1)}\} - \left\| \hat{g} - g^* \right\|_\infty \\ &\geq \min\{g^*(x) : x \in \mathcal{S}_\gamma\} - t(\delta,n_1)\sqrt{d}\mathrm{Lip}(g^*) - \left\| \hat{g} - g^* \right\|_\infty \\ &= \kappa^* - t(\delta,n_1)\sqrt{d}\mathrm{Lip}(g^*) - \left\| \hat{g} - g^* \right\|_\infty . \end{aligned}$$

Putting the preceding two inequalities into the Markov inequality (10), and notice that $h$ is non decreasing by its definition, the theorem is proved. □