Appendix

EC.1. Examples of Applying Theorem 1

EXAMPLE EC.1 (ONE-DIMENSIONAL GAMMA DISTRIBUTION). Suppose $X \sim Gamma(\alpha, \beta)$ where $\alpha, \beta > 0$. The density function is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0.$$

We have that

$$\lambda(s) = \begin{cases} -\alpha \log(1 - s/\beta) & s < \beta; \\ \infty & s \geq \beta. \end{cases}$$

In the one dimensional case, we consider $\mathcal{S}_\gamma = [a^*, \infty)$ where $a^* > \alpha/\beta$ such that $[a^*, a^* + \varepsilon] \subset \mathcal{S}_\gamma$ for $\epsilon > 0$ and $0 < I(a^*) < \infty$. We have that $s^* = \beta - \alpha/a^*$. The dominating set is $\{a^*\}$ by definition and the IS distribution is defined by density function $f^*(x) = \frac{1}{\Gamma(\alpha)}(\alpha/a^*)^\alpha x^{\alpha-1} e^{-\frac{\alpha}{a^*}x}$. That is, the IS distribution is exactly $Gamma(\alpha, \alpha/a^*)$, where we move the mean value from α/β to a^*. If $a^* \leq x \leq a^* + \varepsilon$, then $f^*(x) \geq f^*(a^* + \varepsilon) = \frac{1}{\Gamma(\alpha)}(\alpha/a^*)^\alpha (a^* + \varepsilon)^{\alpha-1} e^{-\alpha \frac{a^*+\varepsilon}{a^*}}$. If $a^* \to \infty$ at most polynomially in γ, then this lower bound is also polynomial in γ. We could easily verify that Assumption 1 and 2 hold and hence the efficiency certificate is achieved.

EXAMPLE EC.2 (MULTIVARIATE GAUSSIAN DISTRIBUTION). Suppose that $X \sim N(m, \Sigma)$ where $m \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite. We have that $\lambda(s) = m^\top s + \frac{1}{2} s^\top \Sigma s$ and $I(y) = \frac{1}{2}(y - m)^\top \Sigma^{-1}(y - m)$. Then $a^* = \arg\min_{y \in \mathcal{S}_\gamma} (y - m)^\top \Sigma^{-1}(y - m)$. We assume that $\{x : a^* \leq x \leq a^* + \varepsilon \mathbf{1}\} \subset \mathcal{S}_\gamma$ for some $\varepsilon > 0$ and also $m \notin \mathcal{S}_\gamma$ such that $I(a^*) > 0$. We have that $s^* = \Sigma^{-1}(a^* - m)$. Then $f^*(x) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x - a^*)^\top \Sigma^{-1}(x - a^*)\}$, which is the density function of $N(a^*, \Sigma)$. Moreover, if $a^* \leq x \leq a^* + \varepsilon \mathbf{1}$, then we can find a constant $\varepsilon' > 0$ such that $f^*(x) > \varepsilon'$. Thus, as long as the components of s^* at most grow polynomially in γ, the efficiency certificate is guaranteed.

EXAMPLE EC.3 (GAUSSIAN MIXTURE DISTRIBUTION). Suppose that X follows a Gaussian mixture distribution with density $f(x) = \sum_{i=1}^k \pi_i (2\pi)^{-d/2} |\Sigma_i|^{-1/2} \exp\{-(x - m_i)^\top \Sigma_i^{-1}(x - m_i)/2\}$ where $m_i \in \mathbb{R}^d$, $\Sigma_i \in \mathbb{R}^{d\times d}$ is positive definite, $\pi_i > 0$ and $\sum_{i=1}^k \pi_i = 1$. Similarly, we assume that $\{x : a^* \leq x \leq a^* + \varepsilon \mathbf{1}\} \subset \mathcal{S}_\gamma$ for some $\varepsilon > 0$ and $m_i \notin \mathcal{S}_\gamma, \forall i$. Suppose that a_{ij}'s are the dominating points for $\mathcal{S}_\gamma$ associated with the distribution $N(m_i, \Sigma_i)$. Denote $a^{(i)} := \arg\min_{a_{ij}} (a_{ij} - m_i)^\top \Sigma_i^{-1}(a_{ij} - m_i)$ and $s^{(i)} := s_{a^{(i)}}$. We also assume that the components of $s^{(i)}$ at most grow polynomially in γ. From Example EC.2, we know that if the input distribution is $N(m_i, \Sigma_i)$, then

the IS distribution $\tilde{f}(x) = \sum_j \alpha_{ij}(2\pi)^{-d/2}|\Sigma_i|^{-1/2}\exp\{-(x-a_{ij})^\top\Sigma_i^{-1}(x-a_{ij})/2\}$ achieves the efficiency certificate for any α_{ij}'s such that $\alpha_{ij} > 0, \forall j$ and $\sum_j \alpha_{ij} = 1$. Now we consider the input distribution $f(x) = \sum_{i=1}^k \pi_i(2\pi)^{-d/2}|\Sigma_i|^{-1/2}\exp\{-(x-m_i)^\top\Sigma_i^{-1}(x-m_i)/2\}$. We will show that the IS distribution $\tilde{f}(x) = \sum_{i=1}^k \pi_i \sum_j \alpha_{ij}(2\pi)^{-d/2}|\Sigma_i|^{-1/2}\exp\{-(x-a_{ij})^\top\Sigma_i^{-1}(x-a_{ij})/2\}$ achieves the efficiency certificate. First, we know that the likelihood ratio function is

$$\begin{aligned} L(x) &= \frac{\sum_{i=1}^k \pi_i(2\pi)^{-d/2}|\Sigma_i|^{-1/2}\exp\{-(x-m_i)^\top\Sigma_i^{-1}(x-m_i)/2\}}{\sum_{i=1}^k \pi_i \sum_j \alpha_{ij}(2\pi)^{-d/2}|\Sigma_i|^{-1/2}\exp\{-(x-a_{ij})^\top\Sigma_i^{-1}(x-a_{ij})/2\}} \\ &\le \sum_{i=1}^k \frac{\exp\{-(x-m_i)^\top\Sigma_i^{-1}(x-m_i)/2\}}{\sum_j \alpha_{ij}\exp\{-(x-a_{ij})^\top\Sigma_i^{-1}(x-a_{ij})/2\}} \end{aligned}$$

and hence

$$L^2(x) \le k\sum_{i=1}^k \left(\frac{\exp\{-(x-m_i)^\top\Sigma_i^{-1}(x-m_i)/2\}}{\sum_j \alpha_{ij}\exp\{-(x-a_{ij})^\top\Sigma_i^{-1}(x-a_{ij})/2\}}\right)^2.$$

Following the proof of Theorem 1, we get that

$$\begin{aligned} &\tilde{\mathbb{E}}\left[I(X\in\mathcal{S}_\gamma)\left(\frac{\exp\{-(X-m_i)^\top\Sigma_i^{-1}(X-m_i)/2\}}{\sum_j \alpha_{ij}\exp\{-(X-a_{ij})^\top\Sigma_i^{-1}(X-a_{ij})/2\}}\right)^2\right] \\ &\le \left(\sum_j \frac{1}{\alpha_{ij}^2}\right)\exp\{-(a^{(i)}-m_i)^\top\Sigma_i^{-1}(a^{(i)}-m_i)\}. \end{aligned}$$

Thus we have that

$$\tilde{\mathbb{E}}[I(X\in\mathcal{S}_\gamma)L^2(X)] \le k\left(\sum_i\sum_j \frac{1}{\alpha_{ij}^2}\right)\exp\{-\min_i(a^{(i)}-m_i)^\top\Sigma_i^{-1}(a^{(i)}-m_i)\}.$$

Moreover, we know that (also see the proof for details)

$$\begin{aligned} \tilde{\mathbb{E}}[I(X\in\mathcal{S}_\gamma)L(X)] &= \mathbb{P}(X\in\mathcal{S}_\gamma) \\ &= \sum_{i=1}^k \pi_i\mathbb{P}(N(m_i,\Sigma_i)\in\mathcal{S}_\gamma) \\ &\sim \exp\{-\min_i(a^{(i)}-m_i)^\top\Sigma_i^{-1}(a^{(i)}-m_i)/2\} \end{aligned}$$

where $\sim$ is up to polynomial factor in γ. Combining the results, we get that this IS distribution achieves the efficiency certificate.

EXAMPLE EC.4 (MULTIVARIATE LAPLACE DISTRIBUTION). Suppose that $X = \sqrt{W}Y$ where $Y \sim N(0, I_d)$ and $W \sim Exp(1)$ is independent of Y. It is known that the density function is

$$f(x) = \frac{2}{(2\pi)^{d/2}}\left(\frac{x^\top x}{2}\right)^{\nu/2}K_\nu(\sqrt{2x^\top x})$$

where $\nu = (2-d)/2$ and K_ν is the modified Bessel function of the second kind. We have that

$$\lambda(s) = \begin{cases} -\log(1 - \frac{1}{2}s^\top s) & s^\top s < 2; \\ \infty & s^\top s \geq 2. \end{cases}$$

We assume that $\{x : a^* \leq x \leq a^* + \varepsilon\mathbf{1}\} \subset \mathcal{S}_\gamma$ for some $\varepsilon > 0$ and $0 \notin \mathcal{S}_\gamma$. By solving $\nabla\lambda(s^*) = a^*$, we get that

$$s^* = \frac{\sqrt{2a^{*T}a^* + 1} - 1}{a^{*T}a^*} a^*.$$

Then

$$\begin{aligned} f^*(x) &= f(x)\exp\{s^{*T}x - \lambda(s^*)\} \\ &= \frac{2}{(2\pi)^{d/2}} \left(\frac{x^\top x}{2}\right)^{\nu/2} K_\nu(\sqrt{2x^\top x}) \exp\left\{\frac{\sqrt{2a^{*T}a^* + 1} - 1}{a^{*T}a^*} a^{*T}x\right\} \frac{\sqrt{2a^{*T}a^* + 1} - 1}{a^{*T}a^*}. \end{aligned}$$

We assume that $a_i^* \to \infty$ polynomially in γ for any i. For $a^* \leq x \leq a^* + \varepsilon\mathbf{1}$, we have that $a^{*T}a^* \leq x^\top x \leq (a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})$ and $a^{*T}x \geq a^{*T}a^*$. Thus, we have that

$$\left(\frac{x^\top x}{2}\right)^{\nu/2} \geq \begin{cases} \left(\frac{a^{*T}a^*}{2}\right)^{\nu/2} & d = 1 \\ \left(\frac{(a^*+\varepsilon\mathbf{1})^\top(a^*+\varepsilon\mathbf{1})}{2}\right)^{\nu/2} & d \geq 2 \end{cases}$$

and

$$\begin{aligned} &K_\nu(\sqrt{2x^\top x}) \exp\left\{\frac{\sqrt{2a^{*T}a^* + 1} - 1}{a^{*T}a^*} a^{*T}x\right\} \\ \geq &K_\nu(\sqrt{2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})}) \exp\left\{\frac{\sqrt{2a^{*T}a^* + 1} - 1}{a^{*T}a^*} a^{*T}a^*\right\} \\ \approx &\frac{1}{(2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1}))^{1/4}} \exp\left\{\sqrt{2a^{*T}a^* + 1} - 1 - \sqrt{2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})}\right\}. \end{aligned}$$

Here we use the fact that $K_\nu(x) \approx e^{-x}/\sqrt{x}$ as $x \to \infty$ where $\approx$ is up to constant factor. Since we have assumed that a_i^* is polynomial in γ, we only need to show that $\exp\left\{\sqrt{2a^{*T}a^* + 1} - 1 - \sqrt{2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})}\right\}$ does not decay exponentially fast in γ. In fact, we have that

$$\begin{aligned} \sqrt{2a^{*T}a^* + 1} - \sqrt{2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})} &= \frac{1 - 4d\varepsilon^2 - 4\varepsilon a^{*T}\mathbf{1}}{\sqrt{2a^{*T}a^* + 1} + \sqrt{2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})}} \\ &\geq \frac{1 - 4d\varepsilon^2 - 4\varepsilon\sqrt{d}\sqrt{a^{*T}a^*}}{\sqrt{2a^{*T}a^* + 1} + \sqrt{2(a^* + \varepsilon\mathbf{1})^\top(a^* + \varepsilon\mathbf{1})}} \end{aligned}$$

and the lower bound converges to a constant. Moreover, it is implied that the components of s^* do not grow exponentially fast in γ. Therefore, all the required assumptions are satisfied.

EC.2. Conservativeness of Standard ERM and Lazy Learner

As in the setup of Theorem 4, suppose that Stage 1 samples are drawn from q. Suppose we use empirical risk minimization (ERM) to train $\hat{g}$, i.e., $\hat{g} := \text{argmin}_{g\in\mathcal{G}}\{R_{n_1}(g) := \frac{1}{n_1}\sum_{i=1}^{n_1}\ell(g(\tilde{X}_i), Y_i)\}$ where ℓ is a loss function and $\mathcal{G}$ is the considered hypothesis class. Correspondingly, let $R(g) := \mathbb{E}\ell(g(\tilde{X}), I(\tilde{X}\in\mathcal{S}_\gamma))$ be the true risk function and $g^* := \arg\min_{g\in\mathcal{G}} R(g)$ its minimizer. Also let $\kappa^* := \min_{x\in\mathcal{S}_\gamma} g^*(x)$ be the true threshold associated with g^* in obtaining the smallest outer rare-event set approximation. Then we have the following result.

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

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

Here, $Lip(g^*)$ *is the Lipschitz parameter of* g^*, *and* $t(\delta,n_1)=3\left(\frac{\log(n_1q_l)+d\log M+\log\frac{1}{\delta}}{n_1q_l}\right)^{\frac{1}{d}}$.

A question is how to give a more refined bound for the false positive rate based on Theorem EC.1, and Theorem 4, that depends on explicit constants of the classification model or training process. Let us examine the terms appearing in Theorem EC.1. The term $\sup_{g_\theta\in\mathcal{G}}\left|R_{n_1}(g_\theta)-R(g_\theta)\right|$ 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 $\mathrm{Lip}(g^*)$ and $R(g^*)$. Future investigations on the expressive power of restricted classes of neural networks would help refining our conservativeness results further.

We provide a corresponding 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 EC.2 **(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 $1-\delta$,*

$$\begin{aligned}
&\mathbb{P}(\tilde{X} \in \mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma) \\
\le& M^{d-1} q_u \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1) \\
=& \sqrt{\frac{e}{\pi(d-1)}} \left(\frac{1}{2}\pi e\right)^{\frac{d-1}{2}} q_u t(\delta, n_1)(1 + O(d^{-1})).
\end{aligned}$$

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.

EC.3. 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 EC.1. We say an estimator $\hat{\mu}_n$ satisfies an lower-bound *relaxed efficiency certificate* to estimate μ if $\mathbb{P}(\hat{\mu}_n - \mu > \epsilon\mu) \le \delta$ 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 μ up to an error of $\epsilon\mu$. We have the following analog to Proposition 2:

COROLLARY EC.1. *Suppose $\hat{\mu}_n$ is downward biased, i.e., $\underline{\mu} := \mathbb{E}[\hat{\mu}_n] \le \mu$. Moreover, suppose $\hat{\mu}_n$ takes the form of an average of n i.i.d. simulation runs Z_i, with* $\mathrm{RE} = \mathrm{Var}(Z_i)/\underline{\mu}^2 = \tilde{O}(\log(1/\underline{\mu}))$. *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 1 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 $\underline{\mathcal{S}}_\gamma$ that has zero false positive rate, i.e.,

$$\mathbb{P}(X \in \underline{\mathcal{S}}_\gamma, Y = 0) = 0. \tag{EC.1}$$

To make sure of (EC.1), 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 $Q(x) := \{x' : x' \ge x\}$. We construct $\mathcal{J}(T_1) := \cup_{x\in T_1} Q(x)$ 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 $\underline{\mathcal{S}}_\gamma$ such that (EC.1) 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 μ 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 $\mathbb{P}(X \in \mathcal{J}(T_1))$ in Stage 2. Since $\mathcal{J}(T_1)$ takes the form $\cup_{x\in T_1} Q(x)$, 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 $\underline{\mathcal{S}}^\kappa_\gamma = \{x : \hat{g}(x) \ge \kappa\}$, where κ is say $1/2$. Then we increase κ minimally above from $1/2$, say to $\hat{\kappa}$, so that $\underline{\mathcal{S}}^{\hat{\kappa}}_\gamma \subset \mathcal{J}(T_1)$, i.e., $\hat{\kappa} = \min\{\kappa \in \mathbb{R} : \underline{\mathcal{S}}^{\hat{\kappa}}_\gamma \subset \mathcal{J}(T_1)\}$. Then $\underline{\mathcal{S}}^{\hat{\kappa}}_\gamma$ 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 $\mathbb{P}(X \in \underline{\mathcal{S}}^{\hat{\kappa}}_\gamma)$ in Stage 2.

As we can see, compared with Algorithm 1, the only difference is how we adjust κ in Stage 1. And similar to Theorem 3, we also have that Algorithm 2 attains the lower-bound relaxed efficiency certificate:

Algorithm 2: Deep-PrAE to estimate $\mu = \mathbb{P}(X \in \mathcal{S}_\gamma)$ (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 density function $f(x)$.

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

Stage 1 (Set Learning):

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

Replace κ by $\hat{\kappa} = \min\{\kappa \in \mathbb{R} : \underline{\mathcal{S}}_\gamma^{\hat{\kappa}} \subset \mathcal{J}(T_1)\}$;

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

The same as Stage 2 of Algorithm 1.

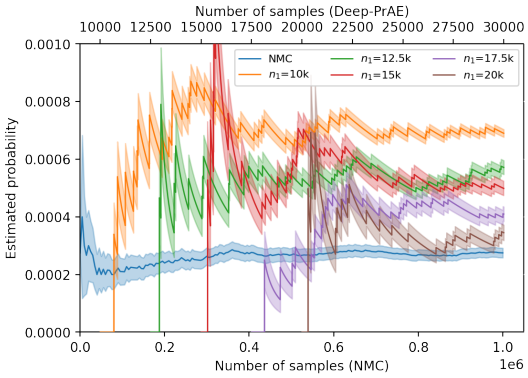
THEOREM EC.3 **(Lower-bound relaxed efficiency certificate)**. *Suppose $\mathcal{S}_\gamma$ is orthogonally monotone, and $\underline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ satisfies the same conditions for $\mathcal{S}_\gamma$ in Theorem 1. Then Algorithm 2 attains the lower-bound relaxed efficiency certificate 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 $\mathbb{P}(X \notin \underline{\mathcal{S}}_\gamma^{\hat{\kappa}}, Y = 1)$. Like in Section 3, we use ERM to train $\hat{g}$, i.e., $\hat{g} := \operatorname{argmin}_{g \in \mathcal{G}}\{R_{n_1}(g) := \frac{1}{n_1}\sum_{i=1}^{n_1} \ell(g(\tilde{X}_i), Y_i)\}$ where ℓ is a loss function and $\mathcal{G}$ is the considered hypothesis class. Let g^* be the true risk minimizer as described in Section 3. For inner approximation, we let $\kappa^* := \max_{x \in \mathcal{S}_\gamma^c} g^*(x)$ 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 EC.1.

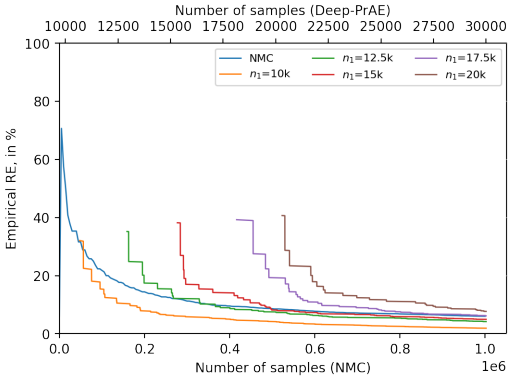
THEOREM EC.4 **(Lower-bound estimation conservativeness)**. *Consider $\underline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ obtained in Algorithm 2 where $\hat{g}$ is trained from an ERM. Suppose the density q has bounded support $K \subset [0, M]^d$ and $0 < q_l \leq q(x) \leq q_u$ for any $x \in K$. Also 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 ℓ is the squared loss, then $h(\kappa)$ could be chosen as $h(\kappa) = (1-\kappa)^2$). Then, with probability at least $1 - \delta$,*

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

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



(a) Estimated rare-event probability (Additional Example 1)



(b) Estimator's empirical relative error (Additional Example 1)

Figure EC.1: Estimated rare-event probability and empirical relative error for Additional Example 1

EC.4. Additional Numerical Experiments

In this section, we study two additional experiments and provide implementation details of the intelligent driving example. The first additional example is a classical setting in rare-event simulation which allows an analytic approach to design IS. The second additional example shows how our approach could be applied when the input distribution is not Gaussian.

EC.4.1. Additional Example 1: Random Walk

In this example, our goal is to estimate the excursion probability of a T-step random walk $\mu = P\left(\max_{t=1,\cdots,T} S_t > \gamma\right)$, where $S_t = \sum_{i=1}^{t} X_i$, X_i for $i = 1, \ldots, T$ are i.i.d. following distribution p, and the rarity parameter is the cross level γ. Note that μ is equivalent to $P(\tau \leq T)$, where $\tau = \min\{t \in \mathbb{N} : S_t > \gamma\}$ is a stopping time. We use $T = 10$, $p = N(0, \sigma^2 I_{T\times T})$, and $\gamma = 11$.

Figure EC.1 compares the estimators and REs obtained using NMC and Deep-PrAE with various Stage 1 sample sizes (n_1) using uniform sampling on $[-4, 4]^T$. We set the maximum total sample

size for any of the Deep-PrAE estimators to be $n = 30,000$, and the Stage 2 sample size is $n_2 = n - n_1$ and train a fully connected neural network with 4 hidden layers (8, 8, 4, 2 hidden nodes) using all n_1 samples, L2 regularization, and 1,000 training iterations, each using a batch size of 1,000 and stochastic gradient descent to update the neural network parameters. In the figure, we show the sample sizes for NMC and Deep-PrAE separately due to their significant difference in magnitude and usage: NMC sample size in the bottom x-axis (all used to drive down estimator variance) and Deep-PrAE sample size in the top x-axis (used for both n_1 and n_2 with different proportions). The starting x-value of Deep-PrAE estimator curve marks the portion of sample budget used for n_1 while the rest used for n_2. Therefore, high variances occur at the beginning of each curve, which then saturate as larger n_2 is allocated to drive down the variance.

We observe that all Deep-PrAE estimators output valid upper bounds for the target probability, up to 2.5 times larger than the target value, with much better relative errors (all converge to less than 5% with total sample size $n = 3 \times 10^4$) compared to NMC which has a relative error of 15% when sample size $n = 10^6$). We also observe that allocating more samples to Stage 1 (using $n_1 = 17.5 \times 10^3$ and $n_2 = 12.5 \times 10^3$) leads to less conservative estimate and more confident estimation in this example. However, it is expected that when n_1 increases too much further, then there could be a deterioration in performances because of a strained n_2.

EC.4.2. Additional Example 2: Non-Gaussian Example

We now consider a problem setting with non-Gaussian random variables. Our Algorithm 1 can be directly applied to non-Gaussian distributions by replacing the rate function in (4) with the corresponding rate function of the new distribution. This requires solving MIP with non-quadratic objectives and, while we can build or look for solvers for this task, for coherence we would like to stay with using the commercial solver Gurobi (which is used in all other examples). Unfortunately, the latter only accepts linear and quadratic objectives (see Section 10.3 in Gurobi Optimization, LLC (2023)). This prompts us to consider a general transformation that allows converting a non-Gaussian problem back to Gaussian which involves only quadratic-objective MIPs that are amenable to Gurobi.

With the above motivation, we describe our transformation approach that can convert a problem with independent random variables following an arbitrary continuous distribution into a Gaussian problem. Moreover, this transformation maintains the desirable orthogonal monotonicity property of the rare-event set. More concretely, suppose we are interested in estimating the probability

$P((X_1, X_2, \cdots, X_d) \in \mathcal{E})$, where $X_1, \cdots, X_d$ are independent and follow arbitrary continuous distributions $F_1, \cdots, F_d$ respectively. By applying the standard inverse transform method, the rare-event probability can be written as $P((Y_1, \cdots, Y_d) \in \tilde{\mathcal{E}})$, where $Y_1, \cdots, Y_d$ are i.i.d. $N(0,1)$ variables and the transformed rare-event set $\tilde{\mathcal{E}}$ is given by $\tilde{\mathcal{E}} = \{(y_1, \cdots, y_d) : (F_1^{-1}(\Phi(y_1)), \cdots, F_d^{-1}(\Phi(y_d))) \in \mathcal{E}\}$. We have the following guarantee:

PROPOSITION EC.1. *Suppose $X_1, \cdots, X_d$ are independent and follow arbitrary continuous distributions $F_1, \cdots, F_d$ respectively and the rare-event set $\mathcal{E}$ is orthogonally monotone. Consider the probability $P((Y_1, \cdots, Y_d) \in \tilde{\mathcal{E}})$, where $Y_1, \cdots, Y_d$ are i.i.d. $N(0,1)$ variables and the transformed rare-event set $\tilde{\mathcal{E}}$ is given by $\tilde{\mathcal{E}} = \{(y_1, \cdots, y_d) : (F_1^{-1}(\Phi(y_1)), \cdots, F_d^{-1}(\Phi(y_d))) \in \mathcal{E}\}$. Then we have $P((X_1, X_2, \cdots, X_d) \in \mathcal{E}) = P((Y_1, \cdots, Y_d) \in \tilde{\mathcal{E}})$, where $\tilde{\mathcal{E}}$ is orthogonally monotone.*

Proposition EC.1 stipulates that both the value of the probability and the orthogonal monotonicity of the rare-event set are retained in the transformation as long as the marginal distribution functions involved in the transformation are continuous. As discussed above, this transformation allows us to convert non-Gaussian problems into Gaussian that facilitates MIP computation. On the other hand, we also note that this transformation technique is very general, applying to not only our Deep-PrAE framework but also classical rare-event problems with more tractable target sets. Moreover, it not only changes the cumulant generating function of the underlying distribution, but also has the ability to change the problem nature from heavy to light tail. The full investigation of such a technique warrants a separate future work, and here we only aim to use this technique to bypass the need to use an alternative solver than Gurobi.

Next, let us consider the example of estimating the probability of the sum of random variables exceeding a certain threshold γ_r, where the random variables are a combination of exponential and generalized Pareto. Specifically, we let $X_i \sim \text{Expo}(\mu_i)$ for $i \in \{1,2,3\}$ and $X_i \sim \text{GeneralizedPareto}(\xi_i, \sigma_i)$ for $i \in \{4,5,6\}$ where $1/\mu_i$ is the rate of exponential distribution and ξ_i, σ_i are the tail and shape parameters of the generalized Pareto distribution, respectively. We aim to estimate

$$\mathbb{P}\left(X_1 + \ldots + X_d > \gamma_r\right). \tag{EC.2}$$

Using the above transformation technique, we can rewrite (EC.2) as

$$\mathbb{P}\left((Y_1, \ldots, Y_d) \in \tilde{\mathcal{E}}\right),$$

with $\tilde{\mathcal{E}}$ defined as

$$\tilde{\mathcal{E}} = \left\{(y_1, y_2, \ldots, y_d) : \sum_{i=1}^{d} F_i^{-1}\left(\Phi\left(y_i\right)\right) > \gamma\right\}.$$

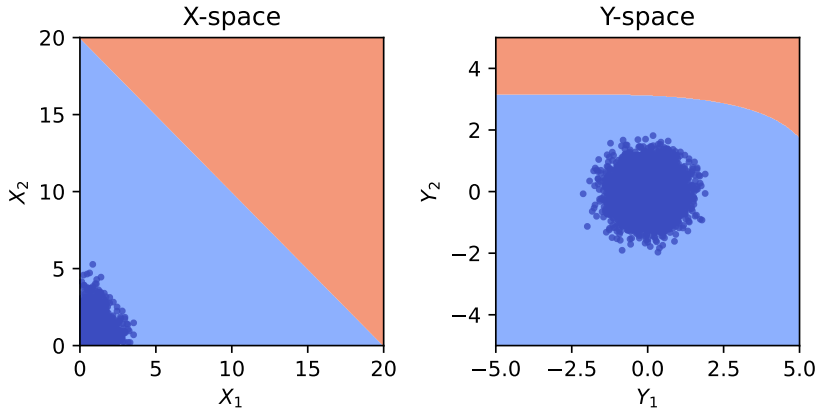


Figure EC.2: Illustration of the rare-event problem setting in Additional Example 2 in a 2-d space. Blue: non-rare-event set, orange: rare-event set.

Note that we have used $d = 6$ in this problem with $\gamma_r = 140$. For better visualization, Fig. EC.2 illustrates the problem for a simplified 2-d example with $X_1 \sim \text{Expo}(1)$ and $X_2 \sim \text{GeneralizedPareto}(\frac{1}{4}, 1)$, and $\gamma_r = 20$. We see that the half-space boundary of the rare-event set in the X-space transforms into a more complex, but still orthogonally monotonic, boundary in the Y-space.

We then apply Deep-PrAE, by using $n_1 = 5,000$ Stage 1 samples from uniform distribution on $[-5, 5]^d$ in the Y-space to train the classifier. Here, we continue to use a feedforward structure as previous examples with 4 hidden layers, each with 4, 8, 4, and 2 nodes and ReLU activation functions except in the last layer. We train the model using L2 regularization with 1,000 iterations, achieving approximately 77% classification accuracy, before applying the threshold tuning procedure. After tuning the threshold parameter, we obtain our outer-approximation, which then allows us to solve for the dominating points, construct our IS proposal distribution, and collect Stage 2 samples to compute our final estimate. Fig. EC.3 illustrates this procedure. Finally, the numerical result for the full 6-d problem with $\mu_i = 1, 1.5, 2$ for $i = 1, 2, 3$ and $\xi_i = 1/4, 1/5, 1/5, \sigma_i = 1, 2, 3$ for $i = 4, 5, 6$ is shown in Fig. EC.4.

In this non-Gaussian example, we assess the performance of four estimation methods NMC, Deep-PrAE, CE and AMS. We first estimate the ground-truth target value as 1.23×10^{-5} by using NMC with a much larger sample size (100 million). Figure EC.4 shows the results. Deep-PrAE overestimates the probability by approximately 44.14%, while both CE and AMS underestimate the probability, with CE exhibiting an underestimation of 65.59% and AMS of 48.39%. Regarding the empirical RE, we observe that using a total sample size of $n = 500,000$, Deep-PrAE achieves a

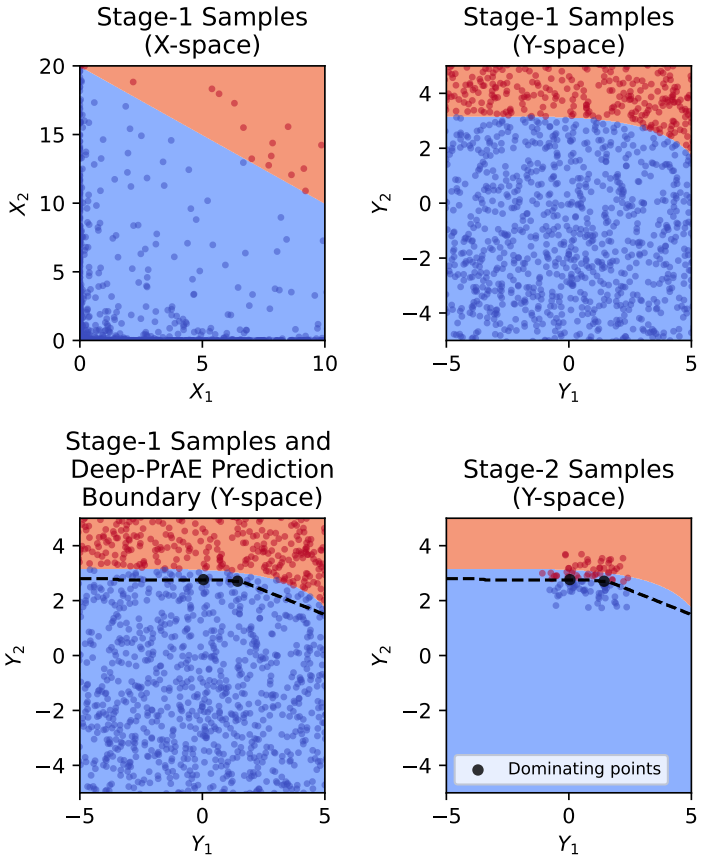


Figure EC.3: Illustration of Deep-PrAE approach applied using uniform Stage 1 samples in the Y-space.

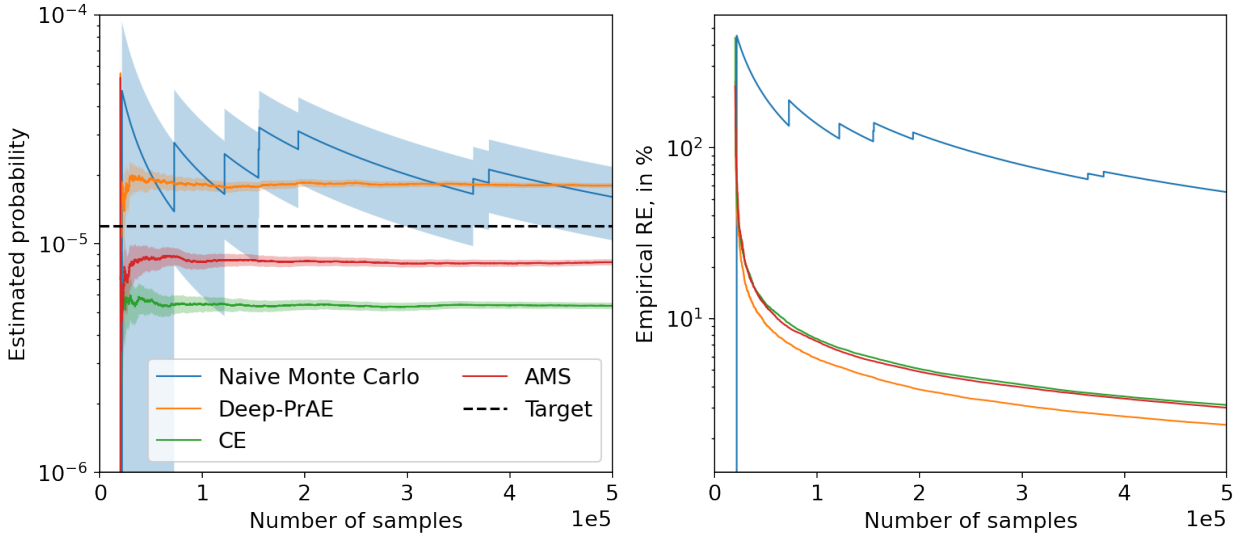


Figure EC.4: Estimated rare-event probability and the empirical RE (Additional Example 2).

reduction in empirical RE compared to NMC by 95.47%, while CE and AMS achieve reductions of 93.38% and 93.51% respectively. These observations appear consistent with previous examples in that Deep-PrAE avoids under-estimation of the target probability experienced by CE and AMS. Moreover, the RE of Deep-PrAE is competitive compared to the latter. Our findings here suggest that Deep-PrAE offers a promising alternative to benchmark methods even for non-Gaussian problems.

EC.4.3. Implementation Details for Intelligent Driving Example

Here we provide implementation details on the example presented in Section 4.4.

EC.4.3.1. LV Actions. The LV action contains human-driving uncertainty in decision making. For every Δt time-steps, a random variable u_t is generated to be appended into 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.

EC.4.3.2. Intelligent Driver Model for AV. The intelligent driver model is governed by the following equations (the subscripts "follow" and "lead" defined in Figure 10 is abbreviated to "f" and "l" for conciseness):

$$\dot{x}_f = v_f \tag{EC.3}$$

$$\dot{x}_l = v_l \tag{EC.4}$$

$$\dot{v}_f = \max\left(a(1 - \left(\frac{v_f}{v_0}\right)^{\delta}\right. \tag{EC.5}$$

$$\left. - \left(\frac{s^*(v_f, \Delta v_f)}{s_f}\right)^2), -d\right) \tag{EC.6}$$

$$s^*(v_f, \Delta v_f) = s_0 + v_f \bar{T} + \frac{v_f \Delta v_f}{2\sqrt{ab}} \tag{EC.7}$$

$$s_f = x_l - x_f - L \tag{EC.8}$$

$$\Delta v_f = v_f - v_l, \tag{EC.9}$$

The parameters are presented in Table EC.1. The randomness of LV actions u_t's propagates into the system and affects the state s_t. The intelligent driver model 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

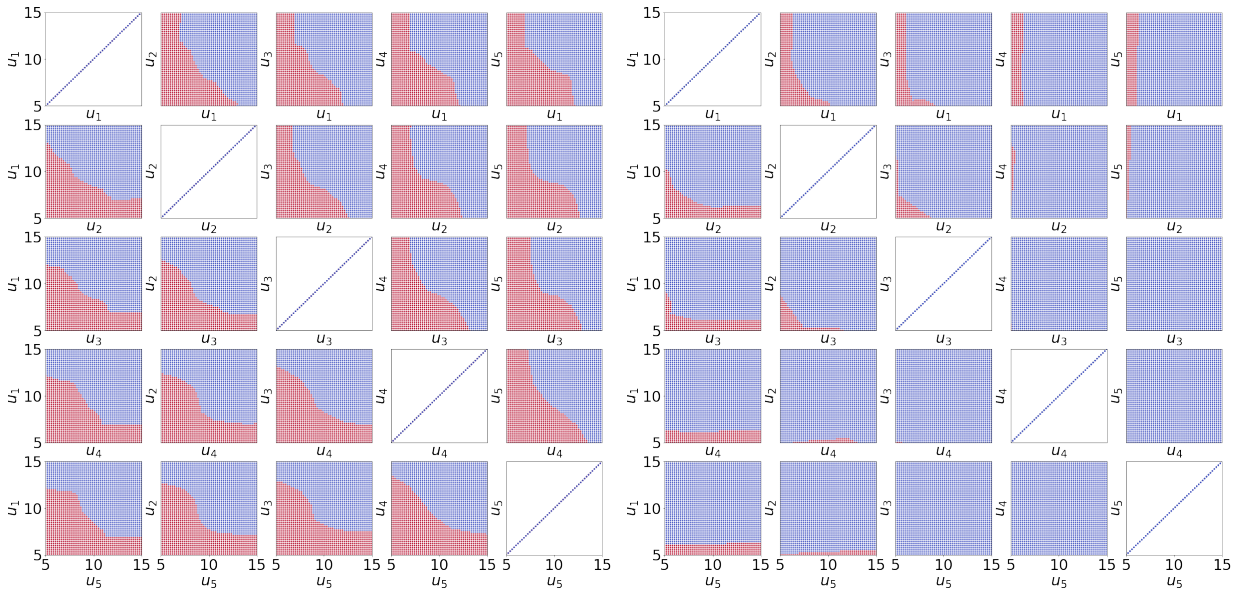
Table EC.1: Parameters of the Intelligent Drivers Model

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

"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.

EC.4.3.3. Rarity Parameter γ. Parameter γ signifies the range invoked by the AV acceleration and deceleration pedals. Increasing γ implies that the AV can have sudden high deceleration and hence avoid crash scenarios better and making crashes rarer. In contrast, decreasing γ 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 EC.5 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.

EC.4.3.4. Sample Trajectories. Figure EC.6 shows two examples of sample trajectories, one successfully maintaining a safe distance, and the other leading to a crash. In Figure EC.6(e)-(h) where we show the crash case, the AV maintains a safe distance behind the LV until the latter starts rapidly decelerating (Figure EC.6(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 EC.5: 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.

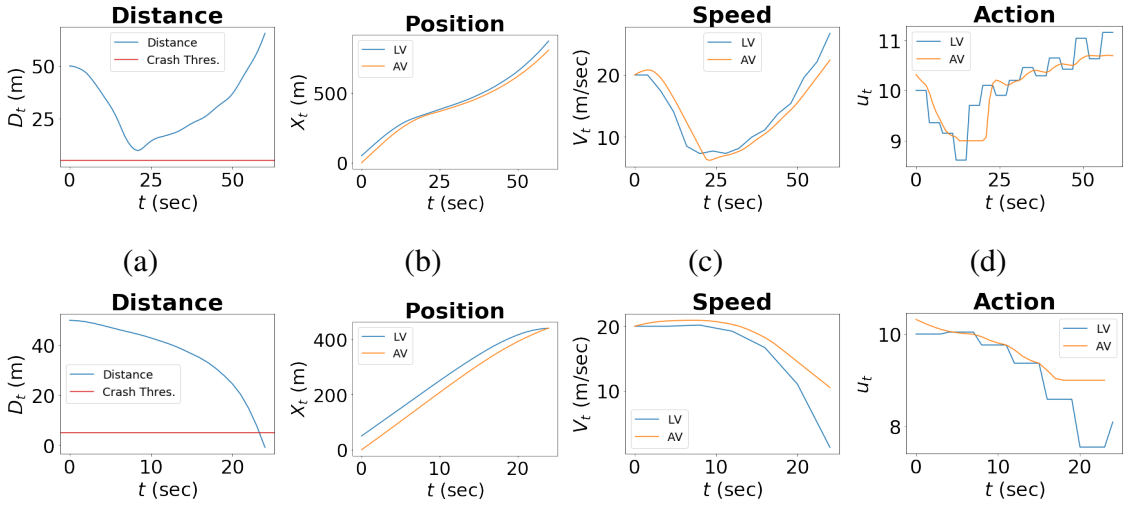


Figure EC.6: Autonomous Car Following Experiment Trajectories. Figures (a) - (d) represent a simulation episode without a crash occurring where the AV follows the LV successfully at a safe distance. Figures (e) - (h) represents a simulation episode where crash occurs at $t = 23$ seconds due to the repeated deceleration of the LV.

EC.5. Proofs

EC.5.1. Proofs for the Dominating Point Methodologies

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

Proof of Lemma 1 See Rockafellar (1970). □

Proof of Theorem 1 For simplicity, we denote $s_j := s_{a_j}$. First, we develop an upper bound for the second moment of the IS estimator. We know that the likelihood ratio function

$$L(x) = \frac{1}{\sum_j \alpha_j \exp\{s_j^\top x - \lambda(s_j)\}} \le \frac{1}{\alpha_j \exp\{s_j^\top x - \lambda(s_j)\}}, \forall j.$$

Since $\mathcal{S}_\gamma^j \subset \{x : s_j^\top (x - a_j) \ge 0\}$, we have that

$$\mathbb{1}(X \in \mathcal{S}_\gamma^j) L(X) \le \frac{1}{\alpha_j \exp\{s_j^\top a_j - \lambda(s_j)\}} = \frac{1}{\alpha_j} \exp\{-I(a_j)\}.$$

As $\mathcal{S}_\gamma^j$'s are disjoint and $\mathcal{S} = \bigcup_j \mathcal{S}_\gamma^j$, we know that $Z = \mathbb{1}(X \in \mathcal{S}_\gamma) L(X) \le \frac{1}{\min_j \alpha_j} \exp\{-\min_j I(a_j)\} = \frac{1}{\min_j \alpha_j} \exp\{-I(a^*)\}$. Therefore,

$$\tilde{\mathbb{E}}[Z^2] \le \frac{1}{(\min_j \alpha_j)^2} \exp\{-2I(a^*)\}.$$

Note that we have assumed $\max_j \frac{1}{\alpha_j}$ at most grows polynomially in γ, so $\frac{1}{(\min_j \alpha_j)^2}$ also at most grows polynomially.

Next, we develop a lower bound for $\tilde{\mathbb{E}}[Z] = \mathbb{P}(X \in \mathcal{S}_\gamma)$. We have assumed that $\{x : a^* \le x \le a^* + \varepsilon\mathbf{1}\} \subset \mathcal{S}_\gamma$ and $f^*(x) := f(x) \exp\{s^{*T} x - \mu(s^*)\} \ge \varepsilon'(a^*)$ for any $a^* \le x \le a^* + \varepsilon\mathbf{1}$ where $\varepsilon > 0$ is a constant and $\varepsilon'(a^*) > 0$ at most decays polynomially in γ. Then we have that

$$\begin{aligned}
\tilde{\mathbb{E}}[Z] &\ge \mathbb{P}(a^* \le X \le a^* + \varepsilon\mathbf{1}) \\
&= \int_{a^* \le x \le a^* + \varepsilon\mathbf{1}} \exp\{\mu(s^*) - s^{*T} x\} f^*(x) \mathrm{d}x \\
&\ge \exp\{\mu(s^*) - s^{*T} a^*\} \varepsilon'(a^*) \int_{a^* \le x \le a^* + \varepsilon\mathbf{1}} \exp\{-s^{*T}(x - a^*)\} \mathrm{d}x \\
&= \exp\{-I(a^*)\} \varepsilon'(a^*) \prod_{i=1}^d \frac{1 - \exp\{-s_i^* \varepsilon\}}{s_i^*}.
\end{aligned}$$

Again, $\{x : a^* \le x \le a^* + \varepsilon\mathbf{1}\} \subset \mathcal{S}_\gamma$ implies that $s_i^* \ge 0, \forall i = 1, \cdots, d$. If $s_i^* = 0$, then we naturally use ε to substitute $\frac{1-\exp\{-s_i^*\varepsilon\}}{s_i^*}$. Since s_i^* at most grows polynomially in γ, finally we get that $\tilde{\mathbb{E}}[Z^2]/\tilde{\mathbb{E}}[Z]^2$ at most grows polynomially in γ. Thus we have proved the efficiency certificate. □

Proof of Theorem 2 We know that $\tilde{\mathbb{E}}[Z] = \bar{\Phi}(\gamma) + \bar{\Phi}(k\gamma)$. Moreover,

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

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

On the other hand, we denote $Z_i' = I(X_i \ge \gamma)L(X_i)$ and we note that $\tilde{\mathbb{E}}[Z'] = \bar{\Phi}(\gamma), \tilde{\mathbb{E}}[Z'^2] = e^{\gamma^2}\bar{\Phi}(2\gamma), \tilde{\mathbb{E}}[Z'^4] = e^{6\gamma^2}\bar{\Phi}(4\gamma)$. We know that

$$\begin{aligned}
&\tilde{\mathbb{P}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \bar{\Phi}(\gamma)\right| > \varepsilon\bar{\Phi}(\gamma)\right) \\
\le&\tilde{\mathbb{P}}(\exists i : X_i \le -k\gamma) + \tilde{\mathbb{P}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i' - \bar{\Phi}(\gamma)\right| > \varepsilon\bar{\Phi}(\gamma)\right).
\end{aligned}$$

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

$$\tilde{\mathbb{P}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i' - \bar{\Phi}(\gamma)\right| > \varepsilon\bar{\Phi}(\gamma)\right) \le \frac{\tilde{\mathbb{E}}[Z'^2]}{n\varepsilon^2\bar{\Phi}^2(\gamma)} = \frac{e^{\gamma^2}\bar{\Phi}(2\gamma)}{n\varepsilon^2\bar{\Phi}^2(\gamma)} = \Theta\left(\frac{\gamma}{n\varepsilon^2}\right).$$

Thus $\mathbb{P}(|\hat{\mu}_n - \bar{\Phi}(\gamma)| > \varepsilon\bar{\Phi}(\gamma)) = O\left(\frac{\gamma}{n\varepsilon^2}\right)$. Similarly, we also know that

$$
\begin{aligned}
&\tilde{\mathbb{P}}\left(\frac{\frac{1}{n}\sum_{i=1}^n Z_i^2}{\left(\frac{1}{n}\sum_{i=1}^n Z_i\right)^2} \geq \frac{1+\varepsilon}{(1-\varepsilon)^2}\frac{e^{\gamma^2}\bar{\Phi}(2\gamma)}{\bar{\Phi}^2(\gamma)}\right)\\
\leq&\tilde{\mathbb{P}}(\exists i : X_i \leq -k\gamma) + \tilde{\mathbb{P}}\left(\frac{\frac{1}{n}\sum_{i=1}^n Z_i'^2}{\left(\frac{1}{n}\sum_{i=1}^n Z_i'\right)^2} > \frac{1+\varepsilon}{(1-\varepsilon)^2}\frac{e^{\gamma^2}\bar{\Phi}(2\gamma)}{\bar{\Phi}^2(\gamma)}\right)\\
\leq&\tilde{\mathbb{P}}(\exists i : X_i \leq -k\gamma) + \tilde{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n Z_i'^2 > (1+\varepsilon)e^{\gamma^2}\bar{\Phi}(2\gamma)\right) + \tilde{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n Z_i' < (1-\varepsilon)\bar{\Phi}(\gamma)\right)\\
=&\tilde{\mathbb{P}}(\exists i : X_i \leq -k\gamma) + \tilde{\mathbb{P}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i'^2 - e^{\gamma^2}\bar{\Phi}(2\gamma)\right| > \varepsilon e^{\gamma^2}\bar{\Phi}(2\gamma)\right) + \tilde{\mathbb{P}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i' - \bar{\Phi}(\gamma)\right| > \varepsilon\bar{\Phi}(\gamma)\right)\\
\leq&\tilde{\mathbb{P}}(\exists i : X_i \leq -k\gamma) + \frac{\tilde{\mathbb{E}}[Z'^4]}{n\varepsilon^2 e^{2\gamma^2}\bar{\Phi}^2(2\gamma)} + \frac{\tilde{\mathbb{E}}[Z'^2]}{n\varepsilon^2\bar{\Phi}^2(\gamma)}\\
=&\Theta\left(\frac{\gamma}{n\varepsilon^2}\right).
\end{aligned}
$$

□

EC.5.2. Proofs for the Relaxed Efficiency Certificate

Proof of Proposition 2 We have

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

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

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

so that the relaxed efficiency certificate is achieved as long as

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

□

Proof of Proposition 3 The proof follows from that of Proposition 2 with a conditioning on D_{n_1}. We have

$$
\begin{aligned}
\mathbb{P}(\hat{\mu}_n - \mu < -\epsilon\mu | D_{n_1}) \leq \mathbb{P}(&\hat{\mu}_n - \overline{\mu}(D_{n_1})\\
&< -\epsilon\overline{\mu}(D_{n_1}) | D_{n_1})
\end{aligned}
$$

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

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

so that

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

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 3 We have assumed that $\overline{\mathcal{S}}_\gamma^{\hat{\kappa}}$ satisfies the assumptions for $\mathcal{S}_\gamma$ in Theorem 1. Then following the proof of Theorem 1, we obtain the efficiency certificate for the IS estimator in estimating its mean. Theorem 3 is then proved by directly applying Corollary 1. □

EC.5.3. Proofs for Conservativeness

Recall that $T_0 = \{\tilde{X}_i : Y_i = 0\}$ where the samples are generated as in Algorithm 1. By some combinatorial 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 EC.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 of Lemma EC.1 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 δ, except for the $d-$th dimension (the δ here is not exactly the δ

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_δ, 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 $B_{3\delta} = \cup_{j\in J} \left(B_{3\delta} \cap Z_j \right) = \cup_{j \in J_0} \left(B_{3\delta} \cap Z_j \right)$. 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

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

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

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

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

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

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 $\mathbb{P}(B_{t(\delta)} \subset \mathcal{H}(T_0)) \geq 1 - \delta$. □

Proof of Theorem EC.2 First, we show the inequality in the theorem, i.e., $\mathbb{P}(\tilde{X} \in \mathcal{H}(T_0)^c \backslash \mathcal{S}_\gamma) \leq M^{d-1} q_u \left(\frac{\sqrt{d}}{2}\right)^{d-1} w_{d-1} t(\delta, n_1)$. 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 EC.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 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

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

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 EC.1 By Markov inequality and the definition of h,$\bar{\mathcal{S}}_\gamma^{\hat{\kappa}}$, we know that

$$\begin{aligned}\mathbb{P}\left(\tilde{X}\in\bar{\mathcal{S}}_\gamma^{\hat{\kappa}}, \tilde{X}\in\mathcal{S}_\gamma^c\right) &= \mathbb{P}(\hat{g}(\tilde{X})\geq\hat{\kappa}, \tilde{X}\in\mathcal{S}_\gamma^c)\\ &\leq \frac{R(\hat{g})}{h(\hat{\kappa})}.\end{aligned} \tag{EC.10}$$

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

$$R(\hat{g}) \leq R_{n_1}(\hat{g}) + \sup_{g_\theta\in\mathcal{G}}\left|R_{n_1}(g_\theta) - R(g_\theta)\right| \tag{EC.11}$$

$$\leq R_{n_1}(g^*) + \sup_{g_\theta\in\mathcal{G}}\left|R_{n_1}(g_\theta) - R(g_\theta)\right|$$

$$\leq R(g^*) + 2\sup_{g_\theta\in\mathcal{G}}\left|R_{n_1}(g_\theta) - R(g_\theta)\right|. \tag{EC.12}$$

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 EC.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 $1-\delta$,

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

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

Proof of Theorem 4 The proof goes the same way as the proof of Theorem EC.1. We only need to replace g and g_θ with f and f_θ in (EC.12) (Note that (EC.10) still holds because of our assumption that $g_\theta(x) \geq \kappa \Rightarrow \ell(f_\theta(x), 0) \geq h(\kappa)$). □

EC.5.4. Proofs for the Number of Dominating Points

Proof of Proposition 4 First, we show that the region $\{x : g(x) \geq \kappa\}$ can be represented as the union of at most $2^{d_1+\cdots+d_{n_g}}$ polyhedrons. This is done by considering all possible combinations of the signs of $LT_i(\cdot)$. Indeed, the polyhedrons can be specified as follows: for each $s = (s_1, \cdots, s_{n_g})$ where $s_i \in \{-1, 1\}^{d_i}$, we define

$$\begin{aligned}
\mathbb{P}_s := \{&x : \mathrm{sign}(LT_1(x)) = s_1, \\
&\mathrm{sign}(LT_2(g_1(x))) = s_2, \\
&\mathrm{sign}(LT_3(g_2 \circ g_1(x))) = s_3, \\
&\cdots, \\
&\mathrm{sign}(LT_{n_g}(g_{n_g-1} \circ \cdots \circ g_1)(x))) = s_{n_g}, \\
&g(x) \geq \kappa\}
\end{aligned}$$

Here, the sign operator is performed elementwise and for the ease of notation, we suppose that sign 0 can be either -1 or 1, so $\mathrm{sign}(x) = 1 \iff x \geq 0$ and $\mathrm{sign}(x) = -1 \iff x \leq 0$. Then, we have that $\{x : g(x) \geq \kappa\} = \cup_{s \in \{-1,1\}^{d_1+\cdots+d_{n_g}}} \mathbb{P}_s$. Moreover, we have that $\mathbb{P}_s$ is a polyhedron since the the max operator can be written as a linear mapping given the signs of LT_i.

Next, we claim that each $\mathbb{P}_s$ can contribute at most one dominating point, which is $\arg\min_{x\in\mathbb{P}_s} I(x)$. To see this, note that $\mathbb{P}_s$ is convex and $I(\cdot)$ is strictly convex, so the only local minimizer of $I(\cdot)$ in $\mathbb{P}_s$ is given by the global minimizer in $\mathbb{P}_s$ (also note that any dominating point must be a local minimizer of $I(\cdot)$). Therefore, the number of dominating points can be bounded by the number of polyhedrons, which is $2^{d_1+\cdots+d_{ng}}$. □

Proof of Proposition 5 We construct a worst-case event for the case where $d = 2$ and $X \sim N(0, I_2)$. In this case, we have that $I(y) = \frac{1}{2}y^\top y$. Suppose that the rare event is given as $\{x : (x_1 + x_2)/\sqrt{2} \geq \gamma, x_1, x_2 \geq 0\}$. Consider the situation where the set $\{x : x^\top x \leq (\gamma - \epsilon)^2\} \subset \mathcal{H}(T_0)$ but $\{x : x^\top x \geq \gamma^2\} \cap \mathcal{H}(T_0) = \emptyset$, which happens with positive probability when n_1 is sufficiently large. Denote $\theta_0 = \arccos\frac{\gamma-\epsilon}{\gamma}$. Note that the dominating points of $\mathcal{H}(T_0)^c$ lie on the boundary of $\mathcal{H}(T_0)$ whose distance to 0 is at least $\gamma - \epsilon$. Therefore, with a bit of geometric argument, we observe that the point that dominates $(\gamma\cos\theta, \gamma\sin\theta)(\in \mathcal{H}(T_0)^c)$ must lie in $S_\theta := \{x = (\gamma_1\cos\theta_1, \gamma_1\sin\theta_1) : \gamma_1 \in [\gamma - \epsilon, r], \theta_1 \in [\theta - \theta_0, \theta + \theta_0]\}$. Therefore, the set of points $\{(\gamma\cos\theta, \gamma\sin\theta) : \theta = 3k\theta_0, k = 0, 1, \cdots, \lfloor\pi/(6\theta_0)\rfloor\}$ would require $\lfloor\pi/(6\theta_0)\rfloor$ different dominating points (as $S_{3k_1\theta_0} \cap S_{3k_2\theta_0} = \emptyset$ when $k_1 \neq k_2$). When ϵ is sufficiently small so that θ_0 is sufficiently small, we have that $\lfloor\pi/(6\theta_0)\rfloor \geq k$, which proves the desired result. □

EC.5.5. Proofs When There Exists Holes

Proof of Proposition 6 The proof actually follows from Lemma EC.1. Indeed, $T_0' \cap \Delta(\mathcal{S}_\gamma) = \emptyset$ is equivalent to say that for every point in $x \in \Delta(\mathcal{S}_\gamma)$, there exists $x' \in T_1$ such that $x' \leq x$. With our assumption, we have that $\Delta(S_\gamma) \subset \{x : x - \rho_t\mathbf{1}_d \in \mathcal{S}_\gamma \cup \Delta(\mathcal{S}_\gamma)\}$ (in other words, every point in the holes has distance at least ρ_t to the boundary of the union of the rare event set and the holes). Therefore, it sufficies to show that every point with sufficient distance to this boundary can be covered by the hall $\bar{\mathcal{H}}(T_1) := \cup_{x'\in T_1}\{x : x \geq x'\}$. This is almost the same as what is proved in Lemma EC.1, except that we are using $\bar{\mathcal{H}}(T_1)$ instead of $\mathcal{H}(T_0)$ and correspondingly $\Delta(\mathcal{S}_\gamma)$ is located in the "northeast" of the boundary instead of the "southwest" of the boundary. □

Proof of Theorem 5 From Proposition 6, we know that the probability that stage 1 provides an outer approximation is at least $1 - \delta$. Therefore, from Theorem 3 and union bound, we get the desired result. □

EC.5.6. Proof for Gaussian Transformation

Proof of Proposition EC.1 We consider the the following random variables $Y_1 = \Phi^{-1}(F_1(X_1)), ..., Y_d = \Phi^{-1}(F_d(X_d))$. Note that $Y_1, ..., Y_d \sim N(0,1)$ (by the continuity of $F_1, ..., F_d$). The joint CDF is given by

$$
\begin{aligned}
&F_{Y_1,...,Y_d}(y_1, ..., y_d) \\
&= P(Y_1 \le y_1, ..., Y_d \le y_d) \\
&= P(\Phi^{-1}(F_1(X_1)) \le y_1, ..., \Phi^{-1}(F_d(X_d)) \le y_d) \\
&= P(X_1 \le F_1^{-1}(\Phi(y_1)), ..., X_d \le F_n^{-1}(\Phi(y_d))) \\
&= F_{X_1,...,X_d}(F_1^{-1}(\Phi(y_1)), ..., F_n^{-1}(\Phi(y_d))) \\
&= F_1(F_1^{-1}(\Phi(y_1)))...F_n(F_d^{-1}(\Phi(y_d))) \\
&= \Phi(y_1)...\Phi(y_d) \\
&= \Phi_{0,I}(y_1, ..., y_d).
\end{aligned}
$$

Hence if $(Y_1, ..., Y_d) \sim N(0, I)$, we have $(F_1^{-1}(\Phi(Y_1)), ..., F_d^{-1}(\Phi(Y_d))) \sim F_{X_1,...,X_d}$.

We note that $P((X_1, X_2, \cdots, X_d) \in \mathcal{E}) = P((Y_1, \cdots, Y_d) \in \tilde{\mathcal{E}})$ with

$$\tilde{\mathcal{E}} = \{(y_1, \cdots, y_d) : (F_1^{-1}(\Phi(y_1)), \cdots, F_n^{-1}(\Phi(y_d))) \in \mathcal{E}\},$$

because $(F_1^{-1}(\Phi(Y_1)), \cdots, F_n^{-1}(\Phi(Y_d)))$ has the same distribution as $(X_1, \cdots, X_d)$. Next we show $\tilde{\mathcal{E}}$ is orthogonal monotonic. Assume $(y_1, \cdots, y_d) \in \tilde{\mathcal{E}}$ and $(y_1', \cdots, y_d') \ge (y_1, \cdots, y_d)$, since $F_1^{-1}, \cdots, F_n^{-1}, \Phi$ are all monotonic function, we have $F_1^{-1}(\Phi(y_1)) \le F_1^{-1}(\Phi(y_1')), \cdots, F_n^{-1}(\Phi(y_d)) \le F_n^{-1}(\Phi(y_d'))$. Since $(y_1, \cdots, y_d) \in \tilde{\mathcal{E}}$, we have $(F_1^{-1}(\Phi(y_1)), \cdots, F_d^{-1}(\Phi(y_d))) \in \mathcal{E}$. By the orthogonal monotonicity of $\mathcal{E}$, it implies $(y_1', \cdots, y_d') \in \tilde{\mathcal{E}}$ and hence $\tilde{\mathcal{E}}$ is also orthogonally monotonic.

□

EC.6. Acronyms used in the paper

We collect the acronyms in this paper in Table EC.2

Table EC.2: Acronyms in the paper

Acronym	Full name
IS	importance sampling
CE	cross entropy
AMS	adaptive multi-level splitting
RE	relative error
MH	Metropolis-Hasting
ReLU	rectified linear unit
MIP	mixed integer program
NN	neural network
UB	upper bound
LB	lower bound
NMC	naive Monte Carlo
LL	lazy learner
GMM-k	Gaussian mixture model with k components
AV	autonomous vehicle
LV	lead vehicle

References for the Appendix

Anil C, Lucas J, Grosse R (2019) Sorting out Lipschitz function approximation. Chaudhuri K, Salakhutdinov R, eds. Proc. 36th Internat. Conf. Machine Learn., Proceedings of Machine Learning Research, vol. 97 (PMLR, Cambridge, MA), 291–301.

Cao Y, Gu Q (2019) Tight sample complexity of learning one-hidden-layer convolutional neural networks. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. Advances in Neural Information Processing Systems 32 (Curran Associates, Red Hook, NY), 10612–10622.

Gurobi Optimization, LLC (2023) Gurobi optimizer reference manual, version 10.0. `https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/10.0/refman.pdf` (accessed July 14, 2026).

Harvey N, Liaw C, Mehrabian A (2017) Nearly-tight VC-dimension bounds for piecewise linear neural networks. Kale S, Shamir O, eds. Proc. Conf. Learn. Theory, Proceedings of Machine Learning Research, vol. 65 (PMLR, Cambridge, MA), 1064–1068.

Lu Z, Pu H, Wang F, Hu Z, Wang L (2017) The expressive power of neural networks: A view from the width. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. Advances in Neural Information Processing Systems 30 (Curran Associates, Red Hook, NY), 6231–6239.

Rockafellar RT (1970) Convex Analysis (Princeton University Press).

van der Vaart AW, Wellner JA (1996) Weak Convergence and Empirical Processes, Springer Series in Statistics (Springer, New York).