Over-Conservativeness of Variance-Based Efficiency Criteria and Probabilistic Efficiency in Rare-Event Simulation

Yuanlu Bai
Columbia University, yb2436@columbia.edu

Zhiyuan Huang
Tongji University, huangzy@tongji.edu.cn

Henry Lam
Columbia University, henry.lam@columbia.edu

Ding Zhao
Carnegie Mellon University, dingzhao@cmu.edu

In rare-event simulation, an importance sampling (IS) estimator is regarded as efficient if its relative error, namely the ratio between its standard deviation and mean, is sufficiently controlled. It is widely known that when a rare-event set contains multiple "important regions" encoded by the so-called dominating points, IS needs to account for all of them via mixing to achieve efficiency. We argue that in typical experiments, missing less significant dominating points may not necessarily cause inefficiency, and the traditional analysis recipe could suffer from intrinsic looseness by using relative error, or in turn estimation variance, as an efficiency criterion. We propose a new efficiency notion, which we call probabilistic efficiency, to tighten this gap. In particular, we show that under the standard Gartner-Ellis large deviations regime, an IS that uses only the most significant dominating points is sufficient to attain this efficiency notion. Our finding is especially relevant in high-dimensional settings where the computational effort to locate all dominating points is enormous.

Key words: rare-event simulation, importance sampling, relative error, large deviations, dominating points

1. Introduction
We study the problem of estimating the probabilities of rare events with Monte Carlo simulation, which falls in the domain of rare-event simulation (Bucklew 2004, Juneja and Shahabuddin 2006, Rubino and Tuffin 2009). Traditionally, rare-event simulation is of wide interest to a variety of areas such as queueing systems (Dupuis et al. 2007, Dupuis and Wang 2009, Blanchet and Mandjes 2007, Blanchet et al. 2009, Blanchet and Lam 2014, Kroese and Nicola 1999, Ridder 2009, Sadowsky 1991, Szechtman and Glynn 2002), highly dependable computer systems and communication networks (Tuffin 2004, Lewis and Böhm 1984, Goyal et al. 1992, Carrasco 1992, Shahabuddin 1994, Kesidis et al. 1993), financial risk management (Glasserman 2003, Glasserman and Li 2005, Glasserman et al. 2008) and insurance modeling (Asmussen 1985, Asmussen and Albrecher 2010). More recently,
with the rapid development of intelligent physical systems such as autonomous vehicles and personal assistive robots (Ding et al. 2021, Arief et al. 2021), rare-event simulation is also applied to assess their risks before deployments in public, where the risks are often quantified by the probabilities of violations of certain safety metrics such as crash or injury rate (Huang et al. 2018, O' Kelly et al. 2018, Zhao et al. 2016, 2018). The latter problems typically involve complex AI-driven underlying algorithms that deem the rare-event structures rough or difficult. The current work is motivated from the importance of handling such type of rare-event problems (e.g., the U.S. National Artificial Intelligence Research and Development Strategic Plan (Kratsios 2019) lists “developing effective evaluation methods for AI” as a top priority) and provides a step towards rigorously grounded procedures in this direction.

The starting challenge in rare-event simulation is that, by its own nature, the target rare events seldom occur in the simulation experiment when using crude Monte Carlo. In other words, to achieve an acceptable estimation accuracy relative to the target probability, the required simulation size could be huge in order to obtain sufficient hits on the target events. Statistically, this issue is manifested as a large ratio between the standard deviation (per run) to the mean, known as the relative error, that determines the order of a required sample size. In the large deviations regime where the target probability can depend exponentially on the rarity parameter, this in particular means the required sample size is exponentially large.

To address the inefficiency of crude Monte Carlo, a range of variance reduction techniques have been developed. Among them, importance sampling (IS) (Siegmund 1976) has been broadly applied to improve the efficiency. IS uses an alternative probability measure to generate the simulation samples, and then reweighs the outputs via likelihood ratio to guarantee unbiasedness. The goal is that by using this alternate estimator than simply counting the frequency of hits in crude Monte Carlo, one can achieve a small relative error with a much smaller sample size.

To this end, it is also widely known that IS is a “delicate” technique, in the sense that the IS probability measure needs to be carefully chosen in order to achieve a small relative error. In the typical large deviations setting, the suggestion is to tilt the probability measure to the “important region”. The delicacy appears when there are more than one important regions, in which case all of them need to be accounted for. More specifically, in the light-tailed regime, these important regions are guided precisely by the so-called dominating points, which capture the most likely scenario in a local region of the rare event. Despite the tempting approach to simply shift the distribution center to the globally most likely scenario, it is well established that if not all the dominating points are included in the IS mixture distribution, then the resulting estimator may no longer be efficient in terms of the relative error (see, for instance, the seminal work Glasserman and Wang (1997)).
Our main goal in this paper is to argue that, in potentially many light-tailed problems, the inclusion of all the dominating points in an IS could be unnecessary. Our study is motivated from high-dimensional settings where finding all dominating points could be computationally expensive or even prohibitive, yet these problems may arise in recent safety-critical applications (e.g., Arief et al. (2021), Webb et al. (2018), Bai et al. (2022)).

To intuitively explain the unnecessity, let us first drill into why all the dominating points are arguably needed in the literature in the first place. Imagine that a rare event set $E$ comprises two disjoint “important” regions, say $E_1$ and $E_2$, and the dominating points are correspondingly $a_1$ and $a_2$, which are sufficiently “far away” from each other, and $a_1$ has say a higher density than $a_2$. Roughly speaking, if an IS scheme only focuses on $E_1$ and tilts the distribution center towards $a_1$, then there is a small chance that the sample from this IS distribution hits $E_2$, so that the contribution from this sample in the ultimate estimator is non-zero and, moreover, may constitute a large likelihood ratio and consequently elicit a large variance. This unfortunate event of falling into a secondary important region is a source of inefficiency according to the relative error criterion.

Now let’s take a step back and think about the following: How likely does the “unfortunate” event above occur? In a typical Monte Carlo experiment, we argue – and we will see clearly in experiments – that this could be very unlikely, to the extent that we shouldn’t be worried at all with a reasonable simulation size. Yet, according to the relative error criterion, it seems necessary to worry about this, because it contributes to the variance of the estimator in each single run. This points to that using variance to measure efficiency in rare-event simulation could be too loose to begin with. This variance measure, in turn, comes from the Markov inequality that converts relative error into a sufficient relative closeness between the estimate and target probability with high confidence. In other words, this Markov inequality itself could be the source of looseness.

This motivates us to propose what we call probabilistic efficiency. Different from all the efficiency criteria in the literature, including asymptotic efficiency (also known as asymptotic optimality or logarithmic efficiency) and bounded relative error (Juneja and Shahabuddin 2006, L’Ecuyer et al. 2010), probabilistic efficiency does not use relative error. Instead, it is a criterion on achieving relative closeness directly. A distinctive element in probabilistic efficiency is the control on the simulation size itself, that we only allow it to grow moderately with the underlying rarity parameter. This moderate simulation size, which is often the only feasible option in experiments, suppresses the occurrence of the unfortunate event of falling into a secondary important region. This way, while the variance could blow up, the high-confidence closeness between the estimate and target probability could still be retained.

With this new framework, we show that under standard assumptions in the widely used Gartner-Ellis large deviations regime, an IS that uses only the most significant dominating points is sufficient
to ensure probabilistic efficiency. The Gartner-Ellis regime has been used across different applications such as queueing (Ridder 2009, Szechtmam and Glynn 2002, Blanchet and Lam 2014), communication systems (Smith et al. 1997, Chen et al. 1993) and finance (Zhang et al. 2009, Blanchet and Glynn 2009, Glasserman and Li 2005, Glasserman et al. 2008). Our results thus stipulate that in all these problems, in order to obtain a good estimate relative to the ground-truth rare-event probability, we only need to exponentially tilt to the most significant dominating point when there is only one such point, without the use of any mixture. This is a sharp contrast to the established IS recipe. Moreover, this makes the construction of IS in closer line with the large deviations theory that governs the rare-event probability asymptotic. More specifically, in large deviations, the most significant dominating point coincides with the minimizer of the so-called rate function that controls the exponential decay rate of the probability. Our theory thus postulates that to attain probabilistic efficiency, it suffices to consider only this rate function minimizer when constructing the IS.

We close this introduction with further discussions of our study in relation to existing works. First, we contrast our probabilistic efficiency with the notion of “well-estimated” in the dependability literature (Definition 3 in Tuffin (2004)). The latter asserts that all the paths which contribute most significantly to the target quantity need to occur sufficiently likely under the IS distribution. This notion is intuitively similar to the scheme of tilting the most significant dominating points which we suggest, but there are substantial differences in the motivations, implications, and also some technicality. First is that Tuffin (2004) uses well-estimated as a sufficient condition to guarantee the efficiency in estimating the variance, not only the rare-event probability itself. In contrast, our work proposes to remove variance as an efficiency criterion, a motive that is essentially orthogonal to Tuffin (2004), and we propose the tilting of only the most significant dominating points as being sufficient to ensure our new criterion of probabilistic efficiency. Our key message is that the existing proposal that suggests using all dominating points, not only the most significant ones, is an over-conservative approach and hence worth our remedy. Regarding technicality, Tuffin (2004) focuses on dependable systems where the target probability decays polynomially instead of exponentially in the rarity parameter as in our setting. Under his framework, there are only countably many points in the sample space, and thus the target probability or variance can be more readily decomposed (i.e., one could explicitly quantify the “contribution” of each point), while in our possibly continuous setting the contribution of each point is less easy to determine. Lastly, well-estimated requires that all the significant paths are no longer rare (i.e., of order $O(1)$) under IS, while we require a less stringent subexponential decay.

Second, we caution that probabilistic efficiency is not meant to replace existing variance-based efficiency criteria, but rather to complement them especially in situations where identifying all
dominating points is infeasible due to problem complexity. In problems where the latter is not an issue, it remains “safer” to use existing criteria, as probabilistic efficiency relies on a more subtle sample size requirement, namely that it is not exponentially large. While this condition is reasonable for realistic problems, it would warrant future experimental diagnostics to detect violations of such a condition. Third, we note that our variance-free approach can potentially be adapted to rare-event estimation problems to streamline IS construction beyond the considered light-tailed large deviations regime. However, for some of these problems (e.g., heavy-tailed problems; Blanchet and Glynn 2008, Hult and Svensson 2012, Blanchet et al. 2012, Chen et al. 2019), the efficiency gain appears less dramatic than ours which possesses an exponential speed-up. Moreover, the Gartner-Ellis paradigm that we consider in this paper is arguably the most widely used and forms the basis of analysis for many rare-event problems.

In the following, we first introduce in more detail the background of rare-event simulation and the established efficiency criteria in the literature, all of which involve estimation variance or relative error (Section 2). Then we show several motivating numerical examples to illustrate how excluding some dominating points in IS appear to give similar and sometimes even better performances than including all these points, the latter suggested predominantly in the literature (Section 3). This motivates our new notion of probabilistic efficiency to explain the observed numerical phenomena (Section 4) and the analysis of efficiency guarantees using this new notion (Section 5). We then show further numerical results to validate our theory and performances of our estimators (Section 6). Finally, we give some cautionary notes about probabilistic efficiency which involve the risk of under-estimation, and suggest some future directions (Section 7).

2. Problem Setting and Existing Framework
We consider an indexed family of rare events \( \{A_\gamma\}_\gamma \), where \( \gamma \) denotes a “rarity parameter” such that as \( \gamma \to \infty \), the event \( A_\gamma \) becomes rarer so that \( P(A_\gamma) \to 0 \). Our goal is to estimate \( p = p(\gamma) := P(A_\gamma) \) using Monte Carlo simulation. Here, the index \( \gamma \) is introduced for modeling purpose so that we can speak of asymptotic rate, which is customary in the rare-event simulation literature.

2.1. Crude Monte Carlo and Relative Error
To motivate the various notions that we would discuss momentarily, let us consider using crude Monte Carlo to estimate \( p \). This means we utilize the unbiased estimator \( Z = I_{A_\gamma} \) for \( p \), where \( I_{A_\gamma} \) denotes the indicator variable on the event \( A_\gamma \). Suppose we generate the output \( Z \) independently \( n \) times, and construct \( \hat{p} \) as their sample mean. Intuitively, when \( p \) is tiny, this estimator \( \hat{p} \) is most likely zero unless \( n \) is a huge number, since a long trial length is needed to land at the rare event \( A_\gamma \).
To describe the above challenge mathematically, we consider the following criterion. For a given tolerance level \( \varepsilon > 0 \) (e.g., 5\%), we would like an estimator \( \hat{p} \) to satisfy

\[
P(|\hat{p} - p| > \delta p) \leq \varepsilon
\]

for a certain \( 0 < \delta < 1 \) when using a simulation size \( n \). In (1), the closeness between \( \hat{p} \) and \( p \), which represents the error of \( \hat{p} \) in estimating \( p \), is measured relative to the magnitude of \( p \) itself. This is because for tiny \( p \), the estimation is only meaningful if the error is small enough relative to this tiny quantity. Linking to crude Monte Carlo, outputting merely \( \hat{p} = 0 \), as likely to happen thereby, would be viewed as incurring a substantial error, i.e., lying in the event \( |\hat{p} - p| > \delta p \) in (1). Put in another way, we argue that a huge sample size \( n \) is needed for crude Monte Carlo to attain (1). Note that by Chebyshev’s inequality, we get that for any \( 0 < \delta < 1 \),

\[
P(|\hat{p} - p| > \delta p) \leq \frac{Var(Z)}{n\delta^2 p^2}.
\]

Hence, \( n \geq \frac{Var(Z)}{\varepsilon^2 p^2} \) implies that \( P(|\hat{p} - p| > \delta p) \leq \varepsilon \) for \( \varepsilon > 0 \). This means that \( \frac{Var(Z)}{\varepsilon^2 p^2} \) is a sufficient size for \( n \) to achieve (1). This quantity depends on the ratio between the standard deviation \( \sqrt{Var(Z)} \) and the mean \( p \), which is known as the relative error. Here, for crude Monte Carlo the relative error is \( \sqrt{\frac{Var(Z)}{p}} = \sqrt{\frac{p(1-p)}{p^2}} = \sqrt{\frac{1-p}{p}} \) which blows up as \( p \to 0 \). Consequently, the sufficient size for \( n \) also blows up as \( p \to 0 \). In particular, if \( p \) decays exponentially in \( \gamma \) – a typical scaling in large deviations, then the required \( n \) to achieve (1) also scales exponentially.

2.2. Importance Sampling and Asymptotic Efficiency

The above challenge motivates variance reduction techniques to reduce the sample size requirement. Among the most popular is IS. In this approach, we generate samples from an alternate measure \( \tilde{P} \) where \( \tilde{P} \) satisfies \( PI_A, \ll \tilde{P} \) (i.e., \( PI_A, \) is absolutely continuous with respect to \( \tilde{P} \)), and use \( Z = I_A, \frac{dP}{d\tilde{P}} \) as an unbiased output for \( p \), where \( \frac{dP}{d\tilde{P}} \) is the Radon-Nikodym derivative, or the so-called likelihood ratio, between \( P \) and \( \tilde{P} \). Though this output is always unbiased thanks to the likelihood ratio adjustment, the performance of the IS estimator in terms of variability heavily depends on the choice of the IS probability measure \( \tilde{P} \). In the literature, several efficiency criteria for IS estimators have been developed. A common criterion is asymptotic efficiency (Asmussen and Glynn 2007, Heidelberger 1995, Juneja and Shahabuddin 2006):

**Definition 1 (Asymptotic Efficiency).** The IS estimator \( Z = I_A, \frac{dP}{d\tilde{P}} \) under \( \tilde{P} \) is said to achieve asymptotic efficiency if \( \lim_{\gamma \to \infty} \frac{\log(E(Z^2))}{\log p} = 2 \) where \( E(\cdot) \) denotes the expectation under \( \tilde{P} \).

For functions \( f, g : \mathbb{R} \to \mathbb{R} \), we say \( g(\gamma) \) is subexponential in \( f(\gamma) \) as \( \gamma \to \infty \) if \( \lim_{\gamma \to \infty} \frac{\log g(\gamma)}{f(\gamma)} = 0 \), i.e. \( g(\gamma) = \exp(f(\gamma) o(1)) \). For functions \( f, g, h \), clearly if \( \lim_{\gamma \to \infty} \frac{\log g(\gamma)}{f(\gamma)} = \lim_{\gamma \to \infty} \frac{\log h(\gamma)}{g(\gamma)} \), then \( \lim_{\gamma \to \infty} \frac{\log g(\gamma)/h(\gamma)}{f(\gamma)} = 0 \) and hence \( g(\gamma)/h(\gamma) \) is subexponential in \( f(\gamma) \) as \( \gamma \to \infty \). By taking \( g(\gamma) \) =
\( \hat{E}(Z^2), h(\gamma) = p^2 \) and \( f(\gamma) = \log p \), we see that the condition in Definition 1 is equivalent to the condition that \( \hat{E}(Z^2)/p^2 \) (or \( \text{Var}(Z)/p^2 \)) is subexponential in \(-\log p\) as \( \gamma \to \infty \).

From (2) and its subsequent discussion, asymptotic efficiency implies that the required simulation size \( n \) to attain a prefixed relative error grows only subexponentially in \(-\log p\). As a stronger requirement, \( Z \) is said to have a bounded relative error if \( \limsup_{\gamma \to \infty} \frac{\text{Var}(Z)}{p^2} < \infty \), which implies that the required simulation size remains bounded no matter how small \( p \) is. The criterion of bounded relative error is sometimes too strict to achieve, so we focus on asymptotic efficiency in this paper. More efficiency criteria could be found in Juneja and Shahabuddin (2006), L’Ecuyer et al. (2010), Blanchet and Lam (2012).

### 2.3. Large Deviations and Dominating Points

In the large deviations setting, the classical notion of *dominating points* is used to guarantee asymptotic efficiency of IS (Sadowsky and Bucklew 1990). To explain, let us first recall the so-called *rate function* in the large deviations theory which, intuitively speaking, measures the likelihood of hitting each point on an exponential scale. More specifically, we consider the standard Gartner-Ellis regime as follows (Dembo and Zeitouni 2009, Bucklew 2004). Without loss of generality consider \( \gamma > 0 \). Suppose that \( \mathcal{A}_\gamma = \{ \frac{1}{\gamma} X_\gamma \in \mathcal{E} \} \) where \( \{ X_\gamma \}_\gamma \) are \( \mathbb{R}^d \)-valued random variables and \( \mathcal{E} \) is a fixed Borel set in \( \mathbb{R}^d \). We define \( \mu_\gamma(x) = \frac{1}{\gamma} \log \mathbb{E}(e^{x^\top X_\gamma}), x \in \mathbb{R}^d \) as the scaled logarithmic moment generating function. We denote \( D(f) = \{ x : f(x) < \infty \} \) as the domain of a function \( f \). With these, we assume the following:

**Assumption 1.** \( \mu_\gamma(x) \) satisfies the following conditions:

1. \( \mu(x) = \lim_{\gamma \to \infty} \mu_\gamma(x) \) exists for any \( x \in \mathbb{R}^d \), where we allow \( \infty \) both as a limit value and as an element of the sequence \( \{ \mu_\gamma(x) \} \);

2. \( 0 \in D(\mu)^o \);

3. \( \mu \) is essentially smooth, i.e., \( D(\mu)^o \) is non-empty, \( \mu \) is differentiable everywhere in \( D(\mu)^o \) and \( \mu \) is steep.

Then we define the rate function \( I(y) = \sup_{x \in \mathbb{R}^d} \{ x^\top y - \mu(x) \}, y \in \mathbb{R}^d \) as the Legendre transform of \( \mu \). For any set \( \mathcal{E} \subset \mathbb{R}^d \), we denote \( I(\mathcal{E}) = \inf_{y \in \mathcal{E}} I(y) \). We make the following assumptions for the set \( \mathcal{E} \):

**Assumption 2.** \( \mathcal{E} \subset \mathbb{R}^d \) is a Borel set such that \( \overline{\mathcal{E}} = \overline{\mathcal{E}^c} \), \( \mathcal{E}^o \cap D(I)^o \neq \emptyset \) and \( I(\mathcal{E}) > 0 \).

Under these assumptions, we have the following result (e.g., adapted from Dembo and Zeitouni (2009) Theorem 3.3.6):

\( \overline{\mathcal{E}} = \overline{\mathcal{E}^c} \), \( \mathcal{E}^o \cap D(I)^o \neq \emptyset \) and \( I(\mathcal{E}) > 0 \).
**Theorem 1 (Gartner-Ellis Theorem).** Suppose that Assumption 1 holds. For any Borel set $\mathcal{E} \subset \mathbb{R}^d$, we have that

$$-I(\mathcal{E}^c) \leq \liminf_{\gamma \to \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_\gamma \in \mathcal{E}\right) \leq \limsup_{\gamma \to \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_\gamma \in \mathcal{E}\right) \leq -I(\mathcal{E}).$$

If additionally Assumption 2 holds, then

$$\lim_{\gamma \to \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_\gamma \in \mathcal{E}\right) = -I(\mathcal{E}).$$

Assumptions 1 and 2 are standard light-tailed conditions on $A_\gamma$, which guarantee the considered probability $P(A_\gamma)$ to decay exponentially in $\gamma$ with decay rate $I(\mathcal{E})$. The rate function $I(y)$ can be viewed as a measurement on the likelihood of hitting $y$ in the exponential scale. The most likely point to hit among $\mathcal{E}$ is hence given by the minimizer of $I(y)$ over $\mathcal{E}$, resulting in the overall exponential decay rate $I(\mathcal{E})$. Note that, by Theorem 1, we have $-\log p = \Theta(\gamma)$ as $\gamma \to \infty$, and hence subexponential (or exponential) in $-\log p$ is equivalent to subexponential (or exponential) in $\gamma$.

Now we present the concept of dominating points and sets:

**Definition 2 (Dominating Set).** Suppose that Assumptions 1 and 2 hold. We call $A = \{a_1, \ldots, a_r\} \subset \partial \mathcal{E}$ a dominating set for $\mathcal{E}$ if

1. For each $i$, $a_i \in \mathcal{D}(I)^\circ$ and there exists a unique $s_{a_i} \in \mathbb{R}^d$ such that $\nabla \mu(s_{a_i}) = a_i$;
2. $\mathcal{E} \subset \bigcup_{i=1}^m \{x \in \mathbb{R}^d : s_{a_i}^\top (x - a_i) \geq 0\}$;
3. For any $i$, $A \setminus \{a_i\}$ does not satisfy Condition 2.

We call any point in $A$ a dominating point. For two dominating points $a$ and $a'$, we say $a$ is more significant than $a'$ if $I(a) < I(a')$.

![Figure 1](image.png)

Figure 1 Illustration of rare-event set $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$ and dominating points.

Dominating points can be understood as the "local minimizers" of the rate function $I$ in the sense that $a_i$ is the minimizer of $I$ in $\mathcal{E} \cap \{x \in \mathbb{R}^d : s_{a_i}^\top (x - a_i) \geq 0\}$. To understand this, first, Condition 1 in Definition 2 stipulates that $s_{a_i}$ is the gradient of $I(x)$ at the point $a_i$. Then, in
Condition 2, $a_i$ can be seen as the minimizer of $I(x)$ over the set $\{x \in \mathbb{R}^d : s_{a_i}^T (x - a_i) \geq 0\}$, a fact that follows from the first-order optimality condition in convex function minimization. Thus, Condition 2 stipulates that any point in the rare-event set $E$ has an $I$ value not less than one of the $a_i$'s. Geometrically, any points in the set $E$ must lie in the half-space tangentially cut by one of the $a_i$'s (i.e., the “backyard” of the $a_i$). Figure 1 is an illustration of a rare-event set $E = E_1 \cup E_2$ and the dominating set $\{a_1, a_2\}$. Here, $a_1$ is the global minimum rate point in $E$, but $E_2$ is not covered by $\{x : s_{a_1}^T (x - a_1) \geq 0\}$, so $a_2$ is included in the dominating set. Finally, we note that Condition 3 in Definition 2 enforces the dominating set to be the minimal set of points such that the geometric properties in Conditions 1 and 2 are satisfied.

Note that dominating points may not be local minimizers of the rate function $I$ in $E$ (even though they are minimizers in $E \cap \{x \in \mathbb{R}^d : s_{a_i}^T (x - a_i) \geq 0\}$ as discussed above). Nonetheless, the most significant dominating points are indeed global minimizers of $I$ in $E$. This is presented in the following theorem:

**Theorem 2.** Suppose that Assumptions 1 and 2 hold, and $A = \{a_1, \ldots, a_r\}$ is a dominating set for $E$. Then $I(E) = \min_{i=1,\ldots,r} I(a_i)$. That is to say, the most significant dominating points are global minimizers of $I$ in $E$.

We should also point out that dominating set defined according to Definition 2 may not be unique. Advantageously, the theory and estimators we present will flexibly apply to any such dominating set.

### 2.4. Asymptotically (In)efficient Importance Samplers

We are now ready to describe the main message of this section, which is the established recipe in constructing efficient IS. The standard proposal is to use a mixture of exponentially tilted distributions, where each exponential tilting is with respect to each dominating point. In particular, suppose that $A = \{a_1, \ldots, a_r\}$ is a dominating set. Then the IS distribution is $\tilde{P}$ such that

$$
\frac{d\tilde{P}}{dP} = \sum_{i=1}^{r} \alpha_i e^{s_{a_i}^T \gamma - \gamma \mu_{\gamma}(s_{a_i})} \tag{3}
$$

with $\sum_{i=1}^{r} \alpha_i = 1, \alpha_i > 0, \forall i$. Here, $e^{s_{a_i}^T \gamma - \gamma \mu_{\gamma}(s_{a_i})} dP$ is the exponential tilting towards the dominating point $a_i$ and $\alpha_i$'s are the mixing weights. The IS (3) is well known to be asymptotically efficient:

**Proposition 1 (Mixture IS is asymptotically efficient).** Suppose Assumptions 1 and 2 hold, and the dominating set has finite cardinality, i.e., $r < \infty$. Then the IS distribution (3) with any fixed $\alpha_i$'s is asymptotically efficient.
While the proof of Proposition 1 is standard, we include it in the Appendix for self-containedness. Here, we describe the key intuition in justifying the necessity of mixture. First, the likelihood ratio in the considered mixture IS is

$$L = \frac{dP}{dP} = \frac{1}{\sum_{i=1}^{\infty} \alpha_i e^{s_{a_i}^T X_\gamma - \gamma \mu_\gamma(s_{a_i})}}$$

and it satisfies that for any $i$,

$$L \leq \frac{1}{\alpha_i e^{s_{a_i}^T X_\gamma - \gamma \mu_\gamma(s_{a_i})}} = \frac{1}{\alpha_i} e^{-s_{a_i}^T (X_\gamma - \gamma a_i) - \gamma (s_{a_i}^T a_i - \mu_\gamma(s_{a_i}))}. \tag{4}$$

In the exponent in the rightmost expression of (4), the second term $\gamma(s_{a_i}^T a_i - \mu_\gamma(s_{a_i}))$ is approximately $\gamma I(a_i)$, and the first term is the “overshoot” of the sampled $X_\gamma$ compared to the dominating point $a_i$. That is, if $X_\gamma$ is in the “backyard” of $a_i$, then this term $s_{a_i}^T (X_\gamma - \gamma a_i)$ is approximately 0. The definition of dominating set, especially Condition 2 in Definition 2, guarantees any $(1/\gamma)X_\gamma$ in $E$ must have $s_{a_i}^T (X_\gamma - \gamma a_i) \geq 0$ for at least one of the $i$'s. Thus, by decomposing the second moment of $Z = I_{a_i} L$ according to the backyards of $a_i$'s, we can ensure that the magnitude of the likelihood ratio, when $X_\gamma$ lies inside the rare event set, is properly controlled. More precisely, write $E = \bigcup_{i=1}^{\infty} E_i$ where each $E_i \subset \{x : s_{a_i}^T (x - a_i) \geq 0\}$. Then the second moment of the IS satisfies

$$\bar{E} \left[ I \left( \frac{1}{\gamma} X_\gamma \in E \right) L^2 \right] \leq \sum_{i=1}^{\infty} \bar{E} \left[ I \left( \frac{1}{\gamma} X_\gamma \in E_i \right) L^2 \right] \leq \sum_{i=1}^{\infty} \bar{E} \left[ I \left( \frac{1}{\gamma} X_\gamma \in E_i \right) \frac{1}{\alpha_i} e^{-2\gamma(s_{a_i}^T a_i - \mu_\gamma(s_{a_i}))} \right]$$

which is approximately bounded by $\sum_{i=1}^{\infty} e^{-2\gamma I(a_i)} / \alpha_i^2$ and hence $e^{-2\gamma I(E)}$ in the exponential scale, thus verifying asymptotic efficiency.

On the other hand, if we miss some dominating points in the construction of the mixture IS, then asymptotic efficiency may fail to be attained. Below we give a simple example to demonstrate this.

**Proposition 2 (Missed dominating point leads to violation of asymptotic efficiency).**

*Suppose that we want to estimate $p = P(\frac{1}{\gamma} X_\gamma \in (-\infty, -2] \cup [1, \infty))$ where $X_\gamma \sim N(0, \gamma)$ under $P$. If the IS distribution is chosen as $X_\gamma \sim N(\gamma, \gamma)$, then $\bar{E}(Z^2)/p^2 = \Theta(\sqrt{\gamma e^{3\gamma/2}})$ grows exponentially in $-\log p = \Theta(\gamma)$, and hence $Z$ is not asymptotically efficient by definition.*

In this example, the dominating points are 1 and $-2$, and $N(\gamma, \gamma)$ is the exponential tilt towards the first dominating point 1 (for Gaussian distribution, exponential tilting amounts to a mean shift). Here, by considering only this point, it is possible that a generated $X_\gamma$ satisfies $(1/\gamma)X_\gamma \in (-\infty, -2]$ while the overshoot $s_{a_i}^T (X_\gamma - \gamma 1)$, as explained for (4), takes a very negative value. This scenario contributes significantly to the overall variance and ultimately violates asymptotic efficiency.

Our main insight in this paper is a rebuke of the above viewpoint. More specifically, we argue that missing inferior dominating point, such as the example in Proposition 2, can still result in a
good IS according to our beginning criterion (1). A core ingredient of this assertion is to question the use of asymptotic efficiency, or more generally variance-based efficiency criteria. Before delving into the theory, let us first present some numerical results to shed light on how much difference it makes to use different numbers of dominating points in the IS mixture. This is the focus of our next section.

3. Motivating Experimental Results

We run three numerical examples to demonstrate that missing dominating points in IS construction, while provably leads to asymptotic inefficiency, could perform well empirically. This thus suggests an inadequacy in using asymptotic efficiency, or more generally variance-based criteria, to measure the performances of rare-event estimators. Besides, the computationally demanding example in Section 3.3 justifies the motivation why we seek to reduce the number of used dominating points in the IS mixture.

3.1. Large Deviations of an I.I.D. Sum

We consider the problem of estimating the tail probability involving a sum of random variables, where \( Y_1, Y_2, \ldots \) are i.i.d and we are interested in

\[
P(|S_m| \geq am),
\]

where \( S_m = \sum_{i=1}^{m} Y_i \). We consider \( m \) as the rarity parameter \( \gamma \) presented in Section 2. Using the notation in the Gartner-Ellis regime, we have \( X_\gamma = S_m \) and \( \mathcal{E} = (-\infty, -a] \cup [a, \infty) \). Then \( \mu(x) = \log E(e^{xY_1}) \) and we suppose Assumption 1 is satisfied. By Theorem 1, when \( |EY_1| < a \), if \( s_a \) and \( s_{-a} \) satisfy \( \nabla \mu(s_a) = a \) and \( \nabla \mu(s_{-a}) = -a \), then we have

\[
-\lim_{m \to \infty} \frac{1}{m} \log P(S_m \geq am) = s_a a - \mu(s_a) = I(a)
\]

and

\[
-\lim_{m \to \infty} \frac{1}{m} \log P(S_m \leq -am) = -s_{-a} a + \mu(s_{-a}) = I(-a).
\]

For this problem, Glasserman and Wang (1997) Section 3 provides two estimators, \( \hat{\alpha}(m) \) and \( \hat{\beta}(m) \). Specifically, we have \( \hat{\alpha}(m) = \exp(-s_a S_m + m\mu(s_a))I_{\{S_m \geq am\}} \) with samples of \( Y_i \) generated from exponentially tilted distribution using \( s_a \). The estimator \( \hat{\beta}(m) = \exp(-s_a S_m + m\mu(s_a))I_{\{S_m \geq am\}} + \exp(-s_{-a} S'_m + m\mu(s_{-a}))I_{\{S'_m \leq -am\}} \) with \( S_m \) and \( S'_m \) constructed from independent sequences of i.i.d. \( Y_i \)'s and \( Y'_i \)'s generated from exponentially tilted distributions using \( s_a \) and \( s_{-a} \) respectively. That is, \( \hat{\beta}(m) \) attempts to estimate \( P(S_m \geq am) \) and \( P(S_m \leq -am) \) separately using different IS samples and sum up these estimates. Here, \( \hat{\alpha}(m) \) only uses one dominating point whereas \( \hat{\beta}(m) \) uses both points (note that even though \( \hat{\beta}(m) \) does not use the mixture IS
scheme in (3), the idea is similar in that it accounts for both dominating points. In our experiment, we follow Glasserman and Wang (1997) to set \( Y_i = A - B \) with \( A \sim N(1.5, 1) \), \( B \sim \text{Exp}(1) \) and \( A, B \) independent, and \( a = 1.5 \) (in this case, \( s_a = (\sqrt{5} - 1)/2, s_{-a} = -2 + \sqrt{2} \), and \( I(a) \approx 0.2902 \), \( I(-a) \approx 0.7044 \)).

We run numerical experiments with \( m = 10, 30, 50 \) and 100. The results using \( 10^4 \) samples are shown in Table 1. By comparing the numbers in the second and third rows, we observe that \( \hat{\alpha}(m) \) and \( \hat{\beta}(m) \) have very similar empirical performances. However, note that:

**Proposition 3.** Under the problem specification above, \( \hat{\beta}(m) \) is asymptotically efficient while \( \hat{\alpha}(m) \) is not. In fact, \( \tilde{E}(\hat{\alpha}(m)^2) \to 0 \) as \( m \to \infty \) where \( \tilde{E} \) denotes the expectation under the exponential tilting towards \( a \).

In view of Proposition 3, \( \hat{\alpha}(m) \) is arguably a very poor estimator as it bears an exploding variance. We therefore see an apparent discrepancy between empirical performances and theoretical guidance – The theoretically bad variance does not result in poor empirical performances. Proposition 3 is proved in Glasserman and Wang (1997), where the asymptotic efficiency of \( \hat{\beta}(m) \) follows from their Proposition 1, while the variance behavior of \( \hat{\alpha}(m) \) appears in their Theorem 1.

| \( m \)  | 10          | 30          | 50          | 100         |
|--------|-------------|-------------|-------------|-------------|
| \( \hat{\alpha}(m) \) | 8.22(±0.26) × 10⁻³ | 1.60(±0.07) × 10⁻⁵ | 3.77(±0.18) × 10⁻⁸ | 1.34(±0.08) × 10⁻¹⁴ |
| \( \hat{\beta}(m) \)  | 8.29(±0.26) × 10⁻³ | 1.60(±0.07) × 10⁻⁵ | 3.77(±0.18) × 10⁻⁸ | 1.34(±0.08) × 10⁻¹⁴ |

### 3.2. Overshoot Probability of Random Walk

We consider the problem of estimating the overshoot probability of the finite-horizon maximum of a random walk. We define the probability of interest as

\[
p = P\left( \max_{m=1,\ldots,d} S_m \geq a \right),
\]

where \( S_m = \sum_{i=1}^{m} Y_i \) and \( Y_i \)'s are Gaussian distributed with mean 0, standard deviation \( \sigma \), and pairwise correlation \(-0.02\), i.e., \( \text{corr}(Y_i, Y_j) = -0.02 \) for any \( i, j \in \{1, \ldots, d\} \) with \( i \neq j \). Suppose that the rarity parameter is \( \gamma = 1/\sigma^2 \to \infty \). We note that we can reformulate this target rare event as \( \left\{ \frac{1}{\gamma} X_\gamma \in \bigcup_{m=1}^{d} \mathcal{H}_m \right\} \) where \( X_\gamma = \gamma(Y_1, \ldots, Y_d)^\top \) and \( \mathcal{H}_m = \{ x \in \mathbb{R}^d : \sum_{i=1}^{m} x_i \geq a \} \) with \( x_i \) denoting the \( i \)th element in \( x \). This decomposition allows us to construct an IS estimator using the \( d \) dominating points corresponding to each half-space \( \mathcal{H}_m \). More specifically, in this example, \( I(y) = \frac{1}{2} y^\top \Sigma^{-1} y \), and the dominating points, ranking from the most to the least significant (i.e., increasing rate function value), are

\[
a_1 = \frac{a_1 \Sigma d}{\epsilon_d \Sigma e_d}, \quad a_2 = \frac{a_2 \Sigma d-1}{\epsilon_d-1 \Sigma e_{d-1}}, \ldots, \quad a_d = \frac{a_d \Sigma 1}{\epsilon_1 \Sigma e_1},
\]

where \( e_i \) denotes the vector with 1 in the first \( i \) elements and 0 for the rest.
In our experiments, we fix $a$ and vary $\sigma$ for different rarity levels. In addition, we set $d = 10$. We generate $10^4$ samples from IS distributions using a varying, partial list of dominating points. That is, we choose the IS distributions for $X_\gamma$ as $\frac{1}{k} \sum_{i=1}^{k} \phi(x; \gamma a_i, \gamma \Sigma)$, where $\phi(x; a, b)$ denotes the Gaussian density with mean $a$ and variance $b$, for $k = 1, \ldots, d$. The performances of these IS estimators are shown in Table 2. We observe that these estimators with different numbers of dominating points all perform similarly. In particular, we present the cases with $\sigma = 0.2$ and $\sigma = 0.3$ respectively in Figure 2. We observe that in both cases the performances of the IS estimators are almost independent of the number of used dominating points, with the probability estimates all comparable while using more dominating points slightly increases the CI width.

On the other hand, Proposition 1 implies that the IS using all dominating points is asymptotically efficient while we have:

**Proposition 4.** Under the problem specification above, the IS estimator that exponentially tilts towards the most significant dominating point $a_1$, i.e., $X_\gamma$ distributed as $N(\gamma a_1, \gamma \Sigma)$, is not asymptotically efficient.

Thus, like in Section 3.1, there appears a mismatch between theoretical guidance and empirical observation. The asymptotic inefficiency of simple exponential tilting towards only the most significant dominating point does not result in a poor experimental performance.

![Figure 2](image-url)  
**Figure 2** Simulation results for the random walk experiment. Point estimates and CI widths for IS estimators using different numbers of dominating points.

### 3.3. Robustness Assessment for an MNIST Classification Model

We consider a rare-event probability estimation problem from an image classification task. Our goal is to estimate the probability of misclassification when the input of a prediction model is perturbed by tiny noise. This probability estimate is of interest as a robustness measure of the
 prediction model (Webb et al. 2018). More specifically, suppose that the prediction model \( g \) is able to predict the label of input \( x_0 \), i.e. \( g(x_0) = c \) where \( c \) is the true label of \( x_0 \). Then \( P(g(x_0 + \varepsilon) \neq c) \) where \( \varepsilon \) is a random perturbation can be used to measure the robustness of \( g \).

In particular, we consider the classification problem on MNIST dataset which contains 70,000 images of handwritten digits and each image consists of 28 \( \times \) 28 pixels. We train a 2-ReLU-layer neural network with 20 neurons in each layer using 60,000 training data, which achieves approximately 95\% of testing data accuracy in predicting the digits. We perturb a fixed input (that is correctly predicted) with a Gaussian noise with mean 0 and standard deviation \( \sigma \) on each of the 784 dimensions to assess the robustness of the prediction. Note that the rarity of this problem is determined by the value of \( \sigma \), and we let the rarity parameter \( \gamma = 1/\sigma^2 \to \infty \). The target rare event can be reformulated as \( \left\{ \frac{1}{\sigma} X_\gamma \in \{ x : g(x_0 + x) \neq c \} \right\} \) where \( X_\gamma = \gamma \varepsilon \).

We apply mixtures of exponential tiltings as IS estimators for this problem, namely by considering the IS distribution \( \frac{1}{k} \sum_{i=1}^k \phi(x; \gamma a_i, \gamma I) \), where \( \phi(x; a, b) \) denotes the Gaussian density with mean \( a \) and variance \( b \) as in Section 3.2, for \( k = 1, 2, \ldots \). Here \( a_i, i = 1, 2, \ldots \) denote the dominating points. In order to compute these points, we apply the scheme introduced in Huang et al. (2018) and Bai et al. (2022), which sequentially searches for dominating points by minimizing the rate function on the rare-event set that excludes the half-spaces cut from previous more significant dominating points. In the Gaussian case with piecewise linear rare-event set boundary as in our current example, each iteration amounts to finding the highest-density point on a piecewise-linear-boundary set, which can be conducted using mixed integer programming (see Algorithm 1 in the Appendix). Due to the high dimensionality of the input space and the complexity of the neural network predictor, the number of dominating points in this problem is huge. We implemented this sequential searching algorithm and it took a week to find the first 100 dominating points. Since we stopped the algorithm prematurely, the actual number of dominating points can be much larger.

We run IS distributions with different numbers of dominating points (ranging from 1 to 41) and

| \( \sigma \) | 0.2 | 0.22 | 0.24 | 0.26 | 0.28 | 0.3 |
|---|---|---|---|---|---|---|
| \# | prob (with CI) | prob (with CI) | prob (with CI) | prob (with CI) | prob (with CI) | prob (with CI) |
| 1 | 9.15(±0.52) \times 10^{-8} | 1.24(±0.19) \times 10^{-6} | 7.96(±0.85) \times 10^{-8} | 3.52(±0.29) \times 10^{-6} | 1.15(±0.08) \times 10^{-6} | 3.11(±0.20) \times 10^{-6} |
| 2 | 9.63(±0.52) \times 10^{-8} | 1.24(±0.08) \times 10^{-6} | 8.38(±0.52) \times 10^{-8} | 3.71(±0.20) \times 10^{-6} | 1.23(±0.06) \times 10^{-6} | 3.31(±0.16) \times 10^{-6} |
| 3 | 9.36(±0.60) \times 10^{-8} | 1.15(±0.07) \times 10^{-6} | 7.87(±0.43) \times 10^{-8} | 3.61(±0.19) \times 10^{-6} | 1.21(±0.06) \times 10^{-6} | 3.21(±0.15) \times 10^{-6} |
| 4 | 9.69(±0.72) \times 10^{-8} | 1.18(±0.08) \times 10^{-6} | 7.98(±0.49) \times 10^{-8} | 3.64(±0.20) \times 10^{-6} | 1.21(±0.06) \times 10^{-6} | 3.26(±0.16) \times 10^{-6} |
| 5 | 9.68(±0.80) \times 10^{-8} | 1.17(±0.09) \times 10^{-6} | 7.92(±0.54) \times 10^{-8} | 3.59(±0.22) \times 10^{-6} | 1.20(±0.07) \times 10^{-6} | 3.30(±0.17) \times 10^{-6} |
| 6 | 9.60(±0.80) \times 10^{-8} | 1.15(±0.10) \times 10^{-6} | 7.79(±0.60) \times 10^{-8} | 3.56(±0.23) \times 10^{-6} | 1.19(±0.08) \times 10^{-6} | 3.17(±0.19) \times 10^{-6} |
| 7 | 9.89(±0.97) \times 10^{-8} | 1.20(±0.11) \times 10^{-6} | 8.13(±0.66) \times 10^{-8} | 3.60(±0.27) \times 10^{-6} | 1.22(±0.08) \times 10^{-6} | 3.26(±0.21) \times 10^{-6} |
| 8 | 9.44(±1.02) \times 10^{-8} | 1.16(±0.11) \times 10^{-6} | 7.93(±0.70) \times 10^{-8} | 3.62(±0.29) \times 10^{-6} | 1.20(±0.09) \times 10^{-6} | 3.21(±0.22) \times 10^{-6} |
| 9 | 8.97(±1.03) \times 10^{-8} | 1.11(±0.12) \times 10^{-6} | 7.66(±0.72) \times 10^{-8} | 3.48(±0.30) \times 10^{-6} | 1.16(±0.09) \times 10^{-6} | 3.11(±0.23) \times 10^{-6} |
| 10 | 9.23(±1.12) \times 10^{-8} | 1.15(±0.13) \times 10^{-6} | 7.87(±0.77) \times 10^{-8} | 3.55(±0.32) \times 10^{-6} | 1.19(±0.10) \times 10^{-6} | 3.20(±0.25) \times 10^{-6} |

Table 2  Point estimates (and 95\% CIs) from IS estimators using different numbers of dominating points for the overshoot probability. \# denotes the number of dominating points used in the IS estimator.
magnitudes of $\sigma$ (ranging from 0.1 to 0.2) and report the estimated probabilities and CIs. We use $10^5$ samples for IS estimators and $10^7$ samples for crude Monte Carlo estimators.

Figures 3a and 3b show the results. Missing less significant dominating points does not seem to make noticeable differences in this problem. As shown in Figure 3a, when we fix the rarity of the problem, the estimate is not sensitive to the number of dominating points. The CI width has an increasing trend as the number of dominating points gets larger, indicating that additional dominating points can in fact even hurt performances.

In Figure 3b, we vary the rarity of the problem and compare the performances of different IS estimators and crude Monte Carlo. Note that estimates using crude Monte Carlo are unavailable for rarer configurations due to its inefficiency. We observe that the estimates from different IS estimators overlap visually in all considered cases, which indicates that the differences among these estimates are negligible. We also note that these estimates are consistent with the crude Monte Carlo estimates (when available), which shows their correctness.

Nonetheless, once again we have an apparent mismatch between theoretical inefficiency and good empirical performances:

**Proposition 5.** Under the problem specification above, the IS estimator that exponentially tilts towards the most significant dominating point, i.e., $X_\gamma$ distributed as $N(\gamma a_1, \gamma I)$, is not asymptotically efficient.

4. **Probabilistic Efficiency**

Section 3 shows that IS estimators that miss some dominating points could perform competitively compared to estimators that consider all of them, thus suggesting a gap between the notion of
asymptotic efficiency and empirical performances. In light of this, we propose the concept of probabilistic efficiency as a relaxation of asymptotic efficiency. The key of probabilistic efficiency is to consider the high-probability relative discrepancy of the estimator from the ground truth directly, instead of using the relative error or equivalently the estimation variance. The latter, as can be seen in the arguments in Section 2, provides a sufficient, but not necessary, condition on the required sample size. In particular, there is an intrinsic looseness brought by the Markov or Chebyshev inequality (2) that converts relative error into the required sample size.

To proceed, we first define the following:

**Definition 3 (Minimal relative discrepancy).** For any estimator \( \hat{p} \) of \( p \) and any \( \varepsilon > 0 \), the minimal relative discrepancy of \( \hat{p} \), at tolerance level \( \varepsilon \), is given by

\[
\delta_\varepsilon(\hat{p},p) := \inf\{\delta > 0 : \tilde{P}(|\hat{p} - p| > \delta p) \leq \varepsilon\}.
\]  

(7)

The minimal relative discrepancy \( \delta_\varepsilon(\hat{p},p) \) measures the relative accuracy of the estimator \( \hat{p} \), in that it gives the smallest relative discrepancy of \( \hat{p} \) from \( p \) that can be achieved with probability \( 1 - \varepsilon \). Thus the smaller is \( \delta_\varepsilon(\hat{p},p) \), the more accurate is \( \hat{p} \). Note that in (7), the probability \( \tilde{P} \) is the one generating the estimator \( \hat{p} \).

We say that \( \hat{p} \) is probabilistically efficient if \( \delta_\varepsilon(\hat{p},p) \) can be made small in some sense, without needing to use a gigantic amount of computation. More precisely, we propose the following notions:

**Definition 4 (Probabilistic Efficiency).** Suppose that \( \{A_\gamma\}_\gamma \) is an indexed family of rare events and \( p = P(A_\gamma) \to 0 \) as \( \gamma \to \infty \). Consider an estimator \( \hat{p} \) obtained from \( n = n(\gamma) \) independent replications of \( Z \). For any \( \varepsilon > 0 \), we define \( \delta_\varepsilon(\hat{p},p) \) as in (7). Then

1. We call \( Z \) strongly probabilistically efficient if we can choose \( n \) subexponential in \( -\log p \) such that, for any \( \varepsilon > 0 \), \( \lim_{\gamma \to \infty} \delta_\varepsilon(\hat{p},p) = 0 \);

2. We call \( Z \) weakly probabilistically efficient if we can choose \( n \) subexponential in \( -\log p \) such that, for any \( \varepsilon > 0 \), \( \limsup_{\gamma \to \infty} \delta_\varepsilon(\hat{p},p) < 1 \).

Note that strong probabilistic efficiency matches the usual notion in statistical estimation. That is, the estimator approaches the target parameter as \( \gamma \to \infty \). In contrast, weak probabilistic efficiency only cares about a correct magnitude. While this may appear less desirable, in rare-event estimation a correct magnitude can be viewed as sufficient as the target quantity is very small, and this weaker notion allows more flexibility in constructing estimators. We also contrast our proposed probabilistic efficiency with a notion named probabilistic bounded relative error proposed in Tuffin and Ridder (2012), where the IS measure is randomly chosen and efficiency is achieved if the resulting random relative error of the IS estimator is bounded by some constant with high probability, which is conceptually different from our notion.

The following shows that probabilistic efficiency is a relaxation of asymptotic efficiency:
**Proposition 6.** If $Z$ is asymptotically efficient, then $Z$ is strongly probabilistically efficient.

**Proof of Proposition 6.** For any unbiased estimator $Z$, we have that for any $\varepsilon > 0$,

$$\hat{P} \left( |\hat{p} - p| > \frac{\sqrt{\text{Var}(Z)}}{\varepsilon np^2} \right) \leq \varepsilon$$

and hence by definition,

$$\delta_{\varepsilon}(\hat{p}, p) \leq \frac{\sqrt{\text{Var}(Z)}}{\varepsilon np^2}.$$

If $Z$ is asymptotically efficient, $\frac{\text{Var}(Z)}{p^2}$ grows at most subexponentially in $-\log p$, so we could choose $n$ subexponentially growing in $-\log p$ such that $\lim_{\gamma \to \infty} \delta_{\varepsilon}(\hat{p}, p) = 0$ for any $\varepsilon > 0$. By definition, $Z$ is strongly probabilistically efficient. \qed

While asymptotic efficiency implies strong probabilistic efficiency, we note that these two notions are not equivalent. In Section 3, Propositions 3–5 show that asymptotic efficiency does not hold for the considered IS estimators in all the presented examples, but Theorem 10 in Section 6.1 will show that strong probabilistic efficiency actually holds for all of them.

Now we explain how probabilistic efficiency helps us understand the influence of missing some dominating points. Recall the example where $A_{\gamma} = \{ \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \}$ and $\mathcal{E}$ comprises two disjoint and faraway pieces $\mathcal{E}_1$ and $\mathcal{E}_2$. The dominating points are respectively $a_1$ and $a_2$ (recall Figure 1). Denote $p_j = P(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_j), j = 1, 2$, and we assume that $p_2$ is exponentially smaller than $p_1$. If we focus on $\mathcal{E}_1$ and simply use the exponential tilting towards $a_1$ as the IS distribution, then we face the risk of having a sample falling into $\mathcal{E}_2$ while the associated likelihood ratio is very high, which leads to asymptotic inefficiency. However, experimentally if we run the simulation with a moderate sample size, then most likely none of the samples fall into $\mathcal{E}_2$. Conditional on not hitting $\mathcal{E}_2$, we actually get an estimate close to $p_1$, which is in turn close to $p$. In other words, even if the resulting IS estimator is not asymptotically efficient, it could still give a good estimate in terms of its distance to $p$, as long as the sample size is not overly big. The latter is precisely the paradigm of probabilistic efficiency.

More concretely, we have the following theorem:

**Theorem 3 (Achieving strong probabilistic efficiency).** Suppose that $\{A_{\gamma}\}_{\gamma}$ is an indexed family of rare events and $p = P(A_{\gamma}) \to 0$ as $\gamma \to \infty$. We write $A_{\gamma} = A_{\gamma}^1 \cup A_{\gamma}^2$ where $A_{\gamma}^1$ and $A_{\gamma}^2$ are two disjoint events. Denote $p_j = P(A_{\gamma}^j), j = 1, 2$. Assume that

1. $\frac{p_1}{p} \to 1$ as $\gamma \to \infty$;

2. We have an asymptotically efficient IS estimator for $p_1$ obtained from $Z_1 = I_{A_{\gamma}^1} \frac{dp}{d\tilde{P}}$ under $\hat{P}$. This implies that there exists $n = n(\gamma)$ growing subexponentially in $-\log p$ such that $\frac{\text{Var}(Z_1)}{np_1} \to 0$ as $\gamma \to \infty$;
3. \( \tilde{p}_2 := \tilde{P}(A_2^\gamma) \) satisfies that \( n\tilde{p}_2 \to 0 \) as \( \gamma \to \infty \).

Let \( \hat{p} \) be the sample mean of \( n \) independent replications of \( Z = I_{A^\gamma} \text{d}P \) under \( \tilde{P} \). For any \( \varepsilon > 0 \), define \( \delta_{\varepsilon}(\hat{p}, p) \) as in (7). Then we have

\[
\delta_{\varepsilon}(\hat{p}, p) \leq \sqrt{\frac{\text{Var}(Z)}{np_1^2(\varepsilon - n\tilde{p}_2)}} + \frac{p_2}{p} \to 0 \text{ as } \gamma \to \infty.
\]

Hence \( Z \) is strongly probabilistically efficient.

**Proof of Theorem 3.** Suppose that we sample \( \omega_1, \ldots, \omega_n \) under \( \tilde{P} \). Let

\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} I_{A_i}(\omega_i) \text{d}P(\omega_i)
\]

and

\[
\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} I_{A_j}(\omega_i) \text{d}P(\omega_i), j = 1, 2.
\]

Clearly \( \hat{p} = \hat{p}_1 + \hat{p}_2 \). For simplicity, denote

\[
\delta := \sqrt{\frac{\text{Var}(Z_1)}{np_1^2(\varepsilon - n\tilde{p}_2)}} + \frac{p_2}{p}, \quad (8)
\]

\[
\tilde{\delta} := \frac{p_2}{p}, \quad (9)
\]

Then we have that

\[
\tilde{P}(|\hat{p} - p| > \delta p) \leq \tilde{P}(|\hat{p}_1 - p_1| > (\delta - \tilde{\delta})p \text{ or } |\hat{p}_2 - p_2| > \tilde{\delta} p)
\]

\[
\leq \tilde{P}(|\hat{p}_1 - p_1| > (\delta - \tilde{\delta})p) + \tilde{P}(|\hat{p}_2 - p_2| > \tilde{\delta} p)
\]

\[
\leq \tilde{P}(|\hat{p}_1 - p_1| > (\delta - \tilde{\delta})p) + \tilde{P}(\hat{p}_2 > 0)
\]

\[
\leq \frac{\text{Var}(Z_1)}{n(\delta - \tilde{\delta})^2 p_1^2} + n\tilde{p}_2
\]

\[
= \varepsilon.
\]

where the second inequality follows from a union bound, the third inequality follows from \( p_1 \leq p \) and that \( |\hat{p}_2 - p_2| > \tilde{\delta} p \) implies \( \hat{p}_2 > 0 \), the fourth inequality follows from Chebyshev’s inequality in the first term and a union bound in the second term, and the last equality follows from the definitions in (8) and (9). Thus \( \delta_{\varepsilon}(\hat{p}, p) \leq \delta \). Finally, \( \delta \to 0 \) by a direct use of the assumptions. \( \square \)

Similarly, if we relax the assumption that \( p_1/p \to 1 \) as \( \gamma \to \infty \), we get sufficient conditions for weak probabilistic efficiency:

**Theorem 4 (Achieving weak probabilistic efficiency).** Suppose that \( \{A_i\}_\gamma \) is an indexed family of rare events and \( p = P(A_\gamma) \to 0 \) as \( \gamma \to \infty \). We write \( A_\gamma = A_1^\gamma \cup A_2^\gamma \) where \( A_1^\gamma \) and \( A_2^\gamma \) are two disjoint events. Denote \( p_j = P(A_j^\gamma), j = 1, 2 \). Assume that
1. \( \liminf_{\gamma \to \infty} \frac{p_1}{p} = c \) where \( 0 < c \leq 1 \);

2. We have an asymptotically efficient IS estimator for \( p_1 \) obtained from \( Z_1 = I_{A_1^\gamma} \frac{dp}{dP} \) under \( \tilde{P} \). This implies that there exists \( n = n(\gamma) \) growing subexponentially in \( -\log p \) such that \( \frac{\text{Var}(Z_1)}{np_1^2} \to 0 \) as \( \gamma \to \infty \);

3. \( \tilde{p}_2 := \tilde{P} (A_2^\gamma) \) satisfies that \( np_2 \to 0 \) as \( \gamma \to \infty \).

Let \( \hat{p} \) be the sample mean of \( n \) independent replications of \( Z = I_{A_\gamma} \frac{dp}{dP} \) under \( \tilde{P} \). For any \( \varepsilon > 0 \), define \( \delta_\varepsilon (\hat{p}, p) \) as in (7). Then we have

\[
\limsup_{\gamma \to \infty} \delta_\varepsilon (\hat{p}, p) \leq \limsup_{\gamma \to \infty} \left( \sqrt{\frac{\text{Var}(Z_1)}{np_1^2 (\varepsilon - \tilde{p}_2)}} + \frac{p_2}{p} \right) = 1 - c < 1.
\]

Hence \( Z \) is weakly probabilistically efficient.

**Proof of Theorem 4.** Following the proof of Theorem 3, we still get

\[
\delta_\varepsilon (\hat{p}, p) \leq \delta := \sqrt{\frac{\text{Var}(Z_1)}{np_1^2 (\varepsilon - \tilde{p}_2)}} + \frac{p_2}{p}.
\]

Under the conditions of Theorem 4, now \( \limsup_{\gamma \to \infty} \delta = 1 - c < 1 \). By the definition, \( Z \) is weakly probabilistically efficient.

We note that in Theorems 3 and 4, \( A_1^\gamma \) and \( A_2^\gamma \) could be very general events. In particular, when we use dominating points to decompose the rare-event set, \( A_1^\gamma \) and \( A_2^\gamma \) are not necessarily each governed by only one dominating point but could be more as long as the assumptions hold. Moreover, there can be multiple ways to split \( \mathcal{A}_\gamma \) into \( A_1^\gamma \) and \( A_2^\gamma \), and as long as one of these ways validates the assumptions in Theorem 3 or 4 then probabilistic efficiency is guaranteed. This provides flexibility in using Theorems 3 and 4; Sections 6.1.2 and 6.1.3 will demonstrate this in some specific examples.

According to the theorems, supposing that we have found some dominating points while the remaining ones are known to be less significant and “far from” the current ones, we could simply use the current mixture IS distribution instead of keep searching. The remaining question is how we could detect that the remaining dominating points are negligible, i.e. the assumptions of the theorem are satisfied. Besides, probabilistic efficiency only implies that the point estimate is reliable in some sense. This raises questions on inference such as the construction of valid CIs. In the next section, we will make these discussions precise and show our answers under minimal assumptions in the standard Gartner-Ellis regime.

5. **Probabilistically Efficient Estimation in the Gartner-Ellis Regime**

We study probabilistically efficient IS in the widely used Gartner-Ellis regime introduced in Section 2. Our key result is that probabilistic efficiency can be readily achieved by using only the
most significant dominating points, under essentially no more assumptions than what is needed to derive the Gartner-Ellis large deviations asymptotic.

We first consider the case where there is only one most significant dominating point, which is a common scenario (e.g., in all the examples in Section 3):

**Theorem 5 (Using the most significant dominating point is probabilistically efficient).** Consider the problem of estimating \( p = P(\frac{1}{r}X_\gamma \in E) \). Suppose that Assumptions 1 and 2 hold. Suppose also that the dominating set \( A \) has finite cardinality with a unique most significant dominating point \( a \), i.e., \( I(a) < I(\tilde{a}) \) for all other \( \tilde{a} \in A \). Then the IS distribution \( \tilde{P} \) given by the exponential tilting towards \( a \), i.e.,

\[
\frac{d\tilde{P}}{dP}(\omega) = e^{s_a^\top X_\gamma - \gamma \mu_\gamma(s_a)}
\]

is strongly probabilistically efficient.

Theorem 5 stipulates that we only need the most significant dominating point in constructing an efficient IS. This result, which is in sharp contrast to the established IS recipe that suggests using all dominating points, explains the good empirical performance of the “poor” estimators in Section 3. Moreover, the proposal in Theorem 5 is in closer line with the Gartner-Ellis asymptotic theory, in that the use of the most significant dominating point, which is also the minimizer of the rate function (recall Theorem 2), governs both the large deviations asymptotic and the construction of efficient IS. Lastly, regarding the assumptions needed, the only additional condition beyond the standard Gartner-Ellis assumptions (i.e., Assumptions 1 and 2) is the finite cardinality of the dominating set. In fact, if we have multiple most significant points, we have a natural generalization:

**Theorem 6 (Mixing most significant dominating points).** Consider the problem of estimating \( p = P(\frac{1}{r}X_\gamma \in E) \). Suppose that Assumptions 1 and 2 hold. Suppose also that the dominating set \( A \) has finite cardinality with \( k \) most significant dominating points \( a_1, \ldots, a_k \), i.e., \( I(a_1) = \cdots = I(a_k) < I(a') \) for all \( a' \in A \setminus \{a_1, \ldots, a_k\} \). Then the IS distribution \( \tilde{P} \) given by the mixture of exponential tiltings towards \( a_1, \ldots, a_k \), i.e.,

\[
\frac{d\tilde{P}}{dP}(\omega) = \sum_{i=1}^{k} \alpha_i e^{s_{a_i}^\top X_\gamma - \gamma \mu_\gamma(s_{a_i})}
\]

where \( \sum_{i=1}^{k} \alpha_i = 1, \alpha_i > 0, \forall i \), is strongly probabilistically efficient.

That is, we use mixture to account for all the most significant dominating points when there are multiple of them. The proofs of Theorems 5 and 6 amount to verifying the assumptions in Theorem 3 using the Gartner-Ellis conditions. In particular, Conditions 1 and 2 in Theorem 3 can be routinely verified, while Condition 3 is checked by showing that \( \tilde{p}_2 \) is in fact exponentially
decaying in $\gamma$, which requires an application of the Gartner-Ellis theorem under the IS distribution.

The verification of Condition 3 especially reveals a key phenomenon that, under the exponential tilting to the most significant dominating point(s), the probability of an IS sample hitting onto the “backyards” of other dominating points is exponentially small, which in turn fulfills the notion of probabilistic efficiency.

**Proofs of Theorems 5 and 6.** We focus on Theorem 6 since Theorem 5 is a special case therein.

It suffices to verify all the assumptions in Theorem 3. For $A_\gamma = \{ \frac{1}{\gamma} X_\gamma \in \mathcal{E} \}$, by Theorem 1, $p = P(A_\gamma)$ satisfies that $\lim_{\gamma \to \infty} \frac{1}{\gamma} \log p = -I(\mathcal{E}) < 0$, so $p \to 0$ as $\gamma \to \infty$. If $k = |A|$, then the IS estimator from (11) already uses all the dominating points and thus is asymptotically efficient by Proposition 1. Hence, from now on, we assume that $k < |A|$. For convenience, we denote $a_{k+1}$ as a next most significant point other than $a_1, \ldots, a_k$, i.e., $I(a_{k+1}) > I(a_1) = \cdots = I(a_k)$ and $I(a_{k+1}) < I(a')$ for all $a' \in A \setminus \{a_1, \ldots, a_k\}$. We split $\mathcal{E}$ into $\mathcal{E}_1 = \mathcal{E} \cap \bigcup_{i=1}^{k} \{ x \in \mathbb{R}^d : s_{a_i}^\top (x - a_i) \geq 0 \}$ and $\mathcal{E}_2 = \mathcal{E} \setminus \mathcal{E}_1$, and define $A_j = \{ \frac{1}{\gamma} X_\gamma \in \mathcal{E}_j \}$, $p_j = P(A_j)$ for $j = 1, 2$.

First, by Theorem 1, we have that $\limsup_{\gamma \to \infty} \frac{1}{\gamma} \log p_2 \leq -I(\mathcal{E}_2)$. Since $\mathcal{E}_2 \subset \bigcup_{j=k+1}^{|A|} \{ x : s_{a_j}^\top (x - a_j) \geq 0 \}$, we know that $I(\mathcal{E}_2) = I(\bigcup_{j=k+1}^{|A|} \{ x : s_{a_j}^\top (x - a_j) \geq 0 \}) = I(a_{k+1})$, and hence $\limsup_{\gamma \to \infty} \frac{1}{\gamma} \log p_2 \leq -I(a_{k+1}) < -I(\mathcal{E})$. As a result, $p_2/p \to 0$ as $\gamma \to \infty$. This verifies Assumption 1 in Theorem 3.

Second, by the definition, $\{a_1, \ldots, a_k\}$ is a dominating set for $\mathcal{E}_1$, so the IS estimator $Z_1 = I(\frac{1}{\gamma} X_\gamma \in \mathcal{E}_1) \frac{dP_i}{dP}(\omega)$ is asymptotically efficient by Proposition 1. This verifies Assumption 2 in Theorem 3.

Third, we would prove that $\tilde{p}_2$ decays exponentially in $\gamma$ (hence also exponentially in $-\log p$), and hence $n\tilde{p}_2 \to 0$ for subexponentially growing $n$ which verifies Assumption 3 in Theorem 3. Indeed, we have

$$\tilde{p}_2 = \sum_{i=1}^{k} \alpha_i \tilde{P}_i \left( \frac{1}{\gamma} X_\gamma \in \mathcal{E}_2 \right)$$

where $\frac{dP_i}{dP}(\omega) = e^{s_{a_i}^\top X_\gamma - \gamma \mu_\gamma(s_{a_i})}$. Denote $\tilde{E}_i$, $\tilde{\mu}_{\gamma,i}$ and $\tilde{I}_i$ as the corresponding expectation, scaled logarithmic moment generating function and rate function under $\tilde{P}_i$. Then, under $\tilde{P}_i$, we have

$$\tilde{\mu}_{\gamma,i}(x) = \frac{1}{\gamma} \log \tilde{E}_i \left( e^{x^\top X_\gamma} \right) = \frac{1}{\gamma} \log E \left( e^{x^\top X_\gamma + s_{a_i}^\top X_\gamma - \gamma \mu_\gamma(s_{a_i})} \right) = \mu_\gamma(x + s_{a_i}) - \mu_\gamma(s_{a_i})$$

and thus $\tilde{\mu}_i(x) = \lim_{\gamma \to \infty} \tilde{\mu}_{\gamma,i}(x) = \mu(x + s_{a_i}) - \mu(s_{a_i})$. Then the rate function is

$$\tilde{I}_i(y) = \sup_{x \in \mathbb{R}^d} \{ x^\top y - \tilde{\mu}_i(x) \}
= \sup_{x \in \mathbb{R}^d} \{ x^\top y - \mu(x + s_{a_i}) + \mu(s_{a_i}) \}
= \sup_{x \in \mathbb{R}^d} \{ (x + s_{a_i})^\top y - \mu(x + s_{a_i}) \} - s_{a_i}^\top y + \mu(s_{a_i})
= I(y) - s_{a_i}^\top y + \mu(s_{a_i}).$$
For any $y \in \mathcal{E}_2$, we have that $I(y) \geq I(a_{k+1})$ and that $s_{a_1}^T (y - a_i) \leq 0$ for $i = 1, \ldots, k$, and thus $\tilde{I}_i(y) = I(y) - s_{a_1}^T y + \mu(s_{a_1}) \geq I(a_{k+1}) - s_{a_1}^T a_i + \mu(s_{a_1}) = I(a_{k+1}) - I(a_i)$. Therefore, $\tilde{I}_i(\mathcal{E}_2^c) \geq I(a_{k+1}) - I(a_1) > 0$. From the above derivations, Assumption 1 still holds for $\tilde{p}_2$. By Theorem 1,

$$\limsup_{\gamma \to \infty} \frac{1}{\gamma} \log \tilde{P}_i \left( \frac{1}{\gamma} X_\gamma \in \mathcal{E}_2 \right) \leq - \tilde{I}_i(\mathcal{E}_2^c) \leq -(I(a_{k+1}) - I(a_1)) < 0.$$ 

Overall, we have $\tilde{p}_2$ decays exponentially in $\gamma$.

Now we have verified all the assumptions in Theorem 3.

We comment that if we know any one of the most significant points, say $a$, among several such points, satisfies $p_1/p \to c$ for some $0 < c \leq 1$, where $p_1 = P(A_\gamma \cap \{s_{a_1}^T X_\gamma - a \geq 0\})$ is the rare-event probability “contributed” from $a$, then using the IS that exponentially tilts only to $a$, i.e., (10), is weakly probabilistically efficient. This can be shown by a similar argument to the proofs of Theorems 5 and 6 above. Such an approach is in contrast to using the IS mixture in (11) suggested by Theorem 6 that achieves strong, instead of only weak, probabilistic efficiency. Nonetheless, knowing $p_1/p \to c$ typically requires information on the multiplicative factor in front of the exponential decay dictated by the large deviations rate function, which in turn requires derivation of exact asymptotic that is only known for a relatively small number of problems.

Next, besides point estimates, we investigate inference using probabilistically efficient IS estimators, in particular how to construct (asymptotically) valid CIs. First, we consider the interval

$$\mathcal{I}_1 = \left[ \hat{p} - \left( \sqrt{\frac{2\hat{V} \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha) M_\gamma}{3(n-1)} \right), \hat{p} + \left( \sqrt{\frac{2\hat{V} \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha) M_\gamma}{3(n-1)} \right) \right] \tag{12}$$

where $\hat{V}$ is the sample variance and $M_\gamma = \max_{i=1,\ldots,k} \left\{ \frac{1}{\alpha_i} e^{-\gamma(s_{a_1}^T a_i - \mu_\gamma(s_{a_1}))} \right\}$ is deterministic. The following theorem provides an asymptotic coverage guarantee for this CI:

**Theorem 7 (Constructing confidence intervals with probabilistically efficient estimators).**

Under the same setting as Theorem 6, suppose we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)} = I(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}) \frac{dp}{d\tilde{P}}$, $i = 1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$'s. If $n$ is subexponentially growing in $-\log p$ (or $\gamma$) as $\gamma \to \infty$, then, for any $0 < \alpha < 1$,

$$\liminf_{\gamma \to \infty} \tilde{P}(p \in \mathcal{I}_1) \geq 1 - \alpha$$

where $M_\gamma = \max_{i=1,\ldots,k} \left\{ \frac{1}{\alpha_i} e^{-\gamma(s_{a_1}^T a_i - \mu_\gamma(s_{a_1}))} \right\}$.

In Theorem 7, note that even if we neglect the higher-order term (in terms of $n$) $\frac{7 \log(4/\alpha) M_\gamma}{3(n-1)}$, the CI half-width is $\sqrt{2\log(4/\alpha)}$ times $\sqrt{\frac{\hat{V}}{n}}$, which is more conservative than the Central Limit Theorem (CLT) based interval

$$\mathcal{I}_2 = \left[ \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{V}}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{V}}{n}} \right] \tag{13}$$
where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. For instance, when $\alpha = 0.05$, we have $\sqrt{2\log(4/\alpha)} \approx 2.96$, while $z_{1-\alpha/2} \approx 1.96$. Our next theorem shows that, under stronger conditions, the CLT-based CI (13) is also asymptotically valid.

**Theorem 8 (Constructing tight confidence intervals with probabilistically efficient estimators).**

Under the same setting as Theorem 6, suppose we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)} = I(\frac{1}{\gamma}X^{(i)} \in \mathcal{E}) \frac{dp}{d\tilde{P}}$, $i = 1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$’s. In this case, we could choose $n$ subexponentially growing in $-\log p$ (or $\gamma$) such that $\frac{M^2_{\gamma}}{n\text{Var}(Z^{(1)})} \to 0$ and $\frac{E^3[Z^{(1)} - p]^3}{n\text{Var}^3(Z^{(1)})} \to 0$ as $\gamma \to \infty$ where $M_{\gamma}$ is as defined in Theorem 7. Then, for any $0 < \alpha < 1$,$$
abla \gamma \to \infty \tilde{P}(p \in I_2) \geq 1 - \alpha.$$

Theorem 8 tightens the interval in Theorem 7 to using the CLT-based critical value $z_{1-\alpha/2}$ with a more careful choice of sample size $n$.

Finally, we prove that if we use all the dominating points in the mixture, so that the estimator satisfies the classical notion of asymptotic efficiency, then, under conditions similar to Theorem 8, the CLT-based interval possesses an even stronger guarantee that the asymptotic coverage probability is exactly $1 - \alpha$.

**Theorem 9 (Asymptotically exact confidence intervals with asymptotically efficient estimators).**

Consider the problem of estimating $p = P(\frac{1}{\gamma}X, \in \mathcal{E})$. Suppose that Assumptions 1 and 2 hold, and the dominating set is finite. The IS estimator is $Z = I(\frac{1}{\gamma}X, \in \mathcal{E}) \frac{dp}{d\tilde{P}}(\omega)$ under $\tilde{P}$ given by (3). We sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)} = I(\frac{1}{\gamma}X^{(i)} \in \mathcal{E}) \frac{dp}{d\tilde{P}}$, $i = 1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$’s. In this case, we could choose $n$ at least subexponentially growing in $-\log p$ (or $\gamma$) such that $\frac{M^2_{\gamma}}{n\text{Var}(Z^{(1)})} \to 0$ and $\frac{E^3[Z^{(1)} - p]^3}{n\text{Var}^3(Z^{(1)})} \to 0$ as $\gamma \to \infty$ where $M_{\gamma} = \max_{i=1, \ldots, r} \left\{ \frac{1}{\alpha_i} e^{-\gamma(s_i, a_i - \mu_i(s_i))} \right\}$. Then, for any $0 < \alpha < 1$,$$
abla \gamma \to \infty \tilde{P}(p \in I_2) = 1 - \alpha.$$

We make several remarks regarding the properties of the CLT-based CI $I_2$ in Theorems 8 and 9. First, it appears that probabilistically efficient samples sacrifice some looseness in terms of CI coverage compared to asymptotically efficient samples, as the guarantee is valid in Theorem 8 but exact in Theorem 9. Second, in Theorem 8, like Theorem 7, the sample size $n$ is required to be not overly big, manifested by the subexponential growth requirement. This is in contrast to Theorem 9 that does not impose any upper bound on $n$. This ties to the key idea of probabilistic efficiency
that, when the sample size is not overly big, there is a negligible chance of any sample hitting the rare-event region not corresponding to the most significant points. Thus the CI constructed from a probabilistically efficient estimator, much like the point estimate, is valid only when the sample size is not overly big, while asymptotically efficient estimators do not impose such a restriction. Lastly, we see the requirement on $n$ given by $\frac{M_2^2}{n\text{Var}(Z^{(1)})} \to 0$ and $\frac{E_2[Z^{(1)} - p]}{n\text{Var}(Z^{(1)})} \to 0$ in Theorems 8 and 9. While these conditions can be difficult to verify in practice, we should note that they are lower bound requirements, and imposed not only for CIs constructed from probabilistically efficient estimators, but also for classical asymptotically efficient estimators as well (to our best knowledge, conditions on the adequacy of sample size to attain CI coverage guarantees for these classical estimators is not known in the literature). In the next section, we will investigate the performances of all these CIs with reasonable sample sizes.

Lastly, to close this section, we briefly note that Algorithm 1 in Appendix A shows generally how to identify and compute dominating points, sequentially starting from the most significant one. Moreover, Appendix B studies parallel results to this section for an alternative asymptotic regime to Gartner-Ellis that could be suitable for some situations involving highly complex systems.

6. Further Numerical Experiments and Discussions

We have shown several examples in Section 3 where IS estimators using only one or a small number of dominating points perform competitively compared with asymptotically efficient IS estimators that use all dominating points. In fact, we have shown in each example in Sections 3.1, 3.2 and 3.3 that the simple estimator using the most significant dominating point is not asymptotically efficient. In this section, we argue that they are all probabilistically efficient, which is a direct consequence of Theorem 5. We then numerically assess the validity of the conditions in Theorem 3, which forms the underlying basis in justifying probabilistic efficiency. Finally, we test the confidence intervals constructed using our probabilistically efficient estimators discussed in Section 5 and compare with intervals constructed from asymptotically efficient estimators.

6.1. Verifying Conditions for Probabilistic Efficiency

We first state the strong probabilistic efficiency of all the proposed estimators that use only the most significant dominating points in Section 3:

**Theorem 10.** Under the problem specifications in Sections 3.1, 3.2 and 3.3, the IS estimators that use only the most significant dominating points, namely $\hat{\alpha}(m)$ in Section 3.1, $X_\gamma$ distributed as $N(\gamma a_1, \gamma \Sigma)$ in Section 3.2 and $N(\gamma a_1, \gamma I)$ in Section 3.3, are all strongly probabilistically efficient.

Next, we validate the underpinning mechanism of how probabilistic efficiency arises in these examples. Note that the main basis of the strong probabilistic efficiency of these estimators, which
Table 3  Estimates of $p_1$ and $p_2$ for the i.i.d sum example in Section 3.1 with $10^4$ samples.

| $m$  | 10       | 30       | 50       | 100      |
|------|----------|----------|----------|----------|
| $p_1$ | $8.33 \times 10^{-5}$ | $1.59 \times 10^{-8}$ | $3.59 \times 10^{-10}$ | $1.33 \times 10^{-14}$ |
| $p_1/p$ | 0.992    | $\approx 1$   | $\approx 1$   | $\approx 1$   |
| $p_2$ | $6.56 \times 10^{-5}$ | $3.15 \times 10^{-11}$ | $1.85 \times 10^{-12}$ | $6.68 \times 10^{-13}$ |
| $p_2/p$ | 0.008    | $1.98 \times 10^{-6}$ | $5.16 \times 10^{-10}$ | $5.00 \times 10^{-19}$ |
| $p$   | $8.40 \times 10^{-3}$ | $1.59 \times 10^{-10}$ | $3.59 \times 10^{-10}$ | $1.33 \times 10^{-14}$ |

follows from Theorem 5, is Theorem 3. In particular, Theorem 3 states three conditions that allow one to conclude strong probabilistic efficiency. Among them, the second condition is a property about asymptotic efficiency for an estimator that applies to a more restrictive rare-event set, which has been well-established in the asymptotic efficiency literature (basically, by mixing the exponential tiltings towards all the dominating points associated with the more restrictive rare-event set). Conditions 1 and 3 are more delicate. In the setting with a unique most significant dominating point, say $a$, the former requires a small proportion of the “contribution” from the less significant dominating points other than $a$ over the total rare-event probability, i.e., $p_2/p \to 0$ where $p_2 = P(A, \{s_a^+(x-a) \geq 0\})$. The latter requires a small probability of sampling any points in the rare-event set that does not belong to the backyard of $a$, i.e., $\hat{P}(some of the n samples hits A, \{s_a^+(x-a) \geq 0\}) \to 0$ or, as a sufficient condition, $n\hat{p}_2 \to 0$ where $\hat{p}_2 = \hat{P}(A, \{s_a^+(x-a) \geq 0\})$.

Our next goal is to assess the smallness and decreasing trends (as rarity grows) of $p_2/p$ and $\hat{P}(some of the n samples hits A, \{s_a^+(x-a) \geq 0\})$ that drive Theorem 3.

6.1.1. Large Deviations of an I.I.D. Sum. For the experiment in Section 3.1, we use the probabilistically efficient estimator $\hat{\alpha}(m)$. Correspondingly, we have $p_1 = P(S_m \geq am)$ and $p_2 = P(S_m \leq -am)$. Table 3 shows these values as $m$ varies, which we approximate respectively by using estimators $\hat{\beta}_1(m) = \exp(-s_a S_m + m \mu(s_a)) I(S_m \geq am)$ and $\hat{\beta}_2(m) = \exp(-s_a S_m + m \mu(s_a)) I(S_m \leq -am)$, with $s_a, s_{-a}$ defined as in Section 3.1, generated by the same IS samples used in $\hat{\beta}(m)$. From Table 3, we observe that the estimate of $p_2/p$ is 0.008 with $m = 10$ and decreases to $5.00 \times 10^{-19}$ as $m = 100$. This shows that $p_2/p$ is small and approaches 0 as $m$ increases, which matches Condition 1 in Theorem 3 ($m$ is the rarity parameter here).

Next, we examine $\hat{p}_2 = \hat{P}(S_m \leq -am)$. We generate $10^7$ samples from the strongly probabilistically efficient IS distribution. We observe that none of the samples fall into $\{S_m \leq -am\}$, which indicates that $\hat{p}_2$ is extremely small so that $\hat{P}(some of the n samples hits A, \{s_a^+(x-a) \geq 0\})$, with $n = 10^4$ in our experiment here, is close to zero. This matches Condition 3 of Theorem 3.

6.1.2. Overshoot Probability of a Random Walk. For the experiment in Section 3.2, we consider the most significant dominating point $a = a_1$ and our probabilistically efficient estimator is the exponential tilting towards the most significant dominating point only, i.e., $X_{\gamma}$ distributed.
as $N(\gamma_1, \Sigma)$. We define $A^1_\sigma = A_\gamma \cap \{s_a^\top (x-a) \geq 0\}$ and $A^2_\sigma = A_\gamma \setminus \{s_a^\top (x-a) \geq 0\}$, and the corresponding probabilities $p_1 = P(A_\gamma \cap \{s_a^\top (x-a) \geq 0\})$ and $p_2 = P(A_\gamma \setminus \{s_a^\top (x-a) \geq 0\})$. We compute $p_1, p_2$, and also the contribution of each of the nine less significant dominating points in $p_2$. More precisely, we define $B^1_\sigma = A_\gamma \cap \{s_{a_2}^\top (x-a_2) \geq 0\} \setminus \{s_{a_1}^\top (x-a_1) \geq 0\}$, $B^2_\sigma = A_\gamma \setminus \{s_{a_3}^\top (x-a_3) \geq 0\} \setminus (\{s_{a_1}^\top (x-a_1) \geq 0\} \cup \{s_{a_2}^\top (x-a_2) \geq 0\})$, ..., $B^{10}_\sigma = A_\gamma \setminus \{s_{a_{10}}^\top (x-a_{10}) \geq 0\} \setminus \left( \bigcup_{j=1}^{9} \{s_{a_j}^\top (x-a_j) \geq 0\} \right)$, and use $p_{a_2} = P(B^2_\sigma), ..., p_{a_{10}} = P(B^{10}_\sigma)$ to denote the contribution of dominating points $a_2, ..., a_{10}$ (with decreasing significance). For each probability $p_{a_1}, p_{a_2}, ..., p_{a_{10}}$, we construct an IS estimator using the “corresponding” dominating points $a_1, ..., a_{10}$, e.g., for $p_{a_2}$ the IS distribution is mean shifted to $a_2$. Then we estimate $p_2$ through $p_2 = p_{a_2} + ... + p_{a_{10}}$ and $p$ through $p = p_1 + p_2$. Table 4 presents the results estimated using independently generated $10^4$ samples from the corresponding IS distributions. We observe that $p_2/p$ has larger values than those in the previous experiment, in that $p_2/p \approx 0.25$ at $\sigma = 0.3$ and $p_2/p \approx 0.15$ at $\sigma = 0.2$. Nonetheless, $p_2/p$’s value decreases rapidly as $\sigma$ decreases, i.e., the problem becomes rarer, which suggests the trend $p_2/p \to 0$ in Condition 1 of Theorem 3. Additionally, we observe from the values of $p_{a_3}/p, ..., p_{a_{10}}/p$ that the contribution of each less significant dominating point vanishes rapidly with decreasing $\sigma$.

Table 4  Estimates of $p_1$, $p_2$ and the contributions of the less significant dominating points for the random walk example in Section 3.2 with $10^4$ samples, where we use $A^1_\sigma = A_\gamma \cap \{s_a^\top (x-a) \geq 0\}$ and $A^2_\sigma = A_\gamma \setminus \{s_a^\top (x-a) \geq 0\}$.

| $\sigma$ | $p_1$ | $p_1/p$ | $p_2$ | $p_2/p$ | $p_2/p$ | $p_2/p$ | $p_2/p$ | $p_2/p$ |
|---------|-------|---------|-------|---------|---------|---------|---------|---------|
| $0.3$   | $2.42 \times 10^{-4}$ | $0.7434$ | $8.36 \times 10^{-5}$ | $0.2566$ | $3.26 \times 10^{-4}$ | $5.29 \times 10^{-3}$ | $0.1625$ | $2.06 \times 10^{-5}$ |
| $0.28$  | $9.07 \times 10^{-5}$ | $0.7512$ | $3.00 \times 10^{-5}$ | $0.24$ | $1.21 \times 10^{-4}$ | $1.95 \times 10^{-3}$ | $0.1611$ | $7.46 \times 10^{-6}$ |
| $0.26$  | $2.82 \times 10^{-5}$ | $0.7776$ | $8.05 \times 10^{-6}$ | $0.2224$ | $3.62 \times 10^{-5}$ | $5.50 \times 10^{-4}$ | $0.1519$ | $1.88 \times 10^{-6}$ |
| $0.24$  | $6.12 \times 10^{-6}$ | $0.7960$ | $1.57 \times 10^{-5}$ | $0.2040$ | $7.66 \times 10^{-6}$ | $1.14 \times 10^{-3}$ | $0.1498$ | $3.32 \times 10^{-7}$ |
| $0.22$  | $9.79 \times 10^{-7}$ | $0.8311$ | $1.99 \times 10^{-7}$ | $0.1689$ | $1.18 \times 10^{-6}$ | $1.56 \times 10^{-5}$ | $0.1323$ | $3.57 \times 10^{-8}$ |
| $0.2$   | $7.92 \times 10^{-8}$ | $0.8549$ | $1.34 \times 10^{-8}$ | $0.1451$ | $9.26 \times 10^{-8}$ | $1.11 \times 10^{-8}$ | $0.1194$ | $2.06 \times 10^{-9}$ |
Next, we present the probabilities \( \tilde{p}_1 = \tilde{P}(A_\gamma \cap \{ s^\top_a (x - a) \geq 0 \}) \) and \( \tilde{p}_2 = \tilde{P}(A_\gamma \setminus \{ s^\top_a (x - a) \geq 0 \}) \) under the probabilistically efficient IS distribution. For \( \tilde{p}_2 \), we also present the contributions from the dominating points \( a_2, \ldots, a_{10} \), denoted as \( \tilde{p}_{a_2}, \ldots, \tilde{p}_{a_{10}} \) with \( \tilde{p}_{a_i} = \tilde{P}\left( A_\gamma \cap \{ s^\top_{a_i} (x - a_i) \geq 0 \} \setminus \bigcup_{j=1}^{i-1} \{ s^\top_{a_j} (x - a_j) \geq 0 \} \right) \) for \( i = 2, \ldots, 10 \). The probabilities are estimated using the proportion of samples falling into the corresponding sets based on \( 10^4 \) samples drawn from the probabilistically efficient IS distribution. The results are presented in Table 5. We observe that \( \tilde{p}_2 \) generally decreases from 0.0129 at \( \sigma = 0.3 \) to 0.0039 with \( \sigma = 0.2 \). Furthermore, the decreasing trends also appear in each individual contribution from the less significant dominating points, where most of the probabilities (e.g. \( \tilde{p}_{a_4}, \ldots, \tilde{p}_{a_{10}} \)) already vanish when \( \sigma = 0.2 \). Based on the value of \( \tilde{p}_2 \), we estimate the probability \( \tilde{P}(\text{some of the } n \text{ samples hits } A_\gamma \setminus \{ s^\top_a (x - a) \geq 0 \}) \) through \( 1 - (1 - \tilde{p}_2)^n \). We denote this probability as \( \tilde{p}_{\text{hit}} \) and present the results with \( n = 10^4 \) (the sample size we use in Section 3) in the last row of Table 5. We observe that there are samples falling into \( A_\gamma \setminus \{ s^\top_a (x - a) \geq 0 \} \) with approximately probability 1 when we use \( n = 10^4 \) samples. This close-to-1 probability, unfortunately, is quite different from what our Condition 3 in Theorem 3 would entail and cannot explain the good performance of our probabilistically efficient estimator.

To this end, we verify the conditions in Theorem 3 using an alternative construction of \( A_1^\gamma \) and \( A_2^\gamma = A_\gamma \setminus A_1^\gamma \). Here, in our previous construction, we have chosen the \( A_1^\gamma \) to be the half-space cut by a dominating point and it turns out that the corresponding \( \tilde{p}_2 \) is not small and thus the condition of Theorem 3 appears to fail. However, as discussed right after Theorem 4, our main theorems allow more flexibility in choosing our \( A_1^\gamma \), and as long as we find a suitable way to construct \( A_1^\gamma \) to satisfy the needed conditions, Theorem 3 can be used to explain our estimator’s good performance.

Here is how we can construct a suitable alternative \( A_1^\gamma \) for Theorem 3. From the proofs of Propositions 4 and 5, we know that our probabilistically efficient estimator is not asymptotically

**Table 5** Estimates of \( \tilde{p}_1, \tilde{p}_2 \) and the contributions of the less significant dominating points under probabilistically efficient IS for the random walk example in Section 3.2 with \( 10^4 \) samples, where we use \( A_1^\gamma = A_\gamma \cap \{ s^\top_a (x - a) \geq 0 \} \) and \( A_2^\gamma = A_\gamma \setminus \{ s^\top_a (x - a) \geq 0 \} \).

| \( \sigma \) | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
|---|---|---|---|---|---|---|
| \( p_1 \) | 0.5057 | 0.4916 | 0.496 | 0.5053 | 0.4969 | 0.4985 |
| \( p_2 \) | 0.0129 | 0.0141 | 0.0086 | 0.0076 | 0.0057 | 0.0039 |
| \( p_{a_2} \) | 0.0111 | 0.0111 | 0.007 | 0.0062 | 0.0048 | 0.0034 |
| \( p_{a_3} \) | 0.0011 | 0.0024 | 0.0013 | 0.0013 | 0.0008 | 0.0005 |
| \( p_{a_4} \) | 0.0006 | 0.0006 | 0.0002 | 0.0001 | 0.0001 | 0 |
| \( p_{a_5} \) | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| \( p_{a_6} \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( p_{a_7} \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( p_{a_8} \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( p_{a_9} \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( p_{a_{10}} \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( \tilde{p}_{\text{hit}} \) | \( \approx 1 \) | \( \approx 1 \) | \( \approx 1 \) | \( \approx 1 \) | \( \approx 1 \) | \( \approx 1 \) |
Table 6  Estimates of $p_2$ with $10^4$ samples, $\tilde{p}_1$ and $\tilde{p}_2$ with $10^7$ samples, and $\tilde{p}_{hit}$ with $n = 10^4$ for random walk example in Section 3.2, where we use $A^1_i = \{A_i \cap \{(x + a_i)^\top \Sigma^{-1}(x + a_i) \geq 4a_i^\top \Sigma^{-1}a_i\}\}$ and $A^2_i = \{A_i \setminus \{(x + a_i)^\top \Sigma^{-1}(x + a_i) \geq 4a_i^\top \Sigma^{-1}a_i\}\}$.

| $\sigma$ | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
|----------|-----|------|------|------|------|------|
| $p_2$    | $3.85 \times 10^{-13}$ | $2.58 \times 10^{-12}$ | $9.77 \times 10^{-10}$ | $8.58 \times 10^{-18}$ | $2.32 \times 10^{-22}$ | $1.90 \times 10^{-39}$ |
| $p_2/p$  | $1.19 \times 10^{-9}$ | $2.15 \times 10^{-8}$ | $2.72 \times 10^{-10}$ | $1.08 \times 10^{-12}$ | $2.00 \times 10^{-16}$ | $1.89 \times 10^{-23}$ |
| $\tilde{p}_1$ | 0.5149 | 0.5121 | 0.5098 | 0.5078 | 0.5059 | 0.5038 |
| $\tilde{p}_2$ | $3 \times 10^{-7}$ | $5 \times 10^{-7}$ | $7 \times 10^{-7}$ | $1 \times 10^{-7}$ | $1 \times 10^{-7}$ | $4 \times 10^{-7}$ |
| $\tilde{p}_{hit}$ | 0.003 | 0.005 | 0.007 | 0.001 | 0.001 | 0.004 |

efficient if and only if $\min_{x \in A_i} (x + a_i)^\top \Sigma^{-1}(x + a_i) < 4a_i^\top \Sigma^{-1}a_i$. In other words, if we split the rare-event set $A_i$ into two parts, say $A^1_i = \{A_i \cap \{(x + a_i)^\top \Sigma^{-1}(x + a_i) \geq 4a_i^\top \Sigma^{-1}a_i\}\}$ and $A^2_i = \{A_i \setminus \{(x + a_i)^\top \Sigma^{-1}(x + a_i) \geq 4a_i^\top \Sigma^{-1}a_i\}\}$, then our probabilistically efficient estimator is asymptotically efficient for estimating $P(A^1_i)$ because $(x + a_i)^\top \Sigma^{-1}(x + a_i) \geq 4a_i^\top \Sigma^{-1}a_i$ for all $x \in A^1_i$. This implies that, with these choices of $A^1_i$ and $A^2_i$, we satisfy Condition 2 in Theorem 3 (since the IS estimator using the most significant dominating point is asymptotically efficient for estimating $P(A^1_i)$).

Next we check Conditions 1 and 3 in Theorem 3. We define $p_1 = P(A^1_i)$, $p_2 = P(A^2_i)$, $\tilde{p}_1 = \tilde{P}(A^1_i)$, and $\tilde{p}_2 = \tilde{P}(A^2_i)$ for our newly constructed $A^1_i$ and $A^2_i$. We first show $p_2/p \to 0$ and $n\tilde{p}_2 \to 0$. We note that $A_i \cap \{s_a^i(x - a) \geq 0\} \subseteq A^1_i$ because $\{s_a^i(x - a) \geq 0\} \subseteq \{(x + a_i)^\top \Sigma^{-1}(x + a_i) \geq 4a_i^\top \Sigma^{-1}a_i\}$. Hence we also have $A^2_i \subseteq A_i \setminus \{s_a^i(x - a) \geq 0\}$, which leads to $p_2 \leq p_2^{old}$ and $\tilde{p}_2 \leq \tilde{p}_2^{old}$ where $p_2^{old}$ and $\tilde{p}_2^{old}$ refer to the $p_2$ and $\tilde{p}_2$ evaluated using our old constructions $A^1_i = A_i \cap \{s_a^i(x - a) \geq 0\}$ and $A^2_i = A_i \setminus \{s_a^i(x - a) \geq 0\}$. By $p_2^{old}/p \to 0$ and $n\tilde{p}_2^{old} \to 0$ as $\gamma \to \infty$ from Theorem 5 we have $p_2/p \to 0$ and $n\tilde{p}_2 \to 0$. That is, our current new construction $A^1_i$ for Theorem 3 would satisfy the conditions therein, and we would like to numerically verify especially Conditions 1 and 3. Indeed, to estimate $p_2$, we construct an IS estimator that mixes the exponential tiltings towards the dominating points for $\mathcal{H}_i \cap \{(x + a_i)^\top \Sigma^{-1}(x + a_i) \leq 4a_i^\top \Sigma^{-1}a_i\}$ with $i = 1, \ldots, 10$. To estimate $\tilde{p}_1$ and $\tilde{p}_2$, we directly generate samples from the probabilistically efficient IS distribution. We define $\tilde{p}_{hit} = \tilde{P}$ (some of the $n$ samples hits $A^2_i$) and estimate $\tilde{p}_{hit}$ through $1 - (1 - \tilde{p}_2)^n$ with $n = 10^4$. The results are presented in Table 6. We observe that the values of $p_2/p$ are now extremely small (smaller than $10^{-8}$ in all cases). Furthermore, the values of $\tilde{p}_2$ are also small, which lead to $\tilde{p}_{hit} < 0.01$ in all cases when $\sigma$ varies from 0.2 to 0.3. These results now justify Conditions 1 and 3 of Theorem 3 and explain the good performance of our probabilistically efficient estimator in the experiment.

6.1.3. MNIST Example. Like Section 3.2, the experiment in Section 3.3 also uses a probabilistically efficient estimator based on the exponential tilting towards the most significant dominating point only, i.e., $X_\gamma$ distributed as $N(\gamma a_i, \gamma I)$. We define $A^1_i = A_i \cap \{s_a^i(x - a) \geq 0\}$ and $A^2_i = A_i \setminus \{s_a^i(x - a) \geq 0\}$, and the corresponding probabilities $p_1 = P(A_i \cap \{s_a^i(x - a) \geq 0\})$ and $p_2 = P(A_i \setminus \{s_a^i(x - a) \geq 0\})$ which are shown in Table 7. Note that in this MNIST example, the
the first 10 dominating points, i.e. \( A_{10} = A \cap \{ s_{10}^\top (x - a_{10}) \geq 0 \} \) and \( A_{2} = A \setminus \{ s_{2}^\top (x - a_{2}) \geq 0 \} \).

| \( \sigma \) | \( p_1 \) | \( p_1/p \) | \( p_2 \) | \( p_2/p \) | \( p \) | \( p/p \) | \( p_{a2}/p \) | \( p_{a3}/p \) | \( p_{a4}/p \) | \( p_{a5}/p \) | \( p_{a6}/p \) | \( p_{a7} \) | \( p_{a8}/p \) | \( p_{a9} \) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 | 9.18 \times 10^{-6} | 0.7846 | 2.52 \times 10^{-6} | 0.2154 | 1.17 \times 10^{-5} | 0.1074 | 4.12 \times 10^{-8} | 0.0035 | 1.24 \times 10^{-7} | 0.0106 | 2.48 \times 10^{-8} | 0.0021 | 3.60 \times 10^{-8} | 0.0031 |
| 0.19 | 3.34 \times 10^{-6} | 0.7920 | 8.77 \times 10^{-7} | 0.2080 | 4.22 \times 10^{-6} | 0.1173 | 1.22 \times 10^{-8} | 0.0029 | 4.45 \times 10^{-8} | 0.0105 | 9.09 \times 10^{-9} | 0.0022 | 1.31 \times 10^{-8} | 0.0031 |
| 0.18 | 1.02 \times 10^{-6} | 0.7892 | 2.72 \times 10^{-7} | 0.2108 | 1.29 \times 10^{-6} | 0.1245 | 2.93 \times 10^{-9} | 0.0023 | 1.26 \times 10^{-8} | 0.0098 | 2.81 \times 10^{-9} | 0.0022 | 3.28 \times 10^{-9} | 0.0025 |
| 0.17 | 2.54 \times 10^{-7} | 0.8069 | 6.09 \times 10^{-8} | 0.1931 | 3.15 \times 10^{-7} | 0.1313 | 1.32 \times 10^{-9} | 0.0042 | 3.22 \times 10^{-9} | 0.0102 | 7.01 \times 10^{-10} | 0.0026 | 8.11 \times 10^{-10} | 0.0026 |

The total number of dominating points is large and unknown. Thus we only present the contribution of the first 10 dominating points, i.e. \( a_1, a_2, ..., a_{10} \) in Table 7, denoted by \( p_{a2} = P(A \cap \{ s_{a2}^\top (x - a_2) \geq 0 \}) \), ..., \( p_{a10} = P(A \cap \{ s_{a10}^\top (x - a_{10}) \geq 0 \} \). Again, we estimate each of the probabilities \( p_1, p_{a2}, ..., p_{a10} \) using the IS estimator with the corresponding dominating point, i.e. the IS distribution is exponentially tilted using dominating points \( a_1, .., a_{10} \) respectively. We borrow the values of \( p \) from Table 15 where each estimate is computed using crude Monte Carlo, and we estimate \( p_2 \) through \( p_2 = p - p_1 \). We observe that the ratio \( p_{a2}/p \) decreases from 0.2154 to 0.1931 as we decrease the value of \( \sigma \) from 0.2 to 0.17, i.e., the problem becomes rarer. We also observe that some individual relative contribution slightly increases in this experiment. However, these increases do not affect the decreasing trend of the total relative contribution of the less significant dominating points. For example, \( p_{a2}/p \) and \( p_{a3}/p \) both increase slightly as \( \sigma \) decreases (from 0.1074 and 0.0035 with \( \sigma = 0.2 \) to 0.1313 and 0.0042 with \( \sigma = 0.17 \) respectively), but the relative contribution of the rest of the less significant dominating points (excluding the first 10) is 0.0812, 0.0658, 0.0633, and 0.0362 for \( \sigma = 0.2, 0.19, 0.18, 0.17 \) respectively, which vanishes fast as \( \sigma \) decreases.

Table 8 presents the estimates of probabilities \( \tilde{p}_1 = \tilde{P}(A \cap \{ s_{a}^\top (x - a) \geq 0 \}) \) and \( \tilde{p}_2 = \tilde{P}(A \setminus \{ s_{a2}^\top (x - a_{2}) \geq 0 \}) \) under the probabilistically efficient IS distribution. The probabilities \( \tilde{p}_{a2}, ..., \tilde{p}_{a10} \) defined by \( \tilde{p}_i = \tilde{P}\left(A \cap \{ s_{a_i}^\top (x - a_i) \geq 0 \} \setminus \bigcup_{j=1}^{i-1}\{ s_{a_j}^\top (x - a_j) \geq 0 \} \right) \) for \( i = 2, ..., 10 \), are also
shown to illustrate the contributions of the dominating points $a_2, \ldots, a_{10}$ for $\tilde{p}_2$. Again, we find that $\tilde{p}_2$ decreases from 0.0090 to 0.0085 as $\sigma$ decreases from 0.2 to 0.17, i.e., the problem becomes rarer. From the individual contribution, we observe that all the probabilities $\tilde{p}_{a_3}, \ldots, \tilde{p}_{a_{10}}$ decrease rapidly, except $\tilde{p}_{a_2}$ that slightly increases as $\sigma$ decreases. We use the value of $\tilde{p}_2$ to estimate probability $\tilde{p}_{hit} = \tilde{P}(\text{some of the } n \text{ samples hits } A \setminus \{s^\top_a (x - a) \geq 0\})$ through $1 - (1 - \tilde{p}_2)^n$. The last row in Table 8 presents the results with $n = 10^4$ (the sample size we use in Section 3). Like in Section 6.1.2, we observe that there are samples falling into $A \setminus \{s^\top_a (x - a) \geq 0\}$ with approximately probability 1 and hence this result cannot explain the good performance of our probabilistically efficient estimator.

Table 8 shows the contributions of the less significant dominating points for the MNIST example under the probabilistically efficient IS in Section 3.3 with $10^4$ samples.

| $\sigma$ | 0.2 | 0.19 | 0.18 | 0.17 |
|----------|-----|------|------|------|
| $\tilde{p}_1$ | 0.4728 | 0.4745 | 0.4760 | 0.4775 |
| $\tilde{p}_2$ | 0.0090 | 0.0087 | 0.0086 | 0.0085 |
| $\tilde{p}_{a_2}$ | 0.0069 | 0.007 | 0.0072 | 0.0072 |
| $\tilde{p}_{a_3}$ | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| $\tilde{p}_{a_4}$ | 0.0006 | 0.0006 | 0.0006 | 0.0006 |
| $\tilde{p}_{a_5}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_6}$ | 0.0003 | 0.0003 | 0.0001 | 0.0001 |
| $\tilde{p}_{a_7}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_8}$ | 0.0006 | 0.0003 | 0.0003 | 0.0003 |
| $\tilde{p}_{a_9}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{10}}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{hit}$ | $\approx 1$ | $\approx 1$ | $\approx 1$ | $\approx 1$ |

Similar to Section 6.1.2, we consider an alternative construction of $A_{\gamma}^1$ and $A_{\gamma}^2 = A_{\gamma} \setminus A_{\gamma}^1$ to explain our performance. From the proofs of Propositions 4 and 5, we know that our probabilistically efficient estimator is not asymptotically efficient if and only if $\min_{x \in A_{\gamma}} (x + a_1)^\top (x + a_1) < 4a_1^\top a_1$. We split the rare-event set $A_{\gamma}$ into two parts, namely $A_{\gamma}^1 = \{A_{\gamma} \cap \{s^\top_a (x + a_1) \geq 4a_1^\top a_1\}\}$ and $A_{\gamma}^2 = \{A_{\gamma} \setminus \{(x + a_1)^\top (x + a_1) \geq 4a_1^\top a_1\}\}$, and our probabilistically efficient estimator is asymptotically efficient for estimating the probabilities of $A_{\gamma}^1$ and $A_{\gamma} \cap \{s^\top_a (x - a) \geq 0\} \subseteq A_{\gamma}^1$. We define $p_1 = P(A_{\gamma}^1)$, $p_2 = P(A_{\gamma}^2)$, $\tilde{p}_1 = \tilde{P}(A_{\gamma}^1)$, and $\tilde{p}_2 = \tilde{P}(A_{\gamma}^2)$. The use of our newly constructed $A_{\gamma}^1$ for Theorem 3 can be theoretically shown to satisfy Conditions 1 and 3 therein like in Section 6.1.2. We now check the numerical values of $p_2$ and $\tilde{p}_2$ to verify these conditions. We use the mixture of all 100 dominating points as the IS distribution for estimating $p_2$. We generate $10^6$ samples for $\sigma$ varying from 0.17 to 0.2 and find no samples falling into $A_{\gamma}^2$, which indicates that $p_2$ (and hence $p_2/p$) is extremely small in all cases. We generate $10^6$ samples from the probabilistically efficient IS distribution to estimate $\tilde{p}_2$ and observe no samples hitting $A_{\gamma}^2$ for the same range of $\sigma$. In this case, $\tilde{p}_{hit} = \tilde{P}(\text{some of the } n \text{ samples hits } A_{\gamma}^2)$ with $n = 10^4$ would be close to zero due to the extremely
small values of $\tilde{p}_2$. These results match Conditions 1 and 3 of Theorem 3 and hence explain the good performance of our probabilistically efficient estimator in the experiment.

6.1.4. Two-sided Overshoot Probability of a Random Walk. So far we have considered examples on strongly probabilistically efficient estimators. Here, we consider an additional example where we use a weakly probabilistically efficient estimator. We follow the problem setting in Section 3.2, where we consider the overshoot probability of the finite-horizon maximum of a random walk. However, we modify the probability of interest as

$$p = P \left( \max_{m=1,\ldots,d} |S_m| \geq a \right),$$

where we replace $S_m = \sum_{i=1}^m Y_i$ by its absolute value. The rest of the settings are the same as in Section 3.2, i.e., we have $Y_i$'s are Gaussian distributed with mean 0, standard deviation $\sigma$, pairwise correlation $-0.02$, and rarity parameter $\gamma = 1/\sigma^2 \rightarrow \infty$. The target rare event is

$$\left\{ \frac{1}{\gamma} X, \in \left( \bigcup_{m=1}^d \mathcal{H}_m^+ \right) \bigcup \left( \bigcup_{m=1}^d \mathcal{H}_m^- \right) \right\}$$

where $X_i = \gamma(Y_1, \ldots, Y_d)^\top$, $\mathcal{H}_m^+ = \{ x \in \mathbb{R}^d : \sum_{i=1}^m x_i \geq a \}$, and $\mathcal{H}_m^- = \{ x \in \mathbb{R}^d : \sum_{i=1}^m x_i \leq -a \}$ with $x_i$ denoting the $i$th element in $x$. In this case, the rate function is still $I(y) = \frac{1}{2} y^\top \Sigma^{-1} y$ and there are two most significant dominating points $a_1 = \frac{a \Sigma e_d}{e_d \Sigma e_d}$ and $-a_1 = -\frac{a \Sigma e_d}{e_d \Sigma e_d}$ where $e_d \in \mathbb{R}^d$ denotes the vector with 1 in all $d$ elements.

For this experiment, we first introduce an asymptotically efficient estimator. From Proposition 1, we know that the IS estimator using dominating points $a_1, \ldots, a_d, -a_1, \ldots, -a_d$ with $a_1, \ldots, a_d$ defined in Section 3.2 is asymptotically efficient. Next, we show the IS estimator using the dominating point $a_1$ is a weakly probabilistically efficient estimator and is not asymptotically efficient:

**Theorem 11.** Under the problem specifications in Section 3.2 and rare-event probability defined in (14), the IS estimator that use only one of the most significant dominating points, namely $a_1$ in Section 3.2, and $X, \gamma$ distributed as $N(\gamma a_1, \gamma \Sigma)$ in Section 3.2, is weakly probabilistically efficient but not asymptotically efficient.

Compared to the one-sided overshoot example in Section 3.2, here the rare-event set has two most significant dominating points $a_1$ and $-a_1$. Because we only use the first one instead of mixing both of the most significant dominating points in our IS, we only have $p_1/p \rightarrow 1/2$ instead of 0 and thus weak probabilistic efficiency instead of strong probabilistic efficiency holds as guided by Theorem 4.

To empirically verify Theorem 4, let us consider a partition of the rare event set $\mathcal{A}_\gamma = \left\{ \left( \bigcup_{m=1}^d \mathcal{H}_m^+ \right) \bigcup \left( \bigcup_{m=1}^d \mathcal{H}_m^- \right) \right\}$, where we have $\mathcal{A}_1 = \{ A \cap \{(x + a_1)^\top \Sigma^{-1} (x + a_1) \geq 4a_1^\top \Sigma^{-1} a_1 \} \}$ and $\mathcal{A}_2 = \{ A \setminus \{(x + a_1)^\top \Sigma^{-1} (x + a_1) \geq 4a_1^\top \Sigma^{-1} a_1 \} \}$. We define $p_1 = P(\mathcal{A}_1)$, $p_2 = P(\mathcal{A}_2)$, $\tilde{p}_1 = \tilde{P}(\mathcal{A}_1)$, and $\tilde{p}_2 = \tilde{P}(\mathcal{A}_2)$, where $\tilde{P}$ is the IS distribution exponentially tilted using the dominating point $a_1$. In our experiments, we set $d = 10$, fix $a_1$ and vary $\sigma$ for different rarity levels. For each case, we
We observe that in all cases the values of \( \tilde{p} \) of \( \tilde{\gamma} \) asymptotically efficient estimator with 10 magnitude of the rare-event probability, i.e., the estimates are around 0.5 of the estimates given in all considered cases, the estimates have relatively tight CIs and provide a good estimation on the weakly probabilistically efficient estimator. The results are presented in Table 9. We observe that generate 10\(^3\) samples from IS distributions using the above asymptotically efficient estimator and weakly probabilistically efficient estimator. The results are presented in Table 9. We observe that although our weakly probabilistically efficient estimator underestimates the rare-event probability in all considered cases, the estimates have relatively tight CIs and provide a good estimation on the magnitude of the rare-event probability, i.e., the estimates are around 0.5 of the estimates given by the asymptotically efficient estimator.

In Table 10, we investigate the numerical values of \( p_1, p_2, \tilde{p}_1 \) and \( \tilde{p}_2 \) and check the values of \( p_1/p \) and \( \tilde{p}_{hit} = \tilde{P}(\text{some of the n samples hits } \mathcal{A}_n^\gamma) \) with \( n = 10^4 \). We estimate \( p_1 \) and \( p_2 \) using the asymptotically efficient estimator with \( 10^7 \) independently generated samples. For the estimation of \( \tilde{p}_1 \) and \( \tilde{p}_2 \), we generate \( 10^7 \) samples using the weakly probabilistically efficient IS distribution. We observe that in all cases the values of \( p_1/p \) are very close to \( 1/2 \). On the other hand, the probabilities of falling into \( \mathcal{A}_n^\gamma \) are all below \( 10^{-6} \), which lead to \( \tilde{p}_{hit} \) valued smaller than 0.01. These results verify the conditions in Theorem 4 that explain the weak probabilistic efficiency of the IS estimator.

### 6.2. Illustration of Confidence Intervals

We investigate the performances of the CIs proposed in Section 5. In particular, we construct CIs (12) and (13) from probabilistically efficient estimators, namely the IS schemes using only the most significant point in Sections 3.1, 3.2 and 3.3. For convenience, we call interval (12) the “loose CI” and interval (13) the “tight CI”, since the latter has a shorter length and matches the CLT-based interval. For comparison, we also construct CI (13) from asymptotically efficient estimators. In particular, for the settings in Sections 3.1 and 3.2, these estimators are built from mixtures of
Table 11  Highly accurate point estimates (and 95% CI) using asymptotically efficient estimators for the problem in Section 3.1. The estimates are computed with $10^7$ samples for $m = 10, 30, 50$ and $5 \times 10^6$ samples for $m = 100$.

| $m$  | $\hat{p}$            | $\hat{q}$            | $\hat{r}$            |
|------|----------------------|----------------------|----------------------|
| 10   | $8.85(\pm0.0084) \times 10^{-9}$ | $1.58(\pm0.0021) \times 10^{-8}$ | $3.76(\pm0.0056) \times 10^{-5}$ |
| 30   |                      |                      | $1.34(\pm0.0034) \times 10^{-14}$ |

exponential tiltings towards all the dominating points. For the setting in Section 3.3, computing all dominating points requires insurmountable resources (as discussed therein), and so we use the mixture of 100 dominating points as a proxy of an asymptotically efficient estimator (100 is the total number of dominating points we discover using one-week’s computation).

In the experiments, we compare the coverage rates of all three intervals described above. These coverage rates are obtained from a large number of experimental repetitions. Since the ground truths of these problems are unknown, we run a gigantic amount of simulation runs using either asymptotically efficient estimators or crude Monte Carlo to obtain highly accurate estimates, which serve as the “truths” when estimating the coverage of the CIs. The exact number of simulation runs used in our ISs, number of experimental repetitions, and number of runs to approximate the ground truths are specified in the discussion of each example below.

6.2.1. Large Deviations of an I.I.D. Sum. For the experiment in Section 3.1, we use the asymptotically efficient estimator $\hat{\beta}(m)$ to obtain highly accurate estimates for all values of $m$ as the ground truths. These estimates are presented in Table 11. Our probabilistically efficient estimator $\hat{\alpha}(m)$ is computed using $10^4$ independently generated samples. From this, we apply CIs (12) and (13). We also construct CIs (13) using asymptotically efficient estimator $\hat{\beta}(m)$ (which is used to approximate the ground truth) with $10^4$ independently generated samples. We approximate the coverage rates using $10^5$ experimental iterations. Moreover, we compute the average CI width for each type of CIs. The experiment results are shown in Table 12.

From Table 12, we observe that the coverage rates of tight CIs by our probabilistically efficient estimator are close to 95% in three out of the four cases, but is 3% below 95% in one case ($m = 10$). On the other hand, the loose CIs are valid but perform conservatively with more than 99% coverage rates and wider average widths in all cases. The tight CIs by asymptotically efficient estimators provide valid coverage in all four cases. In the problems with rarer probabilities (i.e. $m = 30, 50, 100$), the tight CIs by probabilistically efficient estimators perform similarly as the CIs by asymptotically efficient estimators in terms of both CI width and coverage. This shows the competitiveness of CIs using probabilistically efficient estimators for rarer problems.

6.2.2. Overshoot Probability of a Random Walk. For the experiment in Section 3.2, we use the asymptotically efficient estimator that mixes all dominating points to approximate the
Table 12  Coverage rates and average CI widths of the loose confidence intervals ("Loose CI by PE"), the tight confidence intervals ("Tight CI by PE") for probabilistically efficient estimators, and the tight confidence intervals for asymptotically efficient estimators ("Tight CI by AE") in the experiments of Section 3.1.

|                | Coverage Rate | Average Width         |
|----------------|---------------|-----------------------|
| Loose CI by PE | 0.998         | 9.12 × 10⁻⁴           |
|                | 0.999         | 2.31 × 10⁻⁶           |
|                | 0.9994        | 6.37 × 10⁻⁹           |
|                | 0.999         | 2.81 × 10⁻¹⁵          |
| Tight CI by PE | 0.921         | 5.30 × 10⁻⁴           |
|                | 0.951         | 1.31 × 10⁻⁶           |
|                | 0.950         | 3.54 × 10⁻⁹           |
|                | 0.950         | 1.52 × 10⁻¹⁵          |
| Tight CI by AE | 0.950         | 5.30 × 10⁻⁴           |
|                | 0.960         | 1.30 × 10⁻⁶           |
|                | 0.949         | 3.55 × 10⁻⁹           |
|                | 0.950         | 1.52 × 10⁻¹⁵          |

ground truths presented in Table 13. Our probabilistically efficient estimator is computed using $10^4$ independently generated samples. We construct CIs (12) and (13) based on this estimator. For comparison we also construct CI (13) from asymptotically efficient estimator using $10^4$ samples independently generated from the ones used to approximate the ground truth. We use $10^5$ experimental repetitions to estimate the coverage rates of all CIs. The coverage rates and average widths are presented in Table 14.

From Table 14, the loose CIs perform conservatively with near to 99% coverage rates in all cases. On the other hand, the tight CIs constructed from our probabilistically efficient estimators have coverage rates below 95% in most of the cases. Moreover, as $\sigma$ decreases (the probability become rarer), the coverage rates first drop from around 0.93 (with $\sigma = 0.3$) to around 0.89 (with $\sigma = 0.2$), then they improve as $\sigma$ further decreases and reaches around 95% when $\sigma = 0.1$. The tight CIs by asymptotically efficient estimators have more stable coverage rates than the CIs by probabilistically efficient estimators, but also suffer under-coverage in several cases (e.g., 0.86 with $\sigma = 0.28$). We also observe that the tight CIs by the probabilistically efficient estimators have better average widths than the CIs by the asymptotically efficient estimators with smaller $\sigma$ (e.g., $\sigma = 0.1, 0.12$). The results show the validity of the CIs with probabilistically efficient estimators as $\sigma \to 0$, but also that the coverage rate may not always monotonically improve as the problem becomes rarer.

6.2.3. MNIST Example. For the experiment in Section 3.3, we use $10^{10}$ runs of crude Monte Carlo to approximate the ground truths, which are shown in Table 15. Note that the estimate for $\sigma = 0.17$ is relatively less accurate than other estimates, revealed by the CI width in the magnitude of around 0.1 of the probability estimate. We obtain our probabilistically efficient estimator by generating $10^4$ independent samples and construct CIs (12) and (13) based on this estimator. Since locating all dominating points to construct asymptotically efficient estimator is computationally infeasible in this example, we use IS estimators that mix the most significant 100 dominating points (the number of dominating points obtained from our sequential mixed integer programming procedure in Algorithm 2) as a proxy. We construct CI (13) from this estimator using $10^4$ samples.
Table 13  High accurate point estimates (and 95% CI) using asymptotically efficient estimators for the problem in Section 3.2. The estimates are computed with $10^7$ samples.

| $\sigma$ | $p$       |
|---------|-----------|
| 0.1     | $(5.57 \pm 0.039) \times 10^{-26}$ |
| 0.12    | $(1.30 \pm 0.008) \times 10^{-18}$ |
| 0.14    | $(3.82 \pm 0.020) \times 10^{-14}$ |
| 0.16    | $(3.18 \pm 0.015) \times 10^{-11}$ |
| 0.18    | $(3.32 \pm 0.014) \times 10^{-9}$ |
| 0.2     | $9.51(\pm 0.035) \times 10^{-8}$  |
| 0.22    | $7.93(\pm 0.004) \times 10^{-6}$  |
| 0.24    | $1.55(\pm 0.010) \times 10^{-5}$  |
| 0.26    | $3.59(\pm 0.024) \times 10^{-6}$  |
| 0.28    | $1.20(\pm 0.003) \times 10^{-4}$  |
| 0.3     | $3.22(\pm 0.009) \times 10^{-4}$  |

Table 14  Coverage rates and average CI widths of the loose confidence intervals ("Loose CI by PE") and tight confidence intervals ("Tight CI by PE") for probabilistically efficient estimators, and the tight confidence intervals for asymptotically efficient estimators ("Tight CI by AE") in the experiments of Section 3.2.

| $\sigma$ | Loose CI by PE | Tight CI by PE | Tight CI by AE |
|---------|----------------|----------------|---------------|
|         | Coverage | Width         | Coverage | Width         | Coverage | Width         |
| 0.1     | 0.9997   | $1.49 \times 10^{-26}$ | 0.949     | $7.88 \times 10^{-27}$ | 0.926     | $2.41 \times 10^{-26}$ |
| 0.12    | 0.999    | $3.58 \times 10^{-19}$ | 0.938     | $1.99 \times 10^{-19}$ | 0.917     | $5.08 \times 10^{-19}$ |
| 0.14    | 0.997    | $1.07 \times 10^{-14}$ | 0.914     | $6.15 \times 10^{-15}$ | 0.951     | $1.32 \times 10^{-14}$ |
| 0.16    | 0.994    | $9.37 \times 10^{-12}$ | 0.897     | $5.54 \times 10^{-12}$ | 0.908     | $9.67 \times 10^{-12}$ |
| 0.18    | 0.990    | $9.89 \times 10^{-19}$ | 0.892     | $5.95 \times 10^{-19}$ | 0.951     | $8.72 \times 10^{-19}$ |
| 0.2     | 0.988    | $2.86 \times 10^{-8}$  | 0.895     | $1.75 \times 10^{-8}$  | 0.964     | $2.24 \times 10^{-8}$  |
| 0.22    | 0.988    | $3.15 \times 10^{-7}$  | 0.901     | $1.93 \times 10^{-7}$  | 0.936     | $2.45 \times 10^{-7}$  |
| 0.24    | 0.990    | $2.20 \times 10^{-6}$  | 0.913     | $1.35 \times 10^{-6}$  | 0.959     | $1.55 \times 10^{-6}$  |
| 0.26    | 0.991    | $9.67 \times 10^{-8}$  | 0.918     | $6.00 \times 10^{-8}$  | 0.932     | $6.49 \times 10^{-8}$  |
| 0.28    | 0.992    | $2.99 \times 10^{-3}$  | 0.924     | $1.86 \times 10^{-3}$  | 0.861     | $2.01 \times 10^{-3}$  |
| 0.3     | 0.993    | $7.71 \times 10^{-8}$  | 0.927     | $4.80 \times 10^{-8}$  | 0.956     | $4.90 \times 10^{-8}$  |

We use $10^5$ experimental repetitions to estimate the coverage rates and average widths of the CIs from probabilistically efficient estimators and $10^3$ repetitions for the CIs from IS estimators using 100 dominating points (we use repetition size $10^3$ instead of $10^5$ because of the long computational time caused by a large number of mixtures in the IS distribution). The results are presented in Table 16.

From Table 16, we observe that in three out of the four cases, the tight CIs constructed from probabilistically efficient estimators provide coverage rates that are slightly below 95%. Similar to the previous random walk overshoot problem, the coverage rates are closer to 95% for rarer problems (e.g., the coverage is 94.9% for $\sigma = 0.17$). The under-coverage is alleviated when we use more than one dominating point, as shown in the row of “Tight CI by AE” (where we use 100 dominating points). On the other hand, the loose CIs have higher than nominal coverage rates in all cases, but are conservative since the rates are around 97.5% – 99.5%. Again, we observe the
Table 15  Highly accurate point estimates (and 95% CI) using asymptotically efficient estimators for the problem in Section 3.3. The estimates are computed with $10^{10}$ samples.

| $\sigma$ | $p$ | \(\hat{p}_3\) | \(\pm 0.15\times 10^{-7}\) | \(\pm 0.031\times 10^{-6}\) | \(\pm 0.056\times 10^{-6}\) | \(\pm 0.0094\times 10^{-6}\) |
|----------|-----|----------------|----------------|----------------|----------------|----------------|
| 0.17     | 3.15(±0.15)\times 10^{-7} | 1.29(±0.031)\times 10^{-6} | 4.22(±0.056)\times 10^{-6} | 1.17(±0.0094)\times 10^{-6} |
| 0.18     | 4.43(±0.3)\times 10^{-7} |
| 0.19     | 5.42(±0.58)\times 10^{-7} |
| 0.2      | 6.40(±0.76)\times 10^{-7} |

Table 16  Coverage rates and average CI widths of the loose confidence intervals ("Loose CI by PE"), the tight confidence intervals ("Tight CI by PE") for probabilistically efficient estimators, and the tight confidence intervals for asymptotically efficient estimators ("Tight CI by AE") in the experiments of Section 3.3.

| | $\sigma$ | Coverage Rate | Average Width |
|---|----------|---------------|---------------|
| Loose CI by PE | 0.17 | 0.996 | $6.32 \times 10^{-8}$ |
| | 0.18 | 0.978 | $2.61 \times 10^{-7}$ |
| | 0.19 | 0.980 | $1.13 \times 10^{-6}$ |
| | 0.2 | 0.977 | $3.20 \times 10^{-6}$ |
| Tight CI by PE | 0.17 | 0.949 | $4.19 \times 10^{-8}$ |
| | 0.18 | 0.874 | $1.73 \times 10^{-7}$ |
| | 0.19 | 0.885 | $7.50 \times 10^{-7}$ |
| | 0.2 | 0.877 | $2.12 \times 10^{-6}$ |
| Tight CI by AE | 0.17 | 0.958 | $4.77 \times 10^{-8}$ |
| | 0.18 | 0.933 | $1.96 \times 10^{-7}$ |
| | 0.19 | 0.945 | $6.13 \times 10^{-7}$ |
| | 0.2 | 0.951 | $1.90 \times 10^{-6}$ |

validity of the CIs with probabilistically efficient estimators as the rare-event probability decreases, which validates our analysis.

6.2.4. Summary of Experimental Observations on Confidence Interval Construction. From the CI construction for the three examples in Section 3.1 investigated above, we draw several conclusions: 1) Tight CIs by probabilistically efficient estimators appear to have close to the nominal coverage rate when the problem is rare enough; 2) Loose CIs by probabilistically efficient estimators tend to over-cover, and also have correspondingly larger widths than other methods; 3) Tight CIs by asymptotically efficient estimators appear to give more accurate coverage rates for a larger range of rarity levels than the tight CIs by probabilistically efficient estimators, even though they could still under-cover in some cases; 4) When tight CIs by probabilistically efficient and asymptotically efficient estimators both have accurate coverage rates, their widths appear to be comparable. Overall, it appears that tight CIs by asymptotically efficient estimators are more robust with respect to the rarity level of the problem, which is also in line with the comparison between Theorems 8 and 9 (recall the discussion right after Theorem 9). Nonetheless, recall that one motivation of us proposing the notion of probabilistic efficiency is that asymptotically efficient estimators, which require using more dominating points in their mixtures, could be computationally challenging to construct.

7. Future Work

We conclude the paper with further discussions on the potential risk of the current framework and some future directions.
7.1. Developing Diagnosis for Finite-Sample Under-Estimation

Similar to the established notion of asymptotic efficiency in the rare-event simulation literature, our probabilistic efficiency framework is asymptotic. For a fixed rarity parameter $\gamma$ and given simulation size $n$ in practice, more work needs to be investigated to judge whether the obtained estimate is reliable or not. In particular, a risk of missing dominating points is finite-sample under-estimation. To be more specific, we look at the following example.

**Example 1.** Suppose that our goal is to estimate $p = P(\frac{1}{\gamma}X_\gamma \in (-\infty, -k] \cup [1, \infty))$ where $X_\gamma \sim N(0, \gamma)$ and $k > 1$. In this case, the most significant dominating point is 1, so we consider choosing $X_\gamma \sim N(\gamma, \gamma)$ as the IS distribution. That is, the IS estimator is $Z = I(\frac{1}{\gamma}X_\gamma \in (-\infty, -k] \cup [1, \infty))e^{-X_\gamma + \gamma/2}$ with $X_\gamma \sim N(\gamma, \gamma)$. We generate independent samples $Z^{(1)}, \ldots, Z^{(n)}$ and use $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} Z_i$ to estimate $p$. When $1 < k < 3$, the relative error of $Z$ grows exponentially in $\gamma$ (or $-\log p$), so $Z$ is not asymptotically efficient (see the proof of Proposition 2). On the other hand, from our previous derivations, $Z$ is strongly probabilistically efficient, so we can still get a reliable estimate when $\gamma$ is sufficiently large. However, in the finite-sample case where $k$ is close to 1 and $\gamma$ is not large, if $n$ is chosen as a moderate size, then it could happen that $\{X \leq -k\gamma\}$ is not hit and hence we would get an estimate close to $p_1 = P(\frac{1}{\gamma}X_\gamma \geq 1)$, but $p_2 = P(\frac{1}{\gamma}X_\gamma \leq -k)$ is not negligible compared to $p_1$. In other words, we under-estimate $p$. As a specific example, let $k = 1.01$ and $\gamma = 16$. Then $\hat{p}_2 = \hat{P}(\frac{1}{16}X_{16} \leq -1.01) = \Phi(8.04) \approx 4.44 \times 10^{-16}$. In this case, for a moderate $n$ like 1000, with probability almost 1 the set $\{\frac{1}{16}X_{16} \leq -1.01\}$ is never hit. However, $p_2/p_1 = \Phi(4.04)/\Phi(4) \approx 0.84$. This means with probability almost 1, we under-estimate $p$ by about $0.84/(1 + 0.84) \approx 46\%$.

Example 1 shows that under finite parameter value and finite sample, it could be hard to tell whether we can safely drop less significant dominating points. A good aspect about the conclusion in this example, however, is that the under-estimation is arguably acceptable in relative term (i.e., the estimate is still in the same magnitude as the ground truth), pointing to an estimation resembling weak probabilistic efficiency. We should emphasize that a similar concern applies to asymptotically efficient estimators as well. That is, it is difficult to guarantee whether the sample size $n$ is large enough to give a reliable estimate for a given setup and rarity parameter value. Nevertheless, increasing $n$ would improve the performance of asymptotically efficient estimators, but for probabilistic efficiency, we do not have the luxury of increasing $n$ since our framework requires $n$ to be moderate in size. This points to more need of developing diagnostic methods to detect under-estimation due to finite-sample effects in the future.

7.2. Further Developing Theory of IS with Missed Dominating Points

Despite the presence of under-estimation risks described above, we maintain our motivation of probabilistic efficiency as a theory to allow one to use few dominating points in problems where
finding all of them is infeasible. In fact, what we have focused on in this paper is only one theory where dropping dominating points is valid, among other possibilities. To support this, we revisit Example 1:

**Example 2 (Example 1 continued).** Consider the problem setting in Example 1. When $k \geq 3$, the relative error of $Z$ grows only polynomially in $\gamma$ (or $-\log p$), and hence $Z$ is asymptotically efficient (see the proof of Proposition 2). The intuitive explanation is that $\tilde{p}_2 = \hat{P}(\frac{1}{\gamma}X_\gamma \leq -k)$ is extremely small which mitigates the blow-up of the likelihood ratio. More rigorously, when $\frac{1}{\gamma}X_\gamma \leq -k$ we have $Z \geq e^{(k+\frac{1}{2})\gamma}$, but $\hat{P}(\frac{1}{\gamma}X_\gamma \leq -k) = \bar{\Phi}((k+1)/\sqrt{\gamma}) = \Theta(\frac{1}{\sqrt{\gamma}}e^{-(k+1)^2/2\gamma})$. Overall $\tilde{E}(Z^2) = \Theta(\frac{1}{\sqrt{\gamma}}e^{-\gamma})$, and hence $\tilde{E}(Z^2)/p^2 = \Theta(\sqrt{\gamma})$ which grows polynomially in $-\log p = \Theta(\gamma)$.

Example 2 shows that sometimes the missed dominating points are so rare that missing them does not even harm the asymptotic efficiency. Here, explaining the validity of IS with missed dominating points does not require probabilistic efficiency, but instead an alternate analysis of asymptotic efficiency that is tighter than the standard approach in the literature. Nonetheless, this phenomenon does not apply to our proposed estimators in Sections 3 and 6, as we have mathematically verified the asymptotic inefficiency in our considered estimator in each example.

More generally, we conclude our paper with Figure 4, which shows the relations among different efficiency criteria and IS with missed dominating points. In this paper, we have built sufficient conditions to achieve strong probabilistic efficiency with missed dominating points. Example 2, on the other hand, gives an example in achieving asymptotic efficiency with missed dominating points. Our immediate future endeavor is to fill in the regions in Figure 4 that are not covered by the current work, including the scenario depicted by Example 2, relaxing the current sufficient conditions of probabilistic efficiency to allow $\tilde{p}_2$ to be less tiny, achieving weak instead of strong probabilistic efficiency with missed dominating points, and moreover, to understand the “complementary” regions where missing dominating points would be guaranteed to violate the efficiency notions. In summary, this work serves as a first step in a new line of analysis aiming to relax existing variance-based efficiency criteria in rare-event simulation to be applicable to larger-scale and more complex problems.

**Acknowledgments**

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

**References**

Arief M, Huang Z, Koushik Senthil Kumar G, Bai Y, He S, Ding W, Lam H, Zhao D (2021) Deep probabilistic accelerated evaluation: A robust certifiable rare-event simulation methodology for black-box safety-critical systems. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 595–603 (PMLR).
Figure 4  Relationships among different efficiency criteria and IS estimators missing dominating points.

Asmussen S (1985) Conjugate processes and the simulation of ruin problems. Stochastic Processes and their Applications 20(2):213–229.

Asmussen S, Albrecher H (2010) Ruin Probabilities, volume 14 of Advanced Series on Statistical Science & Applied Probability (World Scientific).

Asmussen S, Glynn PW (2007) Stochastic Simulation: Algorithms and Analysis, volume 57 (Springer Science & Business Media).

Bai Y, Huang Z, Lam H, Zhao D (2022) Rare-event simulation for neural network and random forest predictors. ACM Transactions on Modeling and Computer Simulation (TOMACS) 32(3):1–33.

Blanchet J, Glynn P (2008) Efficient rare-event simulation for the maximum of heavy-tailed random walks. The Annals of Applied Probability 18(4):1351–1378.

Blanchet J, Glynn P (2009) Efficient rare event simulation of continuous time markovian perpetuities. Proceedings of the 2009 Winter Simulation Conference (WSC), 444–451 (IEEE).

Blanchet J, Glynn P, Lam H (2009) Rare Event Simulation for a Slotted Time $M/G/s$ Model. Queueing Systems 63(1):33–57.

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

Blanchet J, Lam H (2014) Rare-event simulation for many-server queues. Mathematics of Operations Research 39(4):1142–1178.

Blanchet J, Lam H, Zwart B (2012) Efficient rare-event simulation for perpetuities. Stochastic Processes and their Applications 122(10):3361–3392.

Blanchet J, Mandjes M (2007) Editorial: Rare-event simulation for queues. Queueing Systems 57(2-3):57–59.

Bucklew JA (2004) Introduction to Rare Event Simulation. Springer Series in Statistics (Springer New York).
Carrasco JA (1992) Failure distance-based simulation of repairable fault-tolerant systems. *Computer performance evaluation: modelling techniques and tools: proceedings of the Fifth International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, Torino, Italy, 13-15 February 1991*, 351–365 (Elsevier).

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

Chen JC, Lu D, Sadowsky JS, Yao K (1993) On importance sampling in digital communications. I. fundamentals. *IEEE Journal on Selected Areas in Communications* 11(3):289–299.

Dembo A, Zeitouni O (2009) *Large deviations techniques and applications*, volume 38 (Springer Science & Business Media).

Ding W, Chen B, Li B, Eun KJ, Zhao D (2021) Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters* 6(2):1551–1558.

Dupuis P, Sezer AD, Wang H (2007) Dynamic importance sampling for queueing networks. *The Annals of Applied Probability* 17(4):1306–1346.

Dupuis P, Wang H (2009) Importance sampling for jackson networks. *Queueing Systems* 62(1):113–157.

Glasserman P (2003) *Monte Carlo Methods in Financial Engineering*, volume 53 of *Stochastic Modelling and Applied Probability* (Springer New York).

Glasserman P, Kang W, Shahabuddin P (2008) Fast simulation of multifactor portfolio credit risk. *Operations Research* 56(5):1200–1217.

Glasserman P, Li J (2005) Importance sampling for portfolio credit risk. *Management science* 51(11):1643–1656.

Glasserman P, Wang Y (1997) Counterexamples in importance sampling for large deviations probabilities. *Annals of Applied Probability* 7(3):731–746.

Goyal A, Shahabuddin P, Heidelberger P, Nicola VF, Glynn PW (1992) A unified framework for simulating markovian models of highly dependable systems. *IEEE Transactions on Computers* 41(1):36–51.

Heidelberger P (1995) Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 5:43–85.

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

Huang Z, Lam H, Zhao D (2018) Designing importance samplers to simulate machine learning predictors via optimization. *2018 Winter Simulation Conference (WSC)*, 1730–1741 (Institute of Electrical and Electronics Engineers, Inc).

Hult H, Svensson J (2012) On importance sampling with mixtures for random walks with heavy tails. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 22(2):1–21.
Juneja S, Shahabuddin P (2006) Chapter 11 rare-event simulation techniques: An introduction and recent advances. *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, 291–350 (Elsevier).

Kesidis G, Walrand J, Chang CS (1993) Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources. *IEEE/ACM Transactions on Networking* 1(4):424–428.

Kratsios M (2019) *The national artificial intelligence research and development strategic plan: 2019 update* (National Science and Technology Council (US)).

Kroese DP, Nicola VF (1999) Efficient estimation of overflow probabilities in queues with breakdowns. *Performance Evaluation* 36-37:471–484.

L’Ecuyer P, Blanchet JH, Tuffin B, Glynn PW (2010) Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation* 20(1):1–41.

Lewis EE, Böhm F (1984) Monte carlo simulation of markov unreliability models. *Nuclear engineering and design* 77(1):49–62.

Maurer A, Pontil M (2009) Empirical bernstein bounds and sample variance penalization. *COLT 2009 - The 22nd Conference on Learning Theory*.

O’ Kelly M, Sinha A, Namkoong H, Tedrake R, Duchi JC (2018) Scalable end-to-end autonomous vehicle testing via rare-event simulation. *Advances in Neural Information Processing Systems*, volume 31 (Curran Associates, Inc.).

Ridder A (2009) Importance sampling algorithms for first passage time probabilities in the infinite server queue. *European Journal of Operational Research* 199(1):176–186.

Rubino G, Tuffin B (2009) *Rare Event Simulation using Monte Carlo Methods* (John Wiley & Sons, Ltd).

Sadowsky JS (1991) Large Deviations Theory and Efficient Simulation of Excessive Backlogs in a $GI/GI/m$ Queue. *IEEE Transactions on Automatic Control* 36(12):1383–1394.

Sadowsky JS, Bucklew JA (1990) On Large Deviations Theory and Asymptotically Efficient Monte Carlo Estimation. *IEEE Transactions on Information Theory* 36(3):579–588.

Shahabuddin P (1994) Importance sampling for the simulation of highly reliable markovian systems. *Management Science* 40(3):333–352.

Siegmund D (1976) Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics* 4(4):673–684.

Smith PJ, Shafi M, Gao H (1997) Quick simulation: A review of importance sampling techniques in communications systems. *IEEE journal on selected areas in communications* 15(4):597–613.

Szechtman R, Glynn PW (2002) Rare-event simulation for infinite server queues. *Proceedings of the Winter Simulation Conference*, volume 1, 416–423 (Institute of Electrical and Electronics Engineers, Inc).
Tuffin B (2004) On numerical problems in simulations of highly reliable markovian systems. *First International Conference on the Quantitative Evaluation of Systems, 2004. QEST 2004. Proceedings.*, 156–164 (IEEE).

Tuffin B, Ridder A (2012) Probabilistic bounded relative error for rare event simulation learning techniques. *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12 (Institute of Electrical and Electronics Engineers, Inc).

Webb S, Rainforth T, Teh YW, Kumar MP (2018) A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*.

Zhang XW, Glynn PW, Giesecke K, Blanchet J (2009) Rare event simulation for a generalized hawkes process. *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 1291–1298 (IEEE).

Zhao D, Huang X, Peng H, Lam H, LeBlanc DJ (2018) Accelerated evaluation of automated vehicles in car-following maneuvers. *IEEE Transactions on Intelligent Transportation Systems* 19(3):733–744.

Zhao D, Lam H, Peng H, Bao S, LeBlanc DJ, Nobukawa K, Pan CS (2016) Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques. *IEEE transactions on intelligent transportation systems* 18(3):595–607.
Appendix A: Algorithms

Algorithm 1 shows a procedure to obtain dominating sets. We briefly explain the idea here. First we minimize $I(x)$ over $x \in \mathcal{E}$ and get the first dominating point $a_1$. If $\mathcal{E} \subset \{x \in \mathbb{R}^d : s_{a_1}^\top (x-a_1) \geq 0\}$, then we know that $\{a_1\}$ is a dominating set and hence we could stop. Otherwise, we minimize $I(x)$ over $x \in \mathcal{E} \setminus \{x \in \mathbb{R}^d : s_{a_1}^\top (x-a_1) \geq 0\}$ to get the second dominating point $a_2$. Then we check whether $\mathcal{E} \subset \bigcup_{i=1}^r \{x \in \mathbb{R}^d : s_{a_i}^\top (x-a_i) \geq 0\}$. By repeating this process, we would finally get a dominating set $A = \{a_1, \ldots, a_r\}$ with $I(a_1) \leq \cdots \leq I(a_r)$.

Algorithm 1: Sequentially find all the dominating points.

**Input:** Rarity parameter $\gamma$, rare-event set $\mathcal{E} \subset \mathbb{R}^d$, rate function $I(y)$, function $a \mapsto s_a$.

**Output:**Dominating set $A$.

1. Start with $A = \emptyset$, $r = 0$;
2. **While** $\{x \in \mathbb{R}^d : x \in \mathcal{E}, s_{a_i}^\top (x-a_i) < 0, \forall i = 1, \ldots, r\} \neq \emptyset$ **do**
3. Find a dominating point $a_{r+1}$ by solving the optimization problem
\[
a_{r+1} = \arg \min_x I(x) \quad \text{s.t.} \quad x \in \mathcal{E}, \quad s_{a_i}^\top (x-a_i) < 0, \forall i = 1, \ldots, r
\]
and update $A \leftarrow A \cup \{a_{r+1}\}$, $r \leftarrow r + 1$.
4. **End**

Appendix B: A New Alternative Asymptotic Regime and Theoretical Guarantees

In this section, we investigate a new regime that $A_\gamma = \{X \in \mathcal{E}\}$ and $\mathcal{E} = \{x \in \mathbb{R}^d : g(x) \geq \gamma\}$ where $X \in \mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}$ is a function. We propose this setting as it arises as a generic representation of recent problems in intelligent system safety testing (Bai et al. 2022). There, $g$ could be highly complicated and leads to a gigantic number of dominating points, which in turn motivates the consideration of dropping most of them and our notion of probabilistic efficiency. We note that technically this setting is slightly different from the classical Gartner-Ellis regime in terms of the position of the scaling parameter $\gamma$ (see, e.g. Sadowsky and Bucklew 1990), but conceptually similar.

In Section B.1, we adapt the notions of rate function and dominating points to this new regime. Then we state the assumptions under which we could build on Theorem 3 to obtain reliable point estimates and CIs as in Section 5. We note that this new regime is harder to analyze as the rare-event set $\mathcal{E}$ can change in a complicated way with $\gamma$, and thus inevitably our assumptions are relatively restrictive and need to be verified case by case. In Section B.2, we consider the special (but important) case where $X$ follows a Gaussian distribution and $g$ is piecewise linear, in particular propose a simple stopping strategy to determine whether it is safe to stop searching for the remaining dominating points. Throughout this section, we write $\alpha(\gamma) \sim \beta(\gamma)$ if $\alpha(\gamma)/\beta(\gamma)$ is subexponential in $-\log p(\gamma)$.
B.1. Guarantees for General Input Distribution

We define $\mu(x) = \log E e^{x^\top X}, x \in \mathbb{R}^d$ as the cumulant generating function of $X$ and $I(y) = \sup_{x \in \mathbb{R}^d} \{x^\top y - \mu(x)\}, y \in \mathbb{R}^d$ as its Legendre transform. For any set $\mathcal{E} \subset \mathbb{R}^d$, we denote $I(\mathcal{E}) = \inf_{y \in \mathcal{E}} I(y)$. Parallel to Assumptions 1 and 2, we make the following two assumptions.

**Assumption 3.** $\mu(x)$ satisfies the following conditions:

1. $0 \in \mathcal{D}(\mu) \circ$;
2. $\mu$ is essentially smooth, i.e., $\mathcal{D}(\mu) \circ$ is non-empty, $\mu$ is differentiable everywhere in $\mathcal{D}(\mu) \circ$ and $\mu$ is steep.

**Assumption 4.** For any $\gamma$, $\mathcal{E} = \mathcal{E}(\gamma) \subset \mathbb{R}^d$ is a Borel set such that $\mathcal{E} = \overline{\mathcal{E}} \setminus \mathcal{E} \cap \mathcal{D}(\mu) \circ \neq \emptyset$ and $I(\mathcal{E}) > 0$.

The concepts of dominating set and dominating points are also similar, but now they change with $\gamma$:

**Definition 5 (Dominating Set (New Regime)).** Suppose that Assumptions 3 and 4 hold. We call $A = A(\gamma) \subset \overline{\mathcal{E}}$ a dominating set for $\mathcal{E}$ if

1. For each $a \in A$, there exists a unique $s_a \in \mathbb{R}^d$ such that $\nabla \mu(s_a) = a$;
2. $\mathcal{E} \subset \bigcup_{a \in A} \{x \in \mathbb{R}^d : s_a^\top(x - a) \geq 0\}$;
3. $A \setminus \{a\}$ does not satisfy the first two conditions for any $a \in A$.

We call any point in $A$ a dominating point. For two dominating points $a$ and $a'$, we say $a$ is more significant than $a'$ if $I(a) < I(a')$.

Suppose that $A = \{a_1, \ldots, a_r\}$ is a dominating set. Then the corresponding mixture IS distribution is given by

$$\frac{d\tilde{P}}{dP}(\omega) = \sum_{i=1}^r \alpha_i e^{s_{a_i}^\top X - \mu(s_{a_i})}$$

with $\sum_{i=1}^r \alpha_i = 1, \alpha_i > 0, \forall i$. Like the discussion in Section 2, we could split the rare-event set as $\mathcal{E} = \bigcup_{i=1}^r \mathcal{E}_i$ where $\mathcal{E}_i$'s are disjoint and $a_i \in \overline{\mathcal{E}_i} \subset \{x \in \mathbb{R}^d : s_{a_i}^\top(x - a_i) \geq 0\}$. Note that $s_{a_i}^\top(x - a_i) \geq 0$ still implies that $I(x) \geq I(a_i)$. Now, the likelihood ratio is given by

$$L(\omega) := \frac{dP}{d\tilde{P}}(\omega) = \frac{1}{\sum_{i=1}^r \alpha_i e^{s_{a_i}^\top X - \mu(s_{a_i})}},$$

and it satisfies that for any $i$,

$$L(\omega) \leq \frac{1}{\alpha_i e^{s_{a_i}^\top X - \mu(s_{a_i})}} = \frac{1}{\alpha_i} e^{-s_{a_i}^\top(X - a_i) - I(a_i)}.$$

Hence we have that $I_{\mathcal{E}_i}(X)L(\omega) \leq \frac{1}{\alpha_i} e^{-I(a_i)}$. Then

$$\overline{\mbox{Var}}(I(X \in \mathcal{E})L(\omega)) \leq \bar{E}(I(X \in \mathcal{E})L^2(\omega)) \leq \frac{1}{(\min_i \alpha_i)^2} e^{-2 \min_i I(a_i)}.$$  

(15)

Thus, supposing we have a large deviations asymptotic given by $p = P(X \in \mathcal{E}) \sim e^{-\min_i I(a_i)}$, then combining with (15) will give us that the IS estimator is asymptotically efficient.

Now we consider using a partial list of dominating points. In particular, we can still sequentially fill in the dominating set $A = \{a_1, \ldots, a_r\}$ where $a_i = \arg \min \{I(y) : y \in \mathcal{E}, s_{a_j}^\top(y - a_j) < 0, j = 1, \ldots, i - 1\}$ and hence $I(a_1) \leq \cdots \leq I(a_r)$, and suppose that we have a stopping strategy $k = k(\gamma)$ with $1 \leq k \leq r$ before locating all the dominating points. We choose the mixture IS distribution given by

$$\frac{d\tilde{P}}{dP}(\omega) = \sum_{i=1}^k \frac{1}{k} e^{s_{a_i}^\top X - \mu(s_{a_i})}.$$  

(16)
Unlike in (11), we no longer assign an individual weight \( \alpha_i \) to each dominating point \( a_i \). Instead, we only consider the uniform mixture for simplicity, since now the number of dominating points \( r \) and \( k \) can both potentially change with \( \gamma \).

First of all, we summarize the basic assumptions on the problem setting to ensure that there exists a dominating set with moderate size:

**Assumption 5.** Consider the problem of estimating \( p = P(X \in \mathcal{E}) \) with \( \mathcal{E} = \{ x \in \mathbb{R}^d : g(x) \geq \gamma \} \) where \( X \in \mathbb{R}^d \) is a random vector and \( g \) is a function. Assume that

1. \( p \to 0 \) as \( \gamma \to \infty \);
2. Assumption 3 holds for the cumulant generating function of \( X \) under \( P \);
3. Assumption 4 holds for \( \mathcal{E} \);
4. For any \( \gamma \), there exists a dominating set for \( \mathcal{E} \), denoted as \( \mathcal{A} = \{ a_1, \ldots, a_r \} \) where \( a_i = \text{arg min}\{ I(y) : y \in \mathcal{E}, s_{a_i}^T(y - a_j) < 0, j = 1, \ldots, i - 1 \} \) and hence \( I(a_1) \leq \cdots \leq I(a_r) \). Besides, \( r = r(\gamma) \) grows at most subexponentially in \( -\log p \).

Now consider a stopping strategy \( k = k(\gamma) \) and the corresponding IS distribution (16). Under Assumption 5, for \( k = k(\gamma) \) with \( 1 \leq k \leq r \), denote \( \mathcal{E}_1 = \mathcal{E} \cap \bigcup_{i=1}^{k} \{ x \in \mathbb{R}^d : s_{a_i}^T(x - a_i) \geq 0 \} \) and \( \mathcal{E}_2 = \mathcal{E} \setminus \mathcal{E}_1 \). Corresponding to the settings in Theorem 3, we let \( \mathcal{A}_i = \{ X \in \mathcal{E}_i \}, i = 1, 2 \). We introduce Assumptions 6 and 7.

**Assumption 6.** If \( k = r \), then let \( a_{k+1} = \infty \mathbf{1}_d \). Assume that \( p_1 := P(\mathcal{A}_1) = P(X \in \mathcal{E}_1) \sim e^{-I(a_1)} \) and that \( p_2 := P(\mathcal{A}_2) = P(X \in \mathcal{E}_2) \) is upper bounded by \( e^{-I(a_{k+1})} \) up to subexponential factor in \( -\log p \). Besides, \( e^{I(a_1)-I(a_{k+1})} \) exponentially decays in \( -\log p \).

**Assumption 7.** Assume that \( \hat{p}_2 := \hat{P}(\mathcal{A}_2) = \hat{P}(X \in \mathcal{E}_2) \) exponentially decays in \( -\log p \).

Roughly, Assumption 6 implies that \( p_2 \) is exponentially smaller than \( p_1 \) and Assumption 7 implies that \( \mathcal{E}_2 \) is hardly hit even under \( \hat{P} \). Applying Theorem 3, we have that

**Theorem 12 (Attaining probabilistic efficiency with a partial list of dominating points (new regime)).** Under Assumptions 5, 6 and 7, the IS estimator \( Z = I(X \in \mathcal{E}) \frac{dP}{d\hat{P}}(\omega) \) under \( \hat{P} \) given by (16) is strongly probabilistically efficient.

Similar to Section 5, we can also construct asymptotically valid CIs with the sample mean and sample variance.

**Theorem 13 (Constructing confidence intervals with probabilistically efficient estimators (new regime)).** Assume that Assumptions 5, 6 and 7 hold. The IS estimator is \( Z = I(X \in \mathcal{E}) \frac{dP}{d\hat{P}}(\omega) \) under \( \hat{P} \) given by (16). We sample \( X^{(1)}, \ldots, X^{(n)} \) i.i.d. from \( \hat{P} \) and let \( Z^{(i)} = I(X^{(i)} \in \mathcal{E}) \frac{dP}{d\hat{P}}, Z_1^{(i)} = I(X^{(i)} \in \mathcal{E}_1) \frac{dP}{d\hat{P}}, i = 1, \ldots, n \). Use \( \hat{P} \) and \( \hat{\nu} \) to respectively denote the sample mean and sample variance of \( Z^{(i)} \)'s. In this case, If \( n \) is subexponentially growing in \( -\log p \) as \( \gamma \to \infty \). Then, for any \( 0 < \alpha < 1 \),

\[
\liminf_{\gamma \to \infty} \hat{P} \left( \left| \hat{p} - p \right| \leq \sqrt{\frac{2\hat{\nu}\log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}} \right) \geq 1 - \alpha.
\]
That is,

\[ \hat{p} \pm \left( \sqrt{\frac{2\hat{V} \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}} \right) \]

is an asymptotically valid \( (1 - \alpha) \)-level CI for \( p \).

Similar to Section 5, the CI in Theorem 13 is more conservative than the CLT-based interval \( \hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{V}{n}} \). If additionally the following assumption holds, then the CLT-based CI is also asymptotically valid:

**Assumption 8.** Denote \( Z_1 = I(X \in E_1) \frac{f(X)}{f(X)} \) under \( \hat{P} \). Assume that \( \hat{V} \operatorname{ar}(Z_1) \sim e^{-2I(a_1)} \).

Similar to Lemma 2, Assumption 8 serves to control the Berry-Esseen error bound. To be more concrete, with this additional assumption, we have:

**Theorem 14 (Constructing tight confidence intervals (new regime)).** Assume that Assumptions 5, 6, 7 and 8 hold. The IS estimator is \( Z = I(X \in E) \frac{dP}{dP}(\omega) \) under \( \hat{P} \) given by (16). We sample \( X^{(1)}, \ldots, X^{(n)} \) i.i.d. from \( \hat{P} \) and let \( Z^{(i)} = I(X^{(i)} \in E) \frac{dP}{dP}, Z_1^{(i)} = I(X^{(i)} \in E_1) \frac{dP}{dP}, i = 1, \ldots, n \). Use \( \hat{p} \) and \( \hat{V} \) to respectively denote the sample mean and sample variance of \( Z^{(i)} \)'s. In this case, we could choose \( n \) subexponentially growing in \( -\log p \) such that \( \frac{k^2e^{-2I(a_1)}}{n\hat{V} \operatorname{ar}(Z_1^{(i)})^{1/2}} \rightarrow 0 \) and \( \frac{E[z^{(1)} - p_1]^3}{n\hat{V} \operatorname{ar}(Z_1^{(i)})} \rightarrow 0 \) as \( \gamma \rightarrow \infty \). Then, for any \( 0 < \alpha < 1 \),

\[ \liminf_{\gamma \rightarrow \infty} \hat{P} \left( |\hat{p} - p| \leq z_{1-\alpha/2} \sqrt{\frac{\hat{V}}{n}} \right) \geq 1 - \alpha. \]

That is,

\[ \hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{V}}{n}} \]

is an asymptotically valid \( (1 - \alpha) \)-level CI for \( p \).

In the above, Assumptions 5, 6, 7 and 8 are technical conditions that need to be analyzed case by case. The next Section B.2 focuses on a specific but important problem setting where we show how to verify all these conditions.

**B.2. Guarantees for Gaussian Input Distribution**

Suppose \( X \sim N(\lambda, \Sigma) \) under \( P \) where \( \lambda \in \mathbb{R}^d \) and \( \Sigma \in \mathbb{R}^{d \times d} \) is positive definite. In this case \( \mu(x) = \lambda^\top x + \frac{1}{2} x^\top \Sigma x, s_a = \Sigma^{-1}(a - \lambda) \) and \( I(x) = \frac{1}{2}(x - \lambda)^\top \Sigma^{-1}(x - \lambda) \). Moreover, we suppose that \( g \) is a piecewise linear function. Related theoretical guarantees for piecewise linear function of Gaussian input can be found in Bai et al. (2022).

We use the following natural stopping strategy. First, fix a constant \( C > 1 \). Then, we sequentially find the dominating points where we stop as long as \( (a_{k+1} - \lambda)^\top \Sigma^{-1}(a_{k+1} - \lambda) > C(a_k - \lambda)^\top \Sigma^{-1}(a_k - \lambda) \). If there is no such \( k \), then we let \( k = r \), the number of all dominating points. The IS distribution is chosen as \( X \sim \frac{1}{k} \sum_{i=1}^{k} \phi(x; a_i, \Sigma) \). We summarize these in Algorithm 2.

We have the following theorem suggesting that under the above setting, all the assumptions listed in Section B.1 are satisfied.
Algorithm 2: Estimate $P(g(X) \geq \gamma)$ with Gaussian input $X$ and piecewise linear function $g$.

**Input:** Piecewise linear function $g$, rarity parameter $\gamma$, input distribution $N(\lambda, \Sigma)$, threshold $C > 1$, sample size $n$.

**Output:** IS estimate $\hat{p}$.

1. Start with $k = 0$;
2. While $\{x : g(x) \geq \gamma, (a_i - \lambda)^\top \Sigma^{-1} (x - a_i) < 0, \forall i = 1, \ldots, k\} \neq \emptyset$ do
3. Find a dominating point $a_{k+1}$ by solving the optimization problem

   $$a_{k+1} = \arg \min_x (x - \lambda)^\top \Sigma^{-1} (x - \lambda)$$

   s.t. $g(x) \geq \gamma$, $(a_i - \lambda)^\top \Sigma^{-1} (x - a_i) < 0, \forall i = 1, \ldots, k$.

4. If $k > 0$ and $(a_{k+1} - \lambda)^\top \Sigma^{-1} (a_{k+1} - \lambda) > C(a_k - \lambda)^\top \Sigma^{-1} (a_k - \lambda)$ do
5.   Break
6. Else do
7.   Update $k \leftarrow k + 1$;
8. End
9. Sample $X_1, \ldots, X_n$ from the mixture distribution $\frac{1}{k} \sum_{i=1}^k \phi(x; a_i, \Sigma)$.
10. Compute the IS estimate $\hat{p} = \frac{1}{n} \sum_{i=1}^n I(g(X_i) \geq \gamma) L(X_i)$ where the likelihood ratio function is

    $$L(x) = \frac{e^{-\frac{1}{2} (x-\lambda)^\top \Sigma^{-1} (x-\lambda)}}{\frac{1}{k} \sum_{i=1}^k e^{-\frac{1}{2} (x-a_i)^\top \Sigma^{-1} (x-a_i)}}.$$

---

**Theorem 15** (Verification of assumptions for Gaussian inputs). Suppose that $X \sim N(\lambda, \Sigma)$ under $P$ where $\lambda \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite, $\mathcal{E} = \{x \in \mathbb{R}^d : g(x) \geq \gamma\}$ where $g$ is a piecewise linear function, $C > 1$ is a constant. Assume that for any $\gamma$, $P(X \in \mathcal{E}) > 0$ and $\lambda \notin \overline{\mathcal{E}}$. For any $\gamma$, we use the stopping strategy described above (more precisely Algorithm 2) to sequentially find dominating points $a_1, a_2, \ldots, a_k$ in decreasing significance and set up the IS distribution $X \sim \frac{1}{k} \sum_{i=1}^k \phi(x; a_i, \Sigma)$. Denote $\mathcal{E}_1 = \mathcal{E} \cap \bigcup_{i=1}^k \{x \in \mathbb{R}^d : (a_i - \lambda)^\top \Sigma^{-1} (x - a_i) \geq 0\}$ and $\mathcal{E}_2 = \mathcal{E} \setminus \mathcal{E}_1$. Then Assumptions 5, 6, 7 and 8 hold.

Note that in Theorem 15, all assumptions including Gaussianity and piecewise linear $g$ are all straightfor-ward to verify. With Theorem 15, we thus get the following corollaries from Theorems 12, 13 and 14:

**Corollary 1** (Probabilistic efficiency for Gaussian inputs). Under the settings in Theorem 15, the IS estimator $Z = I(X \in \mathcal{E}) \frac{dP}{d\tilde{P}}$ is strongly probabilistically efficient.
Corollary 2 (Confidence intervals for Gaussian inputs). Under the settings in Theorem 15, we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)} = I(X^{(i)} \in \mathcal{E}) \frac{d\tilde{P}}{dP}$. Use $\tilde{p}$ and $\tilde{V}$ to denote the sample mean and sample variance of $Z^{(i)}$’s. We could choose $n$ subexponentially growing in $-\log p$ such that for any $0 < \alpha < 1$,
\[
\liminf_{\gamma \to \infty} \tilde{P} \left( |\tilde{p} - p| \leq \sqrt{\frac{2V \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)ke^{-\frac{1}{4}(a_1 - \lambda)\Sigma^{-1}(a_1 - \lambda)}}{3(n-1)} \right) \geq 1 - \alpha.
\]

Corollary 3 (Tight confidence intervals for Gaussian inputs). Under the settings in Theorem 15, we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)} = I(X^{(i)} \in \mathcal{E}) \frac{d\tilde{P}}{dP}$. Use $\tilde{p}$ and $\tilde{V}$ to denote the sample mean and sample variance of $Z^{(i)}$’s. We could choose $n$ subexponentially growing in $-\log p$ such that for any $0 < \alpha < 1$,
\[
\liminf_{\gamma \to \infty} \tilde{P} \left( |\tilde{p} - p| \leq z_{1-\alpha/2} \sqrt{\frac{\tilde{V}}{n}} \right) \geq 1 - \alpha.
\]

While the choice of $C$ (as long as it is $>1$) does not affect the guarantee on strong probabilistic efficiency or the asymptotic validity of the CIs, it does affect the accuracy of the estimates for a given, finite $\gamma$. In particular, it is often the case that we only need to solve one problem for a fixed $\gamma$ instead of solving a series of problems with varying $\gamma$. In this scenario, as long as $(a_2 - \lambda)^\top \Sigma^{-1}(a_2 - \lambda) > (a_1 - \lambda)^\top \Sigma^{-1}(a_1 - \lambda)$, we can pick a $C > 1$ such that we would stop our dominating point search at $k = 1$, i.e., use only the first point. Note that this does not imply that for this choice of $C$ we have $k = 1$ for any $\gamma$. Thus, the finiteness of the rarity parameter comes into play in a subtle way.

Appendix C: Proofs

Proof of Theorem 2. First, we know that $I(a_i) = s_i^\top a_i - \mu(s_i)$ for $i = 1, \ldots, r$. For any $x \in \mathbb{R}^d$ such that $s_i^\top (x - a_i) \geq 0$, we have $I(x) \geq s_i^\top x - \mu(s_i) \geq s_i^\top a_i - \mu(s_i) = I(a_i)$. By Definition 2, we have $\mathcal{E} \subset \bigcup_{i=1}^r \{x \in \mathbb{R}^d : s_i^\top (x - a_i) \geq 0\}$. Thus, for any $x \in \mathcal{E}$, there exists $i$ such that $s_i^\top (x - a_i) \geq 0$, and hence $I(x) \geq I(a_i)$. Therefore, $I(\mathcal{E}) = \inf_{x \in \mathcal{E}} I(x) \geq \min_{i=1, \ldots, r} I(a_i)$. On the other hand, $a_1, \ldots, a_r \in \partial\mathcal{E} \cap \partial (\mathcal{E})^c$ and $I$ is differentiable in $\mathcal{E}^c$, so $I(\mathcal{E}) \leq \min_{i=1, \ldots, r} I(a_i)$. Overall we have $I(\mathcal{E}) = \min_{i=1, \ldots, r} I(a_i)$. \hfill \Box

Proof of Proposition 1. We could split the rare-event set as $\mathcal{E} = \bigcup_{i=1}^r \mathcal{E}_i$, where $\mathcal{E}_i$’s are disjoint and $a_i \in \mathcal{E}_i \subset \{x \in \mathbb{R}^d : s_i^\top (x - a_i) \geq 0\}$. Note that the hyperplane $\{x \in \mathbb{R}^d : s_i^\top (x - a_i) = 0\}$ is tangent to the rate function level set $\{x \in \mathbb{R}^d : I(x) = I(a_i)\}$ at $a_i$, so $s_i^\top (x - a_i) \geq 0$ implies that $I(x) \geq I(a_i)$. Now, the likelihood ratio is given by
\[
L(X_\gamma) := \frac{dP}{d\tilde{P}}(X_\gamma) = \frac{1}{\sum_{i=1}^r \alpha_i e^{s_i^\top X_\gamma - \gamma \mu_\gamma(s_i)}}
\]
and it satisfies that for any $i$,
\[
L(X_\gamma) \leq \frac{1}{\alpha_i e^{s_i^\top X_\gamma - \gamma \mu_\gamma(s_i)}} = \frac{1}{\alpha_i} e^{-\frac{1}{\gamma} (\gamma s_i^\top X_\gamma - \gamma \mu_\gamma(s_i))}.
\]
Hence we have that
\[
I \left( \frac{1}{\gamma} X_\gamma \in \mathcal{E}_i \right) L(X_\gamma) \leq \frac{1}{\alpha_i} e^{-\gamma (s_i^\top a_i - \mu_\gamma(s_i))}.
\]  
(17)
and
\[ I \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) L(X_{\gamma}) \leq \max_{t = 1, \ldots, r} \left\{ \frac{1}{\alpha_t} e^{-\gamma(x_t^\top a_t - \lambda y_t(s_{a_t}))} \right\}. \]

Thus, using Theorem 1, we have that
\[
\liminf_{\gamma \to \infty} \frac{\log E \left( I \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) L^2(X_{\gamma}) \right)}{\log E \left( I \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) L(X_{\gamma}) \right)} \geq \liminf_{\gamma \to \infty} \frac{2 \log \left( \max_{t = 1, \ldots, r} \left\{ \frac{1}{\alpha_t} e^{-\gamma(x_t^\top a_t - \lambda y_t(s_{a_t}))} \right\} \right)}{\log \left( \frac{1}{\gamma} \log P \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) \right)}
\]
\[
= \liminf_{\gamma \to \infty} \frac{2 \max_{t = 1, \ldots, r} \left\{ \frac{1}{\alpha_t} e^{-\gamma(x_t^\top a_t - \lambda y_t(s_{a_t}))} \right\}}{\frac{1}{\gamma} \log \left( \frac{1}{\gamma} \log P \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) \right)}
\]
\[
= -2 \min_{t = 1, \ldots, r} I(a_t) = 2,
\]
which verifies that the IS estimator is asymptotically efficient.

Proof of Proposition 2. We derive the growth rate of the relative error for a more general case. Suppose that we estimate \( p = P(\frac{1}{\gamma} X_{\gamma} \in (-\infty, -k] \cup [1, \infty)) \) where \( k > 1, X_{\gamma} \sim N(0, \gamma) \) under \( P \) and the IS distribution is chosen as \( X_{\gamma} \sim N(\gamma, \gamma) \) under \( \hat{P} \). This is the setup of Examples 1 and 2. There are two dominating points, 1 and \(-k\), and 1 is the more significant one. Under the IS distribution, the likelihood ratio is \( L = \exp^{-X_{\gamma}^2/\gamma} \) and the IS estimator is \( I = I \left( \frac{1}{\gamma} X_{\gamma} \in (-\infty, -k] \cup [1, \infty) \right) \exp^{-X_{\gamma}^2/\gamma} \). Then we have that
\[
\hat{E}(Z^2) = E \left( I \left( \frac{1}{\gamma} X_{\gamma} \in (-\infty, -k] \cup [1, \infty) \right) \exp^{-2X_{\gamma}^2/\gamma} \right)
\]
\[
= \int_{x \leq -k \gamma} e^{-2x^2/\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\gamma)^2}{2\gamma}} dx
\]
\[
= \int_{x \leq -k \gamma} e^{-\frac{(x+\gamma)^2}{2\gamma}} dx
\]
\[
= \int_{y \leq (1-k)\sqrt{\gamma}} e^{\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2\pi}} dy \quad (y = (x+\gamma)/\sqrt{\gamma})
\]
\[
= e^{\gamma} (\Phi((1-k)/\sqrt{\gamma}) + \Phi(2\sqrt{\gamma}))
\]
where \( \Phi \) denotes the tail distribution function of standard normal distribution. It is known that \( \Phi(x) = \Theta(\frac{1}{x} e^{-x^2/2}) \) as \( x \to \infty \), and hence \( \hat{E}(Z^2) = \Theta(\frac{1}{\gamma} e^{(1-k^{1/2})\gamma}) \) if \( 1 < k < 3 \) and \( \hat{E}(Z^2) = \Theta(\frac{1}{\gamma} e^{-\gamma}) \) if \( k \geq 3 \). Besides, \( p = \Phi(\sqrt{\gamma}) + \Phi(k\sqrt{\gamma}) = \Theta(\frac{1}{\sqrt{\gamma}} e^{-\gamma/2}) \). Therefore, \( \hat{E}(Z^2)/p^2 = \Theta(\sqrt{\gamma} e^{(2-k^{1/2})\gamma}) \) which grows exponentially in \( \gamma \) if \( 1 < k < 3 \) and \( \hat{E}(Z^2)/p^2 = \Theta(\sqrt{\gamma}) \) which grows polynomially in \( \gamma \) if \( k \geq 3 \).

Proofs of Propositions 4 and 5. We first show a general result on IS estimator that exponentially tilts to the most significant dominating point. Consider a rare-event set \( \mathcal{E} \) with \( a_1 \) as the most significant dominating point. We estimate the target rare event \( \{ \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \} \) using the IS estimator with likelihood ratio
\[
L = \frac{1}{e^{s_{a_1}^\top X_{\gamma} - \gamma \mu_{a_1}(s_{a_1})}}.
\]
That is, we use exponential tilting with respect to \( a_1 \) only. Specifically, we consider \( X_{\gamma} \sim N(\gamma \lambda, \gamma \Sigma) \). In this case we have \( \mu(x) = \mu_{a_1}(x) = x^\top \lambda + \frac{\lambda}{2} x^\top \Sigma x \), \( s_{a_1} = \Sigma^{-1}(a - \lambda) \), \( I(y) = \frac{1}{2} (y - \lambda)^\top \Sigma^{-1} (y - \lambda) \).

The second moment of this IS estimator is
\[
\hat{E} \left( I \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) L^2(\omega) \right)
\]
\[ \begin{align*}
\text{where } \hat{P} \text{ is the probability measure given by the exponential tilting with respect to } 2\lambda - a_1 \text{ and hence } X_{\gamma} \sim N(\gamma(2\lambda - a_1), \gamma \Sigma) \text{ under } \hat{P}. \text{ We correspondingly denote } \hat{I} \text{ as the rate function under } \hat{P}. \text{ Now, since } X_{\gamma} \text{ is Gaussian Assumption 1 holds. Suppose also that Assumption 2 holds. Then by Theorem 1 we know that}

\[ \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \hat{P} \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) = -\hat{I}(\mathcal{E}) = -\frac{1}{2} \min_{y \in \mathcal{E}} (y + a_1 - 2\lambda)^\top \Sigma^{-1} (y + a_1 - 2\lambda). \]

\text{We also know that } \lim_{\gamma \to \infty} \frac{1}{\gamma} \log P \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) = -I(a_1) = -\frac{1}{2} (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda). \text{ Therefore,}

\[ \lim_{\gamma \to \infty} \frac{\log \hat{E} \left( I \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) L^2(\omega) \right)}{\log P \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right)} = \lim_{\gamma \to \infty} \frac{\frac{1}{\gamma} \log \hat{E} \left( I \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right) L^2(\omega) \right)}{\frac{1}{\gamma} \log P \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right)} = \lim_{\gamma \to \infty} \frac{(a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda) - \frac{1}{2} \log \hat{P} \left( \frac{1}{\gamma} X_{\gamma} \in \mathcal{E} \right)}{\frac{1}{2} (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda) - \frac{1}{2} \min_{y \in \mathcal{E}} (y + a_1 - 2\lambda)^\top \Sigma^{-1} (y + a_1 - 2\lambda)}. \]

\text{By definition, the IS estimator is not asymptotically efficient if and only if } \min_{y \in \mathcal{E}} (y + a_1 - 2\lambda)^\top \Sigma^{-1} (y + a_1 - 2\lambda) < 4(a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda). \]

\text{In order to check the existence of } \hat{y} \text{ such that } (\hat{y} + a_1 - 2\lambda)^\top \Sigma^{-1} (\hat{y} + a_1 - 2\lambda) < 4(a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda), \text{ we can formulate the following optimization}

\[ \min_{y \in \mathcal{E}} (y + a_1 - 2\lambda)^\top \Sigma^{-1} (y + a_1 - 2\lambda) - 4(a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda), \quad (18) \]

\text{and check whether the objective is negative for the optimal (or any feasible) solution. Since the objective function is quadratic on the decision vector } y, \text{ the tractability of the optimization problem (18) is determined by the feasible region } \mathcal{E}. \text{ In the example in Section 3.2, we have } \bigcup_{m=1}^{d} \mathcal{H}_m = \bigcup_{m=1}^{d} \{ x \in \mathbb{R}^d : \sum_{i=1}^{m} x_i \geq a \}. \text{ Since each } \mathcal{H}_m \text{ is a half-space and hence convex, we can independently solve}

\[ \min_{y \in \mathcal{H}_m} (y + a_1 - 2\lambda)^\top \Sigma^{-1} (y + a_1 - 2\lambda) - 4(a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda), \]
for \( m = 1, \ldots, d \). Since the feasible region of the above optimization associated with any \( \mathcal{H}_m \) is a subset of the original problem (18), any solution with negative objective indicates that the IS estimator with dominating points \( a_1 \) is not asymptotically efficient. In the example in Section 3.3, we formulate the problem (18) as a mixed integer programming problem following the treatment of the rare-event set formed by machine learning predictors in Bai et al. (2022). Using this approach, we can show that the IS estimators with the most dominating points are not asymptotically efficient for both examples considered in Sections 3.2 and 3.3.

\[ \square \]

To prove Theorem 7, we need the following lemma:

**Lemma 1 (Theorem 4 in Maurer and Pontil 2009).** Let \( Y, Y_1, \ldots, Y_n \) be i.i.d. random variables with values in \([0,1]\) and let \( \delta > 0 \). Then with probability at least \( 1 - \delta \)

\[
EY - \frac{1}{n} \sum_{i=1}^{n} Y_i \leq \sqrt{\frac{2V_n(Y) \log(2/\delta)}{n}} + \frac{7 \log(2/\delta)}{3(n-1)}
\]

where \( Y = (Y_1, \ldots, Y_n) \) and \( V_n(Y) \) is the sample variance of \( Y_i \)'s.

Now we prove Theorem 7:

**Proof of Theorem 7.** Following the notation in Section 4, we split \( \mathcal{E} \) into \( \mathcal{E}_1 = \mathcal{E} \cap \bigcup_{i=1}^{k} \{ x \in \mathbb{R}^d : s_n^T(x - a_i) \geq 0 \} \) and \( \mathcal{E}_2 = \mathcal{E} \setminus \mathcal{E}_1 \), and then we correspondingly define \( A_1 = \{ \frac{1}{\gamma} X_i \in \mathcal{E}_1 \} \) and \( A_2 = \{ \frac{1}{\gamma} X_i \in \mathcal{E}_2 \} \). We denote \( Z_{1}^{(i)} = I(\frac{1}{\gamma} X_i \in \mathcal{E}_1) \) and \( \hat{V}_1 \) respectively denote the sample mean and sample variance of \( Z_{1}^{(i)} \)'s. We also define \( N = \sum_{i=1}^{n} I(\frac{1}{\gamma} X_i \in \mathcal{E}_2) \). Note that conditional on \( N = 0 \), we have \( Z^{(i)} = Z_{1}^{(i)} \). Then

\[
\begin{aligned}
\tilde{P} \left( |\hat{p} - p| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)M_{\gamma}}{3(n-1)} \right) \\
= \tilde{P} \left( |\hat{p} - p| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)M_{\gamma}}{3(n-1)}, N = 0 \right) \\
+ \tilde{P} \left( |\hat{p} - p| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)M_{\gamma}}{3(n-1)}, N > 0 \right) \\
\leq \tilde{P} \left( |\hat{p}_1 - p_1| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)M_{\gamma}}{3(n-1)}, N = 0 \right) + \tilde{P}(N > 0) \\
\leq \tilde{P} \left( |\hat{p}_1 - p_1| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)M_{\gamma}}{3(n-1)} \right) + \tilde{P}(N > 0) \\
\leq \tilde{P} \left( |\hat{p}_1 - p_1| + p_2 > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)M_{\gamma}}{3(n-1)} \right) + np_2.
\end{aligned}
\]

Similar to the derivation of (17), we know that \( 0 \leq Z_{1}^{(i)} \leq M_{\gamma} \), \( \forall i \). By applying Lemma 1 with \( Y_i = Z_{1}^{(i)}/M_{\gamma} \) and \( Y_i = 1 - Z_{1}^{(i)}/M_{\gamma} \) respectively, we get that

\[
\tilde{P} \left( p_1 > \hat{p}_1 + \sqrt{\frac{2\hat{V}_1 \log(4/\delta)}{n}} + \frac{7 \log(4/\delta)M_{\gamma}}{3(n-1)} \right) \leq \delta/2
\]
and
\[ \hat{P}\left(p_1 - \frac{2\hat{V}_1 \log(4/\delta)}{n} - \frac{7\log(4/\delta)M_n}{3(n-1)}\right) \leq \delta/2 \]
for any \( \delta > 0 \). Thus,
\[ \hat{P}\left(|\hat{p}_1 - p_1| > \sqrt{\frac{2\hat{V}_1 \log(4/\delta)}{n} + \frac{7\log(4/\delta)M_n}{3(n-1)}}\right) \leq \delta. \]  
(19)

Find \( \alpha' = \alpha'(\gamma) \) such that
\[ \frac{7\log(4/\alpha')M_n}{3(n-1)} = \frac{7\log(4/\alpha)M_n}{3(n-1)} - p_2. \]
That is,
\[ \alpha' = \alpha \exp\left(\frac{3(n-1)p_2}{7M_n}\right). \]
Clearly \( \alpha' > \alpha \) and \( \log(4/\alpha') < \log(4/\alpha) \). From the proof of Theorem 6, we have that either (i) \( k = r \) and \( p_2 = 0 \) or (ii) \( k < r \) and \( \limsup_{\gamma \to \infty} \frac{1}{\gamma} \log p_2 \leq -I(a_{k+1}) < -I(a_1) \). Moreover, we have
\[ \frac{1}{\gamma} \log M_n = \max_{i=1,...,k} \left\{ -\frac{1}{\gamma} \log \alpha_i - (s_a^\top a_i - \mu_s(s_a)) \right\} \to -I(a_1). \]
Thus, for subexponentially growing \( n \), we have \( \frac{3(n-1)p_2}{7M_n} \to 0 \) and hence \( \alpha' \to \alpha \) as \( \gamma \to \infty \). We replace \( \delta \) with \( \alpha' \) in (19), and then we get
\[ \hat{P}\left(|\hat{p}_1 - p_1| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha')}{n} + \frac{7\log(4/\alpha')M_n}{3(n-1)}}\right) \leq \alpha'. \]
Hence
\[ \hat{P}\left(|\hat{p}_1 - p_1| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)M_n}{3(n-1)} - p_2}\right) \leq \alpha'. \]
Therefore,
\[ \hat{P}\left(|\hat{p} - p| > \sqrt{\frac{2\hat{V}_1 \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)M_n}{3(n-1)}}\right) \leq \alpha' + n\hat{p}_2 \to \alpha \text{ as } \gamma \to \infty. \]

To prove Theorem 8, we first prove a lemma regarding the variance \( \overline{\text{Var}}(Z_1) \). In particular, we have assumed that \( \overline{\text{Var}}(Z_1) \) cannot be “too large”, while this lemma implies that it cannot be “too small” as well. As we will see, this lemma is used to control the error of the normal approximation in analog to the Berry-Esseen theorem in order to argue the validity of \( I_2 \).

**Lemma 2.** Under the same setting as Theorem 6, split \( \mathcal{E} \) into \( \mathcal{E}_1 = \mathcal{E} \cap \bigcup_{i=1}^k \{ x \in \mathbb{R}^d : s_a^\top (x - a_i) \geq 0 \} \) and \( \mathcal{E}_2 = \mathcal{E} \setminus \mathcal{E}_1 \). Denote \( Z_1 = I_1 X, \in \mathcal{E}_1 \) \( \frac{dP}{d\mu}(\omega) \) with \( X, \sim \hat{P} \). Then \( \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \overline{\text{Var}}(Z_1) = -2I(\mathcal{E}) = -2I(a_1) \).

**Proof of Lemma 2.** From Proposition 1, we know that \( \limsup_{\gamma \to \infty} \frac{1}{\gamma} \log \overline{\text{Var}}(Z_1) \leq -2I(\mathcal{E}) \). Thus, we only need to show that \( \liminf_{\gamma \to \infty} \frac{1}{\gamma} \log \overline{\text{Var}}(Z_1) \geq -2I(\mathcal{E}) \). Indeed, we have that \( \overline{\text{Var}}(Z_1) = \hat{E}((Z_1 - p_1)^2) = \hat{E}((Z_1 - p_1)^2 I(X, \notin \mathcal{E}_1)) = p_1^2 \hat{P}(\frac{1}{\gamma} X, \in \mathcal{E}_1) \geq p_1^2 \alpha_1 \hat{P}(\frac{1}{\gamma} X, \in \mathcal{E}_1) \) where \( \hat{P}_1 \) is as defined in the proof of Theorem 6, and hence \( \liminf_{\gamma \to \infty} \frac{1}{\gamma} \log \overline{\text{Var}}(Z_1) \geq \liminf_{\gamma \to \infty} \left( \frac{1}{\gamma} \log p_1 + \frac{1}{\gamma} \log \alpha_1 + \frac{1}{\gamma} \log \hat{P}(\frac{1}{\gamma} X, \in \mathcal{E}_1) \right) \). First, we know that \( \lim_{\gamma \to \infty} \frac{1}{\gamma} \log p_1 = -I(\mathcal{E}) \). Second, we know that \( \frac{1}{\gamma} \log \alpha_1 \to 0 \). Third, by Theorem 1, we know
that \( \liminf_{\gamma \to \infty} \frac{1}{\gamma} \log \hat{P}_1(\frac{1}{\gamma} X_i \in \mathcal{E}_1) \geq -\tilde{I}_1((\mathcal{E}_1)^\circ) \) where \( \tilde{I}_1(y) = I(y) - s_{\gamma}^T y + \mu(s_{\gamma}) \) as in the proof of Theorem 6. As \( a_1 \) is a dominating point, we have required that \( a_1 \in \mathcal{D}(I)^\circ \), and thus \( a_1 \in \mathcal{D}(\tilde{I}_1)^\circ \). We note that \( (\mathcal{E}_1)^\circ \supseteq \bigcap_{i=1}^k \{ x : s_{\gamma}^T (x - a_i) < 0 \} \). Then \( \tilde{I}_1((\mathcal{E}_1)^\circ) \leq \tilde{I}_1(\bigcap_{i=1}^k \{ x : s_{\gamma}^T (x - a_i) < 0 \}) \). We note that since \( a_1 \) is the most significant dominating point, we have \( s_{\gamma}^T (a_1 - a_i) \leq 0 \) (otherwise we get \( I(a_i) < I(a_1) \), which is a contradiction). Thus \( a_1 \) is on the boundary of \( \bigcap_{i=1}^k \{ x : s_{\gamma}^T (x - a_i) < 0 \} \), and there exists a sequence of points \( \{y_n\}_{n=1}^\infty \subset \mathcal{D}(\tilde{I}_1)^\circ \cap \bigcap_{i=1}^k \{ x : s_{\gamma}^T (x - a_i) < 0 \} \) such that \( y_n \to a_1 \) as \( n \to \infty \). Therefore, we get \( \tilde{I}_1(\bigcap_{i=1}^k \{ x : s_{\gamma}^T (x - a_i) < 0 \}) \leq \tilde{I}_1(a_1) = 0 \). Overall, we have proved that \( \liminf_{\gamma \to \infty} \frac{1}{\gamma} \log \text{Var}(Z_1) \geq -2I(\mathcal{E}) \).

We also need a concentration result for the sample variance:

**Lemma 3 (Theorem 10 in Maurer and Pontil 2009).** Let \( n \geq 2 \) and \( \mathbf{Y} = (Y_1, \ldots, Y_n) \) be a vector of independent random variables with values in \([0, 1]\). Then for \( \delta > 0 \) we have

\[
P \left( \sqrt{EV_n(\mathbf{Y})} > \sqrt{V_n(\mathbf{Y})} + \sqrt{\frac{2 \log(1/\delta)}{n - 1}} \right) \leq \delta;
\]

\[
P \left( \sqrt{V_n(\mathbf{Y})} > \sqrt{EV_n(\mathbf{Y})} + \sqrt{\frac{2 \log(1/\delta)}{n - 1}} \right) \leq \delta.
\]

With Lemmas 2 and 3, now we prove Theorem 8:

**Proof of Theorem 8.** We denote \( Z_1^{(i)} = I(\frac{1}{\gamma} X_i \in \mathcal{E}_1) \frac{d\mathbb{P}}{d\mathbb{P}} \). \( \hat{p}_1 \) and \( \hat{V}_1 \) respectively denote the sample mean and sample variance of \( Z_1^{(i)} \)’s. We also define \( N = \sum_{i=1}^n I(\frac{1}{\gamma} X_i^{(i)} \in \mathcal{E}_2) \). Similar to the proof of Theorem 7, we have

\[
\hat{p}(\hat{p} - p) > z_{1-\alpha/2} \sqrt{\frac{\hat{V}_1}{n}}
\]

\[
= \hat{P}(\hat{p} - p > z_{1-\alpha/2} \sqrt{\frac{\hat{V}_1}{n}, N = 0}) + \hat{P}(\hat{p} - p > z_{1-\alpha/2} \sqrt{\frac{\hat{V}_1}{n}, N > 0})
\]

\[
\leq \hat{P}(\hat{p}_1 - p > z_{1-\alpha/2} \sqrt{\frac{\hat{V}_1}{n}}) + \hat{P}(N > 0)
\]

\[
\leq \hat{P}(\hat{p}_1 - p > z_{1-\alpha/2} \sqrt{\frac{\hat{V}_1}{n} - p_2}) + n\hat{p}_2
\]

\[
\leq \hat{P}(\hat{p}_1 - p > z_{1-\alpha/2} \left( \sqrt{\hat{V}_1} - \sqrt{\frac{2 \log(1/\delta)}{n(n-1)} M_1} \right) - p_2) + \hat{P}(\sqrt{\hat{V}_1} > \sqrt{\hat{V}_1} + \sqrt{\frac{2 \log(1/\delta)}{n-1} M_2} + n\hat{p}_2
\]

for any \( \delta = \delta(\gamma) > 0 \), where \( M_1 \) is as defined in Theorem 7. We know that \( 0 \leq Z_1^{(i)} \leq M_1, \forall i \). By Lemma 3 with \( Y_i = Z_1^{(i)}/M_1 \), we get that

\[
\hat{P}(\sqrt{\hat{V}_1} > \sqrt{\hat{V}_1} + \sqrt{\frac{2 \log(1/\delta)}{n-1} M_1}) \leq \delta.
\]
By Berry-Esseen theorem, we know that for any \( x \in \mathbb{R} \)
\[
\left| \hat{P} \left( \frac{\sqrt{n}(\hat{p}_1 - p_1)}{\sqrt{\text{Var}(Z^{(1)}_1)}} \leq x \right) - \Phi(x) \right| \leq \frac{C\hat{E}|Z^{(1)}_1 - p_1|^3}{\text{Var}^{3/2}(Z^{(1)}_1)\sqrt{n}}
\]
where \( \Phi \) is the CDF of standard normal distribution and \( C \) is a universal constant. Let
\[
x = z_{1-\alpha/2} \left( \frac{\sqrt{\hat{E}V_1}}{\sqrt{\text{Var}(Z^{(1)}_1)}} - \sqrt{\frac{2\log(1/\delta)}{(n-1)\text{Var}(Z^{(1)}_1)}M_\gamma} \right) - p_2 \sqrt{\frac{n}{\text{Var}(Z^{(1)}_1)}}.
\]
Then we get that
\[
\hat{P} \left( \frac{\sqrt{n}|\hat{p}_1 - p_1|}{\sqrt{\text{Var}(Z^{(1)}_1)}} > x \right) \leq 2\Phi(-x) + \frac{2C\hat{E}|Z^{(1)}_1 - p_1|^3}{\text{Var}^{3/2}(Z^{(1)}_1)\sqrt{n}}.
\]
Hence,
\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \leq 2\Phi(-x) + \frac{2C\hat{E}|Z^{(1)}_1 - p_1|^3}{\text{Var}^{3/2}(Z^{(1)}_1)\sqrt{n}} + \delta + n\hat{p}_2.
\]
First, we have that \( \hat{E}|Z^{(1)}_1 - p_1|^3 \leq \max(p^3, (M_\gamma - p_1)^3) \). From the proof of Theorems 6 and 7, we know that \(-1/\gamma \log p_1 \to -I(a_1) \) and \(-1/\gamma \log M_\gamma \to -I(a_1) \). Thus \( \limsup_{\gamma \to \infty} 1/\gamma \log \hat{E}|Z^{(1)}_1 - p_1|^3 \leq -3I(a_1) \). By Lemma 2, we get that \( \frac{\hat{E}|Z^{(1)}_1 - p_1|^3}{\text{Var}^{3/2}(Z^{(1)}_1)} \) grows at most subexponentially in \( \gamma \). Hence under the assumptions we could choose \( n \) as required in the theorem. In particular, we have that
\[
\frac{2C\hat{E}|Z^{(1)}_1 - p_1|^3}{\text{Var}^{3/2}(Z^{(1)}_1)\sqrt{n}} \to 0.
\]
Now we analyze \( x \). We know that \( \sqrt{\frac{\hat{E}V_1}{\text{Var}(Z^{(1)}_1)}} = 1 \) and
\[
p_2 \sqrt{\frac{n}{\text{Var}(Z^{(1)}_1)}} = \frac{p_2}{e^{-\gamma I(a_1)}} \sqrt{\frac{n e^{-2\gamma I(a_1)}}{\text{Var}(Z^{(1)}_1)}} \to 0
\]
since \( \frac{p_2}{e^{-\gamma I(a_1)}} \) decays exponentially (proof of Theorem 6) while \( n \) and \( \frac{n e^{-2\gamma I(a_1)}}{\text{Var}(Z^{(1)}_1)} \) grow subexponentially (Lemma 2) in \( \gamma \). Now we consider
\[
\sqrt{\frac{2\log(1/\delta)}{(n-1)\text{Var}(Z^{(1)}_1)}M_\gamma} = \sqrt{\frac{2\log(1/\delta)n}{n-1}} \sqrt{\frac{M^2_{\gamma}}{n\text{Var}(Z^{(1)}_1)}}.
\]
Since we assume that \( \frac{M^2_{\gamma}}{n\text{Var}(Z^{(1)}_1)} \to 0 \), we could set \( \delta \) such that \( \delta \to 0 \) and
\[
\sqrt{\frac{2\log(1/\delta)}{(n-1)\text{Var}(Z^{(1)}_1)}M_\gamma} \to 0.
\]
In this case, \( x \to z_{1-\alpha/2} \) and hence \( \Phi(-x) \to \alpha/2 \). Combining all the results, we get that
\[
\limsup_{\gamma \to \infty} \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \leq \alpha.
\]
□
Proof of Theorem 9. First, following the proof of Lemma 2, it is easy to get that \( \lim_{\gamma \to \infty} \frac{1}{\gamma} \log \hat{\text{Var}}(Z) = -2I(\mathcal{E}) = -2I(a_1) \). Moreover, we have that

\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \leq \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \left( \sqrt{\frac{\hat{E}V}{n}} - \sqrt{\frac{2\log(1/\delta)}{n(n-1)} M_\gamma} \right) \right)
\]

\[
+ \hat{P} \left( \sqrt{\hat{E}V} > \sqrt{V} + \sqrt{\frac{2\log(1/\delta)}{n-1} M_\gamma} \right)
\]

and

\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \geq \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \left( \sqrt{\frac{\hat{E}V}{n}} + \sqrt{\frac{2\log(1/\delta)}{n(n-1)} M_\gamma} \right) \right)
\]

\[
- \hat{P} \left( \sqrt{\hat{E}V} > \sqrt{V} + \sqrt{\frac{2\log(1/\delta)}{n-1} M_\gamma} \right)
\]

for any \( \delta = \delta(\gamma) > 0 \). We know that \( 0 \leq Z^{(i)} \leq M_\gamma, \forall i \). By Lemma 3, we get that

\[
\hat{P} \left( \sqrt{\hat{E}V} > \sqrt{V} + \sqrt{\frac{2\log(1/\delta)}{n-1} M_\gamma} \right) \leq \delta
\]

and

\[
\hat{P} \left( \sqrt{V} > \sqrt{\hat{E}V} + \sqrt{\frac{2\log(1/\delta)}{n-1} M_\gamma} \right) \leq \delta.
\]

By Berry-Esseen theorem, we know that

\[
\hat{P} \left( \frac{\sqrt{n}|\hat{p} - p|}{\sqrt{\text{Var}(Z^{(i)})}} > x_1 \right) \leq 2\Phi(-x_1) + \frac{2C\hat{E}|Z^{(i)} - p|^3}{\text{Var}^{3/2}(Z^{(i)})\sqrt{n}}
\]

and

\[
\hat{P} \left( \frac{\sqrt{n}|\hat{p} - p|}{\sqrt{\text{Var}(Z^{(i)})}} > x_2 \right) \geq 2\Phi(-x_2) - \frac{2C\hat{E}|Z^{(i)} - p|^3}{\text{Var}^{3/2}(Z^{(i)})\sqrt{n}}
\]

where \( \Phi \) is the CDF of standard normal distribution, \( C \) is a universal constant, and

\[
x_1 = z_{1-\alpha/2} \left( \sqrt{\frac{\hat{E}V}{\text{Var}(Z^{(i)})}} - \sqrt{\frac{2\log(1/\delta)}{(n-1)\text{Var}(Z^{(i)})} M_\gamma} \right),
\]

\[
x_2 = z_{1-\alpha/2} \left( \sqrt{\frac{\hat{E}V}{\text{Var}(Z^{(i)})}} + \sqrt{\frac{2\log(1/\delta)}{(n-1)\text{Var}(Z^{(i)})} M_\gamma} \right).
\]

Combining the above derivations, we get that

\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \leq 2\Phi(-x_1) + \frac{2C\hat{E}|Z^{(i)} - p|^3}{\text{Var}^{3/2}(Z^{(i)})\sqrt{n}} + \delta,
\]

\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \geq 2\Phi(-x_2) - \frac{2C\hat{E}|Z^{(i)} - p|^3}{\text{Var}^{3/2}(Z^{(i)})\sqrt{n}} - \delta.
\]

First, we have that \( \hat{E}|Z^{(i)} - p|^3 \leq \max(p^3, (M_\gamma - p)^3) \). From the proof of the previous theorems, we know that \( -\frac{1}{\gamma} \log p \to -I(a_1) \) and \( -\frac{1}{\gamma} \log M_\gamma \to -I(a_1) \). Thus, \( \frac{\hat{E}|Z^{(i)} - p|^3}{\text{Var}^{3/2}(Z^{(i)})\sqrt{n}} \) grows at most subexponentially in \( \gamma \),
and hence we could choose $n$ as required in the theorem. In particular, we have \( \frac{2CE_{I[Z^{(1)}-p]}^3}{\sqrt{\text{Var}(Z^{(1)})n}} \to 0 \). Next, since we assume that \( \frac{M^2_\gamma}{n\text{Var}(Z^{(1)})} \to 0 \), we could set $\delta$ such that $\delta \to 0$ and

\[
\sqrt{\frac{2\log(1/\delta)}{(n-1)\text{Var}(Z^{(1)})}} M_\gamma \to 0.
\]

Hence, \( x_1, x_2 \to z_{1-\alpha/2} \) as $\gamma \to \infty$. Overall, we get that

\[
\limsup_{\gamma \to \infty} \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \leq \alpha,
\]

\[
\liminf_{\gamma \to \infty} \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \geq \alpha.
\]

Therefore, the theorem is proved. \( \square \)

**Proof of Theorem 10.** The conclusion follows from Theorem 5, by checking, for each setting in Sections 3.1, 3.2 and 3.3, Assumptions 1 and 2 hold and the most significant dominating point is unique. Since we have $\mu(x) = 1.5x + 0.5x^2 - \log(1+x)$ for the case in Section 3.1 and $\mu(x) = \lambda^T x + \frac{1}{2} x^T \Sigma x$ for the cases in Sections 3.2 and 3.3, we verify the conditions in Assumption 1. On the other hand, since all rare-event sets in these examples are closed and contain unique optimal solutions for minimizing the corresponding rate function $I(x)$, we can verify Assumption 2 and the uniqueness of most significant dominating point. \( \square \)

**Proof of Theorem 11.** We check the three conditions in Theorem 4 to show the probabilistic efficiency of the IS estimator using the dominating point $a_1$. Using the notation in Section 6.1.4, we note that $A^1_\gamma$ only one most significant dominating point $a_1$ and $A^2_\gamma$ only one most significant dominating point $-a_1$, while $I(a_1) = I(-a_1)$. This indicates that $p_1/p_2 \to 1$ as $\gamma \to \infty$ and hence $p_1/p \to 1/2$ as $\gamma \to \infty$. Then, following the argument in Section 6.1.2, the IS estimator using the dominating point $a_1$ is asymptotically efficient for $A^1_\gamma$. Lastly we have $\hat{p}_2 \to 0$ exponentially fast and hence we have $n\hat{p}_2 \to 0$ with subexponentially growing sample size $n$. We conclude that the IS estimator using the dominating point $a_1$ is weakly probabilistically efficient.

On the other hand, based on Proposition 4, the IS estimator using the dominating point $a_+$ is not asymptotically efficient for $\{\bigcup_{m=1}^d H^+_m\}$ and hence is not asymptotically efficient for $\left( \left( \bigcup_{m=1}^d H^+_m \right) \cup \left( \bigcup_{m=1}^d H^-_m \right) \right)$. \( \square \)

**Proof of Theorem 12.** We only need to verify the conditions in Theorem 3. First, from Assumption 6, we get that $p_2$ is exponentially smaller than $p_1$, and hence $\frac{a_1}{p} \to 1$ as $\gamma \to \infty$. Second, Let $Z_1 = I(X \in E_1) \frac{dC}{dp}$ under $\hat{P}$. Clearly $\{a_1, \ldots, a_k\}$ is a dominating set for $E_1$. Similar to (15), we get that $\sqrt{\text{Var}(Z_1)} \leq k^2e^{-2I(a_1)} \leq r^2e^{-2I(a_1)}$. Assumption 6 gives that $p_1 \sim e^{-I(a_1)}$, and hence $Z_1$ is an asymptotically efficient estimator for $p_1$. Third, Assumption 7 implies that $n\hat{p}_2 \to 0$ for any $n$ subexponentially growing in $-\log p$. Therefore, all the conditions in Theorem 3 hold. \( \square \)

**Proof of Theorem 13.** We denote $\hat{p}_1 = \frac{1}{n} \sum_{i=1}^n Z_1^{(i)}$ and $Z_1 = (Z_1^{(1)}, \ldots, Z_1^{(n)})$. We also define $N = \sum_{i=1}^n I(X^{(i)} \in E_2)$. Note that conditional on $N = 0$, we have $Z = Z_1$. Then

\[
\hat{P} \left( |\hat{p} - p| > \sqrt{\frac{2\hat{V}\log(4/\alpha)}{n}} + \frac{7\log(4/\alpha)k\hat{e}^{-I(a_1)}}{3(n-1)} \right)
\]
\[
\hat{P}
\left( |\hat{p} - p| > \sqrt{\frac{2V \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}, N = 0} \right) \\
+ \hat{P}
\left( |\hat{p} - p| > \sqrt{\frac{2V \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}, N > 0} \right) \\
\leq \hat{P}
\left( |\hat{p}_1 - p_1| > \sqrt{\frac{2V_1 \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}, N = 0} \right) + P(N > 0) \\
\leq \hat{P}
\left( |\hat{p}_1 - p| > \sqrt{\frac{2V_1 \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}) + P(N > 0) \right) \\
\leq \hat{P}
\left( |\hat{p}_1 - p_1| + p_2 > \sqrt{\frac{2V_1 \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}) + n\hat{p}_2. \right)
\]

Similar to the derivation of (15), we know that \(0 \leq Z_1^{(i)} \leq ke^{-I(a_1)}, \forall i\). By applying Lemma 1 with \(Y_i = Z_1^{(i)}/(ke^{-I(a_1)})\) and \(Y_i = 1 - Z_1^{(i)}/(ke^{-I(a_1)})\) respectively, we get that

\[
\hat{P}
\left( p_1 > \hat{p}_1 + \sqrt{\frac{2V_1 \log(4/\delta)}{n} + \frac{7\log(4/\delta)ke^{-I(a_1)}}{3(n-1)}) \right) \leq \delta/2
\]

and

\[
\hat{P}
\left( p_1 < \hat{p}_1 - \sqrt{\frac{2V_1 \log(4/\delta)}{n} - \frac{7\log(4/\delta)ke^{-I(a_1)}}{3(n-1)}) \right) \leq \delta/2
\]

for any \(\delta > 0\). Thus,

\[
\hat{P}
\left( |\hat{p}_1 - p_1| > \sqrt{\frac{2V_1 \log(4/\delta)}{n} + \frac{7\log(4/\delta)ke^{-I(a_1)}}{3(n-1)}) \right) \leq \delta. \tag{20}
\]

Find \(\alpha' = \alpha'(\gamma)\) such that

\[
\frac{7\log(4/\alpha')ke^{-I(a_1)}}{3(n-1)} = \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)} - p_2.
\]

That is,

\[
\alpha' = \alpha \exp\left( \frac{3(n-1)p_2}{7ke^{-I(a_1)}} \right).
\]

Clearly \(\alpha' > \alpha\) and \(\log(4/\alpha') < \log(4/\alpha)\). Moreover, we know that \(\frac{p_2}{e^{-I(a)}}\) decays exponentially in \(-\log p\) since Assumption 6 holds and that \(n\) grows subexponentially in \(-\log p\), and thus \(\alpha' \to \alpha\) as \(\gamma \to \infty\). We replace \(\delta\) with \(\alpha'\) in (20), and then we get

\[
\hat{P}
\left( |\hat{p}_1 - p_1| > \sqrt{\frac{2V_1 \log(4/\alpha')}{n} + \frac{7\log(4/\alpha')ke^{-I(a_1)}}{3(n-1)}) \right) \leq \alpha'.
\]

Hence

\[
\hat{P}
\left( |\hat{p}_1 - p_1| > \sqrt{\frac{2V \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}} - p_2 \right) \leq \alpha'.
\]

Therefore,

\[
\hat{P}
\left( |\hat{p} - p| > \sqrt{\frac{2V \log(4/\alpha)}{n} + \frac{7\log(4/\alpha)ke^{-I(a_1)}}{3(n-1)}} \right) \leq \alpha' + n\hat{p}_2 \to \alpha\ as \ \gamma \to \infty.
\]

□
\textbf{Proof of Theorem 14.} We denote \( \hat{p}_1 = \frac{1}{n} \sum_{i=1}^{n} Z_1^{(i)} \) and \( Z_1 = (Z_1^{(1)}, \ldots, Z_1^{(n)}) \). We also define \( N = \sum_{i=1}^{n} I(X^{(i)} \in E_2) \). We have
\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{\hat{V}}{n}} \right) 
\leq \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V_{\hat{E}}}{n}}, N = 0 \right) + \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V_{\hat{E}}}{n}}, N > 0 \right)
\leq \hat{P} \left( |\hat{p}_1 - p| > z_{1-\alpha/2} \sqrt{\frac{V_{\hat{E}}}{n}} \right) + P(N > 0)
\leq \hat{P} \left( |\hat{p}_1 - p_1| > z_{1-\alpha/2} \sqrt{\frac{V_{\hat{E}}}{n}} \right) + n\tilde{p}_2
\leq \hat{P} \left( |\hat{p}_1 - p_1| > z_{1-\alpha/2} \left( \sqrt{\frac{\hat{E}V_1}{n}} - \sqrt{\frac{2\log(1/\delta)}{n(n-1)}ke^{-I(a_1)}} - p_2 \right) + n\tilde{p}_2
\end{align*}
\]
for any \( \delta = \delta(\gamma) > 0 \). We know that \( 0 \leq Z_1^{(i)} \leq ke^{-I(a_1)}, \forall i \). By Lemma 3 with \( Y_i = Z_1^{(i)}/(ke^{-I(a_1)}) \), we get that
\[
\hat{P} \left( \sqrt{\hat{E}V_1} > \sqrt{\hat{V}} + \sqrt{\frac{2\log(1/\delta)}{n-1}ke^{-I(a_1)}} \right) \leq \delta.
\]
By Berry-Esseen theorem, we know that for any \( x \in \mathbb{R} \)
\[
\left| \hat{P} \left( \frac{\sqrt{n}(\hat{p}_1 - p_1)}{\sqrt{\text{Var}(Z_1^{(i)})}} \leq x \right) - \Phi(x) \right| \leq C\hat{E}|Z_1^{(1)} - p_1|^3 / \sqrt{\text{Var}(Z_1^{(1)}/\sqrt{n})}
\]
where \( \Phi \) is the CDF of standard normal distribution and \( C \) is a universal constant. Let
\[
x = z_{1-\alpha/2} \left( \sqrt{\frac{\hat{E}V_1}{\text{Var}(Z_1^{(1)})}} - \sqrt{\frac{2\log(1/\delta)}{n(n-1)\text{Var}(Z_1^{(1)})}ke^{-I(a_1)}} \right) - p_2 \sqrt{\frac{n}{\text{Var}(Z_1^{(1)})}}.
\]
Then we get that
\[
\hat{P} \left( \frac{\sqrt{n}(\hat{p}_1 - p_1)}{\sqrt{\text{Var}(Z_1^{(1)})}} > x \right) \leq 2\Phi(-x) + 2C\hat{E}|Z_1^{(1)} - p_1|^3 / \sqrt{\text{Var}(Z_1^{(1)}/\sqrt{n})}.
\]
Hence,
\[
\hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{\hat{V}}{n}} \right) \leq 2\Phi(-x) + 2C\hat{E}|Z_1^{(1)} - p_1|^3 / \sqrt{\text{Var}(Z_1^{(1)}/\sqrt{n})} + \delta + n\tilde{p}_2.
\]
First, we have that \( \hat{E}|Z_1^{(1)} - p_1|^3 \leq \max(p_1^2, (ke^{-I(a_1)} - p_1)^3) \sim e^{-3I(a_1)}. \) Since Assumption 8 holds, we get that \( \frac{E|Z_1^{(1)} - p_1|^3}{\text{Var}(Z_1^{(1)})} \) grows at most subexponentially in \( -\log p \). Hence under the assumptions we could choose \( n \) as required in the theorem. In particular, we have that
\[
\frac{2C\hat{E}|Z_1^{(1)} - p_1|^3}{\sqrt{\text{Var}(Z_1^{(1)}/\sqrt{n})}} \to 0.
\]
Now we analyze $x$. We know that $\sqrt{\frac{E V_1}{\text{Var}(Z_1^{(1)})}} = 1$ and

$$p_2 \sqrt{\frac{n}{\text{Var}(Z_1^{(1)})}} = \frac{p_2}{e^{-I(a_1)}} \sqrt{n e^{-2I(a_1)}} \text{Var}(Z_1^{(1)}) \to 0$$

since $\frac{p_2}{n e^{-1(a_1)}}$ decays exponentially (Assumption 6) while $n$ and $\frac{e^{-2I(a_1)}}{\text{Var}(Z_1^{(1)})}$ grow subexponentially (Assumption 8) in $-\log p$. Now we consider

$$\sqrt{\frac{2 \log(1/\delta)}{(n-1)\text{Var}(Z_1^{(1)})}} k e^{-I(a_1)} = \frac{2k \log(1/\delta)e^{-2I(a_1)}}{(n-1)\text{Var}(Z_1^{(1)})}.$$

Since we assume that $\frac{2k \log(1/\delta)}{n \text{Var}(Z_1^{(1)})} \to 0$, we could set $\delta$ such that $\delta \to 0$ and

$$\sqrt{\frac{2 \log(1/\delta)}{(n-1)\text{Var}(Z_1^{(1)})}} k e^{-I(a_1)} \to 0.$$

In this case, $x \to z_{1-\alpha/2}$ and hence $\Phi(-x) \to \alpha/2$. Combining all the results, we get that

$$\limsup_{\gamma \to \infty} \hat{P} \left( |\hat{p} - p| > z_{1-\alpha/2} \sqrt{\frac{V}{n}} \right) \leq \alpha.$$ 

\[ \square \]

**Proof of Theorem 15.** First of all, it is easy to verify that the cumulant generating function $\mu(x) = \lambda^T x + \frac{1}{2} x^T \Sigma x$ satisfies Assumption 3. Moreover, since $g$ is a piecewise linear function, we can express $E$ as the union of finite closed polyhedrons. If $P(X \in E) > 0$ and $\lambda \notin E$, then $0 < \inf_{x \in E} \frac{1}{2} (x - \lambda)^T \Sigma^{-1} (x - \lambda) < \infty$. From Bai et al. (2022), we know that

$$p = P(g(X) \geq \gamma) \sim e^{-\frac{1}{2} (a_1 - \lambda)^T \Sigma^{-1} (a_1 - \lambda)}$$

and for sufficient large $\gamma$, $(a_1 - \lambda)^T \Sigma^{-1} (a_1 - \lambda)$ is a quadratic function of $\gamma$ which goes to $\infty$ as $\gamma \to \infty$. Finally, it is clear that the number of dominating points will not grow exponentially in $-\log p = \Theta((a_1 - \lambda)^T \Sigma^{-1} (a_1 - \lambda))$. Therefore, Assumption 5 is satisfied in this problem setting.

Next, we check Assumptions 6 and 7. Without loss of generality, we may assume that $k < r$, since otherwise $E_2 = \emptyset$ and $p_2 = \hat{p}_2 = 0$. We know that $a_{k+1} = \arg \min_{x \in E_2} (x - \lambda)^T \Sigma^{-1} (x - \lambda)$, and hence

$$E_2 \subset \{ x \in \mathbb{R}^d : (x - \lambda)^T \Sigma^{-1} (x - \lambda) \geq (a_{k+1} - \lambda)^T \Sigma^{-1} (a_{k+1} - \lambda) \}.$$

Denote $X' = \Sigma^{-1/2} (X - \lambda)$ and then $X' \sim N(0, I_d)$ under $P$. We have that

$$p_2 \leq P((X - \lambda)^T \Sigma^{-1} (X - \lambda) \geq (a_{k+1} - \lambda)^T \Sigma^{-1} (a_{k+1} - \lambda))$$

$$\sim P(X'^T X' \geq (a_{k+1} - \lambda)^T \Sigma^{-1} (a_{k+1} - \lambda))$$

$$\sim e^{-\frac{1}{2} (a_{k+1} - \lambda)^T \Sigma^{-1} (a_{k+1} - \lambda)}.$$ 

We know that $(a_{k+1} - \lambda)^T \Sigma^{-1} (a_{k+1} - \lambda) > C(a_k - \lambda)^T \Sigma^{-1} (a_k - \lambda) \geq C(a_1 - \lambda)^T \Sigma^{-1} (a_1 - \lambda)$, Thus, Assumption 6 holds. Moreover,

$$\hat{p}_2 = \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N(a_i, \Sigma)} (X \in E_2)$$

$$\leq \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N(a_i, \Sigma)} ((X - \lambda)^T \Sigma^{-1} (X - \lambda) \geq (a_{k+1} - \lambda)^T \Sigma^{-1} (a_{k+1} - \lambda)).$$
We have that
\[(X - a_i)^\top \Sigma^{-1} (X - a_i) \]
\[= (X - \lambda)^\top \Sigma^{-1} (X - \lambda) + (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda) - 2(a_i - \lambda)^\top \Sigma^{-1} (X - \lambda) \]
\[\geq (X - \lambda)^\top \Sigma^{-1} (X - \lambda) + (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda) - 2\sqrt{(X - \lambda)^\top \Sigma^{-1} (X - \lambda)} \sqrt{(a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda)} \]
\[= \left( \sqrt{(X - \lambda)^\top \Sigma^{-1} (X - \lambda)} - \sqrt{(a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda)} \right)^2. \]

Hence, if \((X - \lambda)^\top \Sigma^{-1} (X - \lambda) \geq (a_{k+1} - \lambda)^\top \Sigma^{-1} (a_{k+1} - \lambda) > C(a_k - \lambda)^\top \Sigma^{-1} (a_k - \lambda) \geq C(a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda),\) then \((X - a_i)^\top \Sigma^{-1} (X - a_i) \geq (\sqrt{C} - 1)^2 (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda).\) Thus we get that
\[\tilde{p}_2 \leq \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N(a_i, \Sigma)}((X - a_i)^\top \Sigma^{-1} (X - a_i) \geq (\sqrt{C} - 1)^2 (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda)) \]
\[\leq \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N(a_i, \Sigma)}((X - a_i)^\top \Sigma^{-1} (X - a_i) \geq (\sqrt{C} - 1)^2 (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda)) = P(X^\top X' \geq (\sqrt{C} - 1)^2 (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda)) \]
\[\sim e^{-\frac{1}{2} (\sqrt{C} - 1)^2 (a_i - \lambda)^\top \Sigma^{-1} (a_i - \lambda)}. \]

Hence, \(\tilde{p}_2\) exponentially decays in \(-\log p.\) That is, Assumption 7 holds.

Finally, we verify Assumption 8. We have \(\tilde{V}ar(Z_1) = \tilde{E}(Z_1 - p_1)^2 \geq \tilde{E}((Z_1 - p_1)^2 I_{E_1}(X)) = p_1^2 \tilde{P}(X \notin E_1).\)

Since \(p_1 \sim e^{-\frac{1}{2} (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)},\) it suffices to justify that \(\tilde{P}(X \notin E_1)\) does not decay exponentially in \(-\log p.\)

Indeed, we have that \(E_1 \subset \{x \in \mathbb{R}^d : (x - \lambda)^\top \Sigma^{-1} (x - \lambda) \geq (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)\}\) and thus
\[\tilde{P}(X \in E_1) = \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N(a_i, \Sigma)}(X \in E_1) \]
\[\leq \frac{k - 1}{k} + \frac{1}{k} P_{X \sim N(a_i, \Sigma)}(X \in E_1) \]
\[\leq \frac{k - 1}{k} + \frac{1}{k} P_{X \sim N(a_i, \Sigma)}((X - \lambda)^\top \Sigma^{-1} (X - \lambda) \geq (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)) \]
\[= \frac{k - 1}{k} + \frac{1}{k} P((X' + \Sigma^{-1/2} (a_1 - \lambda))^\top (X' + \Sigma^{-1/2} (a_1 - \lambda)) \geq (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)). \]

Then we have that
\[\tilde{P}(X \notin E_1) \geq \frac{1}{k} P((X' + \Sigma^{-1/2} (a_1 - \lambda))^\top (X' + \Sigma^{-1/2} (a_1 - \lambda)) < (a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)) \]
\[\geq \frac{1}{r} P \left( X' \in B \left( -\Sigma^{-1/2} (a_1 - \lambda), \sqrt{(a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)} \right) \right) \]
where \(B(x, R) := \{y \in \mathbb{R}^d : \|y - x\| < R\}.\) For sufficiently large \(r,\) \((a_1 - \lambda)^\top \Sigma^{-1} (a_1 - \lambda)\) monotonically grows to \(\infty,\) and hence \(\tilde{P}(X \notin E_1) \geq \frac{c}{r} \) for some constant \(c > 0.\) As a result, Assumption 8 holds. \(\square\)