SISL
Stanford Intelligent
Systems Laboratory

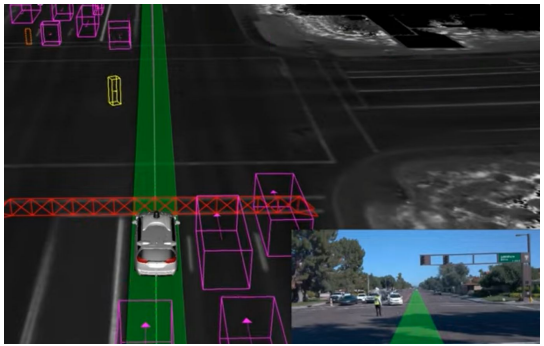
Certifiable Failure Rate Validation using Importance Sampling + Deep Learning

Mansur M. Arief
Research Engineer, Stanford Intelligent Systems Lab and Mineral-X
mansur.arief@stanford.edu

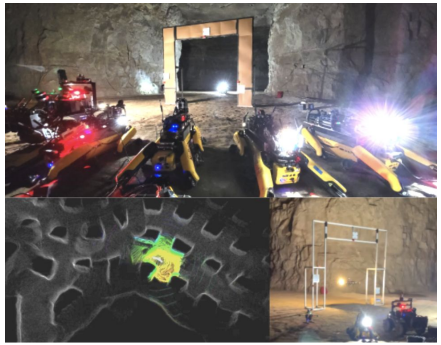
About Me

- **Research Engineer, Stanford Intelligent Systems Lab and Mineral-X**
- **Postdoc, AeroAstro,** Stanford, 2023-2024
- **PhD** in **MechE**, Carnegie Mellon, 2023
  - Dissertation at Safe AI Lab: Certifiable Evaluation for Safe Intelligent Autonomy
- **MSE, Industrial & Operations Engineering**,
  University of Michigan, Ann Arbor
- **BE, Industrial and Systems Engineering,**
  Sepuluh Nopember Institute of Technology, Indonesia

Cyber-Physical Systems (CPSs) are everywhere



Autonomous vehicles



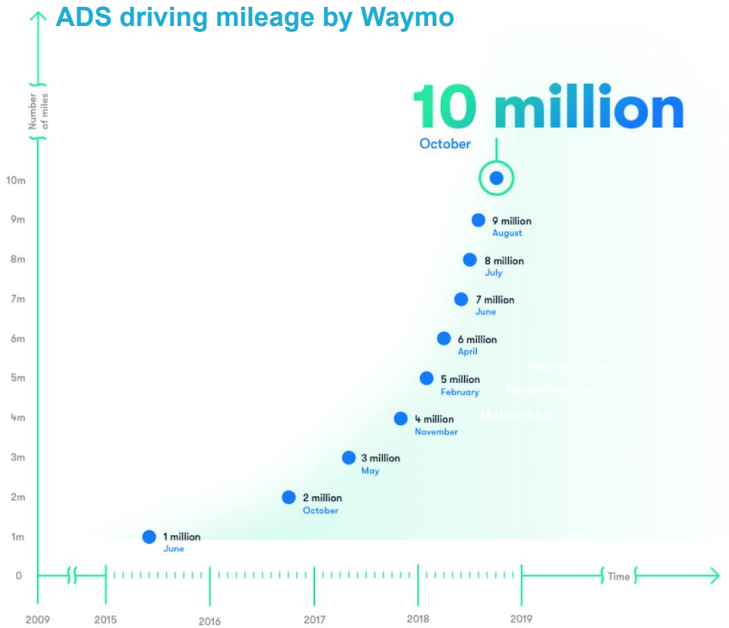
Exploratory robots

Aircraft collision avoidance systems

- Interacting with humans more intensively and collaboratively
- Making more important, even safety-critical, decisions

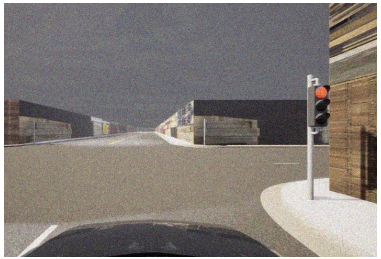
Safe CPS are hard to validate

- If the **failure rate is μ**, we need **$1/\mu$ samples** to observe the first (random) failure
- Monte Carlo sampling estimates have huge relative variance **$\mathbf{Var}(\mu_{est})/\mu$**
- Smaller μ requires larger sample size **(i.e. curse of rarity)**

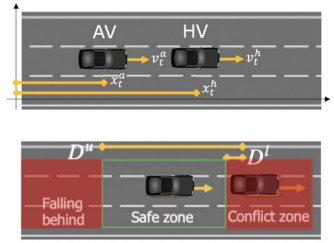


Source: Waymo

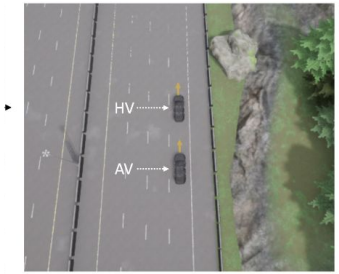
Airplane-level safety requires HUGE simulation runs,



I ran simulations for about a month to compare 99.99% accuracy CV models.



(a) Schematic diagram

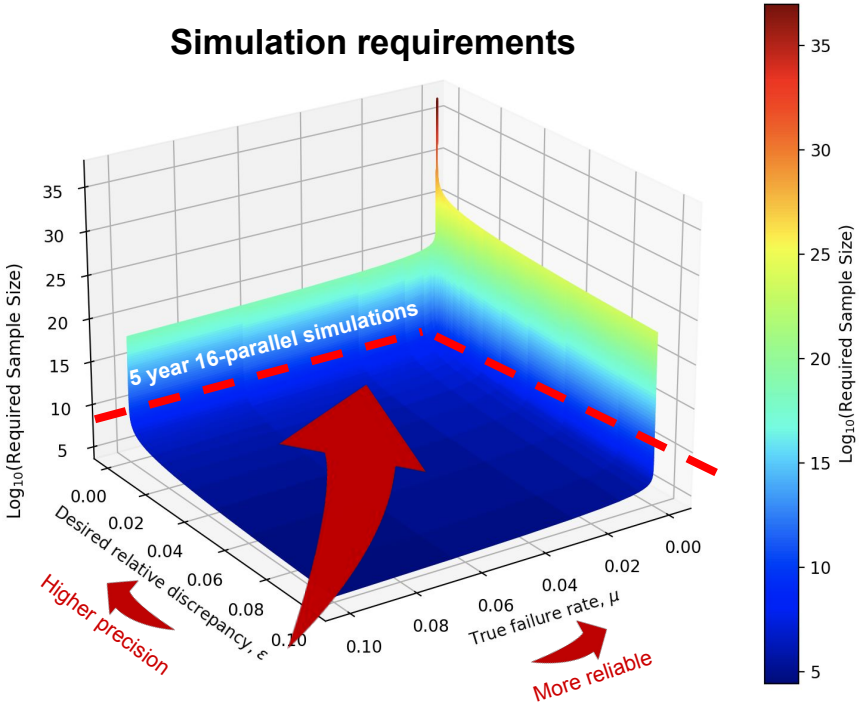


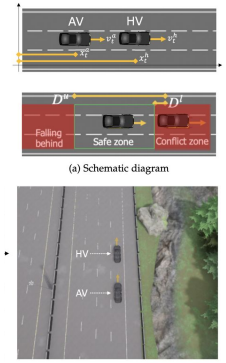
(b) CARLA topview camera

Even more for validating a 10^{-5} failure rate AV model.

Arief, Mansur Maturidi. *Certifiable Evaluation for Safe Intelligent Autonomy*. Diss. Carnegie Mellon University, 2023.

Airplane-level safety requires HUGE simulation runs,





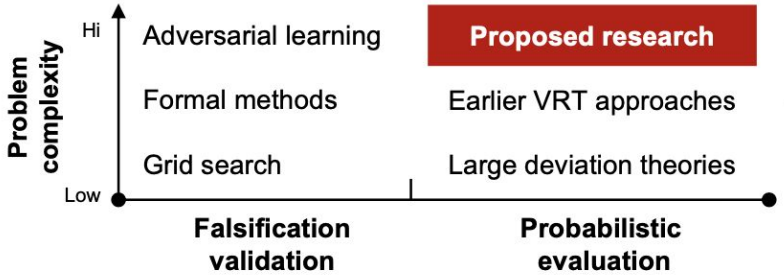
Arief, Mansur Maturidi. *Certifiable Evaluation for Safe Intelligent Autonomy*. Diss. Carnegie Mellon University, 2023.

Major safety validation goals

- **Falsification**: find a failure trajectory that violates specification

- **Most-likely failure**: find failure trajectory with maximum likelihood
- **Failure probability**: infer the violation rate of the specification under the disturbance model

Probabilistic evaluation



Corso, Anthony, et al. "A survey of algorithms for black-box safety validation of cyber-physical systems." *JAIR*, 72 (2021): 377-428.

How to generate validation test cases?

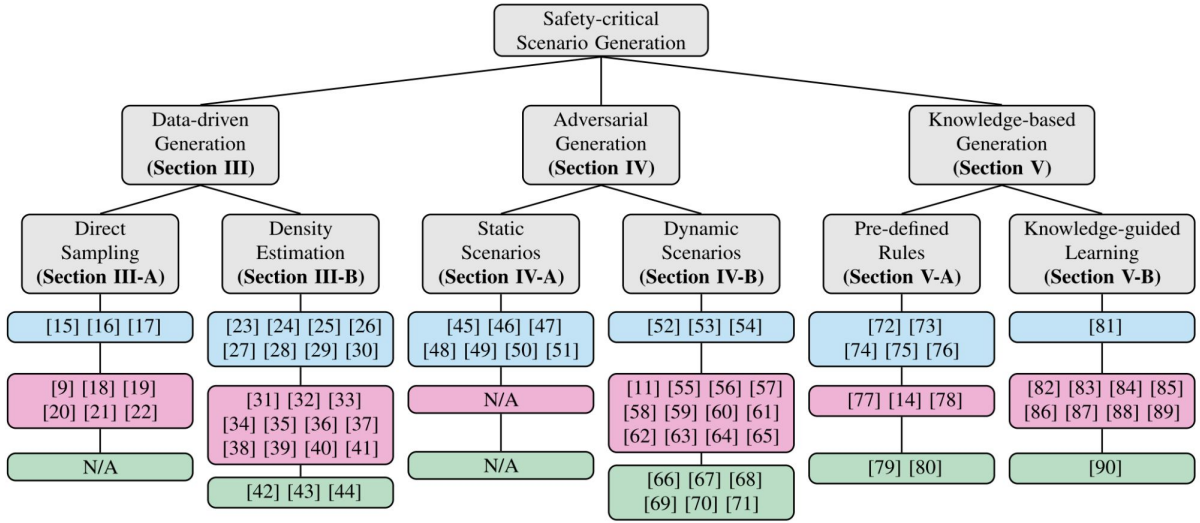
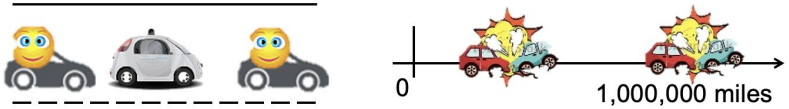


Fig. 2. Taxonomy of safety-critical scenarios generation methods. The colors of boxes denote the modules of the AV system that the generation algorithms target on Perception, Planning, Control.

Ding, Wenhao, et al. "A survey on safety-critical driving scenario generation—A methodological perspective." *IEEE T-ITS* (2023).

If failures are rare, importance sampling is useful

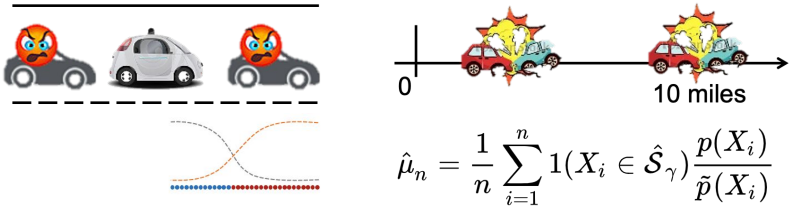
Naturalistic operating conditions:



$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in \hat{\mathcal{S}}_\gamma)$$

Rather than using Monte Carlo estimate,

Aggressive operating conditions:



$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in \hat{\mathcal{S}}_\gamma)\frac{p(X_i)}{\tilde{p}(X_i)}$$

use importance sampling estimate to get an unbiased result.

IS theoretical guarantees

- **Importance Sampling (IS)** uses proposal distribution $\tilde{p}$ and computes

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(X_i \in \mathcal{S}_\gamma\right) L(X_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i L(X_i),$$

$$L(X_i) = \frac{p(X_i)}{\tilde{p}(X_i)}. \Rightarrow \text{called } \textbf{the importance ratio}$$

IS theoretical guarantees

- IS is provably unbiased

$$
\begin{aligned}
\mathbb{E}_{X\sim\tilde{p}}[\hat{\mu}_n] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(X_i \in \mathcal{S}_\gamma\right) L(X_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{1}\left(X_i \in \mathcal{S}_\gamma\right)\frac{p(X_i)}{\tilde{p}(X_i)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\mathbb{1}\left(X_i \in \mathcal{S}_\gamma\right)\frac{p(X_i)}{\tilde{p}(X_i)}\tilde{p}(X_i)dX_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\mathbb{1}\left(X_i \in \mathcal{S}_\gamma\right)p(X_i)dX_i \\
&= \mu.
\end{aligned}
$$

IS theoretical guarantees

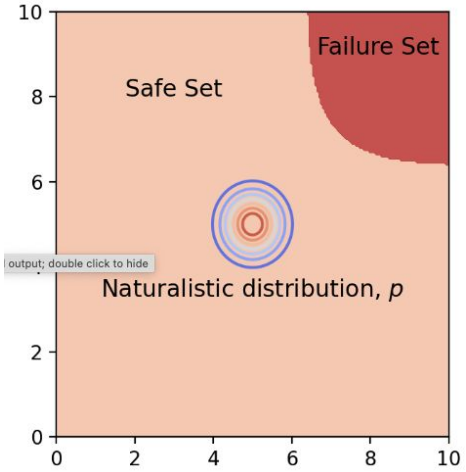
- IS **reduces variance** if the proposal distribution: $\tilde{p}(x) \propto \mathbb{1}(x \in \mathcal{S}_\gamma)\, p(x)$, i.e. the naturalistic distribution conditional on the failure set.
- **Cross Entropy (CE)** minimizes the KL-divergence between the proposal and this theoretically optimal distribution iteratively

$$\max_{\theta \in \Theta} \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{1}(X_i \in \mathcal{S}_\gamma) \frac{p(X_i)}{p_{\theta_j}(X_i)} \ln p_\theta(X_i)$$

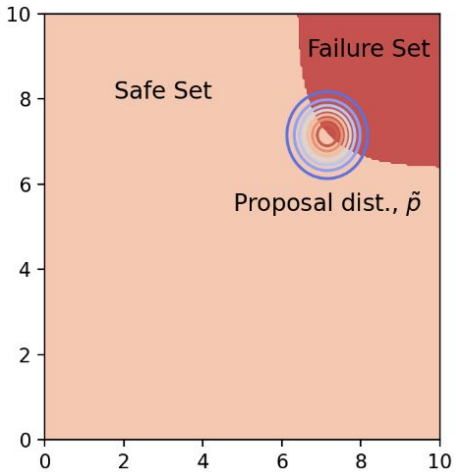
under some parametric class Θ.

IS underlying intuition

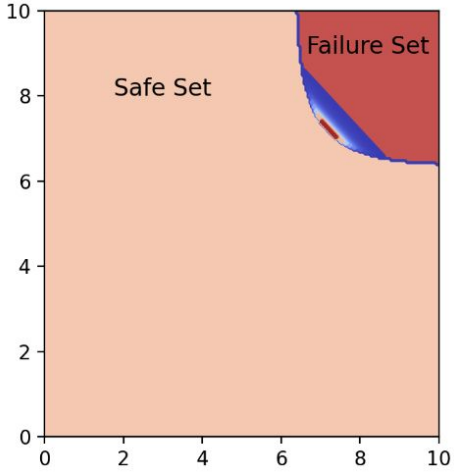
- IS **skews the distribution toward failures** and use likelihood ratio as weights to compute an unbiased estimate.



Naturalistic conditions

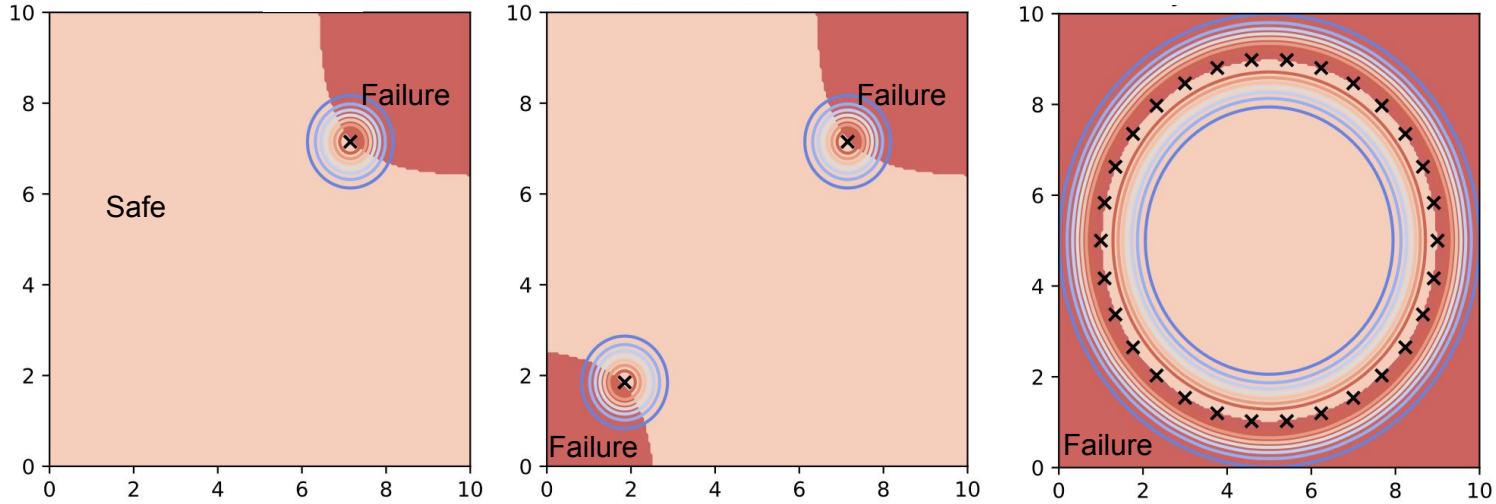


Skewed/aggressive conditions



Likelihood ratio conditional on the failure set

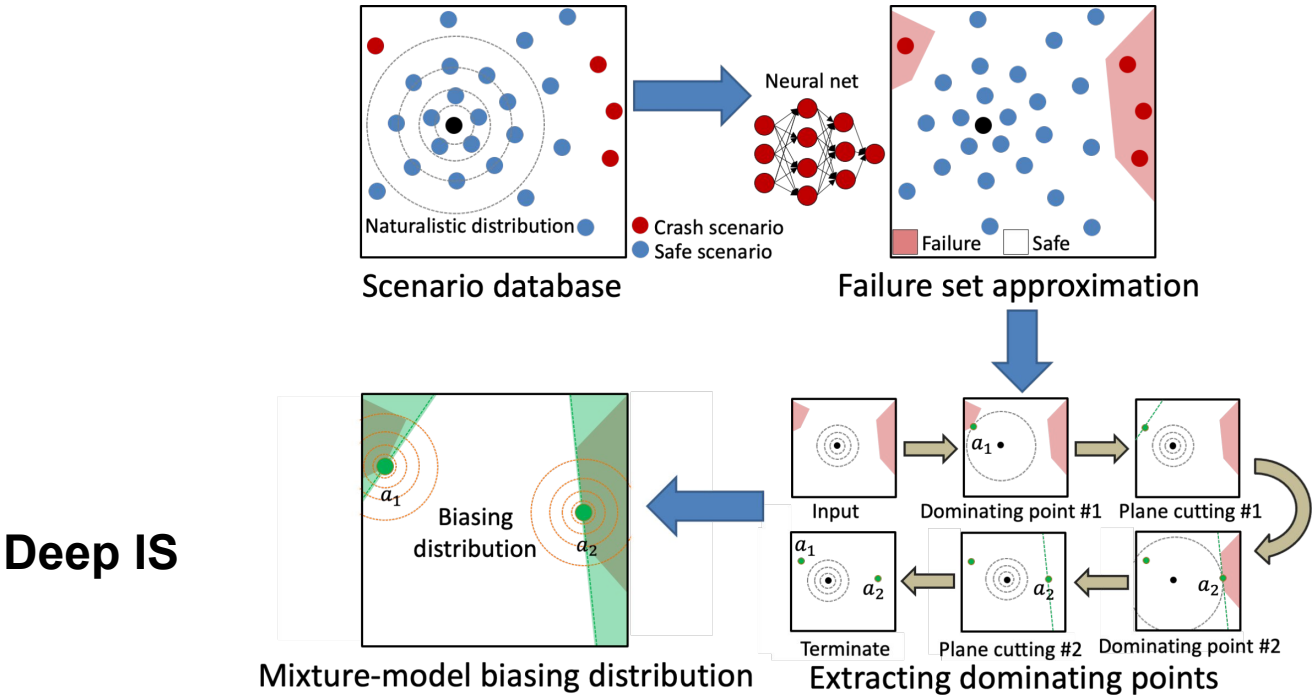
Optimal proposal is centered at all failure modes boundary



Single failure mode | Dual failure mode | Infinitely many failure mode

- Because otherwise, the likelihood ratio $L(X_i) = \frac{p(X_i)}{\tilde{p}(X_i)}$ can blow up!

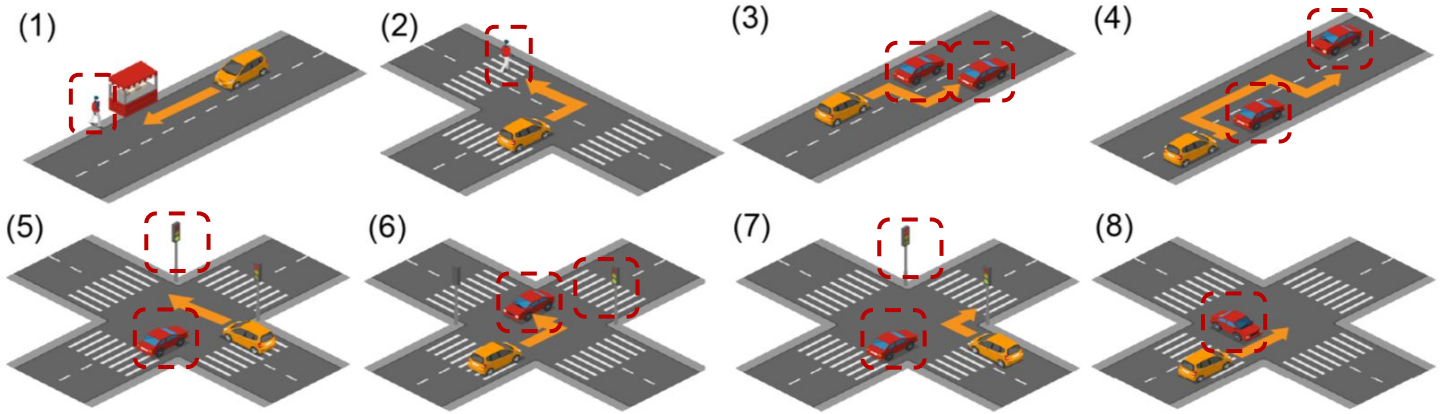
How do we find the failure boundaries?



[1]Arief, Mansur, Zhepeng Cen, Zhenyuan Liu, Zhiyuan Huang, Bo Li, Henry Lam, and Ding Zhao. "Certifiable Evaluation for Autonomous Vehicle Perception Systems Using Deep Importance Sampling (Deep IS)." In *Proceedings of* the *2022 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022. [Link]

Deep IS numerical experiments

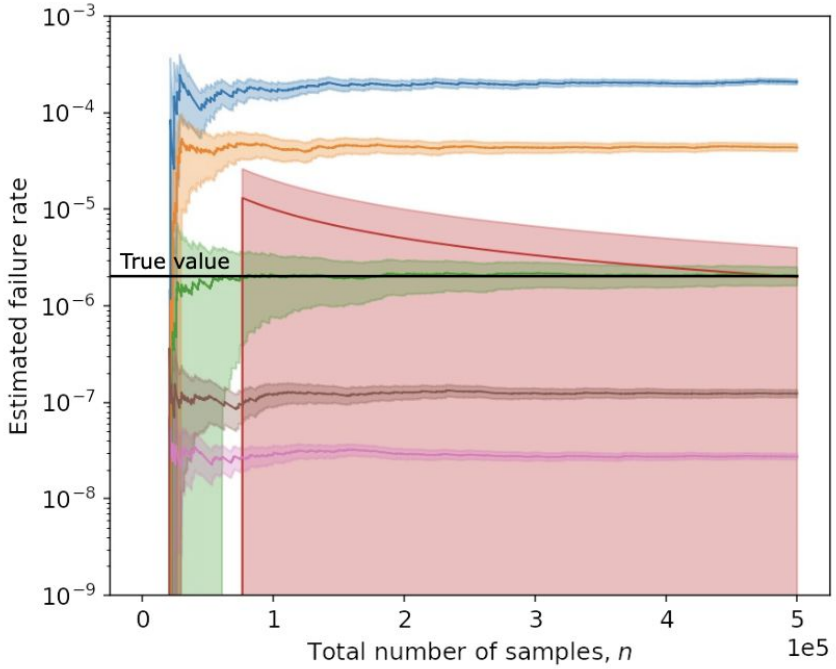
- **Driving scenarios**: Eight driving scenarios as defined in SafeBench [1].

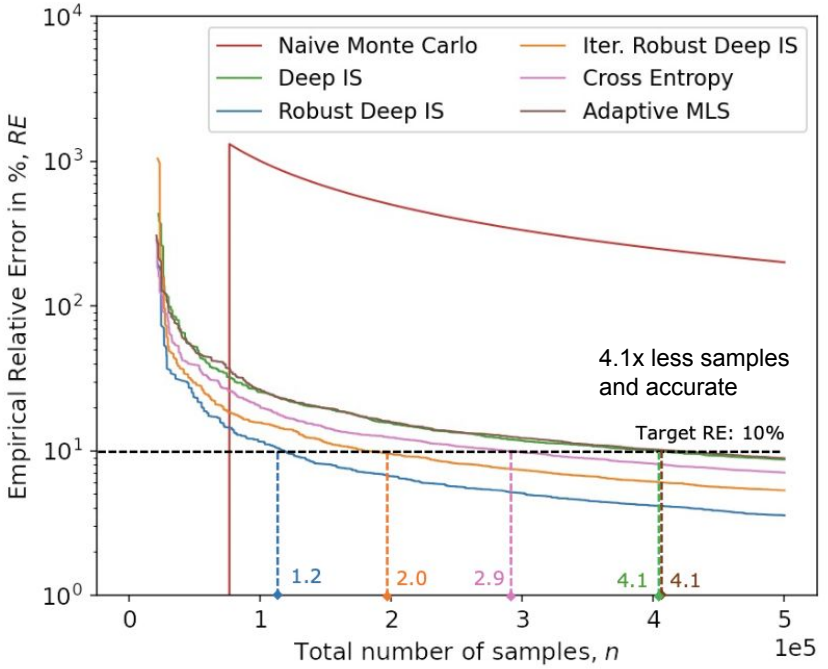


(1) Straight Obstacle, (2) Turning Obstacle, (3) Lane Changing, (4) Vehicle Passing,
(5) Red-light Running, (6) Unprotected Left-turn, (7) Right-turn, and (8) Crossing Negotiation.

Xu, Chejian, et al. "Safebench: A benchmarking platform for safety evaluation of autonomous vehicles." Advances in Neural Information Processing Systems 35 (2022): 25667-25682.

Numerical results



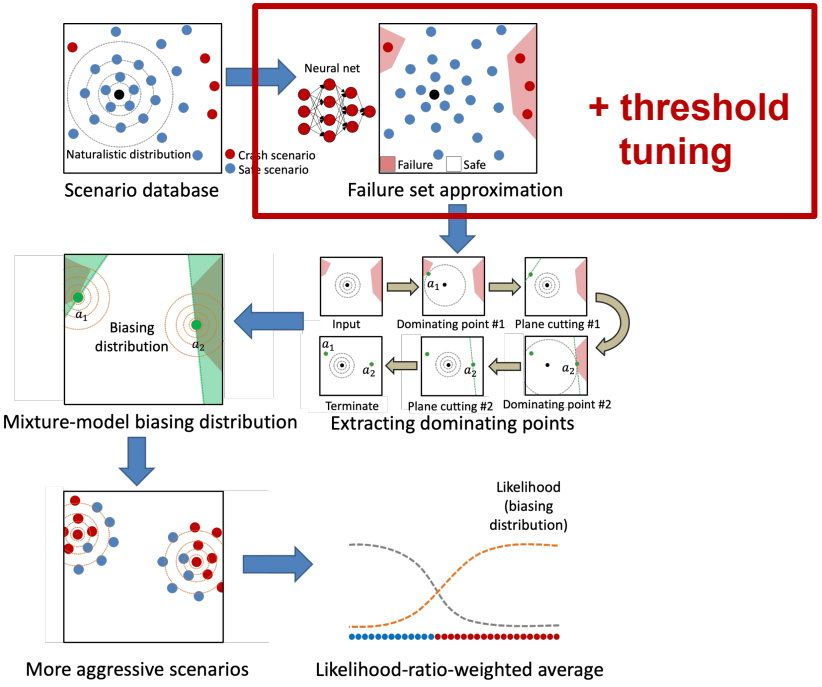


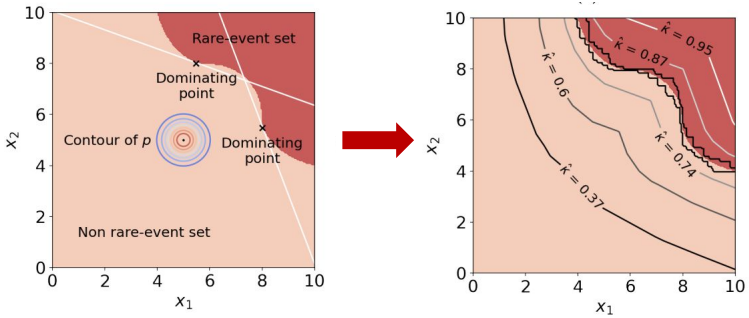
[1]Arief, Mansur, Zhepeng Cen, Zhenyuan Liu, Zhiyuan Huang, Bo Li, Henry Lam, and Ding Zhao. "Certifiable Evaluation for Autonomous Vehicle Perception Systems Using Deep Importance Sampling (Deep IS)." In *Proceedings of* the *2022 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022. [Link]

Highlights

- Looks good in paper, but in practice, we may not know if we have accurate failure set approximation
- What if we are inaccurate…
  - can we avoid underestimating the failure rate at all cost?
  - can we terminate early if an overestimate is allowed?

Deep-PrAE: Using failure set outer-approximation

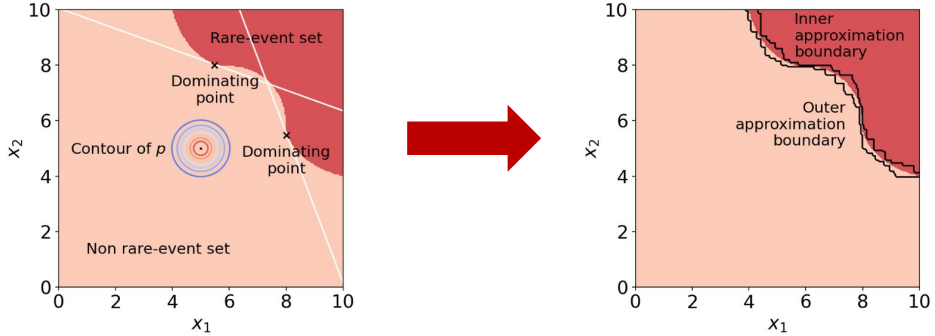




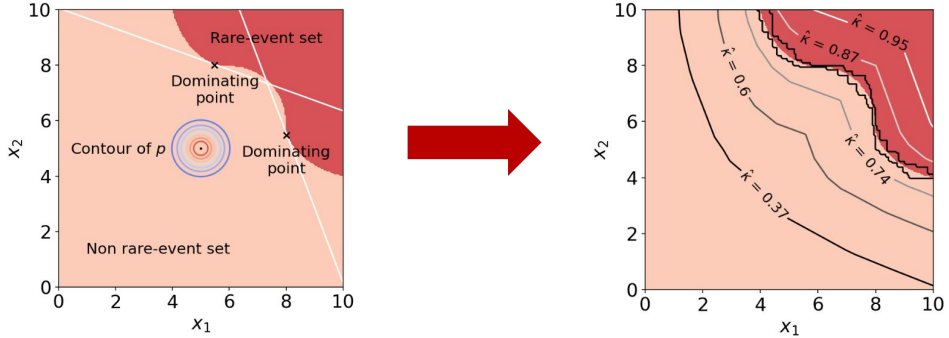
[2]Arief, Mansur, Zhiyuan Huang, Guru Koushik Senthil Kumar, Yuanlu Bai, Shengyi He, Wenhao Ding, Henry Lam, and Ding Zhao. "Deep Probabilistic Accelerated Evaluation: A Certifiable Rare-Event Simulation Methodology for Black-Box Autonomy." In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2021. [Link]

Obtaining outer approximation

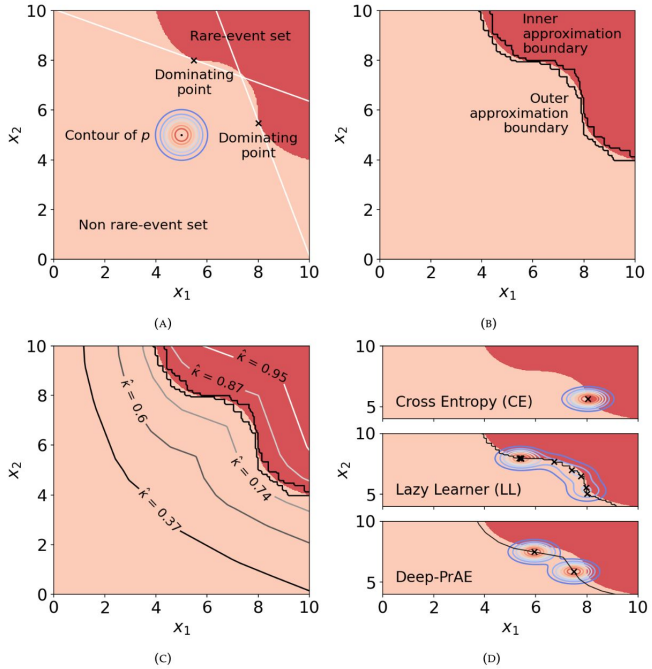
- **Option #1** using an orthogonally monotone hull

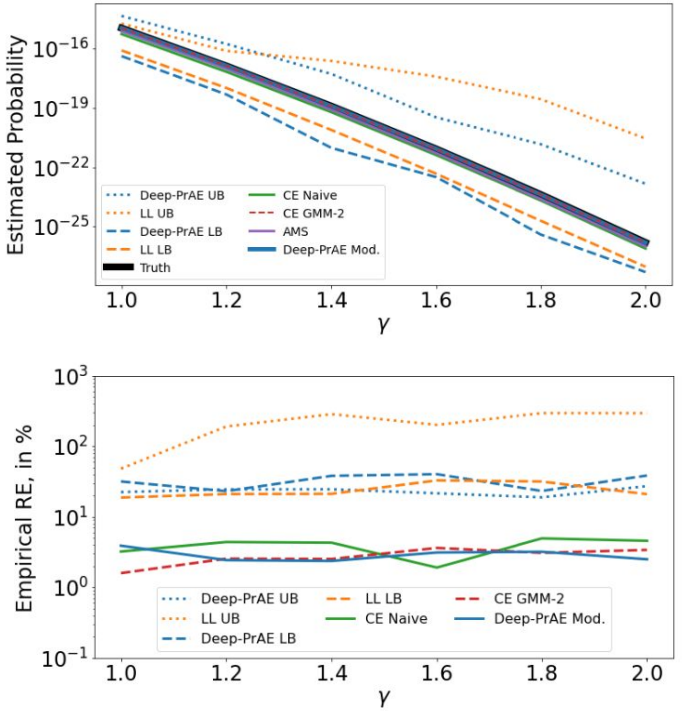


- **Option #2** gradually increase the decision threshold



Deep-PrAE estimates failure rate upper bound efficiently





[2]Arief, Mansur, Zhiyuan Huang, Guru Koushik Senthil Kumar, Yuanlu Bai, Shengyi He, Wenhao Ding, Henry Lam, and Ding Zhao. "Deep Probabilistic Accelerated Evaluation: A Certifiable Rare-Event Simulation Methodology for Black-Box Autonomy." In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2021. [Link]

Limitation

- Threshold tuning gives a loose over-approximation in higher dimensional space.
- Over-approximation might not easy to guarantee.

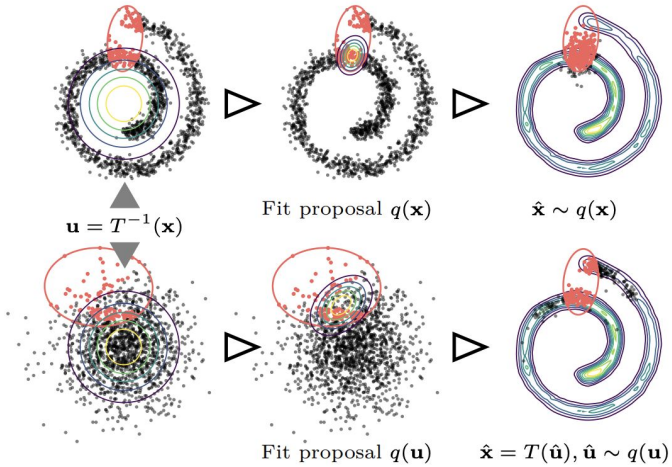
Key Takeaways

- Good to use if an failure rate upper-bound is sufficient for the task.
- Not very useful if accurate failure rate estimation is desired.
- Scaling up to high-dimensional problems is a major challenge.
- Learn more about importance sampling theory:

  Bucklew, James Antonio, and J. Bucklew. Introduction to rare event simulation. Vol. 5. New York: Springer, 2004.

Some more recent work: Normalizing Flow IS

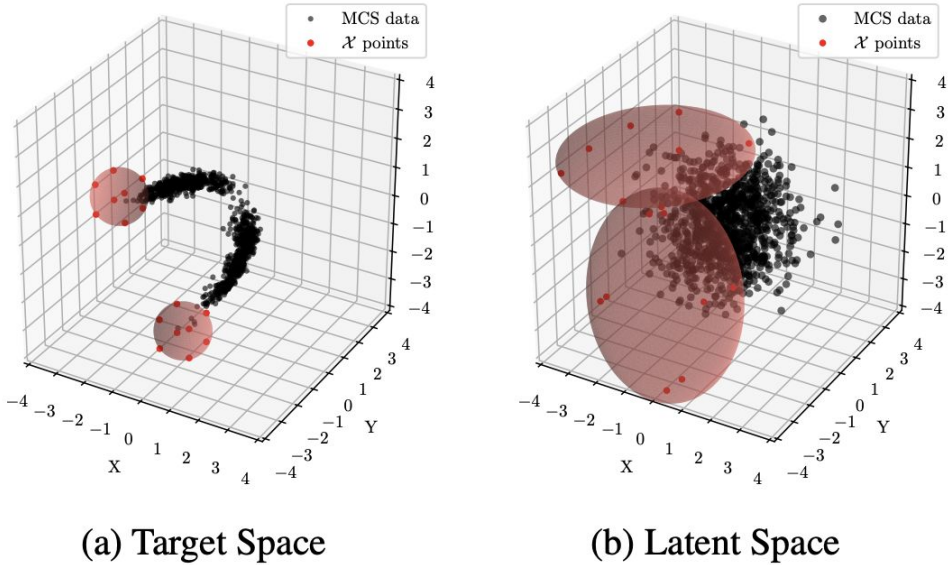
- If the failure set is too complex in input space, but mappable to easier latent space, use normalizing flows.



Kruse, Liam A., et al. "Enhanced Importance Sampling through Latent Space Exploration in Normalizing Flows." arXiv preprint arXiv:2501.03394 (2025).

Some more recent work: Normalizing Flow IS

- An example with non-holonomic robot failure rate estimation:

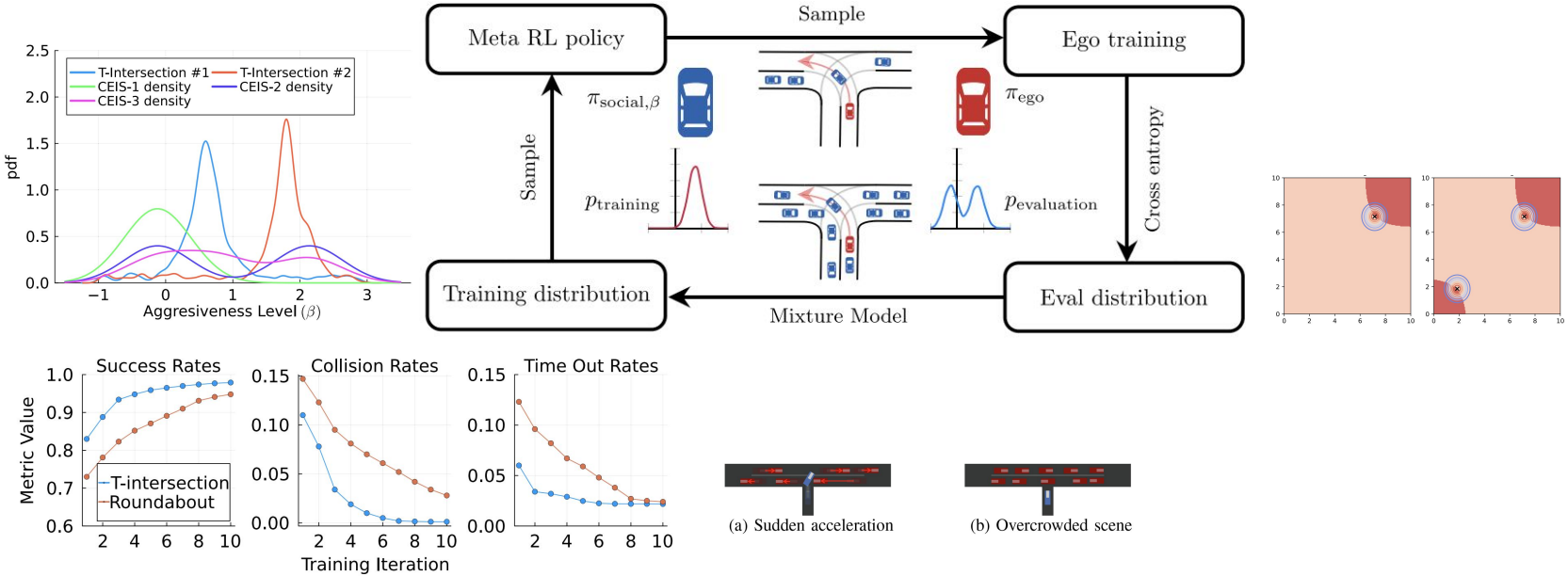


(a) Target Space (b) Latent Space

Kruse, Liam A., et al. "Enhanced Importance Sampling through Latent Space Exploration in Normalizing Flows." arXiv preprint arXiv:2501.03394 (2025).

Some more recent work: Integrated validation+training

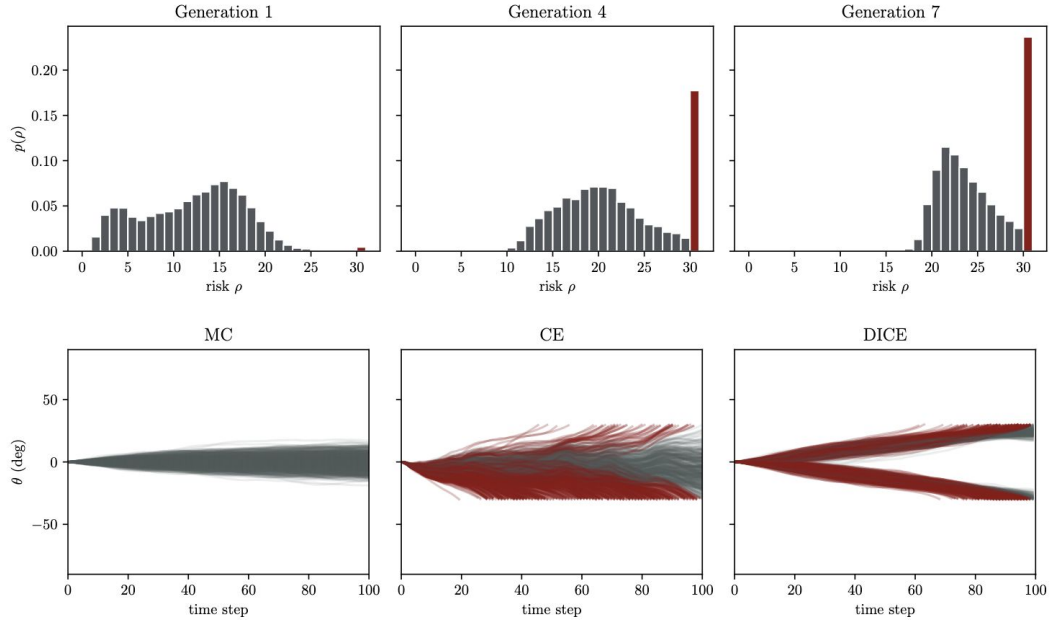
- Can we use the failure modes to generate samples and improve the agent?



Arief, Mansur, et al. "Importance Sampling-Guided Meta-Training for Intelligent Agents in Highly Interactive Environments." IEEE Robotics and Automation Letters (2024).

Some more recent work: Diffusion model + IS

- **Using diffusion models as proposal distribution** (on-going work by Harrison Delecki and Marc Schlichting)
- Extensions to dynamic and partially observable environment



Delecki, Harrison, et al. "Diffusion-Based Failure Sampling for Cyber-Physical Systems." arXiv preprint arXiv:2406.14761 (2024).

Q&A

- **Thanks to:** Mykel Kochenderfer (Stanford), Jef Caers (Stanford), Ding Zhao (CMU), Henry Lam (Columbia), Bo Li (UIUC), Jiachen Li (UCR), David Isele (HRI)
- **Thanks to SISL**: Liam Kruse, Harrison Delecki, Marc Schlichting, Anthony Corso, Robert Moss, Sydney Katz, Licio Romao, Kiana Jafari, Duncan Eddy

Let’s stay in touch

Mansur M. Arief

Research Engineer, SISL and MineralX

Stanford University

Email: mansur.arief@stanford.edu

Web: https://mansurarief.github.io/