Failure Prediction with Statistical Guarantees for Vision-Based Robot Control

Alec Farid,
David Snyder,
Allen Z. Ren,
Anirudha Majumdar *(Princeton University)*

presented by Yancheng Liang
07/26/2022

2.Using some learning theory techniques (PAC-Bayes)

Failure Prediction with Statistical Guarantees for Vision-Based Robot Control

1.Problem setup

1.Problem setup: reduce vision-based failure prediction to a supervised learning problem

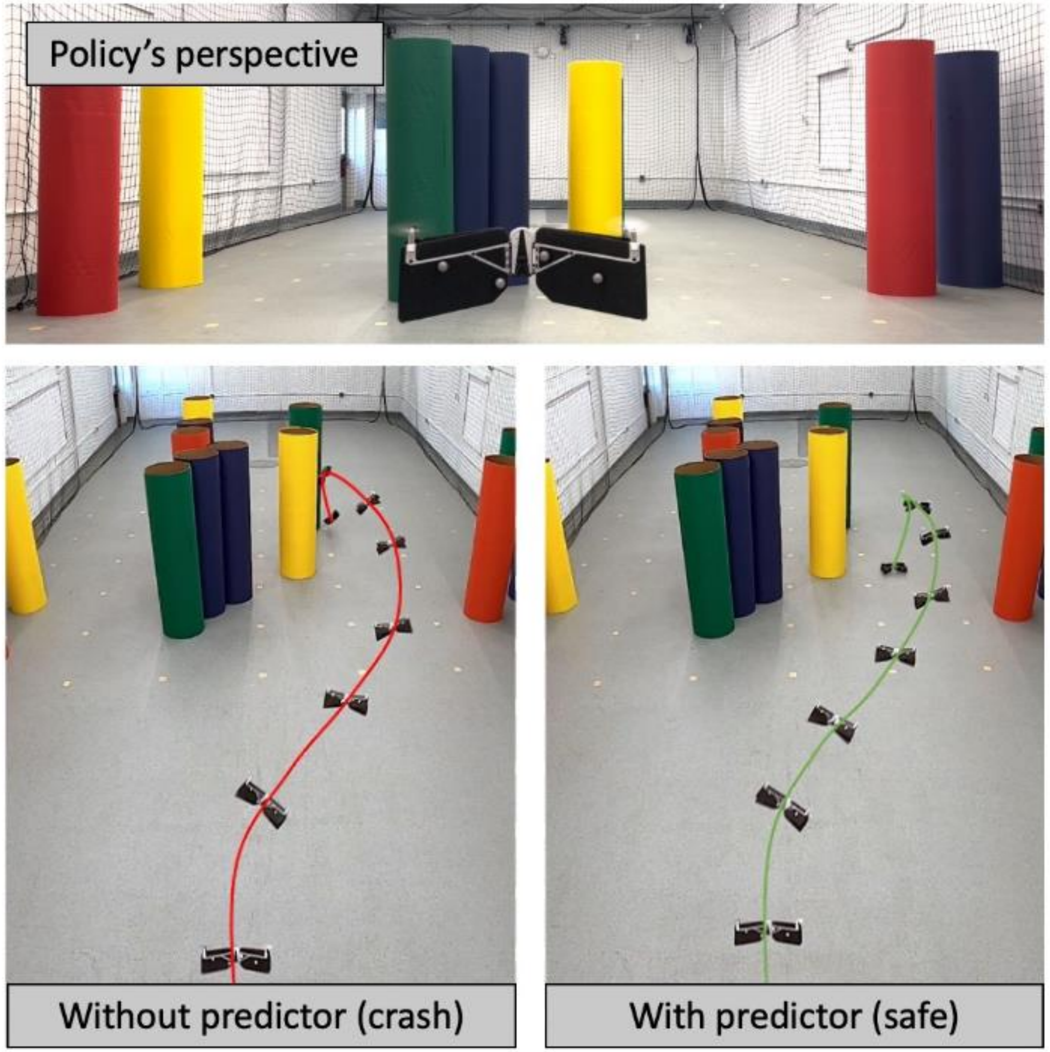
2.They provide a generalization bound for both total error loss and class-conditional loss, a direct result of PAC-Bayes theory

Vision-based Failure Prediction

1. Assume there is a (black-box) policy that takes image as input

2. The predictor also takes images (depth-images) as input, and predicts whether the robot will crash anytime in the future

3. The predictor is trained with a loss from PAC learning theory



Problem Formulation

- Let $\mathcal{E}$ be the space of environments (arrangement of obstacles), Π be the space of control policies.
- Consider the mapping function $r_f: \mathcal{E} \times \Pi \rightarrow X^T \times Y^T$. Here we can sample an environment $E \sim D_{\mathcal{E}}$ and rollout the trajectory $x_{1\ldots T}$ by deploying policy $\pi \in \Pi$ in E. After that, the predictor $f(x_{1\ldots t}) = \hat{y}_t$ predicts whether there is a failure in the whole trajectory.
- Then the error can be formulated as $\mathrm{C}\left(\mathrm{r_f}(\mathrm{E}, \pi)\right) \coloneqq 1[\mathrm{y} \neq \max_{\mathrm{t}<\mathrm{T_{fail}}} \hat{\mathrm{y}}_\mathrm{t}]$, which means the predictor does not make a correct prediction before the first failure time-step.

Class-conditional Loss

- $\mathrm{C}\left(\mathrm{r_f}(\mathrm{E}, \pi)\right) \coloneqq 1[\mathrm{y} \neq \max_{\mathrm{t}<\mathrm{T_{fail}}} \hat{\mathrm{y}}_\mathrm{t}]$
- Total error optimization objective: $\inf_{D_F} \mathbb{E}_{E\sim D_\mathcal{E}, f\sim D_F}\, [\mathrm{C}\left(\mathrm{r_f}(\mathrm{E}, \pi)\right)]$, a supervised learning problem

- Introduce weight $\lambda_0 + \lambda_1 = 1$ and a new optimization objective
$$\inf_{D_F} \mathbb{E}_{E\sim D_\mathcal{E}, f\sim D_F}\left[\tilde{C}\left(\mathrm{r_f}(\mathrm{E}, \pi)\right)\right] := \inf_{D_F} \mathbb{E}_{E\sim D_\mathcal{E}, f\sim D_F}\, [\lambda_y \mathrm{C}\left(\mathrm{r_f}(\mathrm{E}, \pi)\right)]$$
- Motivation: the predictor will be biased when the failed states and the successful states are imbalanced

PAC-Bayes Theory

- PAC theory: an analysis on the generalization error

- Let D_Z be a (unknown to the learner) distribution on input space Z, and $S = \{z_1, \dots, z_N\}$ is $N\ i.i.d.$ samples from D_Z. Then PAC theory tries to give a bound with high probability $1 - \delta$ over the selection of training samples S, that for any hypothesis (predictor/learner) h s.t. there is a bound for the generalization error
- $\mathbb{E}_{z \sim D_Z}[l(z, h)] \leq l_S(h) + Generalization\ Bound$

- PAC-Bayes theory further assumes a prior on hypothesis distribution $p(h) \sim D_H$

PAC-Bayes Theory

- Let $l(h, z) \in [0,1]$ be the loss function of hypothesis h and input z.
- Let D_Z be a distribution on input space Z, $p(h)$ the prior of hypothesis h, then with at least $1 - \delta$ probability over the selection of samples $S = \{z_1, \dots, z_N\}$, the following generalization bounds holds for every hypothesis h:
- $l_{D_Z}(h) \leq l_S(h) + \sqrt{\frac{\ln \frac{1}{p(h)\delta}}{2N}}$

PAC-Bayes Theory

- $l_{D_Z}(h) \le l_S(h) + \sqrt{\frac{2\ln\frac{1}{p(h)\delta}}{N}} = l_S(h) + \epsilon$
- Consider the fraction of samples S that the above bound is violated
- According to Chernoff bound,
- $\Pr\left(\left|l_S(h) - \mathbb{E}_{D_Z}[l(h,z)]\right| \ge \epsilon\right) \le e^{-\frac{N\epsilon^2}{2}} = p(h)\delta$
- Take union bound, the total fraction of S is at most $\sum p(h)\delta = \delta$

- Intuitively, the prior $p(h)$ means how much the model is attended to the hypothesis h and thus the bound for those h with high prior will be tight. Therefore, it’s better to use a prior that is close to the best model.

Back to the Failure Prediction Error

- Learning objective $\inf_{D_F} \mathbb{E}_{E\sim D_{\mathcal{E}}, f\sim D_F} [\mathrm{C}(\mathrm{r_f}(\mathrm{E}, \pi))]$
- PAC-Bayes bound by a predictor prior distribution D_0
- $\mathbb{E}_{E\sim D_{\mathcal{E}}, f\sim D_F}[C(r_{f(E,\pi)})] \leq C_S(D_F) + \sqrt{\frac{KL(D_F||D_0)+\ln(2\sqrt{N}/\delta)}{2N}}$

$$p_F \log \frac{p_F}{p_0} \approx \log \frac{1}{p_0(f)}$$

Supervised Learning	$\leftarrow$	Failure Prediction
Input Data $z \in \mathcal{Z}$		Environment $E \in \mathcal{E}$
Hypothesis $h_w : \mathcal{Z} \to \mathcal{Z}'$		Rollout $r_f : \mathcal{E} \times \Pi \to \mathcal{X}^T \times \mathcal{Y}^T$
Loss $l(w; z)$		Error $C(r_f(E, \pi))$

Class-Conditioned Failure Prediction Error

Any error is generally a weighted class-conditioned error

$$\begin{aligned} p_{\text{error}} &= p_{0\cap 1} + p_{1\cap 0} \\ &= p_{0|1}p_1 + p_{1|0}p_0 \\ &= p_{0|1}(1-\lambda^*) + p_{1|0}(\lambda^*) \\ \rightarrow \text{ generalize to } &= p_{0|1}(1-\lambda) + p_{1|0}(\lambda), \end{aligned} \tag{10}$$

General weighted loss. Problem: $\hat{p}$ depends on sampling.

$$\hat{C}_S(r_f(E,\pi),S) \triangleq (1-\lambda)\hat{p}_{0|1} + \lambda\hat{p}_{1|0}.$$

Solution: a global high-fidelity lower bound estimation $\underline{p}$

$$\begin{aligned} \mathop{\mathbb{E}}_{E\sim\mathcal{D}_\mathcal{E}}\mathop{\mathbb{E}}_{f\sim\mathcal{D}_\mathcal{F}}\left[\tilde{C}(r_f(E,\pi))\right] &\le \hat{C}_S(r_f(E,\pi),S) + R_\lambda, \\ R_\lambda = \frac{5}{3}\sqrt{\frac{(1-\underline{p})\log\frac{2}{\delta}}{N\underline{p}}} &+ C_\lambda R(\mathcal{D}_\mathcal{F},\mathcal{D}_{\mathcal{F},0},\delta). \end{aligned} \tag{13}$$

Implementation: Obstacle Avoidance with a Drone

- Policy: a trained classification DNN that chooses a trajectory out of a pre-defined set. They use a motion capture system to mitigate the sim-to-real gap.
- The predictor takes four recent images as input. 10000 environments are used to first train a prior, and 10000 other environments use PAC-Bayes upper bound to train the failure predictor.
- During test, an emergency policy is activated if a failure is predicted

TABLE II
RESULTS FOR FAILURE PREDICTION ON NAVIGATION TASK

Setting	Standard	Occluded Obstacle
True Expected Failure (Sim)	0.253	0.514
Misclassification Bound	0.128	0.154
True Expected Misclassification (Sim)	0.101	0.125
True Expected Misclassification (Real)	0.067	0.133

Original failure rate

Training error

Test error (15 trails)