On Continuous Monitoring of Risk Violations under Unknown Shift

Alexander Timans*[,1] Rajeev Verma*[,1] Eric Nalisnick[2] Christian A. Naesseth[1]

uai2025

[1] University of Amsterdam — AMLAB Amsterdam Machine Learning Lab

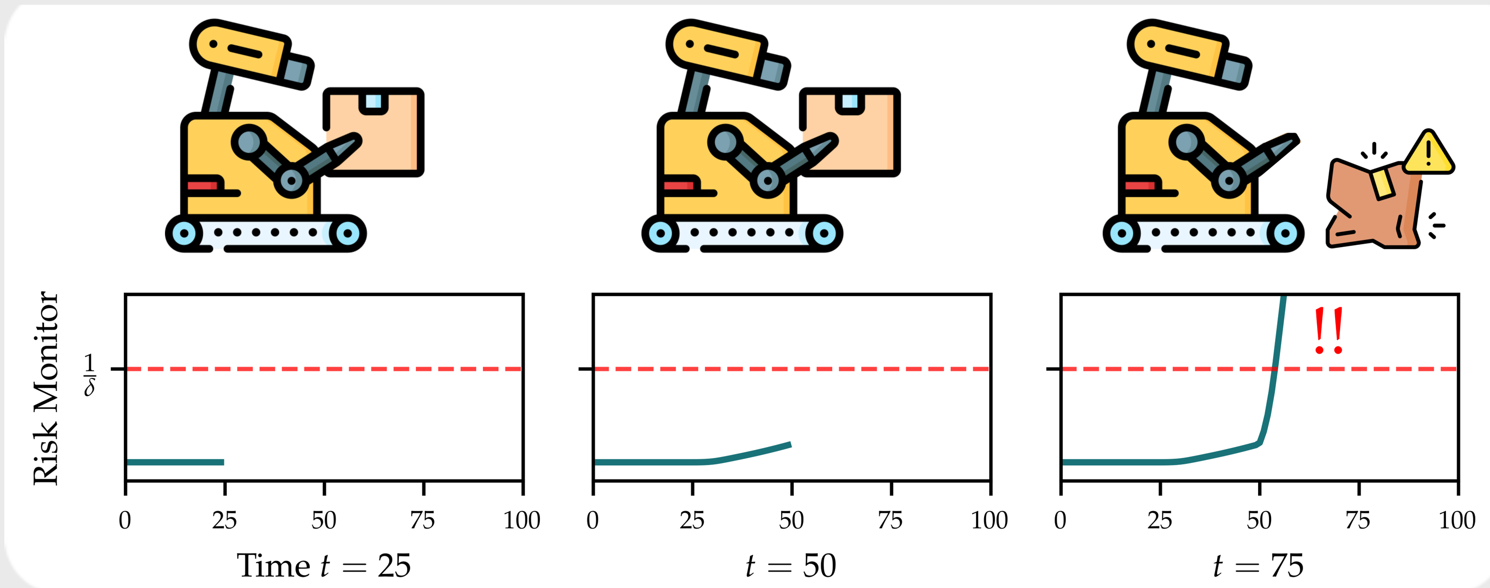
[2] Johns Hopkins University

Motivation

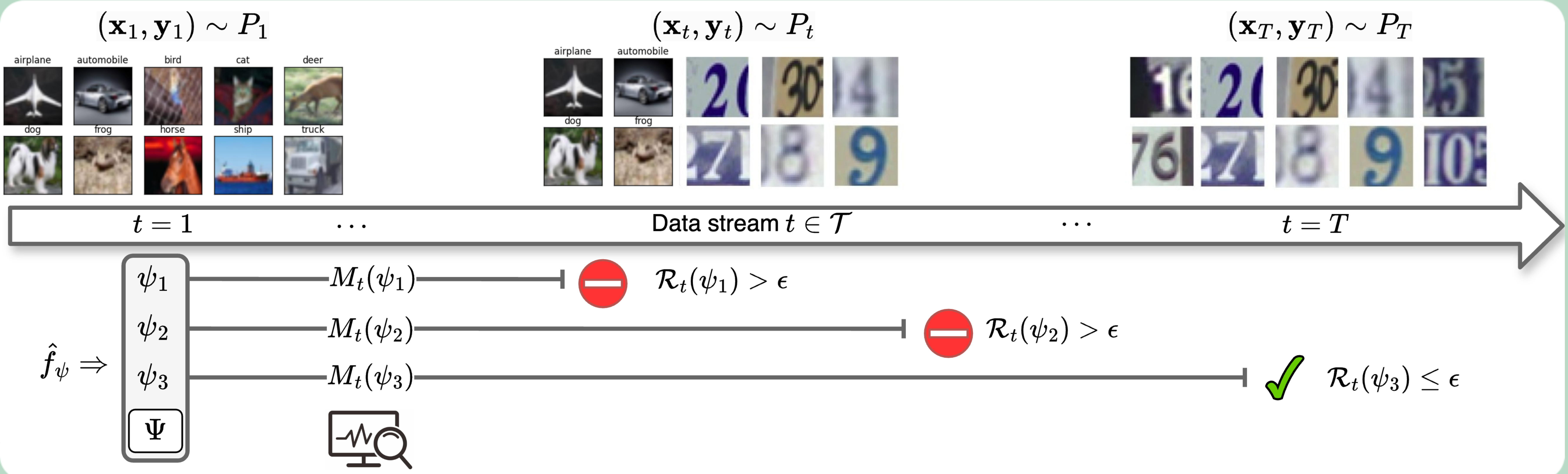
Can we monitor risk development in a deployed model on data streams continuously, with minimal assumptions on the nature of encountered data (under shift), and with statistical reliability?

- Problem: Common frameworks are static or assume i.i.d. data.
- Solution: Risk monitoring based on sequential hypothesis testing (testing-by-betting) with false alarm guarantees.

Example: Production Monitoring



Monitoring as Sequential Testing



- Sequential hypothesis test (no shift assumptions):

$$H_0(\psi) : \mathbb{E}_{P_t}[\mathbf{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon \;\forall t \in \mathcal{T} \quad \text{(risk controlled)}$$

$$H_1(\psi) : \exists t \in \mathcal{T} : \mathbb{E}_{P_t}[\mathbf{z}_t \mid \mathcal{F}_{t-1}] > \epsilon, \quad \text{(risk violated)}$$

- Risk monitor (Test martingale / Wealth / E-Process):

$$M_t(\psi) = \prod_{i=1}^{t} (1 + \lambda_i (\mathbf{z}_i - \epsilon)) \quad \text{with} \quad M_0 = 1,\ \lambda_t \in [0, \frac{1}{\epsilon})$$

- Threshold-based decision model:

$$\hat{f}_\psi(\mathbf{x}) = g(\hat{f}(\mathbf{x}), \psi),\ \psi \in \Psi$$

- Supervised & bounded risk:

$$\mathcal{R}_t(\psi) = \mathbb{E}_{P_t}[\mathbf{z}_t],\ \mathbf{z}_t = \ell(\hat{f}_\psi(\mathbf{x}_t), \mathbf{y}_t) \in [0, 1]$$

- False alarm guarantee (Type-I error control):

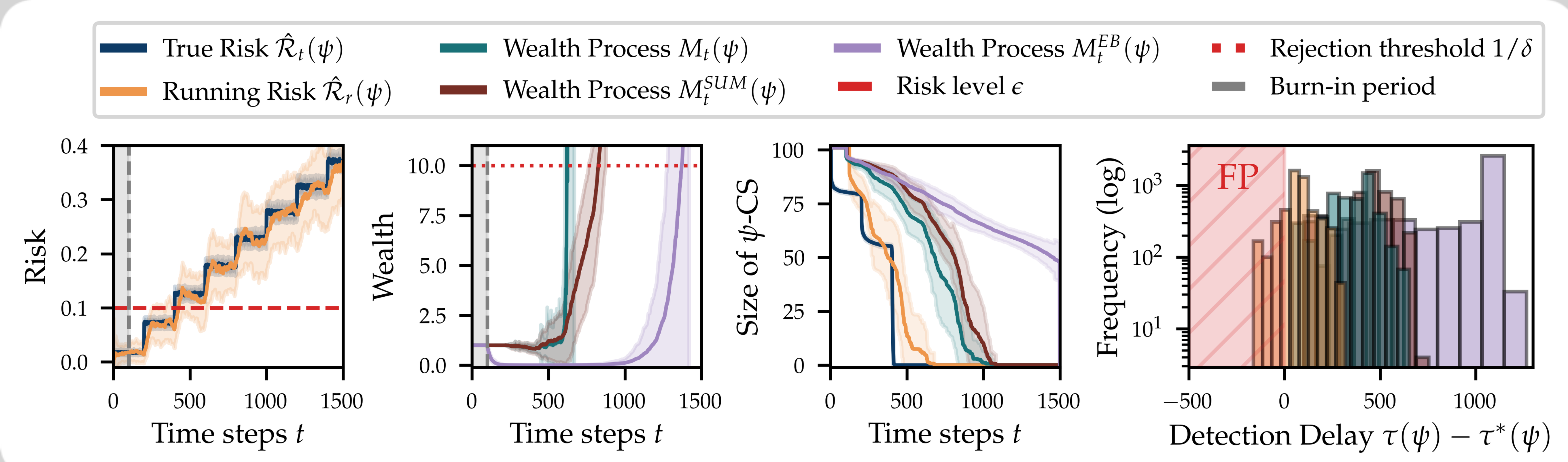
$$\mathbb{P}_{H_0}(\exists t \in \mathcal{T} : M_t(\psi) \geq 1/\delta) \leq \delta$$

Experiments

- Minimize detection delay while ensuring guarantee
- Characterized delay behavior (Prop. 4.5)

Total Error Rate for Outlier Detection

- Outlier labelling: $\hat{f}_\psi(\mathbf{x}) = \mathbf{1}[\text{out}(\mathbf{x}) \geq \psi]$
- TER = FP + FN
- Stepwise shift via mixture sampling
- Monitoring reactive and upholds guarantees



Also in the paper

- Miscoverage rate for set prediction
- Natural temporal shifts (FMoW, Naval)
- Classification & Regression
- More theoretical analysis

References

- Waudby-Smith & Ramdas (2024). Estimating Means of Bounded Random Variables by Betting (J. R. Stat. Soc. B)
- Ramdas et al. (2023). Game-Theoretic Statistics and SAVI (Stat. Science)
- Podkopaev & Ramdas (2022). Tracking the Risk of a Deployed Model and Detecting Harmful Distribution Shifts (ICLR)
- Feldman et al. (2023). Risk Control in Online Learning (TMLR)