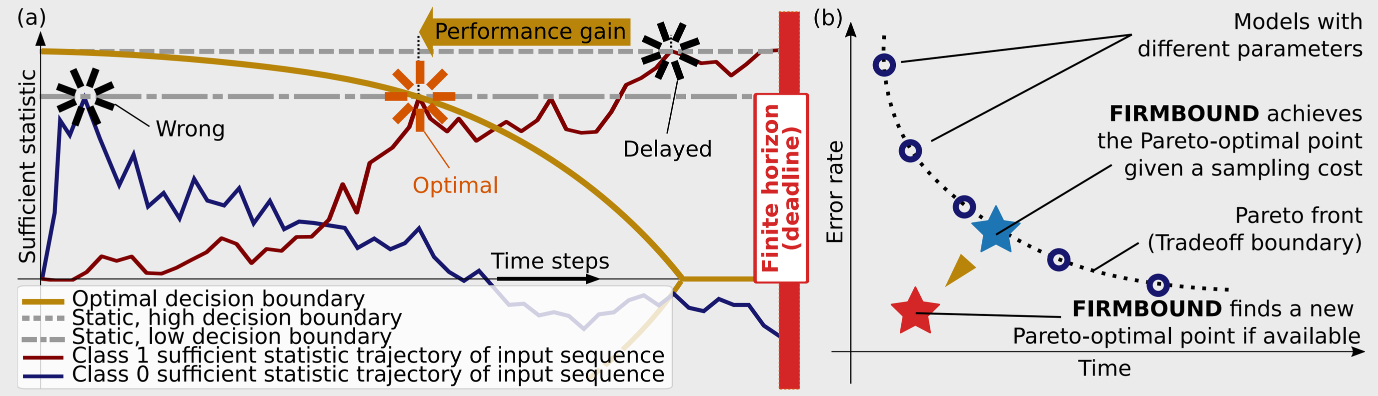
Learning the Optimal Stopping for Early Classification within Finite Horizons via Sequential Probability Ratio Test

Akinori F. Ebihara (aebihara@nec.com), Taiki Miyagawa, Kazuyuki Sakurai, Hitoshi Imaoka

NEC Corporation, Japan

Early Classification Is Pivotal for Time-Sensitive Analysis

Sequential Probability Ratio Test (SPRT) Is Suboptimal Under *Finite Horizon* Constraints



Finding Optimal Thresholds Is Computationally Intensive

Bayes Risk to be minimized with the optimal threshold

Sufficient statistic i.e., Log-likelihood ratio (LLR) or posterior; Decision to classify as class k; Sequential data; Class posterior; Constant; time

$$\mathrm{APR}_t(\mathscr{S}_t, d_t(X^{(1,t)}) = k) := \bar{L}_k(1 - \pi_k(X^{(1,t)})) + ct$$

A posteriori risk (APR); Constant; Misclassification Penalty; Sampling cost

Backward induction equation to minimize the Bayes Risk

Curse of dimensionality at estimating the conditional expectation

Continuation risk
$$\tilde{G}_t(\mathscr{S}_t) = \mathbb{E}\left[G^{\min}_{t+1}(\mathscr{S}_{t+1}) | \mathscr{S}_t\right] + c$$

Stopping risk
$$G^{st}_t(\mathscr{S}_t) = \min_k \left\{\bar{L}_k(1 - \pi_k(X^{(1,t)}))\right\}, \quad \text{Time } t \in \{1, \cdots, T\}$$

Minimum risk
$$G^{\min}_t(\mathscr{S}_t) := \begin{cases} G^{st}(\mathscr{S}_t) & (t = T) \\ \min\left\{G^{st}(\mathscr{S}_t), \tilde{G}_t(\mathscr{S}_t)\right\} & (1 \le t < T). \end{cases}$$

FIRMBOUND Is "Doubly-Consistent"

- providing consistent estimation of likelihood ratio and the thresholds

Noisy convex regression

$$\hat{\tilde{G}}_t(\{X^{(1,T)}_m\}^M_{m=1}) \in \operatorname*{arg\,min}_{f:\,\text{concave}} \left\{\frac{1}{M}\sum_{m=1}^{M}\left(f(\mathscr{S}_t(X^{(1,t)}_m)) - \mathscr{G}^{(t)}_m\right)^2 + \lambda\|f\|\right\} + c$$

Total #data; Concave function; Continuation risk with noise; Mean squared error; Regularization term; Sampling cost; Constant

$$\mathscr{G}^{(t+1)}_m = \hat{\tilde{G}}_t(\mathscr{S}_t(X^{(1,t)}_m)) + \epsilon^{(t)}_m - c$$

Continuation risk; Sequential data; Noise

cf) Siahkamari et al. 2022

(Optional) Gaussian Process (GP) regression for lightweight training

- Significantly reduces training time with potential compromise in statistical consistency
- Assuming $\epsilon^{(t)}_m$ and $\{\tilde{G}_t(\mathscr{S}_{t,m})\}_{m\in[M]}$ are Gaussian noise and GP, respectively, the conditional expectation estimation problem is reduced down to GP regression problem:

$$\mathscr{G}^{(t+1)}_m + c = \tilde{G}_t(\mathscr{S}_{t,m}) + \epsilon^{(t)}_m$$

Response variable; Latent function; Explanatory variable; Gaussian noise

Likelihood ratio estimation with SPRT-TANDEM

cf) Ebihara et al. 2021, Ebihara and Miyagawa, 2021

Trainable parameters; Total #classes; Ground-truth label; Total #data; Total #data per class; Estimated LLR

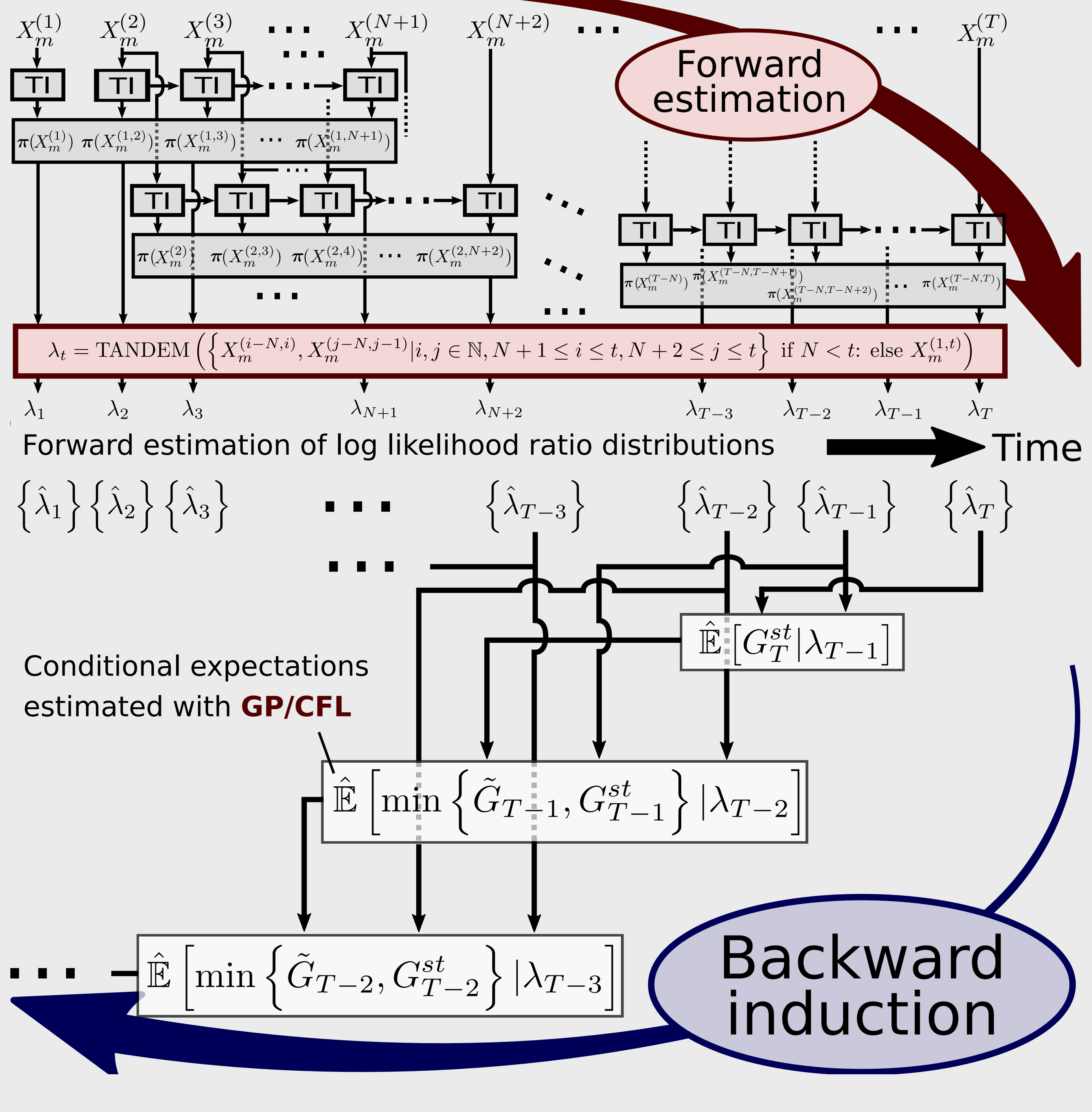
$$\hat{L}_{\mathrm{LSEL}}(\boldsymbol{w}; \{(X^{(1,T)}_m, y_m\}_{m\in[M]}) := \frac{1}{KM}\sum_{k\in[K]}\sum_{t\in T}\frac{1}{M_k}\sum_{i\in I_k}\log(1 + \sum_{l(\neq k)\in[K]} e^{-\hat{\lambda}_{kl}(\boldsymbol{w}, X^{(1,t)})})$$

Loss function LSEL

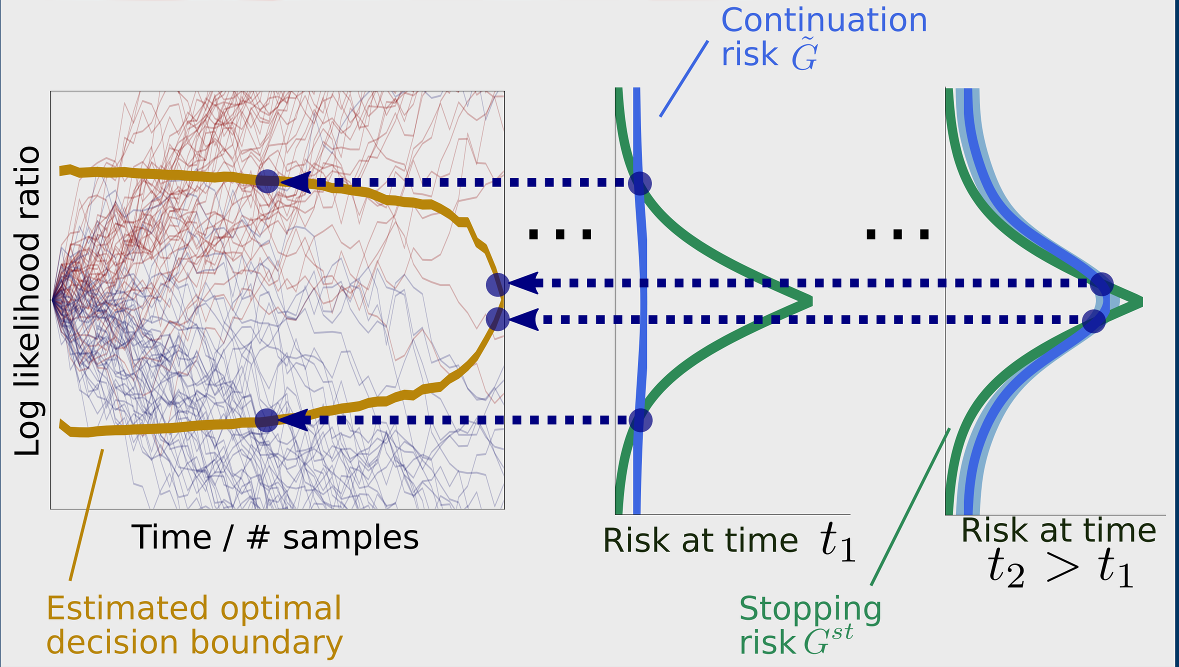
$$\hat{\lambda}_{kl}(X^{(1,t)}) = \sum_{s=N+1}^{t}\log\frac{\pi_k(X^{(s-N,s)})}{\pi_l(X^{(s-N,s)})} - \sum_{s=N+2}^{t}\log\frac{\pi_k(X^{(s-N,s-1)})}{\pi_k(X^{(s-N,s-1)})} - \log\chi_{kl}$$

LLR estimated with TANDEM formula; Class posterior; Order of Markov process; Class prior ratio

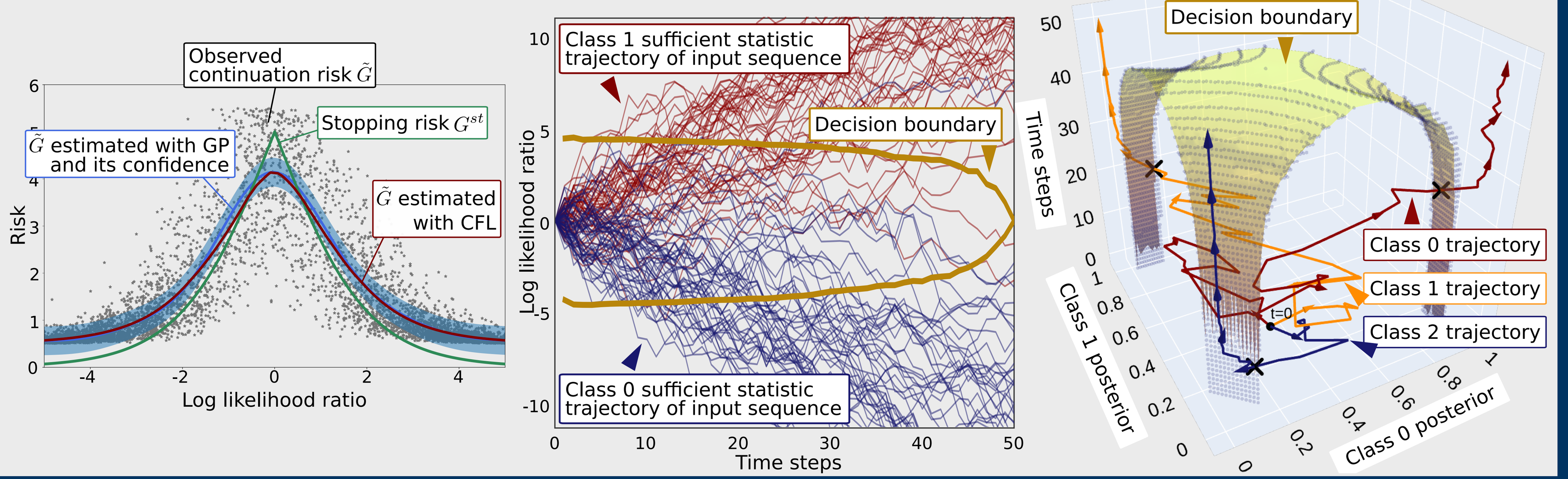
Training Pipeline



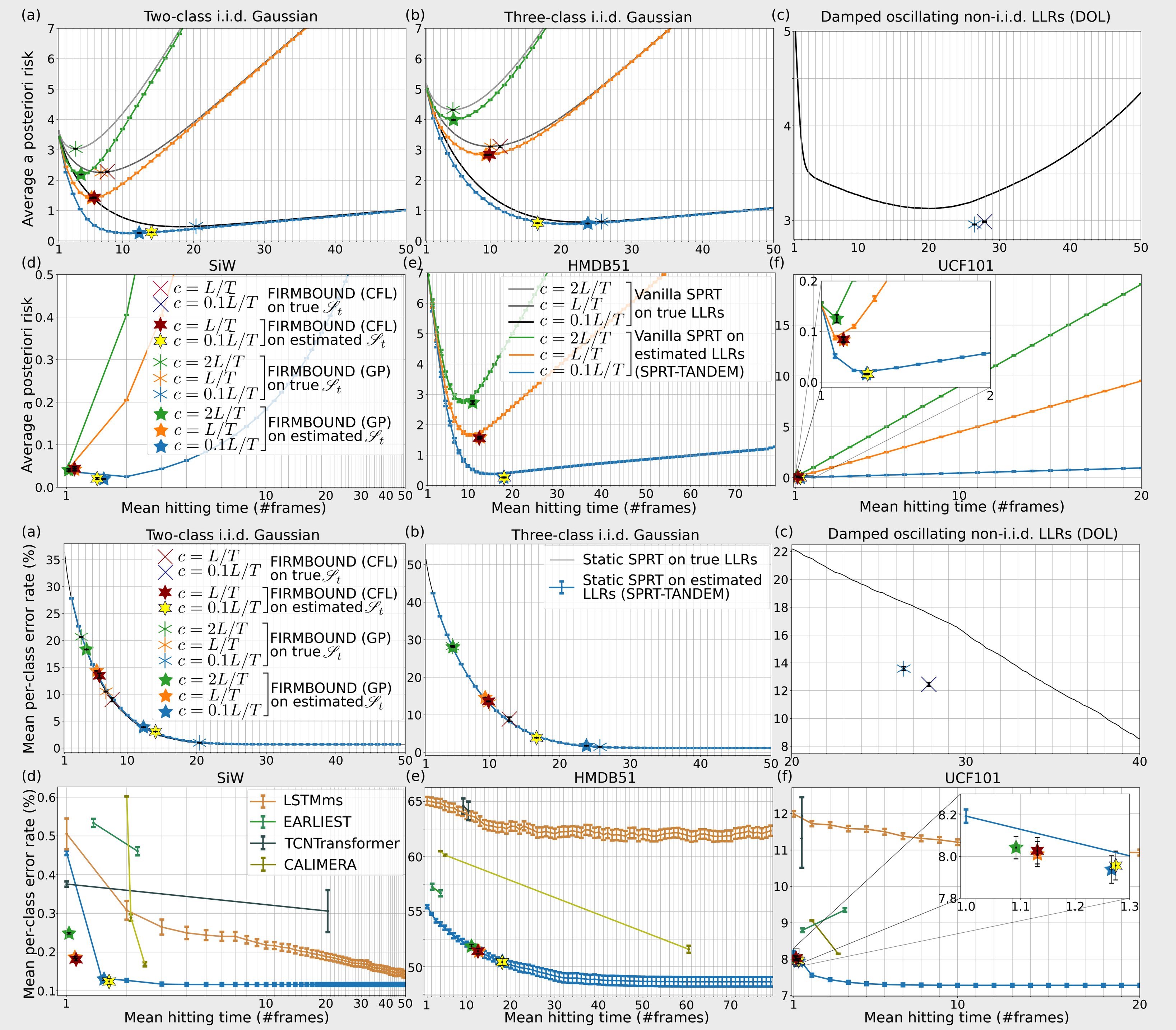
Optimal Threshold Defined by the Intersection of Two Curves



Example Estimated Thresholds on Gaussian Datasets



FIRMBOUND Minimizes the Bayes Risk to Delineate the Pareto Front



Dataset	Gauss2est.	Gauss3est.	DOL	SiW	HMDB	UCF101	FordA
Trial repeats	5	3	3	5	6	10	2
↓MVHT, vanilla SPRT with static threshold	10.47	44.02	489.89	2.87	199.31	0.55	32.15
↓MVHT, FIRMBOUND with CFL	9.01	42.78	405.78	1.97	195.35	0.53	23.39
↑Difference in MVHT (positive is better)	**1.45**	**1.24**	**84.11**	**0.90**	**3.97**	**0.017**	**8.76**

TL;DR

FIRMBOUND is an **SPRT-based early classification** framework that provides a **statistically consistent and computationally efficient estimator of optimal decision boundaries** for time series of finite lengths, tailored for large-scale real-world problems.