# Optimizing Detection Time and Specificity: Early Classification of Time Series with Sensitivity Constraint

**Jiaming Qiu** **Ying-Qi Zhao**

**Yingye Zheng**

Public Health Sciences Division, Fred Hutchinson Cancer Center

## Abstract

From the perspective of sequential decision making, we propose a novel approach for early classification of time series under the Neyman–Pearson paradigm that incorporates a sensitivity constraint. We explicitly formulate the optimal solution, which can be practically obtained utilizing plug-in estimators such as recurrent neural networks. Cast as a constrained multi-objective optimization problem, we establish the Pareto optimality balancing earliness and classification accuracy. Our approach visualizes the inherent trade-off between earliness and specificity, ensuring informed decision making without compromising sensitivity. Experimental validation confirms the feasibility of our approach, demonstrating its potential in various real-world applications.

## 1 Introduction

Early classification of time series (ECTS) is critical for real-time decision-making in domains such as healthcare, finance, and industrial monitoring, where timely interventions can significantly mitigate risks and costs. Unlike traditional time series classification that relies on complete data, ECTS aims to classify based on partial data, allowing for prompt actions.

A wide array of methods has been proposed (see, e.g., Gupta et al., 2020, for a review), focusing on the balance of classification accuracy and earliness. For instance, Mori et al. (2018, 2019) tackle ECTS from the perspective of multi-object optimization to balance earliness and accuracy. However, the consequence of false classification conditioning on actual outcome could be drastically different. For example, in certain medical settings, higher sensitivity is desired even at the cost of lower specificity so as not to miss lethal disease. The classic Neyman–Pearson paradigm (Neyman et al., 1933) has been successful balancing the trade-off between sensitivity and specificity, yet rarely explored in the setting of ECTS.

In sequential analysis where confidence accumulates along increased number of i.i.d. observations, Wald's sequential probability ratio test (SPRT) provides the earliest decision among all tests of the same power (Wald and Wolfowitz, 1948) by thresholding on the cumulative product of probability ratios. However, its optimality weakens given dependent samples (e.g., Schmitz, 1985). The requirement of unlimited observability also makes it less ideal for classifying time series with finite horizon, where decision must be made before certain terminal time.

In this study, we take the perspective of sequential decision making under finite horizon. For each time $t$ before the terminal time $T$, the classifier either conclude positively/negatively based on the series

Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

till $t$ or wait for further information from next time step. We evaluate the classifier by three criteria: sensitivity (true positive rate, TPR), specificity (equivalently 1 - false positive rate, FPR), and earliness (or equivalently cost) of decision. We propose a novel ECTS approach under the Neyman-Pearson paradigm motivated by multi-objective optimization. The proposed method incorporates a sensitivity constraint while seeks to simultaneously maximize the specificity and earliness.

By framing ECTS as a constrained multi-objective optimization problem, we demonstrate the inherent trade-offs among earliness and classification power. Such trade-off is quantified by the Pareto front, where neither earliness or specificity can be improved without sacrificing the other criteria. By decomposing accuracy into two separate metrics (sensitivity and specificity), we offer a more nuanced characterization of the trade-off that is otherwise unavailable in existing ECTS approaches. Based on a straight forward Lagrange dual problem, we explicitly formulate the optimal solution, which is tractable via plugging-in neural networks as showcased by numeric experiments.

## 2 Explicit Form of the Pareto Optimal Classifier

Denote the covariate process (i.e., time series to classify) as $X = X_{1:T}$ with $X_t \in \mathbb{R}$ for $t = 1, \ldots, T$[1] and the associated binary outcome (i.e., class label) as $Y \in \{0, 1\}$.[2] A sequential decision rule is

$$\varphi\left(X_1, \ldots, X_T\right) = \mathbb{1}(f_1 > 0) + \mathbb{1}(f_1 = 0, f_2 > 0) + \cdots + \mathbb{1}(f_{1:T-1} = 0, f_T > 0), \tag{2.1}$$

where $\mathbb{1}(\cdot) \in \{0, 1\}$ takes 1 if the condition is true, and 0 otherwise. We write $f_t = f_t\left(X_{1:t}\right) \in \mathbb{R}$ as the score at time $t$ based on the history $X_{1:t} = (X_1, \ldots, X_t)$; write $f_{1:t}$ as the scores till time $t$ with the convention that $f_{1:T-1} = 0$ means $f_1, \ldots, f_{T-1}$ are all zero. Denote the natural filtration spanned by $X_{1:T}$ as $\mathcal{F}$. The score $f_{1:T}$ is a stochastic process that is *$\mathcal{F}$-adapted*, i.e., $f_t$ is $\mathcal{F}_t$-measurable (equivalently, is measurable function of $X_{1:t}$) for all $t = 1, \ldots, T$.

The decision process could stop early before the terminal time $T$ once receiving a non-zero score. Denote $\tau = \inf\{1 \leq t \leq T : f_t \neq 0\}$ as the *associated stopping time* at which rule (2.1) makes a decision, so that $\varphi$ is the same as an early-stopped rule denoted as $\varphi_\tau{:=}\mathbb{1}(f_\tau > 0)$. To characterize the earliness of decision, we associate time steps with operational cost, either time or monetary. Denote the *accumulated cost* of running the decision process till time $t$ as $C_t$ for some increasing $0 < C_1 < \cdots < C_T$. The actual cost of the early-stopped decision rule is hence $C_\tau$. The introduction of cost allows more flexible modeling: setting $C_t = t/T$ for regularly sampled time series corresponds to the conventional *earliness* notion; one can also encourage earlier or later decision by adjusting cost accordingly.

We say a sequential rule is *efficient under Neyman–Pearson paradigm* if it is the optimum of the multi-objective optimization (MOO) problem:

$$\text{minimize } \left(\mathbb{E}[\varphi_\tau \mid Y = 0], \mathbb{E}[C_\tau]\right), \text{ subject to } \mathbb{E}[\varphi_\tau \mid Y = 1] \geq \beta \tag{2.2}$$

for some $\beta \in (0, 1)$, where $\varphi_\tau{:=}\mathbb{1}(f_\tau > 0)$, with expectations taken over $X_{1:T}$. Here, $\mathbb{E}[\varphi_\tau \mid Y = 1]$ and $\mathbb{E}[\varphi_\tau \mid Y = 0]$ are the true and false positive rate, while $\mathbb{E}[C_\tau]$ is the expected cost.

Usually multiple optima of a multi-objective problem exists, where none of the objectives could be further improved without sacrificing some other objectives, the collection of which is called the *Pareto front* (e.g., Pardalos et al., 2017). One common way to approximate MOO is the $\varepsilon$-constraints approach that translates all but one objectives into constraints, formulated as

$$\text{minimize } \mathbb{E}[\varphi_\tau \mid Y = 0], \text{ subject to } \mathbb{E}[C_\tau] \leq \gamma \text{ and } \mathbb{E}[\varphi_\tau \mid Y = 1] \geq \beta \tag{2.3}$$

for some $0 < \gamma < 1$. Tackling problem (2.3) utilizing Lagrangian multipliers provides explicit form of the optimal score functions. In fact, under mild conditions, the optimum of (2.3) is also Pareto optimal for problem (2.2), which is outlined in the following and detailed in Appendix A.

**Proposition 2.1.** *Denote the dual function of the Lagrangian function $\mathcal{L}\left(f; a, b\right)$ of (2.3) as $\mathcal{G}(a, b){:=}\inf_f \mathcal{L}\left(f; a, b\right)$ with the infimum is taken over all $\mathcal{F}$-adapted process $f_{1:T}$. Then*

$$\mathcal{G}(a, b) = b\beta - a\gamma - \mathbb{E}[S_1]. \tag{2.4}$$

[1] Note that we assume univariate $X_t$ only for convenience yet not required so. The proposed framework does not prohibit multivariate covariates or even different dimensionality of covariates at different $t$.

[2] We refer to 0 as negative and 1 as positive.

*Here, $S_1$ is defined recursively. Writing $p_1 = \mathbb{P}(Y = 1) = 1 - p_0$, we define $\mu_t {:=} \mathbb{E}[Y \mid X_{1:t}]$ and $\eta_t {:=} \left(bp_1^{-1} + p_0^{-1}\right)\mu_t - p_0^{-1}$. Then $S_T {:=} \eta_T^+ - aC_T$ and*

$$S_t {:=} \max\left(\eta_t^+ - aC_t, \nu_t\right), \nu_t {:=} \mathbb{E}[S_{t+1} \mid X_{1:t}] \tag{2.5}$$

*for $t = 1, \ldots, T$. Here $\eta_t^+ {:=} \max(\eta_t, 0)$. Moreover, the infimum over $f$ is achieved at $\tilde{f}_{1:T}(a, b)$ with $\tilde{f}_T(X_{1:T}; a, b) = \eta_T$, and*

$$\tilde{f}_t(X_{1:t}; a, b) = \begin{cases} 1, & \textit{if } \eta_t - aC_t > \max\left(-aC_t, \nu_t\right), \\ 0, & \textit{if } \nu_t \geq \zeta_t, \\ -1, & \textit{if } -aC_t > \max\left(\eta_t - aC_t, \nu_t\right) \end{cases} \tag{2.6}$$

*where $\zeta_t {:=} \eta_t^+ - aC_t$ for $t = 1, \ldots, T - 1$.*

Therefore, the dual problem of (2.3) is

$$\text{maximize } \mathcal{G}(a, b) \text{ for } a, b \geq 0. \tag{2.7}$$

To connect the primal and dual, one key assumption is that $\mu_t$ are continuous functions of $X_{1:t}$ for all $t$, similar to that in a non-randomized likelihood ratio test which ensures any desired sensitivity is achievable. With few other mild assumptions, there always exists some dual optimal $a^*, b^*$ with no duality gap, i.e., *strong duality condition* holds. By plugging-in $f^*(X) = \tilde{f}(X; a^*, b^*)$ and subsequently $\varphi^*$, the primal-dual pair $(f^*, a^*, b^*)$ optimizes problem (2.3) and (2.7) satisfying the constraints by equalities, as shown in Proposition A.1. Further, such optima of problem (2.3) is in fact also Pareto optimal for the MOO, so that we transform the MOO (2.2) into a convex optimization of (2.4). The Pareto front is obtained by solving multiple primal-dual problems looping over $\gamma$.

**Theorem 2.1.** *Suppose for $t = 1, \ldots, T$, the $X_t$ are continuous r.v. admits density, and $X_{1:t} \mapsto \mu_{1:t}$ are continuous functions bounded away from zero, then for any $\beta, \gamma \in (0, 1)$ there exists some $\varphi^*$ whose scores take the form of* (2.6) *such that i) $\varphi^*$ minimizes problem* (2.3) *satisfying the constraints by equality; ii) $\varphi^*$ is a Pareto optimum for problem* (2.2) *as well as the MOO problem that seeks to simultaneously minimize* $(\mathbb{E}[\varphi_\tau \mid Y = 0], \mathbb{E}[C_\tau])$ *and maximize* $\mathbb{E}[\varphi_\tau \mid Y = 1]$.

The preceding optimality guarantee partly generalizes the Neyman–Pearson lemma to sequential setting by providing an optimal form for the classifier that maximizes specificity, earliness, and sensitivity.

## 3 Plug-in Estimation via Recurrent Neural Network

It suffices to recursively estimate the $\mathcal{F}$-adapted processes $\mu_{1:T}$ and $\nu_{1:T-1}$ then plug into (2.6). Recurrent neural networks (RNN) are suitable candidates as they handle sequence-to-sequence tasks by recurrent evaluation without the need of stacking a list of models. Most importantly, their only utilize $X_{1:t}$ at time $t$, making the output sequence inherently $\mathcal{F}$-adapted. For added flexibility (especially for $\nu$), we follow the RNNs by a time-specific layer, essentially one fully connected linear layer for each $t = 1, \ldots, T$ that do not share parameters across time.

Denote $\mathring{\mu}\left(X_{1:t}, \theta_\mu\right)$ and $\mathring{\nu}\left(X_{1:t}, \theta_\nu\right)$ as neural network models for $\mu_t$ and $\nu_t$ respectively, and subsequently $\mathring{\eta}, \mathring{\zeta}, \mathring{S}$, $\mathring{\tau}$, and $\mathring{\varphi}$ by plugging-in $\mathring{\mu}$ and $\mathring{\nu}$ into corresponding definitions. To track conditional expectation, suppose $S$ is known, it suffices to minimize mean squared errors

$$Q_\mu\left(\theta_\mu\right) = \sum_{t=1}^{T} \mathbb{E}[(\mathring{\mu}_t - Y)^2], \quad Q_\nu\left(\theta_\nu | \theta_\mu, a, b, S_{1:T}\right) = \sum_{t=1}^{T-1} \mathbb{E}[(\mathring{\nu}_t - S_{t+1})^2], \tag{3.1}$$

and maximize the dual function $\mathring{\mathcal{G}}\left(a, b | \theta_\mu, \theta_\nu, S\right) = b\beta - a\gamma - \mathbb{E}[S_1]$. Invoking stochastic gradient descent iteratively admits estimated parameters denoted as $\hat{\theta}_\mu, \hat{\theta}_\nu, \hat{a}$, and $\hat{b}$; and subsequently the optimal rule $\hat{\varphi}$. While $Q_\nu$ and $\mathring{\mathcal{G}}$ both rely on $\beta$ and $\gamma$, we suppress those arguments for simpler notation. Note that $Q_\mu$ essentially tracks a many-to-one classification but utilizing the entire output sequence. It is simple to obtain $\hat{\theta}_\mu \in \arg\min Q_\mu$, hence decoupled from the iterative update of $Q_\nu$ and $\mathring{\mathcal{G}}$. Similarly $Q_\nu$ tracks a many-to-many regression if $S$ is available, it suffices to update $\theta_\nu$

**Algorithm 1** Update of $\theta_\nu$, $a$ and $b$ given $\beta$ and $\gamma$ for $K$ iterations

```
Obtain $\hat{\theta}_\mu \leftarrow \arg\min Q_\mu$, then initialize $\theta_\nu^{(0)}$, $a^{(0)}$, and $b^{(0)}$, set $k = 0$.
while $k \leq K$ do
    Compute $\hat{S}_{1:T}$ by plugging-in $\mathring{\mu}, \mathring{\nu}$ under parameters $\hat{\theta}_\mu, \theta_\nu^{(k)}, a^{(k)}$, and $b^{(k)}$ to (2.5).
    Compute $\theta_\nu^{(k+1)}$ from $\theta_\nu^{(k)}$ base on $\partial Q_\nu(\theta_\nu | \hat{\theta}_\mu, a^{(k)}, b^{(k)}, \hat{S}_{1:T})$;
    Compute $\left(a^{(k+1)}, b^{(k+1)}\right)$ from $\left(a^{(k)}, b^{(k)}\right)$ base on $\partial \mathring{\mathcal{G}}(a, b | \hat{\theta}_\mu, \theta_\nu^{(k+1)}, \hat{S}_{1:T})$.
    $k \leftarrow k + 1$.
end while
```

iteratively, leading to Algorithm 1. In practice, it suffices to replace the expectations by sample/batch averages. See Appendix B for additional implementation details.

The proposed algorithm effectively parameterize $\mathring{\varphi}$ by $(\beta, \gamma)$, i.e., the desired sensitivity and expected cost, making it possible to dynamically adjust $(\beta, \gamma)$ according to the sensitivity and average cost on validation data. This seamlessly incorporates into Algorithm 1 by slightly increasing or decreasing the $\beta$ or $\gamma$ used for computation every few epochs.

## 4 Experiments

We now validate the proposed methods as a feasible approach for early classification. Since the proposed method is not directly comparable to existing ECTS approaches that focus on the accuracy combining sensitivity and specificity into one objective, we resort to Wald's SPRT (see Appendix C) for a baseline comparison.

We considered two simulated examples based on stationary autoregressive AR(1) time series, a sensory example (Ford-A, Dau et al., 2018), and a bivariate example from online handwritten recognition (Pendigits, E. Alpaydin, 1996). For simplicity, across all our experiments we set cost $C_t = t/T$ and only targeted sensitivity level of 0.9. All experiments were repeated 100 times, where for each repeat data were randomly split to create training, validation, and testing sets (70%:10%:20% unless otherwise specified). The classifiers were trained on the training set with the $(\beta, \gamma)$ tuned by their performance on the validation set. The final performance were assessed on the testing set and reported. Figure 4.1 pictures the estimated Pareto front, while Figure 4.2 validates the constraints in problem (2.3). We refer to Appendix D for details of the experiments setup.

Apparently the FPR decreases with increased cost, yet interestingly we observe that the reduction of FPR diminished when the average cost exceeded 75% of total sequence length for simulated examples and 50% for the two real data. Recalling that the sensitivity (TPR) were always retained at 0.9, this suggests an oracle possibility for the proposed classifier to achieve performance similar to a full-length classifier (e.g., the horizontal dashed lines in the upper panels of Figure 4.1) but on average utilizing shorter sequences.

It is clear that for the simulated AR(1) examples, the proposed method outperformed the Wald's SPRT by achieving a smaller FPR at the same cost (green curves and the crosses are lower than the white pixels, upper panels of Figure 4.1), yet admittedly not quite for the Ford-A and Pendigits examples. However the key distinction is that the SPRT tended to decide prematurely. Under our setup where sensitivity restricted around 0.9, no SPRT reached an average cost of more than 0.8 in all four examples (even no more than 0.63 for the AR(1)-logistic examples), so that it rushed the decisions and failed to demonstrate the diminishing benefits in FPR of larger cost. Such limitation appears more apparent for harder tasks (e.g., the AR(1) examples) where FPR is higher and/or longer sequences are necessary to achieve an oracle performance.

## 5 Discussion

Several limitations necessitate future research, such as the reliance on the assumptions, the lack of tailored neural network architectures, and the accumulated errors during recursive estimation that potentially caused suboptimality in the presented real data examples. Additionally, generalization across diverse datasets mandates comprehensive validation efforts. Though, instead of extensive

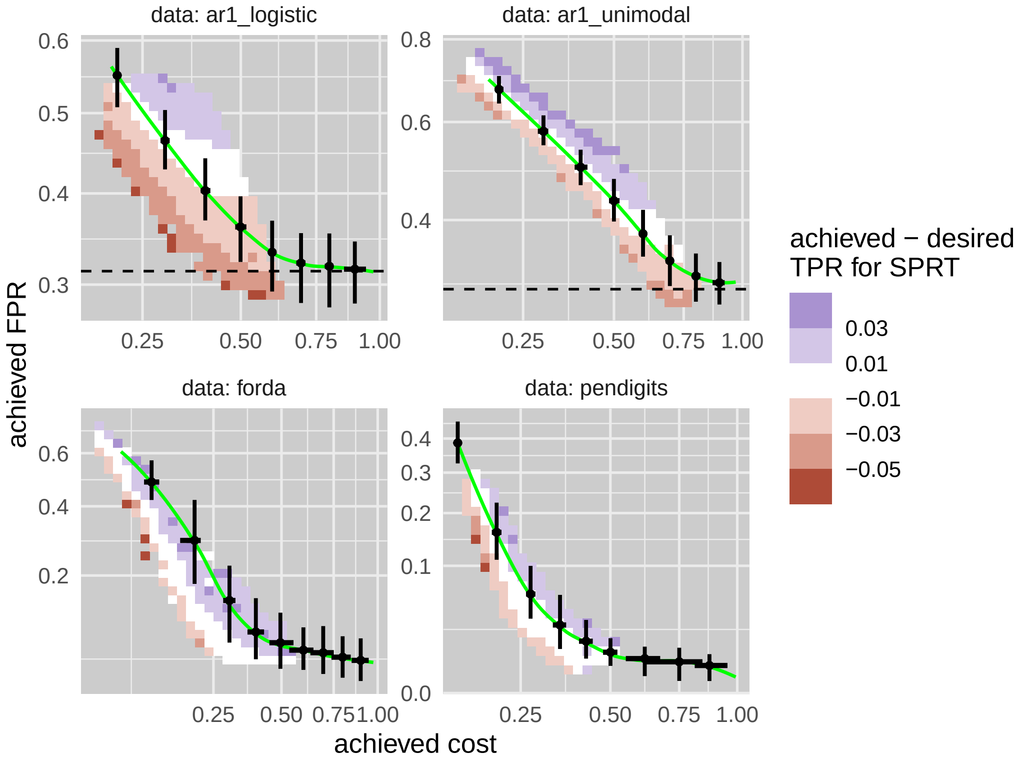

Figure 4.1: Under a desired TPR of 0.9, the estimated Pareto front (green lines and black crosses) in comparison to Wald's SPRT (red-white-blue pixels) on simulated and real data over 100 repeats. The panels from top left to bottom right show: univariate AR(1) of length 10 and autoregressive correlation 0.75 with logistic or unimodal responses probability; the Ford-A data; and the Pendigits data. The black crosses show 90% interval for the achieved FPR (vertical) and cost (horizontal), while the green lines are the average estimated Pareto fronts pooling repeats. The horizontal dashed lines in the upper panels show, under the same 0.9 TPR restriction, the FPR of full-length classifiers when the true class probability (i.e., $\mu_T$) is available.

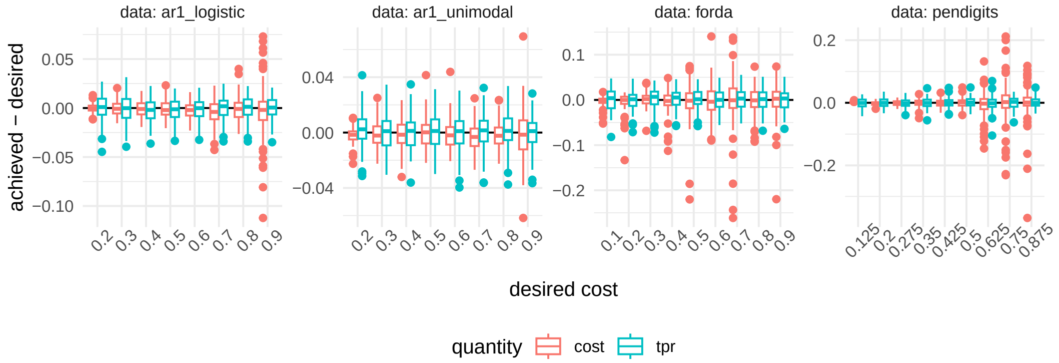

Figure 4.2: The boxplot of the deviance in the achieved v.s. desired cost (varying from 0.1 to 0.9) and TPR (0.9) over the 100 repeated experiments under the setup of Figure 4.1, grouped by the desired cost (x-axis).

benchmarking on diverse data and/or against state-of-the-art methods, our contribution primarily focuses on the feasibility of early decision making under Neyman–Pearson paradigm, which, we hope, could shed new light on early time series classification and its practical applications.

## A Proofs

*Proof of Proposition 2.1.* The Lagrangian function of (2.3) is, with the convention that $f_0 \equiv 0$,

$$\begin{aligned}\mathcal{L}\left(f_{1:T}, \boldsymbol{a}\right) = b\beta - a\gamma &+ \mathbb{E}[\mathbb{1}(f_{0:T-1} = 0)\left(aC_T - \eta_T \mathbb{1}(f_T > 0)\right)] \\ &+ \sum_{t=1}^{T-1} \mathbb{E}[\mathbb{1}(f_{0:t-1} = 0)\left(aC_t \mathbb{1}(f_t \neq 0) - \eta_t \mathbb{1}(f_t > 0)\right)] \end{aligned} \tag{A.1}$$

for non-negative $\boldsymbol{a} = (a, b)$, where $\eta$ is a $\mathcal{F}$-adapted process defined in Section 2.

The infimum over $f$ in $\mathcal{G}$ can be taken recursively as $\inf_{f_1} \inf_{f_2} \dots \inf_{f_T} \mathcal{L}(f, \boldsymbol{a})$. Thus by (A.1), with the convention that $\tilde{f}_0 \equiv 0$ and $M_{T+1} \equiv 0$, denote

$$\begin{aligned} L_t(f_{1:t}, \boldsymbol{a}) &= \mathbb{E}[\mathbb{1}(f_{1:t-1} = 0)\left(\eta_t \mathbb{1}(f_t > 0) - aC_t \mathbb{1}(f_t \neq 0)\right)], \\ M_t(f_{1:t-1}, \boldsymbol{a}) &= \max_{f_{t:T}} \left(L_t + \dots + L_T\right) = \max_{f_t} \left(L_t + M_{t+1}(f_{1:t}, \boldsymbol{a})\right). \end{aligned}$$

Clearly $M_T(f_{1:T-1}, \boldsymbol{a}) = \mathbb{E}[\mathbb{1}(f_{1:T-1} = 0) S_T]$ at $\tilde{f}_T = \eta_T$. Then by induction we claim $M_t(f_{1:t-1}, \boldsymbol{a}) = \mathbb{E}[\mathbb{1}(f_{1:t-1} = 0) S_t]$ at $\tilde{f}_{t:T}$ of (2.6). Indeed,

$$\begin{aligned} &L_t(f_{1:t}, \boldsymbol{a}) + M_{t+1}(f_{1:t}, \boldsymbol{a}) \\ &= \mathbb{E}[\mathbb{1}(f_{1:t-1} = 0)\left(\eta_t \mathbb{1}(f_t > 0) - aC_t \mathbb{1}(f_t \neq 0) + \mathbb{E}[S_{t+1} \mid X_{1:t}] \mathbb{1}(f_t = 0)\right)] \end{aligned}$$

for $t = T-1, T-2, \dots, 1$. In the end, note that $\mathcal{G}(\boldsymbol{a}) = b\beta - a\gamma - M_1$. □

**Proposition A.1** (Feasibility)**.** *Suppose for $t = 1, \dots, T$, the $X_t$ are continuous r.v. admits density, and $X_{1:t} \mapsto \mu_{1:t}$ are continuous functions bounded away from zero, there always exists some $\boldsymbol{a}^*$ that maximize problem (2.7). Define $f^*(X) = \tilde{f}(X; \boldsymbol{a}^*)$ and $\varphi^*$ by plugging-in $f^*$ to (2.1), then $\varphi^*$ optimizes problem (2.3) satisfying the constraints by equality, and that $\mathcal{G}(\boldsymbol{a}^*) = \mathbb{E}[\varphi^* \mid Y = 0]$.*

*proof of Proposition A.1.* First show the concave function $\mathcal{G}$ has stationary point(s), then by definition of $f^*$ there is no duality gap. It is not difficult to conclude that almost surely,

$$\begin{aligned} \frac{\partial S_t}{\partial a} &= \mathbb{E}[C_t \mathbb{1}\left(\tilde{f}_t \neq 0\right) + C_{t+1} \mathbb{1}\left(\tilde{f}_t = 0, \tilde{f}_{t+1} \neq 0\right) + \dots + C_T \mathbb{1}\left(\tilde{f}_{t:T-1} = 0\right) \mid X_{1:t}], \\ \frac{\partial S_t}{\partial b} &= p_1^{-1}\, \mathbb{E}[\mu_t \mathbb{1}\left(\tilde{f}_t = 1\right) + \mu_{t+1} \mathbb{1}\left(\tilde{f}_t = 0, \tilde{f}_{t+1} = 1\right) + \dots + \mu_T \mathbb{1}\left(\tilde{f}_{t:T-1} = 0, \tilde{f}_T = 1\right) \mid X_{1:t}], \end{aligned} \tag{A.2}$$

so that $\partial \mathcal{G}/\partial a = \mathbb{E}[C_\tau] - \gamma$ and $\partial \mathcal{G}/\partial b = \beta - \mathbb{E}[\tilde{\varphi} \mid Y = 1]$, where $\tilde{\varphi}$ is the rule (2.1) with $\tilde{f} = \tilde{f}(X, \boldsymbol{a})$ of (2.6) plugged-in.

Next it suffices to show there exists some $a^*$ and $b^*$ such that zeros the previous two equations utilizing the connectedness of the solution sets $\Psi_\beta(a^*) := \{b : \mathbb{E}[\tilde{\varphi}|_{a^*} \mid Y = 1] = \beta\}$ and $\Psi_\gamma(b^*) := \{a : \mathbb{E}[C_\tau|_{b^*}] = \gamma\}$. Here $\tilde{\varphi}|_{a^*}$ means a plug-in classifier fixing $a = a^*$, similarly $C_\tau|_{b^*}$ fixes $b = b^*$.

For any fixed $b \geq 0$, we can let $\tilde{f}_t \equiv 0$ for all $t$ by setting $a \leq 0$, so that $\mathbb{E}[C_\tau] = C_T$. On the other hand, by $\mathbb{E}[\eta_T^+ \mid X_{1:T-1}] \geq (\mathbb{E}[\eta_T \mid X_{1:T-1}]) = \eta_{T-1}^+$, there is always some $a_{T-1} > 0$ so that $S_{T-1} = \zeta_{T-1}$ for all $a \geq a_{T-1}$. Inductively, there exists $a_t > 0$ such that $S_t = \zeta_t$ for all $a \geq a_t$, i.e., $\tilde{f}_t \neq 0$, so that $\mathbb{E}[C_\tau] \leq C_t$. In other words, one can increase $a$ from zero, and the expected cost would reduce from $C_T$ to $C_1$. Hence, for any $b \geq 0$, there always exists some $a_b$ such that $\mathbb{E}[C_\tau] = \gamma$, i.e., $\Psi_\gamma(b) \neq \varnothing$. Further, by continuity of $X_{1:t} \mapsto \mu_{1:t}$, the mapping $a \mapsto \mathbb{E}[C_\tau]$ is a continuous non-increasing function, implying $\Psi_\gamma(b)$ is a closed interval.

Similarly given $a \geq 0$, note that $\tilde{\varphi} \equiv 0$ for $b < p_1/p_0\left(1/\sup_x \mu_t - 1\right)$ and that $\tilde{\varphi} \equiv 1$ for $b > p_1/p_0\left(1/\inf_x \mu_t - 1\right)$. Thus for any given $a$ the mapping $b \mapsto \mathbb{E}[\tilde{\varphi} \mid Y = 1]$ is a continuous non-decreasing function, hence $\Psi_\beta(a)$ is also closed interval.

Further, the continuity of $(a, b) \mapsto \mathbb{E}[C_\tau]$ implies that $A_\beta := \{(a, b) : b \in \Psi_\beta(a)\}$ is a non-empty connected set, similarly $A_\gamma := \{(a, b) : a \in \Psi_\gamma(b)\}$ is also non-empty and connected. In combine $a^*$ and $b^*$ that zero out (A.2) exist since $A_\beta$ and $A_\gamma$ must intersect.

In sum, $f^* = \tilde{f}(\boldsymbol{a}^*)$ is a feasible solution to problem (2.3) that satisfies the constraints by equality. It then suffices to show strong duality by showing that $(f^*, \boldsymbol{a}^*)$ is a saddle point of $\mathcal{L}$, i.e., $\mathcal{L}(f^*, \boldsymbol{a}) \leq \mathcal{L}(f^*, \boldsymbol{a}^*) \leq \mathcal{L}(f, \boldsymbol{a}^*)$ for any $a, b \geq 0$, and $\mathcal{F}$-adapted $f$. The second inequality is obvious by definition of $f^* = \tilde{f}(\boldsymbol{a}^*)$. For the first inequality, by definition of Lagrangian function, $\mathcal{L}(f, \boldsymbol{a}) \leq \mathbb{E}[\varphi \mid Y = 0]$ for all $a, b$, and any feasible $f$ (and corr. $\varphi$). Moreover, $\mathcal{L}\left(f^*, \boldsymbol{a}^*\right) = \mathbb{E}[\varphi^* \mid Y = 0]$ since $f^*$ satisfies the constraints of problem (2.3) by equality. In combine we have $\mathcal{L}(f^*, \boldsymbol{a}) \leq \mathbb{E}[\varphi^* \mid Y = 0] = \mathcal{L}(f^*, \boldsymbol{a}^*) \leq \mathcal{L}(f, \boldsymbol{a}^*)$, implying that $(f^*, \boldsymbol{a}^*)$ is a saddle point for $\mathcal{L}$.

□

*proof of Theorem 2.1*. This is a direct implication of Proposition A.1 under Theorem 3.2.2 in part II of Miettinen (1999). Technically, it suffices to repeat the previous steps with some reparameterization of the $a^*$ and $b^*$ to see that $\varphi^*$ also optimizes

$$\text{minimize } \mathbb{E}[C_\tau], \text{ subject to } \mathbb{E}[\varphi_\tau \mid Y = 0] \leq \alpha \text{ and } \mathbb{E}[\varphi_\tau \mid Y = 1] \geq \beta \tag{A.3}$$

for $\alpha = \mathbb{E}[\varphi^* \mid Y = 0]$. I.e., the performance of $\varphi^*$ in terms of the objectives in problem (2.3) cannot be further improved in either direction without sacrificing another. □

## B Remarks on the Neural Networks Plug-in

Note that one shall not use the bidirectional RNNs which also utilize future information, violating the $\mathcal{F}$-adapted requirement.

It is important to note that $\theta_\nu, a, b$, and $S$ are treated as constant in $Q_\nu$, so are $\theta_\mu$, $\theta_\nu$, and $S$ in $\mathring{\mathcal{G}}$. In other words, $(a, b)$ are not updated when computing $\theta_\nu$, so is $\theta_\nu$ not updated when computing $(a, b)$. Furthermore, gradient shall not flow through $S_{t+1}$ in $Q_\nu$, and not through $\mathring{\mu}$, $\mathring{\nu}$, and $S$ in $\mathring{\mathcal{G}}$ despite implicit dependency. For backpropagation we utilize the gradients in (A.2) so $\partial\mathring{\mathcal{G}}/\partial a = \mathbb{E}[C_{\mathring{\tau}}] - \gamma$ and $\partial\mathring{\mathcal{G}}/\partial b = \beta - \mathbb{E}[\mathring{\varphi} \mid Y = 1]$; i.e., the average cost and true positive rate of the plugged-in rule.

In the typical task of estimating the Pareto front for some pre-specified sensitivity level of $\beta$, it suffices to initialize Algorithm 1 with small but gradually increasing $\gamma$ (or vice versa), and record the results. While in theory $0 \leq \beta, \gamma \leq 1$, we found it practically convenience to allow greater upper bounds for better performance. Similarly $a, b$ are allowed to be negative.

## C Wald's Sequential Probability Ratio Test

Wald's SPRT was originally proposed based on cutoffs determined by the desired type I and II error rate upon the cumulative product of probability ratio for i.i.d. observations (Wald, 1947). For binary time series classification, it suffices to impose cutoff on $\mu_t$, noting that $P(X_{1:t}|Y = 0)/P(X_{1:t}|Y = 1) \propto (1 - \mu_t)/\mu_t$. We adopted a truncated SPRT (see Wald, 1947, section 3.8) to account for the finite horizon situation. In sum, we write $\varphi_{\text{SPRT}}$ following (2.1) where $f_{T,\text{SPRT}} = 1$ if $\mu_T \geq \mathbb{P}(Y = 1)$ and $-1$ otherwise; while for $t \leq T - 1$,

$$f_{t,\text{SPRT}} = \begin{cases} 1, \text{ if } \mu_t > \bar{\mu}, \\ -1, \text{ if } \mu_t < \underline{\mu}, \\ 0 \text{ otherwise}, \end{cases}$$

for some cutoffs $0 < \underline{\mu} <\leq \bar{\mu} < 1$. To account for the impact of non-iid sequence and truncation, among a grid of $(\underline{\mu}, \bar{\mu})$-pairs, those admit validation TPR closest to the desired level (deviation smaller than 0.01) were selected and evaluated on test sample for cost, TPR, and FPR.

We re-used the existing neural network estimate $\hat{\mu}$ for SPRT.

## D Additional Experiment Details

1. For the two AR(1) examples, $X_1, \ldots, X_{10}$ is a stationary univariate AR(1) process of length 10 with standard Gaussian innovation and a correlation coefficient of 0.75; while $E(Y|X_{1:T})$ equals sigmoid $(c \sum_t X_t)$ (i.e., the logistic example) or $\exp(-c(\sum_t X_t)^2)$ (the unimodal

example). We generated 5120, 1280, and $10^4$ series for training, validating, and testing respectively.

2. The Ford-A dataset includes sensor data with 500 time points and a binary response for 3601 instances. The time series were down-sampled to $T = 50$ by block: every 10 consecutive time points are pooled as a 10-dimensional covariate. It provides an example to showcase the multivariate capability of the proposed framework.
3. The Pendigits data (E. Alpaydin, 1996) consists of 10992 series of 2-dimensional coordinates of pen strokes when writing digits (downsampled to 8 points) and we relabeled the outcome to whether the digit written is 1, 2, 3, 5, 7.

We implemented the proposed method with Pytorch. Both $\mathring{\mu}$ and $\mathring{\nu}$ were modeled by a single layer GRU with 16 hidden features, followed by a time-specific, single-layered fully connected linear layer to provide scalar prediction, totaling around 2000 parameters.

## Acknowledgments and Disclosure of Funding

The authors would like to express their gratitude to the reviewers, Yanqing, and Lucas for invaluable feedback and helpful discussion. The authors are supported by the National Institutes of Health (U24 CA086368 and R01 CA236558).

## References

Dau, H. A., E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, and Hexagon-ML (2018, October). The UCR time series classification archive.

E. Alpaydin, F. A. (1996). Pen-Based Recognition of Handwritten Digits.

Gupta, A., H. P. Gupta, B. Biswas, and T. Dutta (2020, August). Approaches and applications of early classification of time series: A review. *IEEE Transactions on Artificial Intelligence 1*(1), 47–61.

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*. Number 12 in International Series in Operations Research & Management Science. Boston: Kluwer Academic Publishers.

Mori, U., A. Mendiburu, S. Dasgupta, and J. A. Lozano (2018, October). Early classification of time series by simultaneously optimizing the accuracy and earliness. *IEEE Transactions on Neural Networks and Learning Systems 29*(10), 4569–4578.

Mori, U., A. Mendiburu, I. M. Miranda, and J. A. Lozano (2019, August). Early classification of time series using multi-objective optimization techniques. *Information Sciences 492*, 204–218.

Neyman, J., E. S. Pearson, and K. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231*(694-706), 289–337.

Pardalos, P. M., A. Žilinskas, and J. Žilinskas (2017). *Non-Convex Multi-Objective Optimization*, Volume 123 of *Springer Optimization and Its Applications*. Cham: Springer International Publishing.

Schmitz, N. (1985). Sequential probability ratio tests for stochastic processes: A review note. *Banach Center Publications 1*(16), 465–476.

Wald, A. (1947). *Sequential Analysis*. John Wiley & Sons.

Wald, A. and J. Wolfowitz (1948, September). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics 19*(3), 326–339.