The Power of Log-Sum-Exp: Sequential Density Ratio Matrix Estimation for Speed-Accuracy Optimization

Taiki Miyagawa [1] **Akinori F. Ebihara** [1]

Abstract

We propose a model for multiclass classification of time series to make a prediction as early and as accurate as possible. The *matrix sequential probability ratio test* (MSPRT) is known to be asymptotically optimal for this setting, but contains a critical assumption that hinders broad real-world applications; the MSPRT requires the underlying probability density. To address this problem, we propose to solve *density ratio matrix estimation* (DRME), a novel type of density ratio estimation that consists of estimating matrices of multiple density ratios with constraints and thus is more challenging than the conventional density ratio estimation. We propose a log-sum-exp-type loss function (LSEL) for solving DRME and prove the following: (i) the LSEL provides the true density ratio matrix as the sample size of the training set increases (*consistency*); (ii) it assigns larger gradients to harder classes (*hard class weighting effect*); and (iii) it provides discriminative scores even on class-imbalanced datasets (*guess-aversion*). Our overall architecture for early classification, *MSPRT-TANDEM*, statistically significantly outperforms baseline models on four datasets including action recognition, especially in the early stage of sequential observations. Our code and datasets are publicly available[1].

1. Introduction

Classifying an incoming time series as early and as accurately as possible is challenging yet crucial, especially when the sampling cost is high or when a delay results in serious consequences (Xing et al., 2009; 2012; Mori et al., 2015; Mori et al., 2018). Early classification of time series is a multi-objective optimization problem, and there is usually no ground truth indicating when to stop observation and classify a sequence.

The MSPRT is a provably optimal algorithm for early multiclass classification and has been developed in mathematical statistics (Armitage, 1950; Chernoff, 1959; Kiefer & Sacks, 1963; Lorden, 1967; 1977; Pavlov, 1984; Dragalin, 1987; Pavlov, 1991; Baum & Veeravalli, 1994; Dragalin & Novikov, 1999). The MSPRT uses a *matrix* of log-likelihood ratios (LLRs), the (k, l)-entry of which is the LLR of hypothesis H_k to hypothesis H_l and depends on the current time t through consecutive observations of sequential data $X^{(1,t)}$ (Figure 1). A notable property of the MSPRT is that it is *asymptotically optimal* (Tartakovsky, 1998): It achieves the minimum stopping time among all the algorithms with bounded error probabilities as the thresholds go to infinity, or equivalently, as the error probabilities go to zero or the stopping time goes to infinity (Appendix A). Therefore, the MSPRT is a promising approach to early multiclass classification with strong theoretical support.

However, the MSPRT has a critical drawback that hinders its real-world applications in that it requires the true LLR matrix, which is generally inaccessible. To address this problem, we propose to solve *density ratio matrix estimation* (DRME); i.e., we attempt to estimate the LLR matrix from a dataset. DRME has yet to be explored in the literature but can be regarded as a generalization of the conventional density ratio estimation (DRE), which usually focuses on only two densities (Sugiyama et al., 2012). The difficulties with DRME come from simultaneous optimization of multiple density ratios; the training easily diverges when the denominator of only one density ratio is small. In fact, a naive application of conventional binary DRE-based loss functions does not generalize well in this setting, and sometimes causes instability and divergence of the training (Figure 2 Top).

Therefore, we propose a novel loss function for solving DRME, the *log-sum-exp loss (LSEL)*. We prove three properties of the LSEL, all of which contribute to enhancing the performance of the MSPRT. (i) The LSEL is *consistent*; i.e., by minimizing the LSEL, we can obtain the true LLR

[1]NEC Corporation, Kanagawa, Japan. Correspondence to: Taiki Miyagawa <miyagawataik@nec.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]https://github.com/TaikiMiyagawa/MSPRT-TANDEM

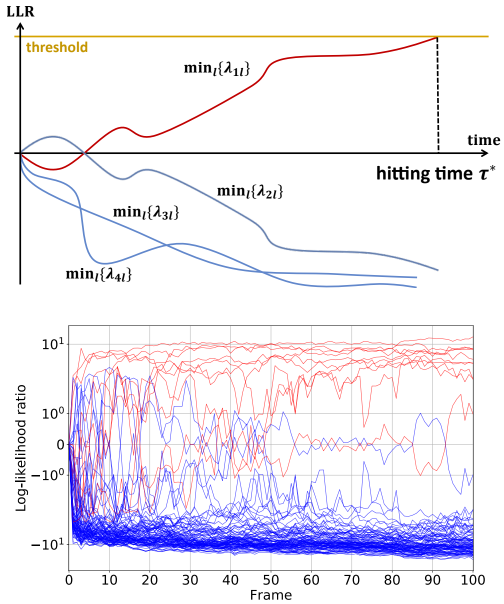

Figure 1. **Top: Early Classification of Time Series with MSPRT.** The figure illustrates how the MSPRT predicts the label y of an incoming time series $X^{(1,t)} = \{x^{(1)}, x^{(2)}, ...x^{(t)}\}$. The MSPRT uses the LLR matrix denoted by $\lambda_{kl}(X^{(1,t)}) := \log(p(X^{(1,t)}|y=k)/p(X^{(1,t)}|y=l))$, where $k,l = 1,2,...,K$. $K (\in \mathbb{N})$ is the number of classes. If one of $\min_l \lambda_{kl} = \log p(X^{(1,t)}|k)/\max_l p(X^{(1,t)}|l)$ ($k \in \{1,2,...,K\}$) reaches the threshold, the prediction is made; otherwise, the observation continues. In this figure, $K = 4$, the prediction is $y = 1$, and the hitting time is τ^*. A larger threshold leads to more accurate but delayed predictions, while a smaller threshold leads to earlier but less accurate predictions. **Bottom: Estimated LLRs of ten sequences.** (See Appendix I.6 for exact settings.)

matrix as the sample size of the training set increases. (ii) The LSEL has the *hard class weighting effect*; i.e., it assigns larger gradients to harder classes, accelerating convergence of neural network training. Our proof also explains why log-sum-exp-type losses, e.g., (Song et al., 2016; Wang et al., 2019; Sun et al., 2020), have performed better than sum-log-exp-type losses. (iii) We propose the cost-sensitive LSEL for class-imbalanced datasets and prove that it is *guess-averse* (Beijbom et al., 2014). Cost-sensitive learning (Elkan, 2001), or loss re-weighting, is a typical and simple solution to the class imbalance problem (Kubat & Matwin, 1997; Japkowicz & Stephen, 2002; He & Garcia, 2009; Buda et al., 2018). Although the consistency does not necessarily hold for the cost-sensitive LSEL, we show that the cost-sensitive LSEL nevertheless provides discriminative "LLRs" (scores) by proving its guess-aversion.

Along with the novel loss function, we propose the first DRE-based model for early multiclass classification in deep learning, *MSPRT-TANDEM*, enabling the MSPRT's practi-

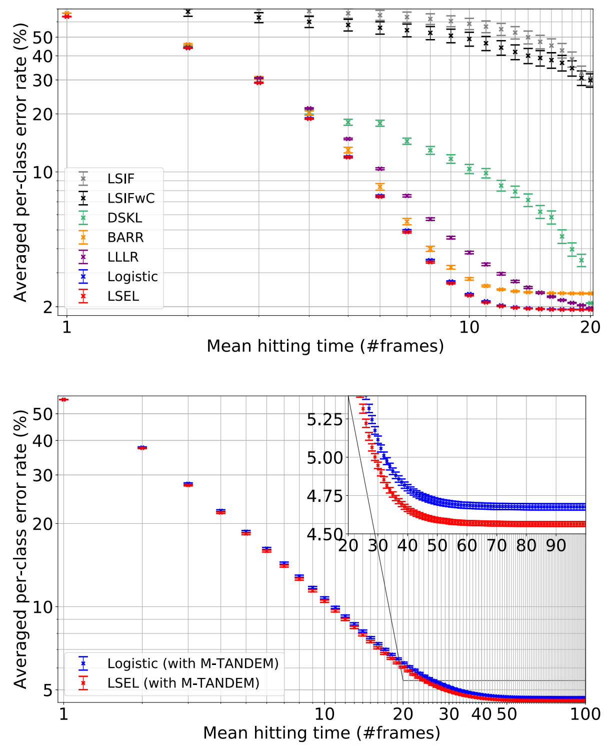

Figure 2. **LSEL v.s. Conventional Losses.** The datasets are NMNSIT-H and NMNIST-100f (Section 4). Curves in the lower left region are better. **Top: LSEL v.s. Binary DRE-based Losses on NMNSIT-H.** The conventional losses do not generalize well in DRME. The MSPRT is run, using the LLR matrices estimated with seven different loss functions: LSIF (Kanamori et al., 2009) minimizes the mean squared error of p and $\hat{r}q$ ($\hat{r} = \hat{p}/\hat{q}$); LSIFwC stabilizes LSIF by adding a normalization constraint of $\hat{r}q$; DSKL (Khan et al., 2019) is based on KLIEP (Sugiyama et al., 2008) and minimizes the Kullback-Leibler divergence between p and $\hat{r}q$; BARR (Khan et al., 2019) stabilizes DSKL by adding the normalization constraint; LLLR (Ebihara et al., 2021) is similar to DSKL but is bounded above and below and is thus more stable; the logistic loss is the standard sum-log-exp-type loss; and the LSEL is our proposed loss. Their formal definitions are summarized in Appendix I.8. Only the logistic loss shows a comparable performance, but the LSEL is consistently better (Tables 35 and 36). **Bottom: LSEL v.s. Logistic Loss on NMNIST-100f.** The M-TANDEM approximation is used, which is introduced in Section 3.4. The error gap is statistically significant (Appendix L).

cal use on real-world datasets. MSPRT-TANDEM can be used for arbitrary sequential data and thus has a wide variety of potential applications. To test its empirical performance, we conduct experiments on four publicly available datasets. We conduct two-way analysis of variance (ANOVA) (Fisher, 1925) followed by the Tukey-Kramer multi-comparison test (Tukey, 1949; Kramer, 1956) for reproducibility and find that MSPRT-TANDEM provides statistically significantly better accuracy with a smaller number of observations than baseline models.

Our contributions are summarized as follows.

1. We formulate a novel problem setting, DRME, to enable the MSPRT on real-world datasets.
2. We propose a loss function, LSEL, and prove its consistency, hard class weighting effect, and guess-aversion.
3. We propose MSPRT-TANDEM: the first DRE-based model for early multiclass classification in deep learning. We show that it outperforms baseline models statistically significantly.

2. Related Work

Early classification of time series. Early classification of time series aims to make a prediction as early and as accurately as possible (Xing et al., 2009; Mori et al., 2015; 2016; Mori et al., 2018). An increasing number of real-world problems require earliness as well as accuracy, especially when a sampling cost is high or when a delay results in serious consequences; e.g., early detection of human actions for video surveillance and health care (Vats & Chan, 2016), early detection of patient deterioration on real-time sensor data (Mao et al., 2012), early warning of power system dynamics (Zhang et al., 2017), and autonomous driving for early and safe action selection (Doná et al., 2019). In addition, early classification saves computational costs (Ghodrati et al., 2021).

SPRT. Sequential multihypothesis testing has been developed in (Sobel & Wald, 1949; Armitage, 1950; Paulson, 1963; Simons, 1967). The extension of the binary SPRT to multihypothesis testing for i.i.d. data was conducted in (Armitage, 1950; Chernoff, 1959; Kiefer & Sacks, 1963; Lorden, 1967; 1977; Pavlov, 1984; Dragalin, 1987; Pavlov, 1991; Baum & Veeravalli, 1994; Dragalin & Novikov, 1999). The MSPRT for non-i.i.d. distributions was discussed in (Lai, 1981; Tartakovsky, 1998; Dragalin et al., 1999; Tartakovsky et al., 2014). The asymptotic optimality of the MSPRT was proven in (Tartakovsky, 1998).

Density ratio estimation. DRE consists of estimating a ratio of two densities from their samples without separately estimating the numerator and denominator (Sugiyama et al., 2012). DRE has been widely used for, e.g., covariate shift adaptation (Sugiyama et al., 2008), representation learning (Oord et al., 2018; Hjelm et al., 2019), mutual information estimation (Belghazi et al., 2018), and off-policy reward estimation in reinforcement learning (Liu et al., 2018). Our proof of the consistency of the LSEL is based on (Gutmann & Hyvärinen, 2012).

We provide more extensive references in Appendix B. To the best of our knowledge, only (Ebihara et al., 2021) and (Moustakides & Basioti, 2019) combine the SPRT with DRE. Both restrict the number of classes to only two. The loss function proposed in (Ebihara et al., 2021) has not been proven to be unbiased; there is no guarantee for the estimated LLR to converge to the true one. (Moustakides & Basioti, 2019) does not provide empirical validation for the SPRT.

3. Density Ratio Matrix Estimation for MSPRT

3.1. Log-Likelihood Ratio Matrix

Let p be a probability density over $(X^{(1,T)}, y)$. $X^{(1,T)} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T} \in \mathcal{X}$ is an example of sequential data, where $T \in \mathbb{N}$ is the sequence length. $\boldsymbol{x}^{(t)} \in \mathbb{R}^{d_x}$ is a feature vector at timestamp t; e.g., an image at the t-th frame in a video $X^{(1,T)}$. $y \in \mathcal{Y} = [K] := \{1, 2, ..., K\}$ is a multiclass label, where $K \in \mathbb{N}$ is the number of classes. The LLR matrix is defined as $\lambda(X^{(1,t)}) := (\lambda_{kl}(X^{(1,t)}))_{k,l\in[K]} := (\log p(X^{(1,t)}|y = k)/p(X^{(1,t)}|y = l))_{k,l\in[K]}$, where $p(X^{(1,t)}|y)$ is a conditional probability density. $\lambda(X^{(1,t)})$ is an anti-symmetric matrix by definition; thus the diagonal entries are 0. Also, λ satisfies $\lambda_{kl} + \lambda_{lm} = \lambda_{km}$ ($\forall k, l, m \in [K]$). Let $\hat{\lambda}(X^{(1,t)}; \boldsymbol{\theta}) := (\hat{\lambda}_{kl}(X^{(1,t)}; \boldsymbol{\theta}))_{k,l\in[K]} := (\log \hat{p}_{\boldsymbol{\theta}}(X^{(1,t)}|y = k)/\hat{p}_{\boldsymbol{\theta}}(X^{(1,t)}|y = l))_{k,l\in[K]}$ be an estimator of the true LLR matrix $\lambda(X^{(1,t)})$, where $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$ ($d_\theta \in \mathbb{N}$) denotes trainable parameters, e.g., weight parameters of a neural network. We use the hat symbol $\hat{(\cdot)}$ to highlight that the quantity is an estimated value. The $\hat{\lambda}$ should be anti-symmetric and satisfy $\hat{\lambda}_{kl} + \hat{\lambda}_{lm} = \hat{\lambda}_{km}$ ($\forall k, l, m \in [K]$). To satisfy these constraints, one may introduce additional regularization terms to the objective loss function, which can cause learning instability. Instead, we use specific combinations of the posterior density ratios $\hat{p}_{\boldsymbol{\theta}}(y = k|X^{(1,t)})/\hat{p}_{\boldsymbol{\theta}}(y = l|X^{(1,t)})$, which explicitly satisfy the aforementioned constraints (see the following M-TANDEM and M-TANDEMwO formulae).

3.2. MSPRT

Formally, the MSPRT is defined as follows (see Appendix A for more details):

Definition 3.1 (Matrix sequential probability ratio test). *Let P and P_k $(k \in [K])$ be probability distributions. Define a threshold matrix $a_{kl} \in \mathbb{R}$ $(k,l \in [K])$, where the diagonal elements are immaterial and arbitrary, e.g., 0. The MSPRT of multihypothesis $H_k : P = P_k$ $(k \in [K])$ is defined as $\delta^* := (d^*, \tau^*)$, where $d^* := k$ if $\tau^* = \tau_k$ $(k \in [K])$, $\tau^* := \min\{\tau_k | k \in [K]\}$, and $\tau_k := \inf\{t \geq 1 | \min_{l(\neq k)\in[K]}\{\lambda_{kl}(X^{(1,t)}) - a_{lk}\} \geq 0\}$.*

In other words, the MSPRT terminates at the smallest timestamp t such that for a class of $k \in [K]$, $\lambda_{kl}(t)$ is greater than or equal to the threshold a_{lk} for all $l(\neq k)$ (Figure 1). By definition, we must know the true LLR matrix $\lambda(X^{(1,t)})$ of the incoming time series $X^{(1,t)}$; therefore, we estimate λ with the help of the LSEL defined in the next section. For simplicity, we use single-valued threshold matrices ($a_{kl} = a_{k'l'}$ for all $k, l, k', l' \in [K]$) in our experiment.

3.3. LSEL for DRME

To estimate the LLR matrix, we propose the *log-sum-exp loss* (LSEL):

$$L_{\text{LSEL}}[\tilde{\lambda}] := \frac{1}{KT} \sum_{k\in[K]} \sum_{t\in[T]} \int dX^{(1,t)} p(X^{(1,t)}|k) \log(1 + \sum_{l(\neq k)} e^{-\tilde{\lambda}_{kl}(X^{(1,t)})}) \,. \quad (1)$$

Let $S := \{(X_i^{(1,T)}, y_i)\}_{i=1}^M \sim p(X^{(1,T)}, y)^M$ be a training dataset, where $M \in \mathbb{N}$ is the sample size. The empirical approximation of the LSEL is

$$\hat{L}_{\text{LSEL}}(\boldsymbol{\theta}; S) := \frac{1}{KT} \sum_{k\in[K]} \sum_{t\in[T]} \frac{1}{M_k} \sum_{i\in I_k} \log(1 + \sum_{l(\neq k)} e^{-\hat{\lambda}_{kl}(X_i^{(1,t)};\boldsymbol{\theta})}) \,. \quad (2)$$

M_k and I_k denote the sample size and index set of class k, respectively; i.e., $M_k = |\{i \in [M] | y_i = k\}| = |I_k|$ and $\sum_k M_k = M$.

3.3.1. Consistency

A crucial property of the LSEL is *consistency*; therefore, by minimizing (2), the estimated LLR matrix $\hat{\lambda}$ approaches the true LLR matrix λ as the sample size increases. The formal statement is given as follows:

Theorem 3.1 (Consistency of the LSEL). *Let $L(\boldsymbol{\theta})$ and $\hat{L}_S(\boldsymbol{\theta})$ denote $L_{LSEL}[\hat{\lambda}(\cdot;\boldsymbol{\theta})]$ and $\hat{L}_{LSEL}(\boldsymbol{\theta}; S)$ respectively. Let $\hat{\boldsymbol{\theta}}_S$ be the empirical risk minimizer of $\hat{L}_S$; namely, $\hat{\boldsymbol{\theta}}_S := \mathrm{argmin}_{\boldsymbol{\theta}} \hat{L}_S(\boldsymbol{\theta})$. Let $\Theta^* := \{\boldsymbol{\theta}^* \in \mathbb{R}^{d_\theta} | \hat{\lambda}(X^{(1,t)}; \boldsymbol{\theta}^*) = \lambda(X^{(1,t)})$ $(\forall t \in [T])\}$ be the target parameter set. Assume, for simplicity of proof, that each $\boldsymbol{\theta}^*$ is separated in Θ^*; i.e., $\exists \delta > 0$ such that $B(\boldsymbol{\theta}^*; \delta) \cap B(\boldsymbol{\theta}^{*\prime}; \delta) = \emptyset$ for arbitrary $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{*\prime} \in \Theta^*$, where $B(\boldsymbol{\theta}; \delta)$ denotes an open ball at center $\boldsymbol{\theta}$ with radius δ. Assume the following three conditions:*

(a) $\forall k, l \in [K]$, $\forall t \in [T]$, $p(X^{(1,t)}|k) = 0 \iff p(X^{(1,t)}|l) = 0$.

(b) $\sup_{\boldsymbol{\theta}} |\hat{L}_S(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \xrightarrow[M\to\infty]{P} 0$; i.e., $\hat{L}_S(\boldsymbol{\theta})$ converges in probability uniformly over $\boldsymbol{\theta}$ to $L(\boldsymbol{\theta})$.

(c) For all $\theta^* \in \Theta^*$, there exist $t \in [T]$, $k \in [K]$ and $l \in [K]$, such that the following $d_\theta \times d_\theta$ matrix is full-rank:

$$\int dX^{(1,t)} p(X^{(1,t)}|k) \times \times \nabla_{\boldsymbol{\theta}^*} \hat{\lambda}_{kl}(X^{(1,t)}; \boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}^*} \hat{\lambda}_{kl}(X^{(1,t)}; \boldsymbol{\theta}^*)^\top \,. \quad (3)$$

Then, $P(\hat{\boldsymbol{\theta}}_S \notin \Theta^*) \xrightarrow{M\to\infty} 0$; i.e., $\hat{\boldsymbol{\theta}}_S$ converges in probability into Θ^*.

Assumption (a) ensures that $\lambda(X^{(1,t)})$ exists and is finite. Assumption (b) can be satisfied under the standard assumptions of the uniform law of large numbers (compactness, continuity, measurability, and dominance) (Jennrich, 1969; Newey & McFadden, 1986). Assumption (c) is a technical requirement, often assumed in the literature (Gutmann & Hyvärinen, 2012). The complete proof is given in Appendix C.

The critical hurdle of the MSPRT to practical applications (availability to the true LLR matrix) is now relaxed by virtue of the LSEL, which is provably consistent and enables a precise estimation of the LLR matrix. We emphasize that the MSPRT is the earliest and most accurate algorithm for early classification of time series, at least asymptotically (Theorem A.1, A.2, and A.3).

3.3.2. Hard Class Weighting Effect

We further discuss the LSEL by focusing on a connection with hard negative mining (Song et al., 2016). It is empirically known that designing a loss function to emphasize hard classes improves model performance (Lin et al., 2017). The LSEL has this mechanism.

Let us consider a multiclass classification problem to obtain a high-performance discriminative model. To emphasize hard classes, let us minimize $\hat{L} := \frac{1}{KT} \sum_{k\in[K]} \sum_{t\in[T]} \frac{1}{M_k} \sum_{i\in I_k} \max_{l(\neq y_i)}\{e^{-\hat{\lambda}_{y_i l}(X_i^{(1,t)};\boldsymbol{\theta})}\}$; however, mining the single hardest class with the max function induces a bias and causes the network to converge to a bad local minimum. Instead of $\hat{L}$, we can use the LSEL because it is not only provably consistent but is a smooth upper bound of $\hat{L}$: Because

$\max_{l(\neq y_i)}\{e^{-\hat{\lambda}_{y_i l}(X_i^{(1,t)};\boldsymbol{\theta})}\} < \sum_{l(\neq y_i)} e^{-\hat{\lambda}_{y_i l}(X_i^{(1,t)};\boldsymbol{\theta})}$, we obtain $\hat{L} < \hat{L}_{\mathrm{LSEL}}$ by summing up both sides with respect to $i \in I_k$ and then $k \in [K]$ and $t \in [T]$. Therefore, a small $\hat{L}_{\mathrm{LSEL}}$ indicates a small $\hat{L}$. In addition, the gradients of the LSEL are dominated by the hardest class $k^* \in \mathrm{argmax}_{k(\neq y)}\{e^{-\hat{\lambda}_{yk}(X^{(1,T)};\boldsymbol{\theta})}\}$, because for all $k(\neq y, k^*)$,

$$\left|\frac{\partial \hat{L}_{\mathrm{LSEL}}}{\partial \hat{\lambda}_{yk}}\right| \propto \frac{e^{-\hat{\lambda}_{yk}}}{\sum_{l\in[K]} e^{-\hat{\lambda}_{yl}}} < \frac{e^{-\hat{\lambda}_{yk^*}}}{\sum_{l\in[K]} e^{-\hat{\lambda}_{yl}}} \propto \left|\frac{\partial \hat{L}_{\mathrm{LSEL}}}{\partial \hat{\lambda}_{yk^*}}\right|,$$

meaning that the LSEL assigns large gradients to the hardest class during training, which accelerates convergence.

Let us compare the hard class weighting effect of the LSEL with that of the logistic loss (a sum-log-exp-type loss extensively used in machine learning). For notational convenience, let us define $\ell_{\mathrm{LSEL}} := \log(1 + \sum_{k(\neq y)} e^{a_k})$ and $\ell_{\mathrm{logistic}} := \sum_{k(\neq y)} \log(1 + e^{a_k})$, where $a_k := -\hat{\lambda}_{yk}(X^{(1,t)};\boldsymbol{\theta})$, and compare their gradient scales. The gradients for $k \neq y$ are:

$$\frac{\partial \ell_{\mathrm{logistic}}}{\partial \hat{\lambda}_{yk}} = -\frac{e^{-\hat{\lambda}_{yk}}}{1 + e^{-\hat{\lambda}_{yk}}} =: b_k\,,$$
$$\frac{\partial \ell_{\mathrm{LSEL}}}{\partial \hat{\lambda}_{yk}} = -\frac{e^{-\hat{\lambda}_{yk}}}{\sum_{l\in[K]} e^{-\hat{\lambda}_{yl}}} =: c_k\,.$$

The relative gradient scales of the hardest class to the easiest class are:

$$R_{\mathrm{logistic}} := \frac{\max_{k(\neq y)}\{b_k\}}{\min_{k(\neq y)}\{b_k\}} = \frac{e^{a_{k^*}}}{e^{a_{k_*}}}\frac{e^{1+a_{k_*}}}{e^{1+a_{k^*}}}\,,$$
$$R_{\mathrm{LSEL}} := \frac{\max_{k(\neq y)}\{c_k\}}{\min_{k(\neq y)}\{c_k\}} = \frac{e^{a_{k^*}}}{e^{a_{k_*}}}\,,$$

where $k_* := \mathrm{argmin}_{k(\neq y)}\{a_k\}$. Since $R_{\mathrm{logistic}} \leq R_{\mathrm{LSEL}}$, we conclude that the LSEL weighs hard classes more than the logistic loss. Note that our discussion above also explains why log-sum-exp-type losses (e.g., (Song et al., 2016; Wang et al., 2019; Sun et al., 2020)) perform better than sum-log-exp-type losses. In addition, Figure 2 (Top and Bottom) shows that the LSEL performs better than the logistic loss—a result that supports the discussion above. See Appendix E for more empirical results.

3.3.3. Cost-Sensitive LSEL and Guess-Aversion

Furthermore, we prove that the cost-sensitive LSEL provides discriminative scores even on imbalanced datasets. Conventional research for cost-sensitive learning has been mainly focused on binary classification problems (Fan et al., 1999; Elkan, 2001; Viola & Jones, 2002; Masnadi-Shirazi & Vasconcelos, 2010). However, in *multiclass* cost-sensitive learning, (Beijbom et al., 2014) proved that random score functions (a "random guess") can lead to even smaller values of the loss function. Therefore, we should investigate whether our loss function is averse (robust) to such random guesses, i.e., *guess-averse*.

Definitions Let $\boldsymbol{s} : \mathcal{X} \to \mathbb{R}^K$ be a score vector function; i.e., $s_k(X^{(1,t)})$ represents how likely it is that $X^{(1,t)}$ is sampled from class k. In the LSEL, we can regard $\log \hat{p}_{\boldsymbol{\theta}}(X^{(1,t)}|k)$ as $s_k(X^{(1,t)})$. A cost matrix C is a matrix on $\mathbb{R}^{K\times K}$ such that $C_{kl} \geq 0$ ($\forall k, l \in [K]$), $C_{kk} = 0$ ($\forall k \in [K]$), $\sum_{l\in[K]} C_{kl} \neq 0$ ($\forall k \in [K]$). C_{kl} represents a misclassification cost, or a weight for the loss function, when the true label is k and the prediction is l. The *support set* of class k is defined as $\mathcal{S}_k := \{\boldsymbol{v} \in \mathbb{R}^K \mid \forall l(\neq k),\ v_k > v_l\}$. Ideally, discriminative score vectors should be in $\mathcal{S}_k$ when the label is k. In contrast, the *arbitrary guess set* is defined as $\mathcal{A} := \{\boldsymbol{v} \in \mathbb{R}^K \mid v_1 = v_2 = ... = v_K\}$. If $\boldsymbol{s}(X_i^{(1,t)}) \in \mathcal{A}$, we cannot gain any information from $X_i^{(1,t)}$; therefore, well-trained discriminative models should avoid such an *arbitrary guess* of $\boldsymbol{s}$. We consider a class of loss functions such that $\ell(\boldsymbol{s}(X^{(1,t)}), y; C)$: It depends on $X^{(1,t)}$ through the score function $\boldsymbol{s}$. The loss $\ell(\boldsymbol{s}(X^{(1,t)}), y; C)$ is *guess-averse*, if for any $k \in [K]$, any $\boldsymbol{s} \in \mathcal{S}_k$, any $\boldsymbol{s}' \in \mathcal{A}$, and any cost matrix C, $\ell(\boldsymbol{s}, k; C) < \ell(\boldsymbol{s}', k; C)$; thus, the guess-averse loss can provide discriminative scores by minimizing it. The empirical loss $\hat{L} = \frac{1}{MT}\sum_{i=1}^{M}\sum_{t=1}^{T} \ell(\boldsymbol{s}(X_i^{(1,t)}), y_i; C)$ is said to be guess-averse, if ℓ is guess-averse. The guess-aversion trivially holds for most binary and multiclass loss functions but does not generally hold for cost-sensitive multiclass loss functions due to the complexity of multiclass decision boundaries (Beijbom et al., 2014).

Cost-sensitive LSEL is guess-averse. We define a cost-sensitive LSEL:

$$\hat{L}_{\mathrm{CLSEL}}(\boldsymbol{\theta}, C; S) := \frac{1}{MT}\sum_{i=1}^{M}\sum_{t=1}^{T} C_{y_i} \log(1 + \sum_{l(\neq y_i)} e^{-\hat{\lambda}_{y_i l}(X_i^{(1,t)};\boldsymbol{\theta})})\,, \tag{4}$$

where $C_{kl} = C_k$ ($\forall k, l \in [K]$). Note that $\hat{\lambda}$ is no longer an unbiased estimator of the true LLR matrix; i.e., $\hat{\lambda}$ does not necessarily converge to λ as $M \to \infty$, except when $C_k = M/M_k(K-1)$ ($\hat{L}_{\mathrm{CLSEL}}$ reduces to $\hat{L}_{\mathrm{LSEL}}$). Nonetheless, the following theorem shows that $\hat{L}_{\mathrm{CLSEL}}$ is guess-averse. The proof is given in Appendix G.1.

Theorem 3.2. *$\hat{L}_{\mathrm{CLSEL}}$ is guess-averse, provided that the log-likelihood vector*

$$\Big(\log \hat{p}_{\boldsymbol{\theta}}(X^{(1,t)}|y=1), \log \hat{p}_{\boldsymbol{\theta}}(X^{(1,t)}|y=2), ..., \log \hat{p}_{\boldsymbol{\theta}}(X^{(1,t)}|y=K)\Big)^\top \in \mathbb{R}^K$$

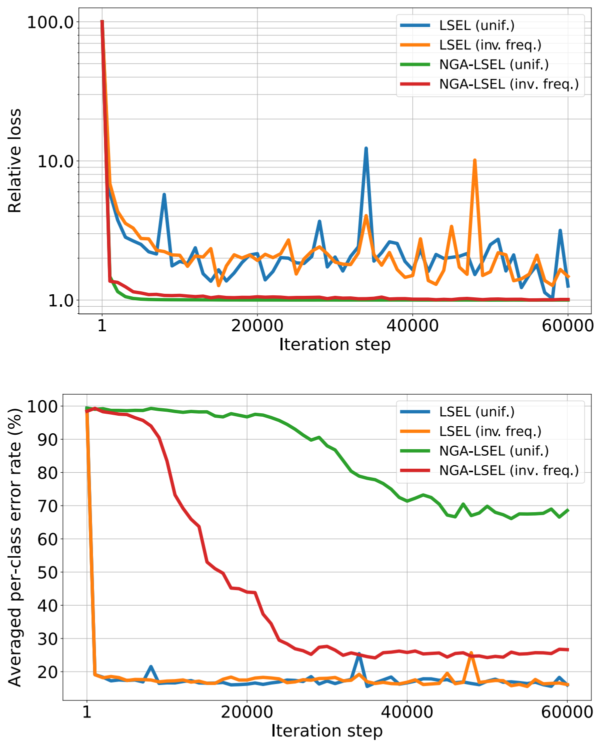

Figure 3. **Top: Relative Loss v.s. Training Iteration of LSEL and NGA-LSEL with Two Cost Matrices for Each. Bottom: Averaged Per-Class Error Rate of Last Frame v.s. Training Iteration.** Although all the loss curves decrease and converge (top), the error rates of the NGA-LSEL converge slowly and show a large gap depending on the cost matrix, while the error rates of the LSEL converge rapidly, and the gap is small (bottom). "unif." means $C_{kl} = 1$ and "inv. freq." means $C_{kl} = 1/M_k$. The dataset is UCF101 (Soomro et al., 2012).

is regarded as the score vector $\boldsymbol{s}(X^{(1,t)})$.

Figure 3 illustrates the risk of non-guess-averse losses. We define the non-guess-averse LSEL (NGA-LSEL) as $\ell(\boldsymbol{s}, y; C) = \sum_{k(\neq y)} C_{y,l} \log(1 + \sum_{l(\neq k)} e^{s_l - s_k})$. It is inspired by a variant of an exponential loss $\ell(\boldsymbol{s}, y; C) = \sum_{k,l \in [K]} C_{y,l} e^{s_l - s_k}$, which is proven to be classification calibrated but is not guess-averse (Beijbom et al., 2014). The NGA-LSEL benefits from the log-sum-exp structure but is not guess-averse (Appendix G.2), unlike the LSEL.

3.4. MSPRT-TANDEM

Although the LSEL alone works well, we further combine the LSEL with a DRE-based model, SPRT-TANDEM, recently proposed in (Ebihara et al., 2021). Specifically, we use the *TANDEM formula* and *multiplet loss* to accelerate the convergence. The TANDEM formula transforms

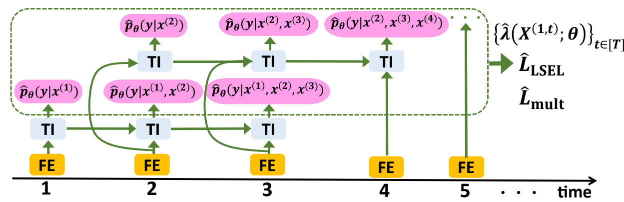

Figure 4. **MSPRT-TANDEM ($N = 2$).** $x^{(t)}$ is an input vector; e.g., a video frame. FE is a feature extractor. TI is a temporal integrator, which allows two inputs: the feature vector and a hidden state vector, which encodes the information of the past frames. We use ResNet and LSTM for FE and TI, respectively, but are not limited to them in general. The output posterior densities are highlighted with pink circles. By aggregating the posterior densities, the multiplet loss is calculated. Also, the estimated LLR matrix $\hat{\lambda}$ is constructed using the M-TANDEM or M-TANDEMwO formulae. Finally, $\hat{\lambda}$ is input to the LSEL. $\hat{L}_{\text{LSEL}} + \hat{L}_{\text{mult}}$ is optimized with gradient descent. In the test phase, $\hat{\lambda}(X^{(1,t)})$ is used to execute the MSPRT (Figure 1 and Definition 3.1).

the output of the network ($\hat{p}(y|X^{(1,t)})$) to the likelihood $\hat{p}(X^{(1,t)}|y)$ under the N-th order Markov approximation, which avoids the gradient vanishing of recurrent neural networks (Ebihara et al., 2021):

$$\hat{\lambda}_{kl}(X^{(1,t)}) \fallingdotseq \sum_{s=N+1}^{t} \log \left(\frac{\hat{p}_{\boldsymbol{\theta}}(k|X^{(s-N,s)})}{\hat{p}_{\boldsymbol{\theta}}(l|X^{(s-N,s)})} \right) - \sum_{s=N+2}^{t} \log \left(\frac{\hat{p}_{\boldsymbol{\theta}}(k|X^{(s-N,s-1)})}{\hat{p}_{\boldsymbol{\theta}}(l|X^{(s-N,s-1)})} \right), \quad (5)$$

where we do not use the prior ratio term $-\log(\hat{p}(k)/\hat{p}(l)) = -\log(M_k/M_l)$ in our experiments because it plays a similar role to the cost matrix (Menon et al., 2021). Note that (5) is a generalization of the original to DRME, and thus we call it the *M-TANDEM formula*.

However, we find that the M-TANDEM formula contains contradictory gradient updates caused by the middle minus sign. Let us consider an example $z_i := (X_i^{(1,t)}, y_i)$. The posterior $\hat{p}_{\boldsymbol{\theta}}(y = y_i|X_i^{(s-N,s-1)})$ (appears in (5)) should take a *large* value for z_i because the posterior density represents the probability that the label is y_i. For the same reason, $\hat{\lambda}_{y_i l}(X^{(1,t)})$ should take a high value; thus $\hat{p}_{\boldsymbol{\theta}}(y = y_i|x^{(s-N)}, ..., x^{(s-1)})$ should take a *small* value in accordance with (5) — an apparent contradiction. These contradictory updates may cause a conflict of gradients and slow the convergence of training, leading to performance deterioration. Therefore, in the experiments, we use either (5) or another approximation formula: $\hat{\lambda}_{kl}(X^{(1,t)}; \boldsymbol{\theta}) \fallingdotseq \log(\hat{p}_{\boldsymbol{\theta}}(k|X^{(t-N,t)})/\hat{p}_{\boldsymbol{\theta}}(l|X^{(t-N,t)}))$, which we call the M-TANDEM with Oblivion (M-TANDEMwO) formula. Clearly, the gradient does not conflict. Note that both the M-TANDEM and M-TANDEMwO formulae are anti-symmetric and do not violate the constraint $\hat{\lambda}_{kl} + \hat{\lambda}_{lm} = \hat{\lambda}_{km}$ $(\forall k, l, m \in [K])$. See Appendix F for a more detailed empirical comparison. Finally, the multiplet loss is a cross-entropy loss combined with the N-th order approximation: $\hat{L}_{\text{mult}}(\boldsymbol{\theta}; S) := \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{N+1}\sum_{t=k}^{T-(N+1-k)}(-\log \hat{p}_{\boldsymbol{\theta}}(y_i | X_i^{(t-k+1,t)}))$. An ablation study of the multiplet loss and the LSEL is provided in Appendix H.

The overall architecture, *MSPRT-TANDEM*, is illustrated in Figure 4. MSPRT-TANDEM can be used for *arbitrary* sequential data and thus has a wide variety of potential applications, such as computer vision, natural language processing, and signal processing. We focus on vision tasks in our experiments.

Note that in the training phase, MSPRT-TANDEM does not require a hyperparameter that controls the speed-accuracy tradeoff. A common strategy in early classification of time series is to construct a model that optimizes two cost functions: one for earliness and the other for accuracy (Dachraoui et al., 2015; Mori et al., 2015; Tavenard & Malinowski, 2016; Mori et al., 2018; Martinez et al., 2020). This approach typically requires a hyperparameter that controls earliness and accuracy (Achenchabe et al., 2020). The tradeoff hyperparameter is determined by heuristics and cannot be changed after training. However, MSPRT-TANDEM does not require such a hyperparameter and enables us to control the speed-accuracy tradeoff *after training* because we can change the threshold of MSPRT-TANDEM without retraining. This flexibility is an advantage for efficient deployment (Cai et al., 2020).

4. Experiment

To evaluate the performance of MSPRT-TANDEM, we use averaged per-class error rate and mean hitting time: Both measures are necessary because early classification of time series is a multi-objective optimization problem. The averaged per-class error rate, or balanced error, is defined as $1 - \frac{1}{K}\sum_{k=1}^{K}\frac{|\{i \in [M] | h_i = y_i = k\}|}{|\{i \in [M] | y_i = k\}|}$, where $h_i \in [K]$ is the prediction of the model for $i \in [M]$ in the dataset. The mean hitting time is defined as the arithmetic mean of the stopping times of all sequences.

We use four datasets: two are new simulated datasets made from MNIST (LeCun et al., 2010) (NMNIST-H and NMNIST-100f), and two real-world public datasets for multiclass action recognition (UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011)). A sequence in NMNIST-H consists of 20 frames of an MNIST image filled with dense random noise, which is gradually removed (10 pixels per frame), while a sequence in NMNIST-100f consists of 100 frames of an MNIST image filled with random noise that is so dense that humans cannot classify any video (Appendix K); only 15 of 28×28 pixels maintain the original image. The noise changes temporally and randomly and is not removed, unlike in NMNIST-H.

4.1. Models

We compare the performance of MSPRT-TANDEM with four other models: LSTM-s, LSTM-m (Ma et al., 2016), EARLIEST (Hartvigsen et al., 2019), and the Neyman-Pearson (NP) test (Neyman & Pearson, 1933). LSTM-s and LSTM-m, proposed in a pioneering work in deep learning-based early detection of human action (Ma et al., 2016), use loss functions that enhance monotonicity of class probabilities (LSTM-s) and margins of class probabilities (LSTM-m). Note that LSTM-s/m support only the fixed-length test; i.e., the stopping time is fixed, unlike MSPRT-TANDEM. EARLIEST is a reinforcement learning algorithm based on recurrent neural networks (RNNs). The base RNN of EARLIEST calculates a current state vector from an incoming time series. The state vector is then used to generate a stopping probability in accordance with the binary action sampled: Halt or Continue. EARLIEST has two objective functions in the total loss: one for classification error and one for earliness. The balance between them cannot change after training. The NP test is known to be the most powerful, Bayes optimal, and minimax optimal test (Borovkov, 1998; Lehmann & Romano, 2006). The NP test uses the LLR to make a decision in a similar manner to the MSPRT, but the decision time is fixed. The decision rule is $d^{\text{NP}}(X^{(1,t)}) := \text{argmax}_{k \in [K]} \min_{l \in [K]} \lambda_{kl}(X^{(1,t)})$ with a fixed $t \in [T]$. In summary, LSTM-s/m have different loss functions from MSPRT-TANDEM, and the stopping time is fixed. EARLIEST is based on reinforcement learning, and its stopping rule is stochastic. The only difference between the NP test and MSPRT-TANDEM is whether the stopping time is fixed.

We first train the feature extractor (ResNet (He et al., 2016a;b)) by solving multiclass classification with the softmax loss and extract the bottleneck features, which are then used to train LSTM-s/m, EARLIEST, and the temporal integrator for MSPRT-TANDEM and NP test. Note that all models use the same feature vectors for the training. For a fair comparison, hyperparameter tuning is carried out with the default algorithm of Optuna (Akiba et al., 2019) with an equal number of tuning trials for all models. Also, all models have the same order of trainable parameters. After fixing the hyperparameters, we repeatedly train the models with different random seeds to consider statistical fluctuation due to random initialization and stochastic optimizers. Finally, we test the statistical significance of the models with the two-way ANOVA followed by the Tukey-Kramer multi-comparison test. More detailed settings are given in Appendix I.

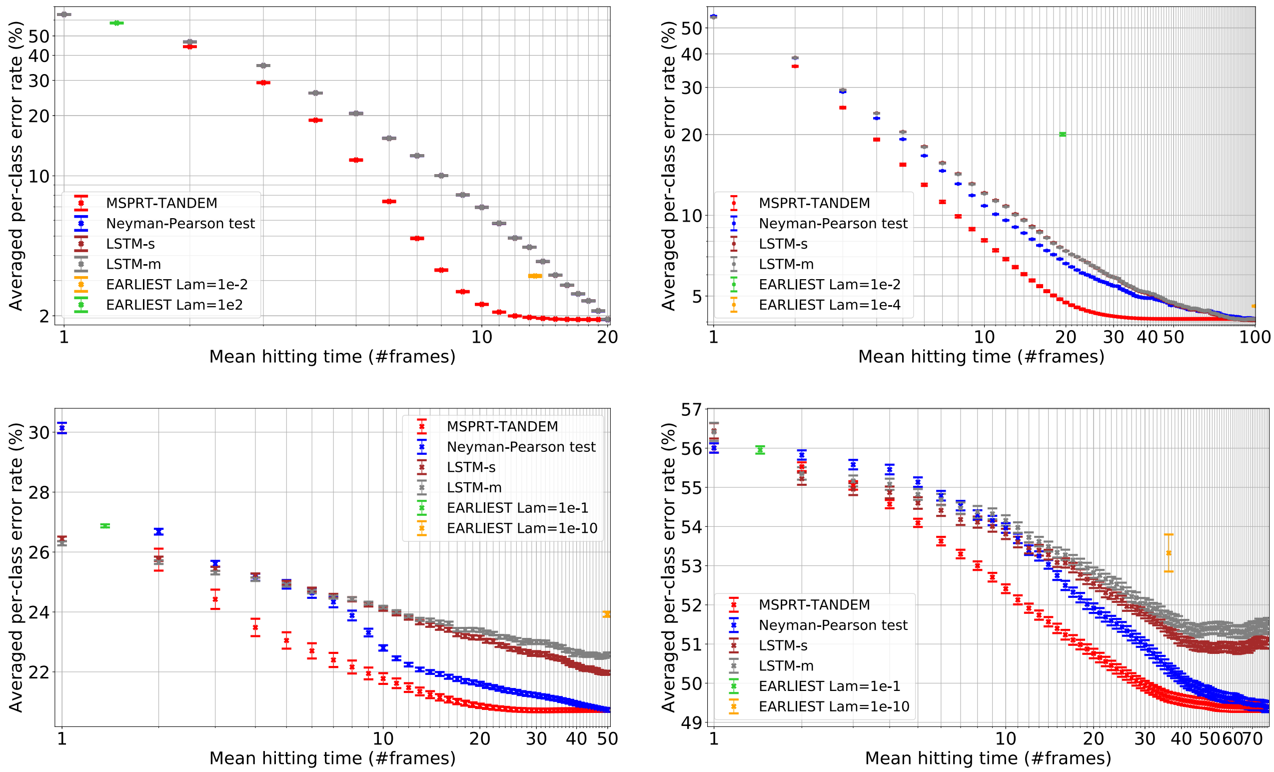

Figure 5. **Speed-Accuracy Tradeoff (SAT) Curves.** The vertical axis represents the averaged per-class error rate, and the horizontal axis represents the mean hitting time. Early and accurate models come in the lower-left area. The vertical error bars are the standard error of mean (SEM); however, some of the error bars are too small and are collapsed. **Upper left: NMNIST-H.** The Neyman-Pearson test, LSTM-s, and LSTM-m almost completely overlap. **Upper right: NMNIST-100f.** LSTM-s and LSTM-m completely overlap. **Lower left: UCF101**. **Lower right: HMDB51**.

4.2. Results

The performances of all the models are summarized in Figure 5 (The lower left area is preferable). We can see that MSPRT-TANDEM outperforms all the other models by a large margin, especially in the early stage of sequential observations. We confirm that the results have statistical significance; i.e., our results are reproducible (Appendix L). The loss functions of LSTM-s/m force the prediction score to be monotonic, even when noisy data are temporally observed, leading to a suboptimal prediction. In addition, LSTM-s/m have to make a decision, even when the prediction score is too small to make a confident prediction. However, MSPRT-TANDEM can wait until a sufficient amount of evidence is accumulated. A potential weakness of EARLIEST is that reinforcement learning is generally unstable during training, as pointed out in (Nikishin et al., 2018; Kumar et al., 2020). The NP test requires more observations to attain a comparable error rate to that of MSPRT-TANDEM, as expected from the theoretical perspective (Tartakovsky et al., 2014): In fact, the SPRT was originally developed to outperform the NP test in sequential testing (Wald, 1945; 1947).

5. Conclusion

We propose the LSEL for DRME, which has yet to be explored in the literature. The LSEL relaxes the crucial assumption of the MSPRT and enables its real-world applications. We prove that the LSEL has a theoretically strong background: consistency, hard class weighting, and guess-aversion. We also propose MSPRT-TANDEM, the first DRE-based model for early multiclass classification in deep learning. The experiment shows that the LSEL and MSPRT-TANDEM outperform other baseline models statistically significantly.

Acknowledgments

The authors thank the anonymous reviewers for their careful reading to improve the manuscript. We would like to thank Jiro Abe, Genki Kusano, Yuta Hatakeyama, Kazuma Shimizu, and Natsuhiko Sato for helpful comments on the

proof of the LSEL's consistency.

References

Achenchabe, Y., Bondu, A., Cornuéjols, A., and Dachraoui, A. Early classification of time series. cost-based optimization criterion and algorithms. *arXiv preprint arXiv:2005.09945*, 2020.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. License: MIT License.

Armitage, P. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(1):137–144, 1950.

Baum, C. W. and Veeravalli, V. V. A sequential procedure for multihypothesis testing. *IEEE Transactions on Information Theory*, 40(6):1994–2007, Nov 1994.

Beijbom, O., Saberian, M., Kriegman, D., and Vasconcelos, N. Guess-averse loss functions for cost-sensitive multiclass boosting. In *International Conference on Machine Learning*, pp. 586–594, 2014.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540, 2018.

Borovkov, A. *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Burkholder, D. L. and Wijsman, R. A. Optimum properties and admissibility of sequential tests. *The Annals of Mathematical Statistics*, 34(1):1–17, 1963.

Cai, H., Gan, C., Wang, T., Zhang, Z., and Han, S. Once-for-all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HylxE1HKwS`.

Chernoff, H. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Dachraoui, A., Bondu, A., and Cornuéjols, A. Early classification of time series as a non myopic sequential decision making problem. In Appice, A., Rodrigues, P. P., Santos Costa, V., Soares, C., Gama, J., and Jorge, A. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 433–447, Cham, 2015. Springer International Publishing. ISBN 978-3-319-23528-8.

Doná, R., Papini, G. P. R., and Valenti, G. MSPRT action selection model for bio-inspired autonomous driving and intention prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop*, 2019. URL `https://www.dreams4cars.eu/sites/default/files/2020-01/IROS2019_Riccardo.pdf`.

Dragalin, V. Asymptotic solution of a problem of detecting a signal from k channels. *Russian Mathematical Surveys*, 42(3):213, 1987.

Dragalin, V. and Novikov, A. Adaptive sequential tests for composite hypotheses. *Survey of Applied and Industrial Mathematics*, 6:387–398, 1999.

Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V. V. Multihypothesis sequential probability ratio tests. i. asymptotic optimality. *IEEE Transactions on Information Theory*, 45(7):2448–2461, November 1999. doi: 10.1109/18.796383.

Ebihara, A. F., Miyagawa, T., Sakurai, K., and Imaoka, H. Sequential density ratio estimation for simultaneous optimization of speed and accuracy. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Rhsu5qD36cL`.

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Fan, W., Stolfo, S. J., Zhang, J., and Chan, P. K. Adacost: Misclassification cost-sensitive boosting. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 97–105, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Ferguson, T. S. *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press, 2014.

Fisher, R. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

Ghalwash, M. F. and Obradovic, Z. Early classification of multivariate temporal observations by extraction of interpretable shapelets. *BMC bioinformatics*, 13(1):195, 2012.

Ghalwash, M. F., Radosavljevic, V., and Obradovic, Z. Extraction of interpretable multivariate patterns for early diagnostics. In *2013 IEEE 13th International Conference on Data Mining*, pp. 201–210, 2013. doi: 10.1109/ICDM.2013.19.

Ghalwash, M. F., Radosavljevic, V., and Obradovic, Z. Utilizing temporal patterns for estimating uncertainty in interpretable early decision making. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 402–411, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623694. URL https://doi.org/10.1145/2623330.2623694.

Ghodrati, A., Bejnordi, B. E., and Habibian, A. FrameExit: Conditional early exiting for efficient video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Gutmann, M. and Hirayama, J.-i. Bregman divergence as general framework to estimate unnormalized statistical models. *arXiv preprint arXiv:1202.3727*, 2012.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 13(1):307–361, 2012.

Hartvigsen, T., Sen, C., Kong, X., and Rundensteiner, E. Adaptive-halting policy network for early classification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 101–110, New York, NY, USA, 2019. ACM.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pp. 630–645, 2016b.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In Kremer, S. C. and Kolen, J. F. (eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, October 2002. ISSN 1088-467X.

Jennrich, R. I. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643, 04 1969. doi: 10.1214/aoms/1177697731. URL https://doi.org/10.1214/aoms/1177697731.

Kanamori, T., Hido, S., and Sugiyama, M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

Karim, F., Darabi, H., Harford, S., Chen, S., and Sharabiani, A. A framework for accurate time series classification based on partial observation. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pp. 634–639, 2019. doi: 10.1109/COASE.2019.8843256.

Khan, H., Marcuse, L., and Yener, B. Deep density ratio estimation for change point detection. *arXiv preprint arXiv:1905.09876*, 2019.

Kiefer, J. and Sacks, J. Asymptotically optimum sequential inference and design. *The Annals of Mathematical Statistics*, pp. 705–750, 1963.

Kramer, C. Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 12(3):307–310, 1956.

Kubat, M. and Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186. Morgan Kaufmann, 1997. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4487.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. URL https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/. License: Creative Commons Attribution 4.0 International License.

Kumar, A., Gupta, A., and Levine, S. DisCor: Corrective feedback in reinforcement learning via distribution correction. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2020.

Lai, T. L. Asymptotic optimality of invariant sequential probability ratio tests. *The Annals of Statistics*, pp. 318–333, 1981.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun. com/exdb/mnist*, 2, 2010. License: Creative Commons Attribution-Share Alike 3.0 license.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Lorden, G. Integrated risk of asymptotically bayes sequential tests. *The Annals of Mathematical Statistics*, 38(5): 1399–1422, 1967.

Lorden, G. Nearly-optimal sequential tests for finitely many parameter values. *Annals of Statistics*, 5:1–21, 01 1977. doi: 10.1214/aos/1176343737.

Ma, S., Sigal, L., and Sclaroff, S. Learning activity progression in lstms for activity detection and early detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1942–1950, 2016.

Mao, Y., Chen, W., Chen, Y., Lu, C., Kollef, M., and Bailey, T. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1140–1148, 2012.

Martinez, C., Ramasso, E., Perrin, G., and Rombaut, M. Adaptive early classification of temporal sequences using deep reinforcement learning. *Knowledge-Based Systems*, 190:105290, February 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.105290. URL http://www.sciencedirect.com/science/article/pii/S0950705119305829.

Masnadi-Shirazi, H. and Vasconcelos, N. Cost-sensitive boosting. *IEEE Transactions on pattern analysis and machine intelligence*, 33(2):294–309, 2010.

Matches, T. K. On the optimality of sequential probability ratio tests. *The Annals of Mathematical Statistics*, 34:18, 1963.

McGovern, A., Rosendahl, D. H., Brown, R. A., and Droegemeier, K. K. Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction. *Data Mining and Knowledge Discovery*, 22(1-2): 232–258, 2011.

Menon, A. and Ong, C. S. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pp. 304–313, 2016.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Mori, U., Mendiburu, A., Dasgupta, S., and Lozano, J. A. Early classification of time series from a cost minimization point of view. In *Proceedings of the NIPS Time Series Workshop*, 2015.

Mori, U., Mendiburu, A., Keogh, E. J., and Lozano, J. A. Reliable early classification of time series based on discriminating the classes over time. *Data Mining and Knowledge Discovery*, 31:233–263, 04 2016. doi: 10.1007/s10618-016-0462-1.

Mori, U., Mendiburu, A., Dasgupta, S., and Lozano, J. A. Early classification of time series by simultaneously optimizing the accuracy and earliness. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (10):4569–4578, Oct 2018. ISSN 2162-237X. doi: 10.1109/TNNLS.2017.2764939.

Moustakides, G. V. and Basioti, K. Training neural networks for likelihood/density ratio estimation. *arXiv preprint arXiv:1911.00405*, 2019.

Newey, W. and McFadden, D. Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. (eds.), *Handbook of Econometrics*, volume 4, chapter 36, pp. 2111–2245. Elsevier, 1 edition, 1986. URL https://EconPapers.repec.org/RePEc:eee:ecochp:4-36.

Neyman, J. and Pearson, E. S. 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*, 231:289–337, 1933. ISSN 02643952. URL http://www.jstor.org/stable/91247.

Nikishin, E., Izmailov, P., Athiwaratkun, B., Podoprikhin, D., Garipov, T., Shvechikov, P., Vetrov, D., and Wilson, A. G. Improving stability in deep reinforcement learning with weight averaging. In *Uncertainty in artificial intelligence workshop on uncertainty in Deep learning*, 2018.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Paulson, E. A sequential decision procedure for choosing one of k hypotheses concerning the unknown mean of a normal distribution. *The Annals of Mathematical Statistics*, pp. 549–554, 1963.

Pavlov, I. Sequential decision rule for the case of many complex hypotheses. *ENG. CYBER.*, (6):19–22, 1984.

Pavlov, I. V. Sequential procedure of testing composite hypotheses with applications to the kiefer–weiss problem. *Theory of Probability & Its Applications*, 35(2):280–292, 1991.

Shiryaev, A. N. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007.

Simons, G. Lower bounds for average sample number of sequential multihypothesis tests. *The Annals of Mathematical Statistics*, pp. 1343–1364, 1967.

Sobel, M. and Wald, A. A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Statist.*, 20(4):502–522, 12 1949. doi: 10.1214/aoms/1177729944. URL https://doi.org/10.1214/aoms/1177729944.

Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004–4012, 2016. doi: 10.1109/CVPR.2016.434.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. URL https://www.crcv.ucf.edu/data/UCF101.php. License: Unknown.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012. doi: 10.1017/CBO9781139035613.

Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., and Wei, Y. Circle loss: A unified perspective of pair similarity optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tartakovsky, A. Sequential methods in the theory of information systems, 1991.

Tartakovsky, A., Nikiforov, I., and Basseville, M. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman & Hall/CRC, 1st edition, 2014.

Tartakovsky, A. G. Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case. *Statistical Inference for Stochastic Processes*, 1(3):265–295, 1998.

Tavenard, R. and Malinowski, S. Cost-aware early classification of time series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 632–647. Springer, 2016.

Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 5 2:99–114, 1949.

Vats, E. and Chan, C. S. Early detection of human actions—a hybrid approach. *Applied Soft Computing*, 46:953 – 966, 2016. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2015.11.007. URL http://www.sciencedirect.com/science/article/pii/S1568494615007188.

Viola, P. and Jones, M. Fast and robust classification using asymmetric adaboost and a detector cascade. In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems*, volume 14, pp. 1311–1318. MIT Press, 2002. URL https://proceedings.neurips.cc/paper/2001/file/0b1ec366924b26fc98fa7b71a9c249cf-Paper.pdf.

Wald, A. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945.

Wald, A. *Sequential Analysis*. John Wiley and Sons, 1st edition, 1947.

Wald, A. and Wolfowitz, J. Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, 19(3):326–339, 09 1948.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Xing, Z., Pei, J., and Yu, P. S. Early prediction on time series: A nearest neighbor approach. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pp. 1297–1302, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

Xing, Z., Pei, J., Yu, P. S., and Wang, K. Extracting interpretable features for early classification on time series. In *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*, pp. 247–258. SIAM, 2011.

Xing, Z., Pei, J., and Yu, P. S. Early classification on time series. *Knowledge and Information Systems*, 31(1):105–127, April 2012.

Zhang, Y., Xu, Y., Dong, Z. Y., Xu, Z., and Wong, K. P. Intelligent early warning of power system dynamic insecurity risk: Toward optimal accuracy-earliness tradeoff. *IEEE Transactions on Industrial Informatics*, 13(5):2544–2554, 2017. doi: 10.1109/TII.2017.2676879.