Stopping Criterion Design for Recursive Bayesian Classification: Analysis and Decision Geometry

Aziz Koçanaoğulları, Murat Akcakaya, Deniz Erdoğmuş

Abstract—Systems that are based on recursive Bayesian updates for classification limit the cost of evidence collection through certain stopping/termination criteria and accordingly enforce decision making. Conventionally, two termination criteria based on pre-defined thresholds over (i) the maximum of the state posterior distribution; and (ii) the state posterior uncertainty are commonly used. In this paper, we propose a geometric interpretation over the state posterior progression and accordingly we provide a point-by-point analysis over the disadvantages of using such conventional termination criteria. For example, through the proposed geometric interpretation we show that confidence thresholds defined over maximum of the state posteriors suffer from stiffness that results in unnecessary evidence collection whereas uncertainty based thresholding methods are fragile to number of categories and terminate prematurely if some state candidates are already discovered to be unfavorable. Moreover, both types of termination methods neglect the evolution of posterior updates. We then propose a new stopping/termination criterion with a geometrical insight to overcome the limitations of these conventional methods and provide a comparison in terms of decision accuracy and speed. We validate our claims using simulations and using real experimental data obtained through a brain computer interfaced typing system.

Index Terms—Active Learning, Sequential Decision Making, Recursive Bayesian Classification, Optimal Stopping Criterion Design

1 INTRODUCTION

Recursive Bayesian inference for classification (RBC) is beneficial in gradually increasing decision quality by incorporating more evidence into the decision process in situations where data or evidence is acquired sequentially over time. The most recent belief, represented by the latest label posterior probability distribution, is obtained by incorporating new evidence in a Bayesian manner at each update step [1], [2], [3]. [4]. The trade-off between decision confidence and evidence acquisition cost is controlled by a stopping criterion that controls when to terminate evidence collection and return an estimated class label. Fundamental components of RBC include: (S) A stopping criterion based on the posterior probability to stop evidence collection; (Q) a querying step to decide how to collect further evidence from relevant sources to benefit speed and accuracy objectives of RBC; (C) an classification objective based on the posterior distribution and loss values attributed to each true label and decision option pair to determine the optimal decision once the stopping criterion has been satisfied. The iterative process can be summarized as follows:

\[ \text{while } (S) \text{ not satisfied} \{ \text{do } (Q) \}; \text{return state with } (C) \]

In this paper, we focus on designing stopping criteria for (S) based on the latest label posterior \( p = [p_1, p_2, \ldots, p_n] \), which is a categorical probability distribution. In the remainder of this paper, in order to keep the illustrations and derivations simple, we assume that the objective in (C) is to minimize probability of error (therefore 0-1 loss is assumed), thus when a decision is made, it will be based on the maximum a posteriori (MAP) classification rule; that is the then-most-likely-label in the latest posterior will be selected as the decision. The presented approach can be generalized to the more general expected loss minimization classification setting, but is outside the scope of this paper. We also do not discuss different querying strategies, as they too are outside the scope.

The RBC procedure will terminate when the stopping criterion in (S) (namely \( S_C \)) is met by the current label posterior distribution \( S_C(p) = \text{true} \). The most commonly used approach is to require that a confidence threshold has been exceeded \( S_C(p) = \text{true} \) if \( \max_i p_i > \tau \). All label probability distributions \( p \) that satisfy \( S_C \) form a set called the stopping region \( S_R = \{p | S_C(p) = \text{true} \} \). If the posterior distribution falls into this set, the system terminates and a classification-decision is made based on (C). The design of an optimal \( S_R \) has been referred to as “optimal stopping criterion design” [5].

In this paper, we focus on the analysis of \( S_C \). Specifically identifying the limitations of conventional confidence thresholding and uncertainty based methods, with the aid of analytical representations that arise from the geometry of the probability simplex, we propose a new perspective for stopping criterion design that enables trading off accuracy for speed in ways not possible with conventional uncertainty...
methods. In particular, as illustrated in Fig. 1(c), we propose to bend the stopping region to enable early decisions, especially in situations where there is no strong contender to the top decision choice (posterior moving from the middle region in the simplex), but not when there is a strong second contender option (posterior moving from near the edges in the simplex). In numerical experiments, we demonstrate that this strategy enables significant inference speed-up with negligible increase in expected loss.

**Problem Formulation and Related Work**

Let class label \( \sigma \) be an element of a finite set \( A \). We refer to each iteration in RBC as a **sequence**, indexed by \( s \in \mathbb{N} \). Each sequence may include a set of measurements acquired in response to queries \( \Phi_s \triangleq \{ \phi_1^s, \ldots, \phi_N^s \} \), where \( N \) denotes number of queries. These measurements provide evidence (raw data or processed features) \( \varepsilon_s \triangleq \{ \varepsilon_1^s, \ldots, \varepsilon_N^s \} \) conditioned on the queries, and the true label. RBC posterior updates use this evidence. For notation simplicity, we use \( \mathcal{H}_s \triangleq \{ \varepsilon_1^s, \Phi_1^s, \mathcal{H}_0^s \} \) to represent the combination of all evidence collected in sequences 1 to \( s \), as well as the prior, \( \mathcal{H}_0^s \). RBC steps (assuming MAP classification rule) are:

\[
\begin{align*}
(S) & : \quad p(\sigma|\mathcal{H}_s) \in \{p|\mathcal{S}_C(p) = \text{true}\} \\
(C) & : \quad \hat{\delta} = \arg \max_{\sigma \in A} p(\sigma|\mathcal{H}_s) \\
(Q) & : \quad \Phi_{s+1} \to \varepsilon_{s+1} \text{ with the anticipated joint} \\
& \quad p(\mathcal{H}_{s+1}) = p(\mathcal{H}_s|\varepsilon_{s+1}) p(\varepsilon_{s+1}|\Phi_{s+1})
\end{align*}
\]

In this paper we focus on (C) The most common \( \mathcal{S}_C \) is thresholding the highest value in the latest posterior by \( \tau \) \([6, 10]\). \( \mathcal{S}_C(p) = \text{true} \text{ if } \max_i p_i > \tau \). This rule directly enforces that a decision is made with a prespecified confidence level, and does not consider the distribution of probability mass among other options. To minimize false alarm rate, the threshold is usually set to a large number. Furthermore, this entangles early stopping options and leads to further redundant evidence collection \([11]\).

Relaxation of stiffness of posterior thresholding is possible with information theoretic objectives. Uncertainty measures, including Shannon entropy \([12]\), can be used to have termination based on the spread of posterior probability mass. Golovin \([9]\) uses Shannon entropy as a stopping criterion: \( \mathcal{S}_C(p) = \text{true} \text{ if } H(p) < c_H \). Yingzhen \([13]\) and Igal \([7]\) use Renyi entropy: \( \mathcal{S}_C(p) = \text{true} \text{ if } H_\alpha(p) < c_{H_\alpha} \). Uncertainty based stopping criteria based on Renyi entropy \( H_\alpha(p) \) family with \( \alpha \geq 0 \) includes, as special cases, Shannon entropy (in the limit as \( \alpha \to 1 \)), and confidence thresholding (in the limit as \( \alpha \to \infty \)) methods. In contrast to these uncertainty based stopping criteria, With a diminishing-returns perspective, Weinshall \([14]\) and Geisser \([8]\) apply a threshold to the Kullback-Leibler (KL) divergence between two consecutive posterior distributions to terminate evidence collection: \( \mathcal{S}_C(p) = \text{true} \text{ if } D_{KL}(p_s||p_{s-1}) < c_{D_{KL}} \). Pavlichin \([15]\) proposes a chained-KL divergence to monitor posterior progression in the probability simplex. Banerjee \([16]\) proposes using Bregman divergences in the context of clustering. All of these stopping criteria focus on how far the posterior is from the uniform distribution that is at the center of the probability simplex, as opposed to assessing how close the posterior to a vertex of the simplex (corresponding to a one-hot distribution) \([17]\). In this paper, we discuss the limitations of these criteria summarized in Table 1 and propose a new family of stopping criteria, then illustrate theoretically, and with numerical examples and experiments, the benefits of the proposed approach.

**Contributions**

We introduce a geometrical representation for recursive Bayesian classification and show that: (i) uncertainty based methods are sensitive to the number of possible classes, and (ii) they ignore the posterior update trajectory in the recursive classification task. Using (i) and (ii) together, we show that the stopping regions defined based on uncertainty methods diverge from the region formed by confidence threshold defined over the posterior distribution resulting in decrease in classification accuracy. We then propose a new stopping criterion design perspective that is not only robust to number of classes but also defines a stopping region that is in line with the progression of the posterior update. We show that such a perspective not only increases the classification speed by decreasing redundant evidence collection caused by the stopping regions defined through the confidence level thresholds, but also maintains a high classification accuracy by limiting false positive rate. We also provide a practical use case for the design. Specifically, we consider a brain computer interface typing system. Detailed demonstrations and proofs of the analytical results are provided in the appendix Sec. 7.4 for neat presentation.

**2 Problem geometry**

To promote the representative power of visualization, in this paper we use information geometric representation of recursive classification problem \([18]\). As will be discussed,
this representation allows us to represent progression of the posterior probability distribution \( p(\sigma|H_s) \) as the evidence collection steps \( s \) increases and to introduce a mean for analytical reasoning for stopping criterion \( SC \) design. We use simplex as a domain for probability distributions:

\[
\Delta_n = \{(p_1, p_2, \cdots, p_n) \in \mathbb{R}^n | p_i > 0 \forall i, \sum_i p_i = 1\} \quad (2)
\]

Here \( \Delta_n \) represents the set that includes categorical probability distributions, where the simplex is an \( n-1 \) dimensional geometrical object in \( \mathbb{R}^n \). We visualize an example simplex for a category with 3 elements in Fig.1(a). To preserve neat visualization, we interpret our reasoning using the triangle throughout the paper, but all the results can be generalized to \( n \)-dimensional case without loss of generality. Moreover, we use the following addition operation that allows us to represent posterior updates:

**Addition:** Given \( p, q \in \Delta_n \), the addition operation is defined as:

\[
p \oplus q = \left[\frac{p_1q_1 + p_2q_2 + \cdots + p_nq_n}{\sum_i p_iq_i}\right]
\]

Observe that addition satisfies the Bayes’ theorem: \( p(\sigma|\varepsilon) = p(\sigma) \oplus p(\varepsilon|\sigma) \) which allows us to represent the evolution of the posterior distribution using the addition operation within the simplex. For algebra to work \( p(\varepsilon|\sigma) \) does not need to be unit \( \ell_1 \) norm and hence \( \notin \Delta_n \), but what we interpret here is the normalized vector over \( \sigma \) even algebraically \( \oplus \) is applicable. Rigorous definition of this simplification is shown in the appendix Sec.7.2. The zero element in \( \Delta_n \) is denoted by \( u_n \) being the uniform distribution as shown in Fig.1(d).

**Proposition 1 (19).** Simplex \( \{\Delta_n|\oplus, \otimes\} \) forms a vector space.

In the presence of a prior distribution, recursive classification starts with a distribution probably different than uniform distribution as presented in Fig.1(a). To make a correct decision, the posterior for respective element is required to be the most likely (e.g. the region for \( a \) is highlighted in Fig.1(b)). Through evidence collections, the posterior distribution moves within the simplex visualized in Fig.1(c) that is algebraically denoted with \( \oplus \) above. \( (S) \) criterion forms the decision lines within the simplex and the inequality condition covers the area of evidence collection termination that are visualized by dashed areas in Fig.1(a). Once the posterior probability distribution reaches that area, the system terminates evidence collection, makes an inference outputs the classification decision.

### 3 Conventional Stopping Criteria

In this section we discuss the limitations of the conventional methods that are used for stopping criteria \( SC \) design. Consider the following 2 motivating examples assuming 10 class classification problems:

**Example 1:** \( p, q \in \Delta_{10} \) with \( p = [0.6, 0.4 - 8\varepsilon, \varepsilon, \cdots] \) where \( 0 < \varepsilon \ll 1 \) and \( q = [0.7, 0.03, \cdots] \). It is apparent that \( q \) is a better stopping point to decide on the 1st element as the classification result. However, if Shannon’s entropy is used to measure the distance from uniform distribution \( u_n \), and accordingly utilized as a stopping criterion, then \( 0.97 \approx H(p) < H(q) \approx 1.82 \). This means that the evidence collection would have been terminated when \( p \) is reached. Hence, the uncertainty based \( (S) \) suffers from not taking the probability mass index of interest into consideration.

**Example 2:** \( p, q \in \Delta_{10} \) with \( p_1 = [0.5, 0.5 - 8\varepsilon, \varepsilon, \cdots] \), \( p_2 = [0.6, 0.4 - 8\varepsilon, \varepsilon, \cdots] \) where \( 0 < \varepsilon \ll 1 \) and \( q_1 = [0.5, 0.05, \cdots] \), \( q_2 = [0.6, 0.04, \cdots] \) and we compare \( ps \) then \( qs \). Even the confidence level on a particular class differs the same amount \( 0.1 \), for \( ps \) the 2nd best class is still a legit competitor, whereas for \( qs \) there was no other competitor. Here even though \( qs \) are in a more central position in the simplex than \( ps \) they provide a higher confidence in classification. However, \( ps \) are close to the edge of the simplex and accordingly may result in the termination of the evidence collection before more confident posterior \( qs \) can be achieved. This means that the location of posteriors in \( \Delta_n \) matters and with an appropriate design of stopping criterion, the system evidence collection could be stopped when the confidence on the classification is higher.

In this section we analytically show what these examples mean for the stopping criterion \( (SC) \) design.

#### 3.1 Frail Confidence

The confidence behavior of uncertainty based methods are similar. Shannon entropy on the other hand is the most commonly used and hence in this section we specifically...
To analyze these, we introduce the relationship between the confidence in classification and entropy based stopping region $S_R(H(.)):=\{p|H(p)<c_H\}$ (M3) we define two special probability points in $\Delta_n$:

$$v_n(\tau) = \left[ \frac{1-\tau}{n-1}, \ldots, \frac{1-\tau}{n-1} \right]$$

$$w_n(\tau) = \left[ \tau, 1-\tau, 0, \ldots, 0 \right]$$

(3)

Here, $v_n(\tau)$ corresponds to distributions where one class has likelihood of $\tau$ and the others share the remaining probability uniformly, $w_n(\tau)$ corresponds to distributions where only two class exist with probability values $\tau, 1-\tau$. These definitions are also visualized in Fig.1(d).

In this paper, we bundle the uncertainty and confidence $\tau$ via $\tau' = H(v_n(\tau))$. We show the following weak points: $H(.)$ is sensitive to $n$ and $\tau$, and $H(.)$ conflicts with the inference step (C) of Eqn. (1) for a given set of parameters. To analyze these, we introduce the relationship between confidence and entropy $S_{\tau}$ with the help of (3). We state for a confidence $\tau$ that entropy achieves its maximum value at the point $v_n(\tau)$ and for an equi-entropy contour with value of $H(v_n(\tau))$ maximum achievable confidence is $\tau$. See the following proposition.

**Proposition 2.** For a defined confidence level $\tau$:

$$C_\tau = \{ p | \max p_i = \tau \} \implies \max_{p \in C_\tau} H(p) = H(v_n(\tau))$$

$$S_\tau = \{ p | H(p) = H(v_n(\tau)) \} \implies \max_{p \in S_\tau} \max p_i = \tau$$

**Observation 1.** Following descriptions in Proposition 2 confidence line for $\tau$, $C_\tau$, intersects with the equi-entropy contours, $S_\tau$, only at $v_n(\tau)$ points.

Following this observation, we state that the set between the corners of the simplex $\Delta_n$ and $C_\tau$ is a subset of the set between the corners and $S_\tau$ as presented in the following observation:

**Observation 2.** Define $S_1 = \{ p | \max p_i \geq \tau \}$ and $S_2 = \{ p | H(p) \leq \tau' \}$, $\tau' = H(v_n(\tau)), \forall \tau \in [1/n, 1]$ then $S_1 \subset S_2$. Therefore the $(S)$ region designed by entropy $S_1$ is larger than the region designed with the confidence threshold $S_2$.

It is apparent that the stopping region defined by entropy (M3) is larger than the stopping region defined through confidence level (M1) and hence enlarging the $R_S$ in [1].

This increase in the region, on the other hand, decreases confidence. To analyze it we need to find the minimum confidence in an equi-entropy contour. It is shown that the minimum confidence is attained at $w_n(\tilde{\tau})$ that satisfies the following:

**Observation 3.**

$$H(w_n(\tau)) < 1, S_\tau = \{p|H(p) = H(w_n(\tau))\}$$

$$\text{then } \min_{p \in S_\tau} \max p_i = \tilde{\tau}$$

$$\text{s.t. } -\tilde{\tau} \log_2(\tilde{\tau}) - (1 - \tilde{\tau}) \log_2(1 - \tilde{\tau}) = H(w_n(\tau))$$

This observation states that the entropy based $(S)$ boundary attains the minimum required max-probability value $\tilde{\tau}$ at $w_n(\tilde{\tau})$, $\tilde{\tau}$ vs $\tau$ and entropy values difference between $w_n(\tau)$ and $w_n(\tilde{\tau})$ are presented in Fig.2 for changing number of categories in the classification. Observe from Fig.2 that entropy is fragile with respect to the number of classes and the difference between $\tilde{\tau}$ and $\tau$ increases by decreasing values of $\tau$. Also observe that $H(w_n(\tilde{\tau})) - H(w_n(\tau))$ decreases linearly with $\tau$, and increases exponentially (dotted-line) with $n$ as presented in Fig.2(b).

By nature, uncertainty based $(S)$ is capable of returning a class label when the confidence of the class is low as shown in Obs.2. If there exists no $\tilde{\tau}$ that satisfies Obs.3 equi-entropy contours do not intersect with the simplex boundary. This condition might result in immediate stopping if one of the classes is already unfavorable as shown in Fig.3(a) with the trajectories close to $[a, c]$ edge.

The above Proposition 2, and Observations 1-3 show that under certain conditions, the regions defined by uncertainty criteria may significantly diverge from the region formed by the confidence level threshold that was defined over the posterior distribution. To avoid such drawbacks, the system might be set to a high confidence level (e.g. $\tau = .95$), such termination already mandates system continue with redundant recursions of evidence collection to achieve high confidence. Accordingly, despite their differences, uncertainty based stopping (M3) and constant confidence threshold (M1) yield similar performances in recursive classification tasks. To analyze these similarities in the classification performances, in the next section, we investigate the behavior of the posterior motion over the probability simplex. Through such an analysis, we then gain an insight into designing stopping criterion that will avoid high confidence levels and corresponding redundant recursions/evidence collections.

### 3.2 Posterior Motion

In this section we assume the posterior probability of the true state/class increases on average [21] and with such an assumption we show that there exists a particular trajectory for the posterior motion. With such an insight, we reason the similar operation behavior of constant threshold and uncertainty based stopping criteria. Moreover, this analysis will further provide us insight into relaxing stopping region $S_R$ considering the trajectory for early stopping.
The trajectory of the posterior distribution is determined by the norm projection of point \(p_{\ell c}\) given pairs of queries and evidences \((\Phi, \varepsilon)\) is by the following:

\[
p(\sigma|\Phi_{0:s}, \varepsilon_{0:s}) = p(\sigma) \otimes p(\varepsilon_1|\sigma, \phi_1) \otimes \cdots \otimes p(\varepsilon_s|\sigma, \phi_s)
\]

The trajectory of the posterior distribution is determined by the evidence likelihood \(p(\varepsilon_s|\sigma, \phi_s)\) at each step \(s\) and the evidence collected through query \(\phi_s\). In this section we assume two noisy information channels for true class and incorrect class respectively. Once the true class is queried, evidence is sampled from the “positive” distribution and evidence is sampled from “negative” distribution otherwise. In this section we show that the trajectory of the posterior distribution follows a central path as it gets closer to the corner of interest as shown in Fig. 3(b). To show this relationship we first analyze the behavior of the posterior evolution with evidence is sampled from “positive” distribution otherwise.

**Lemma 1.** Let \(a \in A\) and let \(\phi(a)\) denote the query related with state \(a\) then \(p(\sigma|p_1 = p(a), p_2, \cdots, p_n), p(\sigma), [1, 0, \cdots, 0]\) and \(p(\sigma) \otimes p(\varepsilon|\sigma, \phi(a)) \forall \varepsilon, \forall p(\sigma)\) are collinear.

Hence, once the system queries the environment, posterior probability for classification takes a step on the line that passes through the current position and the corner addressed by query. Moreover, on average, if the query addresses true state, the posterior moves towards the respective corner, if the query addresses an incorrect estimate the posterior moves away from the respective corner. To show that in general posterior gets closer to a central position, we need the projection of a posterior point to the line that passes through the center \((u_n)\) and one of the corners.

**Lemma 2.** Given \(p(\sigma) = p = [p_1, p_2, \cdots, p_n] \in \Delta_n\) and given the line \(\ell_{n,i=1} = \{[r, 1 - r, \cdots, 1 - r] | \forall r \in [0, 1]\}\) then \(\ell_2\) norm projection of point \(p(\sigma)\) onto line \(\ell_{n,i=1}\) is the following:

\[
\text{proj}_{\ell_{n,i=1}}(p(\sigma)) = \arg \min_{p \in \ell_{n,i=1}} \|p(\sigma) - p\|_2
\]

\[
= \left[ \frac{1 - p_1}{n - 1}, \cdots, \frac{1 - p_1}{n - 1} \right]
\]

To give an example, line \(\ell_{n=3, i=1}\) is visualized in Fig. 3(d) where 1st location is a. With this projection operation we show that \(\ell_2\)-norm distance between the projection and the actual point \(d_n\) at sequence \(s\) decreases quadratically with respect to the posterior as \(s\) increases. This statement is given in the following proposition:

**Proposition 3.** Following Lemma 2 given \(p(\sigma) \in \Delta_n:\)

\[
\|p(\sigma) - \text{proj}_{\ell_{n,i=1}}(p(\sigma))\|_2^2 \propto (1 - p_1)^2
\]

Following this, one defines the reduction between two sequences,

\[
\exists s \text{ s.t. } \|p(\sigma|\mathcal{H}_s) - \text{proj}_{\ell_{n,i=1}}(p(\sigma|\mathcal{H}_s))\|_2 = d_s
\]

where \(p(\sigma|\mathcal{H}_s + 1) = p(\sigma|\mathcal{H}_s) \oplus E_s(\varepsilon(\sigma, \phi)) \forall \phi\)

\[
\|p(\sigma|\mathcal{H}_s + 1) - \text{proj}_{\ell_{n,i=1}}(p(\sigma|\mathcal{H}_s + 1))\|_2 = d_{s+1}
\]

Using Lemma 7 \(\Rightarrow d_s > d_{s+1}\) \(\forall s \geq s\).

In recursive classification, on average, probability of the true state/class increases sequentially and following the proposition, the posterior probability gets closer to the line that passes through the respective corner and the center. We visualize examples of average trajectories in Fig. 3(a). Such a behavior of the posterior distribution describes why the uncertainty and confidence level based stopping criteria behave similarly for high confidence thresholds. Note that both of these methods may suffer from redundant recursions/evidence collections in such cases. On the other hand, observe that, the design of \(\mathcal{R}_S\) should be towards the corner to capture the motion to enable possible early stopping as can be seen from Fig. 3(b). Next, based on this observation, we propose a new stopping criterion perspective.

## 4 Proposed Perspective

In this section we propose another insight for \(\mathcal{R}_S\) design. In the previous section we argued equi-entropy contours formed for \((M3)\) are centered around \(u_n\). Moreover \((M3)\) is sensitive to number of categories and stagger in cases where some of the classes are already unfavorable. Additionally we presented in Figure 3(b) \((M3)\) bends \((M1)\) from the edges to provide early stopping. However, our insight from posterior motion guides us to bend \((M1)\) from the center towards \(u_n\) an example is shown in Fig. 3(b) by red boundary. Trivially, it is possible to form this by equi-distance points to the respective corners. However, by definition, edges and corners are \(\notin \Delta_n\) (represents \(\approx [19]\)) which prevents measuring the distance with conventional information theoretic approaches. To avoid this, exploiting \(\Delta_n \subset \mathbb{R}^n\), one can use a distance measure \(\delta\) defined over \(\mathbb{R}^n\) (e.g. \(\ell_p\) norms) and intersect the \(\bar{\tau}\) ball around \(\bar{p}\) \(B_{\bar{\tau}}(\bar{p}) = \{x|\delta(p, x) < \bar{\tau}\}\) with \(\Delta_n\) to obtain \(\bar{B} = B \cap \Delta_n\) centered around a corner and \(\in \Delta_n\) \(\bar{B}_{\bar{\tau}}(\bar{p}) = \{x|\delta(p, x) < \bar{\tau}, x \in \Delta_n\}\). Let \(c_k \in \Delta_n\) be the \(k\)th corner (e.g. \(c_1 = [1, 0, \cdots, 0]\)) decision region with \(\delta\) and \(\bar{\tau}\) is defined as:

\[
\mathcal{S}_R : p \in \bigcup_{k \in \{1, 2, \ldots, n\}} \bar{B}_{\bar{\tau}}(c_k)
\]

We use a distance measure influenced by Kittler’s work [22]: the delta divergence. We use the definition of \(\delta\)-
The closest point of the decision boundary to \( p, q \in \Delta_n \) is denoted as \( \delta_{\text{MP}} \) (MP indicating 'method proposed') defined as the following:

\[
\delta_{\text{MP}}(p, q) = \sum_{i \in I} |p_i - q_i|, \quad I = \{j_1, j_2, k_1, k_2\}
\]

\[
j_1 = \arg \max_i p_i, \quad j_2 = \arg \max_{i \neq j_1} p_i,
\]

\[
k_1 = \arg \max_i q_i, \quad k_2 = \arg \max_{i \neq k_1} q_i.
\]

Observe the measure satisfies non-negativity, identity of indiscernibles and symmetry. Using this distance to obtain balls, we can state the following proposition about the stopping region and criterion, \( \mathcal{S}_R \) and \( \mathcal{S}_C \) respectively;

**Proposition 4** (Proposed \( \mathcal{S}_R \) (MP)). \( \epsilon \) being \( k^{th} \) corner, \( p \in \Delta_n, \delta = \delta_{\text{MP}}, \bar{\tau} \in [1/n, 1] \) then;

\[
\mathcal{S}_R : p \in \bigcup_{k \in \{1, 2, \ldots, n\}} \bar{B}_{\bar{k}}^\delta(c^k)
\]

\[
\mathcal{S}_C : p_{j_1} - p_{j_2} > 1 - \bar{\tau}
\]

where \( j_1 = \arg \max_i p_i \) and \( j_2 = \arg \max_{i \neq j_2} p_i \)

We denote this method by (MP). We visualize an example decision boundary in Figure 3(b). As can be observed from the figure, respective decision boundary is inline with the motion of the probability distribution. To have a decision boundary bending the boundary at \( \tau \) for (M1) from its center the following condition is required:

**Observation 4.** Given \( \bar{\tau} = 2 - 2\tau \) and WLOG for \( p \in \Delta_n, \arg \max_i p_i = 1 \). Define (MP) and (M1) decision boundaries;

\[
C_\tau = \{p | p_1 = \tau\}
\]

\[
B_\tau = \{p | p_1 - p_m = 1 - \bar{\tau}, m = \arg \max_{i \neq 1} p_i\}
\]

\[
\Rightarrow C_\tau \cap B_\tau = w_n(\tau) \text{ and } \max_{p \in B_\tau} \max_i p_i = \tau
\]

**Observation 5.**

\[
B_\tau = \{p | p_1 - p_m = 1 - \bar{\tau}, m = \arg \max_{i \neq 1} p_i\}
\]

\[
\min_{p \in B_\tau} \max_i p_i = \frac{1 + (n - 1)(1 - \bar{\tau})}{n} = \psi, p = v_n(\psi)
\]

The closest point of the decision boundary to \( v_n \) is \( v_n(1 + \bar{\tau}((1/n))\) and hence on the line \( \ell_{C_{\tau}, i} \) for each respective corner \( i \). This implies, unlike uncertainty methods, proposed boundary does not interfere with the inference, \((c)\) function. By definition (5) is robust to number of categories and the cases where one of the classes is already unfavorable. We omit derivations and refer the reader to [22].

**Proposition 5.** Given \( \tau, p = [p_1, p_2, \ldots, p_n] \in \Delta_n \) s.t. \( \bar{\tau} = \arg \max_{i \neq 1} p_i \) and \( p_2 \geq (1 - p_1)\bar{\tau} \) and evidence \( \varepsilon = [\varepsilon_1, 1, \cdots, 1] \text{ where } \varepsilon_1 \sim \lognorm(\mu, \varepsilon^2) \), we define the posterior at sequence \( s \) \( \psi^s = p^0 \oplus \varepsilon \oplus \cdots \oplus \varepsilon \). With \( \bar{\tau} = 2 - 2\tau \) and \( \bar{\tau} = ((2\tau - 1)(n - 1) + 1)/n \) define;

\[
S_{\tau}(M1) = \{p^s | p_i^s \geq \tau\}
\]

\[
s'_{\tau}(M1) = \{p | p \geq \tau\}
\]

\[
S_{\tau}(RMP) = \{p^s | p_i^s - p_j^s \geq 1 - \bar{\tau}\}
\]

\[
s'_{\tau}(RMP) = \{p^s | p_i^s - p_j^s \geq 1 - \bar{\tau}, j = \arg \max p_k^s \Rightarrow
\]

\[
p(p^s \in S_{\tau}(RMP)(p_1, p_2, \bar{\tau}, s)) \geq p(p^s \in S_{\tau}(RMP)(p_1, \tau, s)) \&
\]

\[
p(p^s \in s'_{\tau}(RMP)(p_1, p_2, \bar{\tau}, s)) \geq p(p^s \in s'_{\tau}(RMP)(p_1, \bar{\tau}, s)) \forall s \in \{1, 2, \cdots\}
\]

**Proposition 5** demonstrates that compared to a stopping criterion based on posterior distribution thresholding (M1), the proposed method (MP) always has higher probability of entering the stopping region with correct decision, while the probability of entering an incorrect region resulting in incorrect decision is constrained within certain probability values. Therefore, a system can be designed to achieve a desired true positive probability while limiting false alarm probability by using the proposed stopping criterion. Supporting numerical examples are in the appendix Sec 3.1.

5 EXPERIMENTS AND RESULTS

In this section we run experiments to support our findings. Throughout this section we denote the proposed method introduced in Section 4 by (MP) and compare it with the methods presented in Table 1. Specifically, we designate a confidence level \( \tau \) for (M1). We select respective \( c \) for (M2:4) such that (M1:4) intersect at \( v_n(\tau) \). We select \( \tau \) for (MP) following Obs 4. For (M5) \( c_{\text{MP}} \) is selected as \( 10^{-2} \). We demonstrate results for the two following cases: (i) Synthetic experiments (ii) A letter decision for electroencephalography (EEG)-based brain computer interface (BCI) typing system.

5.1 Synthetic Experiments

In this section, we present synthetic toy examples to support our propositions in Section 4 that states: proposed method (MP) is robust to prior distributions with disfavored classes and (MP) stops earlier by sacrificing marginal accuracy as it is coherent with the motion of the posterior.

1) In the presence of disfavored classes, where the recursive classification is happening in a lower dimension than the cardinality of the class space (e.g. at least one of the class probabilities \( \approx 0 \)), uncertainty based methods (especially Renyi Entropy with orders \( \alpha \leq 1 \)) suffer from immediate or rushed stopping (this was discussed in Section 3.1).

2) As the cardinality of the class space increases the case described in 1 becomes more drastic.

3) Proposed method (MP) in the case of 1 and 2, behaves similar to the confidence thresholding methods. Moreover, (MP) provides an early stopping with marginal accuracy loss when the trajectory follows a central path. (this was discussed in Section 3.2).
In addition to the methods presented in the main paper, we also use a confidence lower bound with a confidence level \(\bar{\tau}\) that is derived in Observation 5 (e.g., \(\bar{\tau}\) for which the minimum accuracy can be achieved as (MP)) and we represent this method with (M1). To be precise, for (M1) and (MP) presented in Fig. 4(b), (M1) is the confidence lines each intersect with (MP) at the respective peak points. By design, we expect to show that (MP) is as robust as (M1) and (MP) reaches the speed of (M1) with marginal accuracy decrease. In our experiments we present scenarios with pre-defined prior information and evidence \(\varepsilon = [\varepsilon_+, \varepsilon_-]\) such that \(\varepsilon_+\) and \(\varepsilon_-\) are sampled from lognormal distributions (specifications are listed under each Table 2, 3). In these tables, we report two measures for each method given number of sequences such that \(p_{\text{stop}}\) represents the probability of stopping, \(p_{\text{true stop}}\) represents the accuracy among all terminations (e.g. \(p_{\text{true stop}} = 0.5\), \(p_{\text{stop}} = 0.6\) means the system correctly selected 300 out of 600 terminated in 1000 Monte Carlo simulations). We highlight the first sequence that achieves \(p_{\text{stop}} \geq 0.5\) indicating a reasonable stopping operational point of a method. Table 2 and Table 3 collectively explain items (1)-(2) such that, in both these tables, it is apparent that (M3) - (M4) perform poorly due to their behavior with disfavored classes. Priors with disfavored classes lead to a trajectory that follows a path closer to sides of the simplex and hence result in rushed and immediate stopping yielding probability of stopping \(\approx 1\) even in the first few sequences with extremely low accuracy. Observe that (M3) characteristics gets closer to (M4) in Table 3 compared to Table 2, because of an increase in class space cardinality (as also discussed in the manuscript Figure 2(a)). Table 4 demonstrates a case where the prior is more centric and hence the posterior follows a more central path. Analogous with the previous discussions, uncertainty based methods and confidence thresholding behave similar. In such scenarios (MP) achieves similar accuracy performance compared to competitors but stops 1 sequence earlier.

To summarize the findings, we visualize time accuracy trade-off for each method for the cases presented in Tables 5 and 6.
We first filter the EEG signal to remove drifts and artifact-related high frequency components with a band pass filter for [1,50]Hz. After filtering, EEG is windowed to extract the signal of interest components [26], [27].

In EEG-BCIs the primary interest of filtering is to extract the signal of interest components [26], [27]. The system relies on reducing EEG time series into one dimensional feature vector. Filtered multi-channel EEG data time windows are passed through channel-wise principal component analysis where the outputs are concatenated to an intermediate feature vector. We assume in each class, feature vectors are drawn from a multivariate Gaussian distribution and hence Regularized discriminant analysis (regularized quadratic discriminant analysis [29]) is a plausible choice that results in one dimensional representation of the signal. Each positive and negative sample in the calibration EEG data is reduced to a single dimensional feature and positive and negative feature distributions are learned accordingly.

**Experimental Task:** For our experiments we use these distributions in Copy-Letter task such that each user’s data is used to type a target letter within a pre-determined phrase for multiple phrases [29]. More specifically, for each letter typing scenario the user is tasked to respond to the system, and the decision is made when the cumulative evidence matches the correct letter after multiple recursions. Therefore, in such a setting, the decision chance level is 0.03%. To make a decision the system recursively queries the user with multiple letter flashes. We designate the number of queries to be presented at each sequence to be $N \in \{15, 10, 5\}$. We present the results for two conditions. First one is a typing scenario with no language model (uniform prior information). And in the second scenario there exists a language model for the requested typing and the candidate letter is in top 16. The choice of two scenarios illustrated in the numerical results is to represent the following: (i) Class priors are uniform (in the 28-vertex simplex, RBC starts from the center of the simplex labeled as $u$ in Figure 1-(d). (ii) The prior probability for the correct class label (desired letter of the user) is selected to be significantly lower than in the first case (in the 28-vertex simplex, RBC starts further away from the target vertex). These cases represent typical situations with uninformative prior and adversarial prior, additionally these are the challenging cases for RBC.

**5.2 Real Data Experiments**

**5.2.1 Experimental Details**

In our experiments we use a BCI typing system called RSVP Keyboard presented in Orhan’s work [23] and the implementation BCIPy [24]. The system is visualized for the stimulus screen and an actual healthy participant performing a task in Fig. 5.

The system visually stimulates series of letters rapidly to the user. Once intended symbol appears on the computer screen, subject’s recorded brain signal evokes a distinguishable response [25]. The system utilizes EEG to record brain signals during presentation where p300 evoked potential presence corresponds to a positive response and absence corresponds to a negative response. The user has an intended letter in mind which indicates the class label among other candidates in the alphabet. The system recursively collects noisy EEG to increase the confidence before making a decision. Subjects approximately allocate [10,30] seconds for typing a single letter and hence an increase in classification speed is of a significant value in typing sentences.

**Data Collection:** Ten healthy participants (six females), 20-35 years old were recruited under IRB-130107 protocol approved by Northeastern University. A DSI-24 Wearable Sensing EEG Headset was used for data acquisition, at a sampling rate of 300 Hz with active dry electrodes. All participants performed the calibration session containing 100 sequences; each sequence includes 5 trials; and one trial in each sequence is the target symbol which is displayed on the screen prior to each sequence (RSVP paradigm). A sequence contains randomly ordered ten symbols with a pre-defined target symbol. EEG is acquired from 16 channels using the International 1020 configuration (Fp1, Fp2,F3, F4, Fz, Fc1, Fc2, Cz, P1, P2,C1, C2, Cp3, Cp4, P5, P6). Recorded EEG are used to learn class conditional EEG evidence distributions.

**Pre-Processing:** In EEG-BCIs the primary interest of filtering is to extract the signal of interest components [26], [27]. Therefore, dimensionality reduction using ICA or PCA is also needed [26]. The system categorizes the users based on their calibration performances. The measure of performance is the area under receiver operating characteristics curve (AUC) of classification based on the features extracted during calibration. We specifically selected user data with AUC performances $\{0.67, 0.72, 0.76, 0.81, 0.84, 0.87\}$ to have a spectrum of ranging performances.

We visualize our findings in Figure 6 and 7 respectively. It is observed that (MP) switches the operation point to a location such that faster results are obtained with the cost of small amount of decreases in accuracy. To show the significance of the (MP), we present here a scenario that includes correctly typing 100 letters on a computer screen. We refer the reader to appendix Sec.7.3.2 for the complete results that allowed us to generate visualizations.

Comparing with the conventional method (M1), when uniform prior is used, (MP) outperforms (M1) in terms of

---

**Fig. 5:** EEG driven Rapid Serial Visual Presentation (RSVP) keyboard typing interface. (a) The stimuli is flashed in the middle of the screen while the user is informed with the text above. (b) The user is conducting copying the phrase task (multiple copy letter tasks). The user is informed about the required phrase. EEG is collected on top of the scalp non-invasively.
Fig. 6: Number of recursion spent and accuracy plots for recursive classification in BCI typing system. Each scenario is generated using human-in-the-loop calibration data trained generative models. In each figure results are presented in ascending order of performance measures (area under receiver operation characteristics curve (AUC)). Number of queries in each recursion from top to bottom, 5, 10, 15 respectively. Legend covers methods from left to right and dots on the figures represent respective accuracy values. The users tried to type “A” without any language model (uniform prior information). Top to bottom legend order is from left to right for each block.

speed, i.e., (MP) and (M1) complete the same task with 1735 and 1945 sequences respectively. Accordingly (MP) saves 210 sequences. This corresponds to saving \(3(m)30(s)/32(m)\) lifetime during typing. Additionally, if a language model prior is used for the same task, (MP) still outperforms (M1), 1580 sequences vs 1728 sequences saving \(2(m)24(s)/28(m)48(s)\) lifetime during typing. These reported amount of time are computed under the condition that at the end of the task 100 letters are completely correctly typed including corrections of the wrongly typed letters (i.e., email is completely correct). Saving time is very crucial for practical BCI typing as these systems are designed for individuals with limited speech and physical abilities. Therefore, fatigue and discomfort caused by the BCI system are important factors that significantly affect the BCI typing performance, and limiting the time to complete the tasks accurately will significantly improve practicality.

6 Conclusion

In this paper, our focus was on the analysis and design of stopping criterion for recursive Bayesian classification. Stopping criterion based on thresholding posterior distribution may result in redundant recursions/evidence collection. To overcome these shortcomings, uncertainty based methods were proposed. Through a geometrical representation of the posterior probability progression in a probability simplex, we demonstrated that such uncertainty methods increase the stopping region to decrease the redundancy of the posterior probability thresholding, but they are sensitive to the number of classes and the increase in the stopping region is not inline with the posterior probability progression resulting in significant decrease in classification accuracy when the stopping region is enlarged. Accordingly, we proposed a new method to overcome the limitation of the existing uncertainty methods and showed that through such a method true positive probability is always larger than the true positives obtained through a stopping criterion based on thresholding posterior probability. We also showed analytically that under certain conditions true positive probability of the proposed method can be designed to be above the true positive of the method that depends on posterior thresholding while the the false alarm probability of the proposed method remains within certain range. We also validated the proposed method using a real-case use on a brain computer interfaced typing system. This work can be extended by considering true positive maximization and false negative minimization as a multi-objective optimization.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley & Sons, 2012.
[2] H. L. Van Trees and K. L. Bell, Detection estimation and modulation theory. Wiley, 2013.
[3] S. Särkkä, Bayesian filtering and smoothing. Cambridge University Press, 2013, vol. 3.
[4] A. M. Alaa and M. Van Der Schaar, “Balancing suspense and surprise: Timely decision making with endogenous information acquisition,” in Advances in Neural Information Processing Systems, 2016, pp. 2910–2918.
[5] A. N. Shiryaev, Optimal stopping rules. Springer Science & Business Media, 2007, vol. 8.
[6] S. K. Jha, E. M. Clarke, C. J. Langmead, A. Legay, A. Platzer, and P. Zuliani, “A bayesian approach to model checking biological systems,” in International Conference on Computational Methods in Systems Biology. Springer, 2009, pp. 218–234.

[7] I. Sason and S. Verdú, “Arimoto-rényi conditional entropy and its application to hypothesis testing,” IEEE Transactions on Information Theory, vol. 64, no. 1, pp. 4–25, 2017.

[8] S. Geisser, Predictive inference. Routledge, 2017.

[9] D. Golovin, A. Krause, and D. Ray, “Near-optimal bayesian active learning with noisy observations,” in Advances in Neural Information Processing Systems, 2010, pp. 766–774.

[10] T. L. Lai, “On optimal stopping problems in sequential hypothesis testing,” Statistica Sinica, vol. 7, no. 1, pp. 33–51, 1997.

[11] ——, “Sequential analysis: some classical problems and new challenges,” Statistica Sinica, pp. 303–351, 2001.

[12] C. E. Shannon, “A mathematical theory of communication,” Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.

[13] Y. Li and R. E. Turner, “Rényi divergence variational inference,” in Advances in Neural Information Processing Systems, 2016, pp. 1073–1081.

[14] D. Weinshall, H. Hermansky, A. Zweig, J. Luo, H. Jimison, F. Ohl, and M. Pavel, “Beyond novelty detection: Incongruent events, when general and specific classifiers disagree,” in Advances in Neural Information Processing Systems, 2009, pp. 1745–1752.

[15] D. S. Pavlíchin and T. Weissman, “Chained kullback-leibler divergences,” in 2016 IEEE International Symposium on Information Theory (ISIT). IEEE, 2016, pp. 580–584.

[16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with bregman divergences,” Journal of Machine Learning Research, vol. 6, no. Oct, pp. 1705–1749, 2005.

[17] V. Pavlowsky-Glahn and J. J. Egozcue, “Geometric approach to statistical analysis on the simplex,” Stochastic Environmental Research and Risk Assessment, vol. 15, no. 5, pp. 384–398, 2001.

[18] S.-i. Amari and H. Nagaoka, Methods of information geometry. American Mathematical Soc., 2007, vol. 191.

[19] J. Aitchison, “The statistical analysis of compositional data,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 44, no. 2, pp. 139–160, 1982.

[20] S.-W. Ho and S. Verdú, “On the interplay between conditional entropy and error probability,” IEEE Transactions on Information Theory, vol. 56, no. 12, pp. 5930–5942, 2010.

[21] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero, “Collaborative 20 questions for target localization,” IEEE Transactions on Information Theory, vol. 60, no. 4, pp. 2233–2252, 2014.

[22] J. Kittler and C. Zor, “Delta divergence: A novel decision cognizant measure of classifier incongruence,” IEEE transactions on cybernetics, vol. 49, no. 6, pp. 2331–2343, 2018.

[23] U. Orhan, K. E. Hild, D. Ergodmus, B. Roark, B. Oken, and M. Fried-Oken, “Rsvp keyboard: An eeg based typing interface,” in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 645–648.

[24] T. Mémont, A. Koçanaoğlu, M. Lawhead, D. Klee, S. Dudy, M. Fried-Oken, and B. Oken, “Bcipy: Brain-computer interface software in python,” arXiv preprint arXiv:2002.06642, 2020.

[25] G. F. Woodman, “A brief introduction to the use of event-related potentials in studies of perception and attention,” Attention, Perception, & Psychophysics, vol. 72, no. 8, pp. 2031–2046, 2010.

[26] M. Akcakaya, B. Peters, M. Moghadamfalahi, A. Moorey, U. Orhan, B. Oken, D. Ergodmus, and M. Fried-Oken, “Noninvasive brain computer interfaces for augmentative and alternative communication,” IEEE Reviews in Biomedical Engineering, vol. 7, no. 1, pp. 31–49, 2014.

[27] M. Moghadamfalahi, U. Orhan, M. Akcakaya, H. Nezamfar, M. Fried-Oken, and D. Ergodmus, “Language-model assisted brain computer interface for typing: A comparison of matrix and rapid serial visual presentation,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 23, no. 5, pp. 910–920, 2015.

[28] J. H. Friedman, “Regularized discriminant analysis,” Journal of the American statistical association, vol. 84, no. 405, pp. 165–175, 1989.

[29] U. Orhan, H. Nezamfar, M. Akcakaya, D. Ergodmus, M. Higiger, M. Moghadamfalahi, A. Fowler, B. Roark, B. Oken, and M. Fried-Oken, “Probabilistic simulation framework for eeg-based bci design,” Brain-Computer Interfaces, vol. 3, no. 4, pp. 171–185, 2016.

[30] C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn, “Mathematical foundations of compositional data analysis,” in Proceedings of IAMG, vol. 1, 2001, pp. 1–20.

[31] A. Rényi et al., “On measures of entropy and information,” in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California, 1961.
Addition: Given \( p, q \in \Delta_n \)
\[
p \oplus q = \left[ \frac{p_1 q_1, p_2 q_2, \ldots, p_n q_n}{\sum_i p_i} \right]
\]

Given \( p(\sigma) \in \Delta_A \) where \( A \) is the state space and
\[
p(\varepsilon|\sigma, \phi) = [p(\varepsilon|\sigma_1, \phi), p(\varepsilon|\sigma_2, \phi), \ldots, p(\varepsilon|\sigma_A, \phi)]
\]
and hence one cannot rigorously define \( p(\sigma) \oplus p(\varepsilon|\sigma, \phi) \). To be able to do such \( p(\varepsilon|\sigma, \phi) \) should be unit \( \ell_1 \) norm wrt. \( \sigma \) and hence we require \( p(\varepsilon|\sigma, \phi)/\sum_{\sigma} p(\varepsilon|\sigma, \phi) \) which is not practical. However the closure operator (normalization as defined in [19]) that maps an arbitrary point to \( \Delta_n \) allows algebraically \( p(\sigma) \oplus p(\varepsilon|\sigma, \phi) \) as the following;
\[
p(\sigma) \oplus \frac{p(\varepsilon|\sigma, \phi)}{\sum_{\sigma} p(\varepsilon|\sigma, \phi)} = \frac{[p(\sigma_1)p(\varepsilon|\sigma_1, \phi)]}{\sum_{\sigma} p(\varepsilon|\sigma, \phi)} \sum_{\sigma} p(\varepsilon|\sigma, \phi)
\]
\[
= \frac{[p(\sigma_1)p(\varepsilon|\sigma_1, \phi)]_{\sum_{\sigma} p(\varepsilon|\sigma, \phi)} \simeq p(\sigma) \oplus p(\varepsilon|\sigma, \phi)}
\]

Following this equation we represent a posterior with the following update; \( p(\sigma|\varepsilon, \phi) = p(\sigma) \oplus p(\varepsilon|\sigma, \phi) \) which is analogous with posterior=likelihood.

7.2 Uncertainty Decision Boundaries / The reason behind analytically analyzing only (M3)

In this section we reason the decision on only analyzing Shannon’s entropy in Section 5. In information theory, Renyi entropy generalizes Shannon entropy [31]. The definition for Renyi entropy, parameterized over \( \alpha \) is the following;
\[
p \in \Delta_n, \ H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha \tag{7}
\]

Observe that the limit case \( \lim_{\alpha \to 1} H_\alpha(p) = H(p) \) results in Shannon entropy measure. In this paper we only propose analytical derivations for Shannon entropy as a special case. However the findings can directly be applied to Renyi measures. The generalization can be analytically shown, but to have a neat presentation we omit the derivations here. However, in Fig. 8 we present decision boundaries for Renyi entropy and Shannon as a special case to demonstrate their similar decision geometry.

![Fig. 8: Decision boundaries formed using \( S_2 \) for Renyi using \( \alpha \in \{2, 1.5, 0.5\} \) and Shannon entropy \( (2, 1.5, \text{Shannon}, 0.5 \text{ in the outer to inner order in the figures}) \). Confidence lines are plotted for reference. Corresponding values for confidence lines are \( \tau = 0.8 \) for left and \( \tau = 0.65 \) for right figure.](image)

7.3 Proposed Method Supplementary

7.3.1 True Positive - False Alarm Guarantee

In this section we present true selection and error probabilities for simple examples that can be visualized on a three class simplex \( \Delta_3 \). We present these result to support our claims in Section 4 of the manuscript. We pick several points on the simplex and with a predefined evidence distribution sampling from a lognormal, we visualize the bounds on error in Figure 9 below. We plot the analytical bounds derived in the proof of Proposition 5 using lognormal distribution assumptions. Lower bound is represented with (M1)(red) and the upper bound is represented with (M1) (black). Instead of plotting the analytic values of (MP) we plot the average probability values calculated over 5000 Monte Carlo simulations. We compute \( p(p_s \in S_M) \) and \( p(p_s \in S_M') \) values for different starting points that are color coded in the figure. To generate the figures we use \( \varepsilon = [\varepsilon_+, 1, \ldots, 1] \) where \( \varepsilon_+ \sim \text{lognorm}(0.8, 0.6^2) \) and the true class is \( a \). Given this evidence model, all posteriors follow a straight path to corner \( a \) (this behavior is discussed in Lemma 7 in the manuscript).

![Fig. 9: In this figure we compare the probability of correct selection (reaching the correct stopping region) by \# of sequences and incorrect selection in \( \Delta_3 \) of (MP) (blue) with (M1) (red) and (M1) (black). 1st column represents the prior points of the recursive classification with different colors in \( \Delta_3 \). 2nd column represents the probability of posterior at sequence \( s \) lying on the correct region for stopping. 3rd column represents the probability of incorrect stopping by sequences. Observe that as described in Proposition 5 probability of correct decision is always above (M1) and probability of error is sandwiched with lower (M1) and upper (M1) curves.](image)
times. We visualize our findings for a range of parameters in Figure 10. As expected, as we increase the standard deviation (c) of the exponentiated Gaussian distribution, the system is more likely to make errors and less likely to decide correctly. On the other hand as mean (µ) increases, the system yields more accurate decisions and less error.

Fig. 10: In this figure we visualize the effects of the distribution variance and mean in recursive classification for the represented points in the figure. We consider c+ ~ lognorm(µ, c2) and calculate the probabilities for a set sequence number s = 5. We represent 4 figures to indicate the effects (on the right side of the Δs simplex with corners a, b and c). Top row represents the effects of the standard deviation of the exponentiated Gaussian distribution on probability of correct selection and incorrect selection for left and right respectively where µ = 0.8. Whereas bottom row visualizes the effects of the mean where c = 0.6. It is observed as mean increases (since the evidence gets higher values) probability of correct selection increases where error chance decreases. On the other hand results are the opposite as expected for the standard deviation increments.

Through the Figures 9 and 10, we summarize our claims that are presented in Section 4 especially about the bounds on error and correct selection probability given in Proposition 5.

7.3.2 BCI Typing Supplementary

In experiments section (Sec. 5), we report time required for a subject to successfully complete a 100 letter typing scenario only for the conventional method (M1). We use the following computation to report required number of sequences to complete the task;

\[
\text{acc, seq from user data: rem } = 100, \text{ #seq } = 0
\]

while rem > 0 do:
\[
\text{rem } \leftarrow \text{rem } - \text{ceil(rem } \times \text{acc/100)}
\]
\[
\text{#seq } \leftarrow \text{#seq } + \text{rem } \times \text{seq}
\]
end

return #seq

In Table 5, we report the performance values for all the methods that are compared to (MP). We specifically report the number sequences required for each user to successfully complete the task in addition to the number of sequences averaged across all participants. Moreover, we report average typing accuracy and sequence required to type 1 letter (correct or incorrect) for each method averaged across users. It is apparent (MP) allows faster typing as reported in the experiments by an accuracy loss of 5% on average. (M4) seems to outperform (MP) in the uniform case. However, (M4) results in frequent incorrect decisions that require corrections which irritates the target population to use such BCI systems. Therefore, such methods similar to (M4) may overall perform worse due to fatigue effects.

| LM | Perf | Seq | Sequence Difference from # (MP) |
|----|------|-----|--------------------------------|
| (M1) | (M2) | (M3) | (M4) | (M5) |
| 67 | 3628 | +1257 | +106 | +143 | -183 | +3687 |
| 72 | 2290 | +295 | +279 | +142 | -130 | +5200 |
| 76 | 1627 | +197 | +186 | +85 | -59 | +3391 |
| 81 | 1180 | +186 | +173 | +93 | +18 | +2524 |
| 84 | 947 | +129 | +124 | +677 | +14 | +2173 |
| 87 | 739 | +126 | +122 | +71 | +35 | +1921 |
| avg | 17.35 | +211 | +196 | +101 | -30 | +3482 |

E(seq) = 15.44, 18.07, 17.96, 16.05, 10.70, 51.95

Table 5: Supplementary for BCI experiments Fig. 6, Fig. 7. For all experiments the confidence threshold is set to τ = 0.85 for (M1) and the constants for the other objectives are calculated accordingly. We report number of sequences spent for (MP) for different performing users and corresponding differences with other methods. avg represents the average sequence values for each method conditioned on LM respectively. acc and E(seq) represent accuracy in typing and average number of sequences spent for 1 letter respectively. For the language model case, based on prior location for “O” after “IT”, the difference between (M4) vanishes.

7.4 Proofs

Proof Proposition 7. Addition: Given \( p, q \in \Delta_n \), the addition operation is defined as the following:

\[
p \oplus q = \left[ \frac{p_1 q_1}{\sum p_i q_i}, \frac{p_2 q_2}{\sum p_i q_i}, \ldots, \frac{p_n q_n}{\sum p_i q_i} \right]
\]

Multiplication with scalar: Given \( p \in \Delta_n, \lambda \in \mathbb{R} \) the multiplication with scalar is defined as the following:

\[
p \otimes \lambda = \left[ \frac{p_1^\lambda, p_2^\lambda, \ldots, p_n^\lambda}{\sum p_i^\lambda} \right]
\]

Let \( p, q, r \in \Delta_n; p \oplus q = \left[ \left[ p \oplus q \right]_{k \in \{1, 2, \ldots, n\}} \right] \in \mathbb{R}^n \) and \( p \otimes q = \left[ \left[ p \otimes q \right]_{k \in \{1, 2, \ldots, n\}} \right] \in \mathbb{R}^n \).

Additive Closure: \( p_k > 0, q_k > 0 \implies p \oplus q \in \Delta_n \).

Scalar Closure: \( p_k > 0, \lambda \in \mathbb{R} \implies p_k^\lambda > 0 \implies [p \otimes \lambda]_k > 0 \).

Commutativity: scalar multiplication is commutative \( \implies p \otimes \lambda = \lambda \otimes p \).

Additive Associativity: scalar multiplication is associative \( \implies (p \otimes q) \otimes r = p \otimes (q \otimes r) \).

Zero Element: \( u_n \in \Delta_n \) as shown above.

Multiplicative 1: \( [p \otimes 1]_k = p_k / \sum_i p_k = p_k \implies p \otimes 1 = p \).

The following properties can be shown using the definitions above easily and hence omitted; additive inverses, scalar multiplication associativity, distributivity across vector addition, distributivity across scalar addition.
**Proof Proposition** \[2\] \[S_\tau = \{ p | \max_i p_i = \tau \} \implies \max_{p \in S_\tau} H(p) = H(v_n(\tau))\]

Let \( p \in S_\tau \). WLOG pick \( p_1 = \tau \) where \( |p| = n \);

\[
\begin{align*}
\max_p H(p) &= \max_{p \in \mathcal{P}} -p_1 \log(p_1) - \sum_{i \in \{2, \ldots, n\}} p_i \log(p_i) \\
&\quad \text{from principle of maximum entropy} \\
&= -\tau \log(\tau) - \sum_{i \in \{2, \ldots, n\}} \frac{1}{n-1} \log \left( \frac{1 - \tau}{n-1} \right) \\
&= -\tau \log(\tau) - (1 - \tau) \log \left( \frac{1 - \tau}{n-1} \right) \\
&= H(v_n(\tau))
\end{align*}
\]

\( S_\tau = \{ p | H(p) = H(v_n(\tau)) \} \implies \max_{i \in C_\tau} p_i = \tau \)

Let \( p \in C_\tau \). WLOG pick \( i = 1, \ p \in C_\tau \implies p_1 \log(p_1) = -H(v_n(\tau)) - \sum_{i \in \{2, \ldots, n\}} p_i \log(p_i) \); \[
\frac{d \log(x)}{dx} \bigg|_{x \in (1/e, \infty)} > 0
\]

\[
\implies \max_x = \arg \max_{x \in (1/e, \infty)} \log(x) \implies \max_{p \in C_\tau} \text{ max } p_1 = \arg_{p \in C_\tau} \max_{p \in C_\tau} p_1 \log(p_1) \]

\[
= \arg_{p \in C_\tau} -H(v_n(\tau)) - \sum_{i \in \{2, \ldots, n\}} p_i \log(p_i) \]

\[
= \arg_{p \in C_\tau} -H(v_n(\tau)) - \sum_{i \in \{2, \ldots, n\}} p_i \log(p_i) \]

\[
\text{where}
\]

\[
\sum_{i \in \{2, \ldots, n\}} p_i = 1 - p_1 \]

Using the identity \( \hat{x} = \arg \max_x -\log(x)^T x = [1/e, \ldots, 1/e]^T \) then project to \( (1 - p_1) \)-unit \( \ell_1 \) norm ball again from principle of maximum entropy we observe \( p_{1:n} = \left[ \frac{1 - p_1}{n-1}, \ldots, \frac{1 - p_1}{n-1} \right] \implies p = [p_1, \frac{1 - p_1}{n-1}, \ldots, \frac{1 - p_1}{n-1}] = v_n(p_1) \).

Since \( H(p) = H(v_n(p_1)) = H(v_n(\tau)) \), \( p_1 = \tau \).

**Proof Observation** \[1\] Derivation is trivial given Proposition \[2\].

**Proof Observation** \[2\] Let \( p \in \Delta_n \). WLOG \( p_1 = \max_i p_i \);

\[
H(p) = -\sum_i p_i \log(p_i) - \sum_{i \in \{2, \ldots, n\}} p_i \log(p_i) \\
\leq -p_1 \log(p_1) - (1 - p_1) \log \left( \frac{1 - p_1}{n-1} \right) \tag{8}
\]

\[
-H(p) \geq p_1 \log(p_1) + (1 - p_1) \log \left( \frac{1 - p_1}{n-1} \right) \\
\geq \tau \log(\tau) - (1 - \tau) \log \left( \frac{1 - \tau}{n-1} \right) \tag{9}
\]

Therefore, \( \forall p_1 \geq \tau \geq 1/n \implies H(p) \leq \tau' \). Hence \( \forall p \in S_1 \implies p \in S_2 \) which concludes \( S_1 \subseteq S_2 \).

Moreover, given \( n = 5 \) where \( \tau' = 2.16 \), therefore corresponding \( \tau = 0.5 \). Pick a distribution with probabilities \( p = [4, 2, 2, 0, 0] \). Obviously, \( p \notin S_1 \). However, \( H(p) = 1.5219 < \tau' \) therefore \( p \in S_2 \) which shows \( S_2 \not\subseteq S_1 \). Bu the counterexample \( S_1 \subseteq S_2 \) reduces to \( S_1 \subseteq S_2 \).

**Proof Observation** \[3\] The edge case for equi-entropy contours intersecting with the edges of the probability domain is WLOG \( p = [0.5, 0.5, 0, \ldots, 0] \) hence \( H(p) = 1 \implies \) the equi-entropy contour \( S_\tau \) intersects with the borders of the simplex. As explained over truncated distributions in \[20\], we use the confidence line definition \( C_\tau = \{ p | \max_i p_i = \tau \} \) and use the identity \( \arg \max_{p \in C_\tau} H(p) = v_n(\tau) \). Equivalently, for \( v_n(\tau) \) to hold the entropy condition, one can write;

\[
-\tau \log(\tau) - (1 - \tau) \log(1 - \tau) = H(v_n(\tau))
\]

**Proof Lemma** \[2\] Pick \( a_1, a_2, a_n \in A \) being \( 1^{\text{st}}, 2^{\text{nd}}, n^{\text{th}} \) class respectively and \( a_1 \) being queried \( \phi(a_1) \) with initial probability distribution over the entire state space \( p(\sigma) = [p_{a_1}, \ldots, p_{a_n}] \) yielding the evidence \( \varepsilon \). Observe the following relation between the posterior and prior using the label assignment for the queried candidate \( \ell \) where for a query \( \phi(a_\ell) \), and a candidate \( a', a = a_\ell \implies \ell = 1 \);

\[
p(\sigma|\varepsilon, \phi(a_1)) = p(\sigma) \oplus p(\varepsilon|\sigma, \phi(a_1))
\]

\[
= p(\sigma) \oplus [p(\varepsilon|a_1, \phi(a_1)), p(\varepsilon|a_2, \phi(a_1)), \ldots, p(\varepsilon|a_n, \phi(a_1))]
\]

\[
= p(\sigma|\varepsilon = 1, p(\varepsilon|\ell = 0, \varepsilon|\ell = 0)]
\]

The following vectors map to the same point on simplex;

\[
[p(\varepsilon|\ell = 1), p(\varepsilon|\ell = 0, \varepsilon|\ell = 0)] \sim [p(\varepsilon|\ell = 1), 1,
\]

\[
\kappa \in [p(\varepsilon|\ell = 1), p(\varepsilon|\ell = 0)]
\]

First we write the posterior using the previous equation;

\[
p(\sigma|\varepsilon, \phi(a_1)) = \left[ \frac{p_1 \times k}{m}, \frac{p_2}{m}, \ldots, \frac{p_n}{m} \right] \in \Delta_N
\]

where \( m = p_1 \times (k - 1) + 1 \)

Trivially, \( p(\sigma) \) and \( [1, 0, \ldots, 1] \) form a line. To show collinearity, we show \( p(\sigma|\varepsilon, \phi(a_1)) \) also lies on that line, in other the point should satisfy the line equation;

\[
(x_1 - 1) / (p_1 - 1) = x_2 / p_2 = \ldots = x_n / p_n
\]

If we insert \( x_1 = p_1 \times k/m \) and respective remaining probabilities;

\[
(p_1 \times k/m - 1) / (p_1 - 1) = p_2/m \times p_1 = \ldots = p_n/m \times p_1 = 1 / m
\]
We show collinearity with the following algebraic manipulation:

\[ m = p_1 (k - 1) + 1 \implies \frac{(p_1 \times k)/m - 1}{p_1 - 1} = \frac{p_1 \times k - p_1 \times k + p_1 - 1}{(p_1 - 1) \times m} = \frac{1}{m} \]
Therefore using the ball definition for $\delta = \delta_{mp}$ and $q = [1, 0, \cdots, 0] \Rightarrow B_{\tilde{\delta}}(q) = \{x|\delta(q, x) < \tilde{\tau}, x \in \Delta_n\} \Rightarrow B_{\tilde{\tau}}(u) = \{x|x_j - x_k > 1 - \tilde{\tau}, j = \arg\max_i x_i, k = \arg\max_{i \neq j} x_i, x \in \Delta_n\}

**Proof Observation 4** Define the following two sets:

\[ C_\tau = \{p|\arg\max_i p_i = 1, \max_i p_i = \tau\} \]

\[ B_\tau = \{p|p_1 - p_m = 1 - \tilde{\tau}, 1 = \arg\max_i p_i, m = \max_{i \neq 1} p_i\} \]

Observe, $\max_{p \in B_\tau} p_i = p_1$. We seek the following:

\[ \max_{p \in B_\tau} 1 \quad \text{s.t.} \quad 1 = \arg\max_i p_i \]

\[ \Rightarrow \max_{p \in B_\tau} 1 - \tilde{\tau} + p_m \quad \text{s.t.} \quad 1 = \arg\max_i p_i, m = \max_{i \neq 1} p_i \]

\[ \Rightarrow \max_{p \in B_\tau} m \quad \text{s.t.} \quad 1 = \arg\max_i p_i, m = \max_{i \neq 1} p_i \]

Trivially, $\max_{p \in B_\tau} m = 1 - p_1$ $\Rightarrow$ $p_1 = 1 - \tilde{\tau} + 1 - p_1 \rightarrow p_1 = 1 - \tilde{\tau}/2$. Observe that $p_i = 0 \forall i \neq 1, m$ and hence $p = w_n(1 - \tilde{\tau}/2)$. If $\tilde{\tau} = 2\tau - 2$ then $p = w_n(\tau)$ and $\arg\max_{p \in B_{\tilde{\tau}}} p_1 = w_n(\tau)$.

Observe $w_n(\tau) \in C_\tau, w_n(\tau) \in B_{\tilde{\tau}}$. and by the equality condition in both sets, $w_n(\tau) = (C_\tau \cap B_{\tilde{\tau}})$. □

**Proof Observation 5** Define the following:

\[ B_\tau = \{p|p_1 - p_m = 1 - \tilde{\tau}, 1 = \arg\max_i p_i, m = \max_{i \neq 1} p_i\} \]

By definition, $\min_{p \in B_\tau} p_i = p_1, p \in B_\tau$ $\Rightarrow p_1 - p_m = 1 - \tilde{\tau}$. Following a similar procedure as described in proof for Observation 4, we follow the stopping region $s.t. (S): p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S$. WLOG, let $i^{th}$ location belong to the target class. Given $p(\varepsilon_{s} \mid \sigma) \propto [\varepsilon, 1, \cdots, 1]$ then the following relations hold:

1. $S_{RM}(M1) = \{p|p_1 \geq \tau, p \in \Delta_n\}$

\[ \Rightarrow \arg\min_{p} p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S \Rightarrow s > \log_{\varepsilon} (1 - p_1)^{\tau}/(1 - \tau)p_1 \]

2. $S_{RM}(MP) = \{p|p_1 - p_i \geq \tau, i = \arg\max_i p_i, p \in \Delta_n\}$

\[ \Rightarrow \arg\min_{p} p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S \Rightarrow s > \log_{\varepsilon} (1 - p_1)^{\tau} + p_i/(1 - \tau)p_1 \]

3. $S_{RM}(M2) = \{p|p_1 \geq \tau, p \in \Delta_n\}$

\[ \Rightarrow \arg\min_{p} p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S \Rightarrow s > \log_{\varepsilon} (1 - p_1)^{\tau} / (1 - \tau)^2 p_1 \]

\[ \delta = 4(2\tau(1 - p_1)^2)_{\Delta n} \leq 4p_1^2(1 - \tau) + 2\tau(1 - p_1)^2 \]

\[ \Rightarrow \frac{2\tau(1 - p_1)^2}{p_1^2(1 - \tau)^2} \leq \frac{2\tau(1 - p_1)^2}{p_1(1 - \tau)^2} \]

\[ s \in \mathbb{N} \Rightarrow \delta \in \mathbb{R} \Rightarrow \tau(1 - 0.5) \Rightarrow \delta \geq 2p_1^2(1 - \tau)^2 \]

From this point one can state the following:

**Lemma 3.** $p(\sigma|\varepsilon_{0:s}) \in \Delta_n \forall s \in \{0, 1, \cdots\}$ with $p(\sigma|\varepsilon_{0:s}) = p(\sigma) \oplus p(\varepsilon_0|\sigma) \oplus \cdots \oplus p(\varepsilon_s|\sigma)$, and $\mathcal{R}_S$ denotes the stopping region s.t. (S): $p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S$. WLOG, let $i^{th}$ location belong to the target class. Given $p(\varepsilon_{s} \mid \sigma) \propto [\varepsilon, 1, \cdots, 1]$ then the following relations hold:

1. $S_{RM}(M1) = \{p|p_1 \geq \tau, p \in \Delta_n\}$

\[ \Rightarrow \arg\min_{p} p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S \Rightarrow s > \log_{\varepsilon} (1 - p_1)^{\tau}/(1 - \tau)p_1 \]

2. $S_{RM}(MP) = \{p|p_1 - p_i \geq \tau, i = \arg\max_i p_i, p \in \Delta_n\}$

\[ \Rightarrow \arg\min_{p} p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S \Rightarrow s > \log_{\varepsilon} (1 - p_1)^{\tau} + p_i/(1 - \tau)p_1 \]

3. $S_{RM}(M2) = \{p|p_1 \geq \tau, p \in \Delta_n\}$

\[ \Rightarrow \arg\min_{p} p(\sigma|\varepsilon_{0:s}) \in \mathcal{R}_S \Rightarrow s > \log_{\varepsilon} (1 - p_1)^{\tau} / (1 - \tau)^2 p_1 \]

\[ \delta = 4(2\tau(1 - p_1)^2)_{\Delta n} \leq 4p_1^2(1 - \tau) + 2\tau(1 - p_1)^2 \]

\[ \Rightarrow \frac{2\tau(1 - p_1)^2}{p_1^2(1 - \tau)^2} \leq \frac{2\tau(1 - p_1)^2}{p_1(1 - \tau)^2} \]

\[ s \in \mathbb{N} \Rightarrow \delta \in \mathbb{R} \Rightarrow \tau(1 - 0.5) \Rightarrow \delta \geq 2p_1^2(1 - \tau)^2 \]

From this point one can state the following:

**Lemma 4.** Following Lemma 3, $p(\varepsilon_{s} \mid \sigma) \propto [\varepsilon, 1, \cdots, 1]$ \implies $p(\sigma|\varepsilon_{0:s}) \propto [\varepsilon, 1, \cdots, 1]$ where
\[ \varepsilon \sim \lognormal(\mu, \sigma^2) \]. Let \( \text{erf}(.) \) denote the error function, then the following relations hold:

\[
\begin{align*}
(1) S_{R(M1)} = \{ p | p_1 \geq \tau, p \in \Delta_n \} \\
\implies p(p(\varepsilon_{0,s}) \in S_{R(M1)}) &= \frac{1}{2} - \frac{1}{2} \text{erf} \left( \frac{\log(k_1) - s\mu}{\sqrt{2s\sigma^2}} \right) \\
\text{where } k_1 &= \frac{((1 - p_1)\tau)/((1 - \tau)p_1)}{(2) S_{R(MP)} = \{ p | p_1 - p_i \geq \tau(2), i = \arg \max_{i \neq 1} p_i, p \in \Delta_n \} \\
\implies p(p(\varepsilon_{0,s}) \in S_{R(MP)}) &= \frac{1}{2} - \frac{1}{2} \text{erf} \left( \frac{\log(k_2) - s\mu}{\sqrt{2s\sigma^2}} \right) \\
\text{where } k_2 &= \frac{((1 - p_1)\tau(2) + p_i)/((1 - \tau(2))p_1)}{(3) S_{R(M2)} = \{ p | p \geq \tau(3), p \in \Delta_n \} \\
\implies p(p(\varepsilon_{0,s}) \in S_{R(M2)}) &= \frac{1}{2} - \frac{1}{2} \text{erf} \left( \frac{\log(k_3) - s\mu}{\sqrt{2s\sigma^2}} \right) \\
\text{where } k_3 &= \frac{((1 - p_1)\tau(3) + (2\tau(3) - 1)^{1/2})}{(1 - \tau(3))^2p_1} \\
\end{align*}
\]

Proof. Observe the following trivial steps;

\[ \varepsilon_i \sim \lognormal(\mu, \sigma^2) \implies \left( \prod_{i=1}^{\Delta_n} \varepsilon_i \right) \sim \lognormal(s\mu, sc^2) \]

\[ p(x > y) = 1 - p(x < y) \]

Following the numbers obtained in Lemma 3, it is trivial to calculate respective values using lognormal cdf.

---

**Proof Proposition** Following Lemma 4

This is trivially a comparison of \( k_1, k_2 \) presented in Lemma 4. Observe the following:

\[ p(p(\varepsilon_{0,s}) \in S_{R(MP)}) - p(p(\varepsilon_{0,s}) \in S_{R(M1)}) = \left( -\text{erf} \left( \frac{\log(k_1) - s\mu}{\sqrt{2s\sigma^2}} \right) \right) - \left( -\text{erf} \left( \frac{\log(k_2) - s\mu}{\sqrt{2s\sigma^2}} \right) \right) \]

Observe that \( \log(.) \) and \( \text{erf}(.) \) are monotonically increasing. Hence \( f(x) = -\text{erf} \left( \frac{\log(x) - s\mu}{\sqrt{2s\sigma^2}} \right) \) is monotonically decreasing wrt. the argument \( x \). Therefore, it is sufficient to compare \( k_1 \) \& \( k_2 \) to conclude the order between the identities. We calculate \( k_i \) wrt. \( \tau \) only;

\[
\begin{align*}
k_1 &= \frac{(1 - p_1)^\tau}{(1 - \tau)p_1} \\
k_2 &= \frac{(1 - p_1)(1 - \tau) + p_i}{(1 - \tau(2)p_1)} = \frac{(1 - p_1)(2\tau - 1) + p_i}{(1 - 2\tau + 1)p_1} = \frac{2(1 - p_1)\tau - (1 - p_1 - p_i)}{2(1 - \tau)p_1} \\
\end{align*}
\]

\[ (1 - p_1)^\tau - \frac{1 - p_1 - p_i}{(1 - \tau)p_1} \rightarrow p_1 + p_i \leq 1 \implies k_2 < k_1 \]

Hence this concludes \( f(k_1) < f(k_2) \implies p(p(\varepsilon_{0,s}) \in S_{R(MP)}) > p(p(\varepsilon_{0,s}) \in S_{R(M1)}) \) implying that the probability of true selection in (MP) is higher.

Accordingly probabilities of stopping with an incorrect decision is calculated as the following; WLOG choose \( i = 2 \),

\[
\begin{align*}
p(p(\varepsilon_{0,s}) \in S_{R(M1)}) &\rightarrow \frac{p_2}{p_1\varepsilon + (1 - p_1)} > \tau \\
\implies \varepsilon < \frac{p_2 - (1 - p_1)\tau}{p_1\varepsilon} &\implies p(p(\varepsilon_{0,s}) \in S_{R(M1)}) \\
&= p \left( \varepsilon < \frac{p_2 - (1 - p_1)\tau}{p_1\varepsilon} \right) \\
\end{align*}
\]

False alarm probability for (MP) is calculated as the following, probability of a competitor \( (p_2) \) exceeding the probability of the target class \( (p_1) \) by the margin. We use the relation \( \hat{\tau} = 2 - 2\tau \) and write the following:

\[
\begin{align*}
p(p(\varepsilon_{0,s}) \in S_{R(MP)}) &\rightarrow \frac{p_2}{p_1\varepsilon + (1 - p_1)} - \frac{p_1\varepsilon}{p_1\varepsilon + (1 - p_1)} > (1 - \hat{\tau}) \rightarrow \frac{p_2 - (1 - p_1)(2\hat{\tau} - 1)}{p_1(1 + (2\hat{\tau} - 1))} > \varepsilon \\
\implies p_2 - (1 - p_1)(2\hat{\tau} - 1) &> \varepsilon \implies p(p(\varepsilon_{0,s}) \in S_{R(MP)}) \\
&= p \left( \varepsilon < \frac{p_2 - (1 - p_1)(2\hat{\tau} - 1)}{p_1(2\hat{\tau})} \right) \\
\end{align*}
\]

Observe that, this happens due to \( p_2 \) being the second highest competitor in the classification task. Following previous statements, one can compare values for \( \varepsilon \) cdf's for (M1), (MP) and (MP). The probabilities are listed in the following using the definitions \( \hat{\tau} = 2 - 2\tau \) and \( \hat{\tau} = \frac{(n-1)(2\tau-1)+1}{n} \);

\[
\begin{align*}
p(p(\varepsilon_{0,s}) \in S'_{R(M1)}(\hat{\tau})) &= p \left( \varepsilon < \frac{p_2 - (1 - p_1)\tau}{\tau p_1} \right) \\
p(p(\varepsilon_{0,s}) \in S'_R(MP)(\hat{\tau})) &= p \left( \varepsilon < \frac{p_2 - (1 - p_1)(2\tau - 1)}{p_1(2\tau)} \right) \\
&\implies p(p(\varepsilon_{0,s}) \in S'_{R(M1)}(\hat{\tau})) \\
&= p \left( \varepsilon < \frac{np_2 - (1 - p_1)((n-1)(2\tau - 1) + 1)}{((n-1)(2\tau - 1) + 1)p_1} \right) \\
\end{align*}
\]

Furthermore we define the following;

\[
\begin{align*}
k_1' &= \frac{p_2 - (1 - p_1)\tau}{\tau p_1} \\
k_2' &= \frac{p_2 - (1 - p_1)(2\tau - 1)}{p_1(2\tau)} \\
k_1' &= \frac{np_2 - (1 - p_1)((n-1)(2\tau - 1) + 1)}{((n-1)(2\tau - 1) + 1)p_1} \\
\end{align*}
\]

Observe the following relations;

\[ k_2' - k_1' = \frac{1 - p_1 - p_2}{2\tau p_1} \geq 0 \implies k_2' > k_1' \]
and,

\[
\frac{\delta k_1'}{\delta n} = \frac{2p_2(1-\tau)}{p_1(2\tau(n-1) - n + 2)^2} \geq 0 \implies \bar{k}_1' \bigg|_{n=2}
\]

\[\implies p_2 = 1 - p_1 \implies k_2' = \frac{p_2 - (1 - p_1)(2\tau - 1)}{p_1(2\tau)} = \frac{(1 - p_1)(2 - 2\tau)}{(1 - p_1)(1 - \tau)}
\]

\& \[\bar{k}_1' = \frac{2p_2 - (1 - p_1)(2\tau)}{(2\tau)p_1} = \frac{\tau p_1}{(1 - p_1)(1 - \tau)}\]

\[\implies k_2' = \bar{k}_1' \bigg|_{n=2} \implies \bar{k}_1' > k_2' \text{ from monotonic increasing.}
\]

Therefore we conclude \(\bar{k}_1' > k_2' > k_1'\). Similarly with the first part of the proof this implies; \(p(p(\sigma|\varepsilon_{0:s}) \in S_{R(M1)}')((\bar{\tau})) > p(p(\sigma|\varepsilon_{0:s}) \in S_{R(MP)}')((\bar{\tau})) \implies p(p(\sigma|\varepsilon_{0:s}) \in S_{R(M1)}')((\tau)) \) implying that the incorrect selection can be sandwiched. \(\square\)