Early Time Classification with Accumulated Accuracy Gap Control

Liran Ringel [1] **Regev Cohen** [2] **Daniel Freedman** [2] **Michael Elad** [2] **Yaniv Romano** [1 3]

Abstract

Early time classification algorithms aim to label a stream of features without processing the full input stream, while maintaining accuracy comparable to that achieved by applying the classifier to the entire input. In this paper, we introduce a statistical framework that can be applied to any sequential classifier, formulating a calibrated stopping rule. This data-driven rule attains finite-sample, distribution-free control of the accuracy gap between full and early-time classification. We start by presenting a novel method that builds on the Learn-then-Test calibration framework to control this gap marginally, on average over i.i.d. instances. As this algorithm tends to yield an excessively high accuracy gap for early halt times, our main contribution is the proposal of a framework that controls a stronger notion of error, where the accuracy gap is controlled conditionally on the accumulated halt times. Numerical experiments demonstrate the effectiveness, applicability, and usefulness of our method. We show that our proposed early stopping mechanism reduces up to 94% of timesteps used for classification while achieving rigorous accuracy gap control.

1. Introduction

The goal of early time series classification (ETSC) is to predict the label of a given input data stream as quickly as possible. Such methods are especially advantageous in scenarios requiring prompt predictive inference. For example, consider the problem of reading comprehension, illustrated in Figure 1. Suppose we employ an autoregressive large language model (LLM) to analyze a given document (context) and select an answer to the provided question. Given that the inference time of LLMs increases with the number of processed tokens (or sentences), we wish to terminate the processing of the context retrieved from the document as soon as the necessary information is found, rather than processing the entire document naively. Other tasks for which ETSC is highly desired include real-time song identification (think of the `Shazam` application) and reducing radiation exposure in computational tomography (CT) systems, among many others. In all of these applications, the objective is to stop the inference process early while preserving accuracy, as if the predictive model had been applied to the entire data stream.

Question: What was the nationality of Ronald Fisher?
Options: **(1)** American **(2)** British **(3)** Canadian **(4)** Australian
Context: Sir Ronald Aylmer Fisher FRS (17 February 1890 – 29 July 1962) was a British polymath who was active as a mathematician, statistician, biologist, geneticist, and academic. For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". In genetics, his work used mathematics to combine Mendelian genetics and natural selection...
Answer: (2) British.

Figure 1. **An illustration of a reading comprehension task.** An LLM sequentially processes the given document to find the answer to the question provided and, ideally, should stop scanning the document immediately after the required information is found. The context is taken from Wikipedia.

Consider labeled pairs of the form (X, Y) sampled i.i.d. from P_{XY}, where $X = (X^1, X^2, \ldots, X^{t_{\text{max}}}) \in \mathcal{X}$ represents an observed input sequence with a maximum length of t_{max}, e.g. a sequence of tokens representing sentences in a document and a question. The variable $Y \in \mathcal{Y} = \{1, \ldots, K\}$ is the unknown label we wish to predict, e.g. the correct answer to the given question. Suppose we are handed a pre-trained classifier $\hat{f} : \mathcal{X} \to [0,1]^K$ that processes the input X sequentially and, at each timestep t, maps $X^{\leq t} = (X^1, \ldots, X^t)$ to an estimated probability distribution over the labels. We employ a stopping rule function that, at each timestep t, decides whether to stop the inference process *only based on the data observed up to timestep* t. Denote the stopping time by $\hat{\tau}(X) \in \{1, \ldots, t_{\text{max}}\}$ and let $\hat{Y}_{\text{early}}(\hat{\tau})$ and $\hat{Y}_{\text{full}}$ be the predicted labels obtained by $\hat{f}(X^{\leq \hat{\tau}(X)})$ and $\hat{f}(X)$, respectively. With these notations

[1]Department of Computer Science, Technion—Israel Institute of Technology, Haifa, Israel [2]Verily AI, Israel [3]Department of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Haifa, Israel. Correspondence to: Liran Ringel <liranringel@cs.technion.ac.il>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

in place, we define the *accuracy gap* as the proportion of samples for which the classifier's prediction is correct when applied to the entire sequence but incorrect when the same classifier is applied only up to the early timestep $\hat{\tau}(X)$.

Let $\alpha \in (0,1)$, e.g., 10%, be the tolerable accuracy gap, representing the acceptable trade-off for early stopping. Denote by $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$ a *holdout calibration set*, with samples drawn i.i.d. from P_{XY}. Our initial objective is to leverage $\mathcal{D}_{\text{cal}}$ to identify an early stopping rule $\hat{\tau}(X)$ that minimizes the halt time while ensuring the accuracy gap remains below α with a probability of at least $1-\delta$:

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(R_{\text{gap}}^{\text{marginal}}(\hat{\tau}) \leq \alpha\right) \geq 1-\delta, \tag{1}$$

where

$$R_{\text{gap}}^{\text{marginal}}(\hat{\tau}) = \mathbb{E}_{P_{XY}}\left[L_{\text{gap}}(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}(\hat{\tau}))\right], \tag{2}$$

and

$$L_{\text{gap}}(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}(\hat{\tau})) = \left(\mathbb{I}_{Y=\hat{Y}^{\text{full}}} - \mathbb{I}_{Y=\hat{Y}^{\text{early}}(\hat{\tau})}\right)_+. \tag{3}$$

Notably, the probability in (1) is taken over the randomness in $\mathcal{D}_{\text{cal}}$, and δ is a user-defined level, e.g., 1%. The operator $(z)_+$ in (3) returns the value z if $z \geq 0$ and zero otherwise, and the indicator function $\mathbb{I}_{a=b}$ equals 1 when $a=b$ and zero otherwise. In simpler terms, the expected value of $L_{\text{gap}}(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}(\hat{\tau})) \in \{0,1\}$ reflects the proportion of samples in which the decision to stop early increases the error rate. We refer to (1) as *marginal risk control* as it states that the accuracy gap will not exceed α, *on average* over future observations and stopping times. In Section 3, we present an algorithm that rigorously attains (1), termed the *marginal method*.

While the marginal guarantee in (1) provides a controlled mechanism for early classification, it may not be entirely satisfying in most practical settings. This is because an algorithm that controls the accuracy gap over all possible sequences is permitted to perform poorly on sequences with early halt times while excelling on sequences with late halt times. This, in turn, can undermine the reliability of predictions for sequences with early halt times. Recognizing this limitation, our main contribution is a novel algorithm that aims to *control the accuracy gap conditional on the halt time being less or equal to* t. More formally, let

$$R_{\text{gap}}^{\leq t}(\hat{\tau}) = \mathbb{E}_{P_{XY}}\left[L_{\text{gap}}(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}(\hat{\tau})) \mid \hat{\tau}(X) \leq t\right]. \tag{4}$$

Our goal is to formulate a stopping rule that achieves

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(R_{\text{gap}}^{\leq t}(\hat{\tau}) \leq \alpha \text{ for all } t \geq t_0\right) \geq 1-\delta, \tag{5}$$

where t_0 is defined as the first timestep for which $P(\hat{\tau}(X) \leq t_0) > 0$, as otherwise (4) is undefined. In particular, controlling (5) implies that we also control the accuracy gap marginally, as $R_{\text{gap}}^{\text{marginal}} = R_{\text{gap}}^{\leq t_{\max}}$. Throughout this work, we refer to (5) as *conditional risk control on the accumulated halt time*, or simply, *conditional risk control*. In Section 4 we present an algorithm that achieves this goal, which we refer to as the *conditional method*.

It is crucial to distinguish (5) from the stronger time- or instance-conditional guarantee, where the objective is to control the accuracy gap for a specific timestep t or for a specific X. Unfortunately, attainment of non-trivial stopping rules with time- or instance-conditional risk control is infeasible without resorting to unrealistic assumptions (Vovk, 2012; Lei & Wasserman, 2014; Foygel Barber et al., 2021), which we aim to avoid: we pursue distribution-free, finite-sample guarantees. As a consequence, we posit that the risk in (5) strikes a reasonable compromise between controlling the relatively weak marginal risk and the unattainable time- or instance-conditional risk.

1.1. A Motivating Example: Reading Comprehension

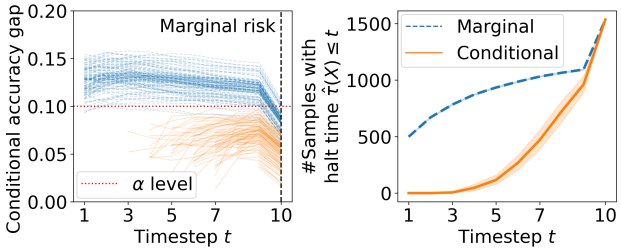


Figure 2. **Comparison between the marginal and conditional methods for the reading comprehension task.** Nominal accuracy gap level is $\alpha = 10\%$ and $\delta = 1\%$. Left: empirical conditional accuracy gap, $\hat{R}_{\text{gap}}^{\leq t}$, across 100 trials; each curve corresponds to a different random split of the calibration and test data. Right: accumulated halt times as a function of t, averaged over 100 random splits; the shaded area represents a 95% confidence interval.

To emphasize the importance of the transition from the marginal (1) to the conditional guarantee (5), we now return to the reading comprehension problem discussed earlier. The `QuALITY` dataset (Pang et al., 2022) consists of 4609 triplets, containing (i) a question, (ii) multiple choice answers, and (iii) a long context, with each triplet accompanied by the correct labeled answer. We utilize a pre-trained autoregressive LLM as the base predictive model. This classifier sequentially processes the context and selects an answer from the four possibilities. For the calibration of the early stopping rule, we employ 3073 labeled samples to form $\mathcal{D}_{\text{cal}}$ while reserving the remaining 1536 samples for testing. Following this, we compare the performance of the proposed marginal and conditional calibration methods presented in Sections 3 and 4, respectively. Specifically, we report two performance metrics: (i) $\hat{R}_{\text{gap}}^{\leq t}$, defined as the empirical accuracy gap of samples with a halt time $\hat{\tau}(X)$ equal to or less than t; and (ii) the cumulative number of

samples on which the model halted until timestep t.

The results are presented in Figure 2. Following the left panel in that figure, we can see that while the two approaches control the marginal risk, the conditional accuracy gap $\hat{R}^{\leq t}_{\text{gap}}$ tends to be higher than the desired 10% level for the marginal method. This implies that the marginal stopping rule tends to halt too early, as evidenced in the right panel of Figure 2, where we can see the relatively large number of samples halted at timestep $t = 1$. In contrast, the conditional approach maintains the conditional accuracy gap below α across all timesteps (left panel) while attaining an effective early stopping mechanism (right panel).

1.2. Preview of our methods

The crux of this work is the formulation of a stopping rule $\hat{\tau}(X)$ that attains valid risk control. Denote by $\hat{\pi} : \mathcal{X} \to [0, 1]$ a score that heuristically reflects how confident the classifier is in its prediction based on $X^{\leq t}$. For example, $\hat{\pi}(X^{\leq t})$ can be the largest softmax value of a neural net classifier. With this in place, we can formulate

$$\hat{\tau}(X) = \tau_{\hat{\underline{\lambda}}}(X) = \min\{t : \hat{\pi}(X^{\leq t}) \geq \hat{\underline{\lambda}}_t \text{ or } t = t_{\max}\}, \tag{6}$$

where $\hat{\underline{\lambda}}_t$ is a hyperparameter, being the t-th element in a vector of thresholds $\hat{\underline{\lambda}} = (\hat{\underline{\lambda}}_1, \hat{\underline{\lambda}}_2, \ldots, \hat{\underline{\lambda}}_{t_{\max}})$. In plain words, we choose to halt the inference process for the first time t that the classifier is "confident enough" in its prediction. But how can we properly choose the vector of hyperparameters $\hat{\underline{\lambda}}$ that attains a valid risk control? Notably, this task becomes particularly challenging when dealing with a large number of hyperparameters that require tuning; in our case, we have $t_{\max}$ parameters. An improper choice of hyperparameters can fail to achieve the desired accuracy gap on future test data, and this problem is especially pronounced when the accuracy gap is a non-monotone function of the hyperparameters, which may occur in our setting due to the complex nature of the pre-trained classifier at hand.

To tackle this challenge, we build on the *Learn then Test* (LTT) framework (Angelopoulos et al., 2021) that formulates the problem of finding hyperparameters that yield risk control as a multiple hypothesis testing problem, where each hypothesis corresponds to a different choice of hyperparameters. However, in situations with a vast array of parameters that need to be tuned, this method faces two practical obstacles (Laufer-Goldshtein et al., 2022). First, the sheer volume of potential configurations, which grows exponentially with $t_{\max}$, makes an extensive search of hyperparameters infeasible. Second, the LTT method may experience a loss of power when confronted with such an exponential number of tests. This drawback can result in our algorithm stopping too late, potentially missing the opportunity to select a more refined set of hyperparameters for the downstream task.

To alleviate these limitations, we propose a two-stage calibration framework that exploits the special structure of the underlying ETSC problem. In the first stage, we find a candidate set of hyperparameters using a novel computationally efficient procedure. Then, we apply a multiple testing procedure on the candidate set to select a valid set of hyperparameters that yields risk control. Overall, the novel algorithm we introduce can efficiently handle long sequences, while selecting a data-adaptive threshold vector $\hat{\underline{\lambda}}$ that formulates a statistically valid early stopping rule. In turn, the contributions of this work are the following:

1. **A novel application for LTT**: we introduce, for the first time, methodologies that support ETSC algorithms with rigorous distribution-free, finite-sample risk-controlling guarantees.

2. **Marginal risk control**: we present a flexible framework that allows predictive models to stop early the inference process while controlling the average accuracy gap.

3. **Conditional risk control**: next, we introduce a novel algorithm for early stopping that controls the accuracy gap conditional on the accumulated halt times.

4. **Theory precisely holds in practice**: we illustrate the effectiveness of our algorithms by applying them to diverse tasks. These include standard time series classification datasets and a novel application in natural language processing (NLP). Our methods controls the risk while saving up to 94% of the timesteps available to make predictions. A software package implementing the proposed methods is publicly available at GitHub.[1]

2. Related Work

There is active research in developing machine learning models for ETSC with stopping rules that aim to balance accuracy and early termination (Hartvigsen et al., 2019; Gupta et al., 2020; Ebihara et al., 2020; Miyagawa & Ebihara, 2021; Ghodrati et al., 2021; Sabet et al., 2021; Tang et al., 2022; Hartvigsen et al., 2022; Chen et al., 2022; Shekhar et al., 2023; Ebihara et al., 2023). While these tools are effective in practice, they often lack statistical assurance. Our proposal enriches this important line of research by introducing versatile tools, compatible with any state-of-the-art ETSC model, which rigorously control the accuracy gap, be it in a marginal or conditional sense.

Our proposal is closely related to calibrated predictive inference techniques, including conformal prediction, risk-controlling methods, and selective classification (Vovk et al., 2005; Papadopoulos & Haralambous, 2011; Lei et al., 2018; Tibshirani et al., 2019; Romano et al., 2020; Bates et al.,

[1] https://github.com/liranringel/etc

2021; Angelopoulos & Bates, 2023; Gibbs & Candes, 2021; Lin et al., 2022; Angelopoulos et al., 2022; Fisch et al., 2022; Feldman et al., 2023; Lee et al., 2023b; Cauchois et al., 2023; Barber et al., 2023). Specifically, we expand the toolbox of risk-controlling tools, particularly when facing situations with high dimensional hyperparameter space. The pioneering LTT work by Angelopoulos et al. (2021) offers an approach to find a data-driven configuration of parameters that, for example, can be used to simultaneously control multiple risks. However, this approach can mostly handle *low dimensional hyperparameter space* and becomes intractable when the search space is large. Recognizing this limitation, Laufer-Goldshtein et al. (2022; 2023) utilize Bayesian optimization tools to find Pareto optimal candidate configurations across various risks, which, in turn, improve the computational and statistical efficiency of LTT. This line of work shares similarities with the challenges we face in this paper; however, instead of utilizing a general purpose Bayesian optimization tool for parameter tuning, or using exhaustive search, we design a specialized procedure that builds upon the structure of the ETSC problem. Our proposal results in a computationally efficient technique to identify plausible configurations among the potentially enormous search space of $\underline{\lambda}$, that controls the accuracy gap with meaningful statistical power.

Our approach is also aligned with recent efforts to design calibration methods that aim to reduce the computational complexity of LLMs (Schuster et al., 2021; 2022). These methods involve formulating an early exit mechanism with, for example, marginal accuracy gap control. A key difference between the above methods and our proposal is that we apply early exit over the time horizon rather than over the intermediate transformer layers. Furthermore, a crucial conceptual and technical difference is our transition from marginal to conditional guarantees, departing from the contributions mentioned above.

3. Warm-up: Marginal Accuracy Gap Control

To set the stage for our framework for conditional risk control, we start by presenting a method that achieves the modest marginal guarantee in (1). The development of this method also exposes the reader to the statistical principles of LTT. Later, in Section 4, we will build on the foundations of the method presented here and introduce our main contribution—a methodology that attains the conditional guarantee of (5). To further simplify the exposition of the proposed marginal approach, consider tuning a single parameter $\hat{\lambda} \in [0,1] \cup \{\infty\}$ for all timesteps so that the stopping rule $\tau_{\hat{\lambda}}(X) = \min\{t : \hat{\pi}(X^{\leq t}) \geq \hat{\lambda} \text{ or } t = t_{\text{max}}\}$ achieves (1).

To start with, suppose we handed a candidate parameter λ, e.g., $\lambda = 0.7$, and we are interested in testing whether it controls the accuracy gap. Following the LTT (Angelopoulos et al., 2021) approach, we define the null hypothesis induced by λ as follows:

$$H_{0,\lambda} : R^{\text{marginal}}_{\text{gap}}(\tau_\lambda) > \alpha. \tag{7}$$

That is, if the null is false, our candidate λ controls the marginal accuracy gap. With this in place, we formulate a statistical test that utilizes the observed labeled data—the calibration set—to decide whether we can reject $H_{0,\lambda}$ and report that λ is likely to control the risk or accept $H_{0,\lambda}$ if there is not enough evidence to reject the null. To formulate such a test, we compute a *p-value* p_λ, where a valid p-value satisfies the following property under the null:

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(p_\lambda \leq u \mid H_{0,\lambda}\right) \leq u, \;\; u \in [0,1]. \tag{8}$$

In plain words, if $H_{0,\lambda}$ is true, the p-value is stochastically greater than or equal to uniform distribution on $[0,1]$. Hence, *considering a single hypothesis*, when observing $p_\lambda \leq \delta$ we can safely reject $H_{0,\lambda}$, knowing that the probability of falsely rejecting the null (type I error) is at most δ.

To compute such a p-value, we leverage the fact that the loss L_{gap} is binary, and thus we can employ the exact tail bound from Bates et al. (2021) (Appendix B); see also Brown et al. (2001). In more detail, denote the cumulative distribution function of the binomial distribution by $\text{CDF}_{\text{bin}}(\hat{k}; n, \alpha)$ where $\hat{k}$ is the number of successes, n is the number of independent Bernoulli trials, and α is the probability of success. Thus, in our case, the p-value is $\hat{p}_\lambda = \text{CDF}_{\text{bin}}\left(n\hat{R}_{\text{gap}}(\tau_\lambda); n, \alpha\right)$, where $\hat{R}_{\text{gap}}(\tau_\lambda)$ is the empirical accuracy gap obtained by the stopping rule τ_λ, evaluated on $n = |\mathcal{D}_{\text{cal}}|$ i.i.d. samples. Put simply, this formula transforms the empirical risk, evaluated on the calibration set $\mathcal{D}_{\text{cal}}$, into a p-value that satisfies (8).

Thus far we have discussed the problem of testing for a single hypothesis, i.e., testing whether a specific candidate λ does not control the accuracy gap. However, naturally, the task of finding $\hat{\lambda}$ that promotes early stopping while controlling the risk involves *testing for multiple hypotheses*: each hypothesis H_{0,λ_i} corresponds to a different $\lambda_i \in \Lambda,\ 1 \leq i \leq |\Lambda|$, where $\Lambda = \{0, \Delta, 2\Delta, \ldots, 1\} = \{\lambda_1, \lambda_2, \ldots, \lambda_{|\Lambda|}\}$ is a discretized grid of possible values and $\Delta \in (0,1)$ defines the resolution of the grid.

The challenge that arises is that we must test all hypotheses simultaneously. To clarify, a naive rejection rule $p_{\lambda_i} \leq \delta$ can lead to a high probability that some of the true null hypotheses are rejected by chance alone, and this probability increases with the number of true nulls that are tested (Miller, 2012). To tackle this issue, we follow Angelopoulos et al. (2021) and formulate a multiple testing procedure that controls the *family-wise error rate* (FWER). Formally, let V be the number of true nulls that are falsely rejected

by the testing procedure, and define FWER $= \mathbb{P}(V \geq 1)$ as the probability of falsely rejecting at least one true null hypothesis. Therefore, to control (1), we should design a testing procedure that ensures the FWER does not exceed δ.

To rigorously control the FWER, we adopt the *fixed sequence testing* procedure (Bauer, 1991) used in LTT, as follows. First, we order the hypotheses from most plausible to least *without looking at the calibration data*. In our context, higher thresholds are more likely to control the risk in (1), and therefore we order the hypotheses from the largest $\lambda_{|\Lambda|}$ to the smallest λ_1. Then, we arrange the p-values according to this ordering and sequentially compare each p-value to the desired level δ. This sequential testing procedure terminates the first time j that $p_{\lambda_j} > \delta$, resulting in a set of valid thresholds $\mathcal{R} = \{\lambda_i : i > j\} \cup \{\infty\}$. Importantly, any threshold in the set $\mathcal{R}$ is guaranteed to control (1), including the trivial choice for which $\lambda = \infty$. (When $\lambda = \infty$ the model will never stop early and thus trivially achieves zero accuracy gap.) Since our goal is to formulate a rule that stops as early as possible, we set the final $\hat{\lambda}$ to be the smallest λ among the rejected ones, i.e., $\hat{\lambda} = \lambda_{j+1}$, or $\hat{\lambda} = \infty$ if $p_{\lambda_{|\Lambda|}} > \delta$. For ease of reference, this procedure is summarized in Algorithm A.3, presented in Appendix A, and its validity is a direct consequence of using fixed sequence testing to control the FWER at level δ.

Proposition 1. *Assuming the calibration and test samples are i.i.d., with $\hat{\lambda}$ selected as outlined in Algorithm A.3, the stopping rule $\tau_{\hat{\lambda}}(X)$ satisfies (1).*

All proofs are presented in Appendix B. In plain words, the above proposition implies that Algorithm A.3 formulates a stopping rule that achieves marginal accuracy gap control given a finite calibration set, no matter what the data distribution is, and regardless of the choice of the "black-box" classifier. While Proposition 1 is appealing, the usefulness of the marginal guarantee in real-world scenarios may be limited, as discussed and demonstrated in Section 1.1. This limitation prompts our exploration in the next section.

4. Conditional Accuracy Gap Control

We now turn to present the focal point of this work: a framework designed to control the conditional accuracy gap (5). Beyond the transition from marginal to conditional guarantee, in this section we utilize a more general formulation of the stopping rule, in which $\hat{\tau}(X) = \tau_{\underline{\hat{\lambda}}}(X) = \min\{t : \hat{\pi}(X^{\leq t}) \geq \hat{\underline{\lambda}}_t\}$ with $\hat{\underline{\lambda}} = (\hat{\underline{\lambda}}_1, \hat{\underline{\lambda}}_2, \ldots, \hat{\underline{\lambda}}_{t_{\text{max}}})$. This choice adds additional flexibility to the proposed framework compared to tuning a single parameter (as in Section 3), allowing us to formulate more effective early stopping rules.

Analogously to Section 3, we will adopt the fixed sequence testing procedure to construct a rejection set $\mathcal{R}$ that contains the configurations of $\underline{\lambda}$ that control the conditional risk. In the view of multiple testing, now each null hypotheses is formulated as

$$H_{0,\underline{\lambda}} : R^{\leq t}_{\text{gap}}(\tau_{\underline{\lambda}}) > \alpha \;\text{ for at least one } t \geq t_0, \qquad (9)$$

where t_0 is the first timestep at which the probability for an early stopping event is not zero, i.e., $P(\tau_{\underline{\lambda}}(X) \leq t_0) > 0$.

In striking contrast to Section 3, the formulation of a FWER-controlling procedure in this case is far more challenging due to the following.

1. There are $(|\Lambda| + 1)^{t_{\text{max}}}$ possible configurations for $\underline{\lambda}$ and thus it is infeasible to sweep over this exponential number of hypotheses. Given this sheer volume, computing a p-value for each hypothesis exceeds reasonable computational limits.

2. To achieve good statistical power with fixed sequence testing, careful ordering of hypotheses is essential: inadequate ordering may lead to a rejection set $\mathcal{R}$ that includes less effective threshold vectors. As discussed in Section 3, there is a natural ordering of the hypotheses when considering the tuning of a *single* threshold; we can simply order the hypotheses from the largest λ to the smallest one. However, it is unclear how to order the hypotheses when working with a vector $\underline{\lambda}$.

3. When faced with a small sample size, the p-value may be too high even if the risk is lower than α. This is attributed to the fact that the p-value produced by CDF_{bin} takes into account the number of samples used to calculate the empirical risk. Importantly, this is not an abstract concern; in practice, as we strive for conditional risk control, situations with a small sample size become prevalent, particularly for the very early timesteps.

In what follows, we present a method that alleviates these issues, taking inspiration from the principle of *split fixed sequence testing* proposed in LTT. In this approach, we first split the calibration set $\mathcal{D}_{\text{cal}}$ into two disjoint sets: $\mathcal{D}_{\text{cal-1}}$ and $\mathcal{D}_{\text{cal-2}}$. Then, we proceed with a two-stage algorithm, described below at a high level.
Stage 1: Candidate Screening: Use $\mathcal{D}_{\text{cal-1}}$ to heuristically find a data-adaptive threshold vector $\hat{\underline{\eta}}$, with an eye towards early stopping with conditional risk control.
Stage 2: Testing: Apply fixed sequence testing to configurations derived from $\hat{\underline{\eta}}$. Here, we use the independent holdout set $\mathcal{D}_{\text{cal-2}}$ to ensure the validity of the test.

4.1. Stage 1: Candidate Screening

We present a greedy algorithm that takes as input a predictive model and calibration data $\mathcal{D}_{\text{cal-1}}$ and returns a candidate threshold vector $\hat{\underline{\eta}}$. This procedure, summarized in Algorithm 1, sequentially updates the elements in the vector $\hat{\underline{\eta}}$ as follows. It starts by updating the first ele-

Algorithm 1 Candidate Screening (Stage 1)

Input: Calibration set $\mathcal{D}_{\text{cal-1}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cal-1}}}$, tolerable accuracy gap α, grid resolution Δ.
$\underline{\hat{\eta}} \leftarrow \{\infty, \ldots, \infty\}$
// Find $\underline{\hat{\eta}}_t$ greedily during the t-th iteration.
for $t = 1, \ldots, t_{\max}$ **do**
  $\underline{\eta} \leftarrow \underline{\hat{\eta}}$
  // Find the lowest $\underline{\eta}_t \in \Lambda$ s.t. $\hat{R}^{\leq t}_{\text{gap}} \leq \alpha$.
  for $\xi = 0, \Delta, 2\Delta, \ldots, 1$ **do**
    $\underline{\eta}_t \leftarrow \xi$
    // Find samples with a halt time $\leq t$.
    $I \leftarrow \{i : \tau_{\underline{\eta}}(X_i) \leq t\}$
    if $I = \emptyset$ **then**
      // Cannot calculate the empirical risk.
      Break inner loop and set $\underline{\hat{\eta}}_t = \infty$
    end if
    // Calculate the empirical risk.
    $\hat{R}^{\leq t}_{\text{gap}} \leftarrow \frac{1}{|I|} \sum_{i \in I} L_{\text{gap}}(Y_i, \hat{Y}_i^{\text{full}}, \hat{Y}_i^{\text{early}}(\tau_{\underline{\eta}}))$
    if $\hat{R}^{\leq t}_{\text{gap}} \leq \alpha$ **then**
      // Found the lowest $\underline{\eta}_t$ s.t. $\hat{R}^{\leq t}_{\text{gap}} \leq \alpha$.
      Break inner loop and set $\underline{\hat{\eta}}_t \leftarrow \xi$
    end if
  end for
end for
Output: $\underline{\hat{\eta}}$

Algorithm 2 Testing (Stage 2)

Input: Calibration set $\mathcal{D}_{\text{cal-2}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cal-2}}}$, candidate thresholds $\underline{\hat{\eta}}$, tolerable accuracy gap α, significance level δ, grid resolution Δ.
// Start with the most conservative stopping rule.
$\underline{\hat{\lambda}} \leftarrow \{\infty, \ldots, \infty\}$
// Gradually reveal another $\underline{\hat{\eta}}_t$ from the end and test it.
for $t = t_{\max}, \ldots, 1$ **do**
  $\underline{\lambda} \leftarrow \underline{\hat{\lambda}}$
  $\underline{\lambda}_t \leftarrow \underline{\hat{\eta}}_t$ // Set $\underline{\lambda}$ to $\underline{\lambda}^t$.
  // Test $H^{t'}_{0,\underline{\lambda}^t}$ for all $t' \geq t$.
  for $t' = t, \ldots, t_{\max}$ **do**
    // Find samples with a halt time $\leq t'$.
    $I \leftarrow \{i : \tau_{\underline{\lambda}}(X_i) \leq t'\}$
    if $I = \emptyset$ **then**
      // No evidence to reject the null, stop testing.
      Break both loops
    end if
    // Calculate the empirical risk.
    $\hat{R}^{\leq t'}_{\text{gap}} \leftarrow \frac{1}{|I|} \sum_{i \in I} L_{\text{gap}}(Y_i, \hat{Y}_i^{\text{full}}, \hat{Y}_i^{\text{early}}(\tau_{\underline{\lambda}}))$
    // Compute a p-value.
    $\hat{p}^{t'}_{\underline{\lambda}^t} \leftarrow \text{CDF}_{\text{bin}}\left(\hat{R}^{\leq t'}_{\text{gap}} \cdot |I|; |I|, \alpha\right)$
    if $\hat{p}^{t'}_{\underline{\lambda}^t} > \delta$ **then**
      // Failed to reject the null, stop testing.
      Break both loops
    end if
  end for
  $\underline{\hat{\lambda}} \leftarrow \underline{\lambda}$ // $H_{0,\underline{\lambda}^t}$ was rejected, update the chosen $\underline{\hat{\lambda}}$.
end for
Output: $\underline{\hat{\lambda}}$

ment $\underline{\hat{\eta}}_1$ that corresponds to the timestep $t = 1$, then proceeds to $\underline{\hat{\eta}}_2$ for $t = 2$, and continues until reaching $\underline{\hat{\eta}}_{\max}$ at $t = t_{\max}$. Specifically, at timestep t, we are handed the vector $\underline{\hat{\eta}} = (\underline{\hat{\eta}}_1, \ldots, \underline{\hat{\eta}}_{t-1}, \infty, \ldots, \infty)$, and set its t-th element $\underline{\hat{\eta}}_t$ to be the smallest $\underline{\eta}_t$ such that $\hat{R}^{\leq t}_{\text{gap}}(\tau_{\underline{\hat{\eta}}}) \leq \alpha$, or keep $\underline{\hat{\eta}}_t = \infty$ if there is no $\underline{\eta}_t$ that satisfies this constraint. Above, $\hat{R}^{\leq t}_{\text{gap}}(\tau_{\underline{\hat{\eta}}})$ is the empirical accuracy gap of the samples with halt time that is less than or equal to t (see line 16).

Before moving to the next stage, we pause to discuss the properties of this greedy method. First, the computational complexity of the proposed algorithm is $\mathcal{O}(t_{\max} \cdot |\Lambda| \cdot |\mathcal{D}_{\text{cal-1}}|)$, which is attributed to the fact that we choose to sequentially update the vector $\underline{\hat{\eta}}$. Second, by design, the choice of $\underline{\hat{\eta}}_{t'}$ for $t' > t$ does not affect $\hat{R}^{\leq t'}_{\text{gap}}$ for $t' \leq t$. Third, this greedy method seeks a vector $\underline{\hat{\eta}}$ that yields a stopping rule whose empirical conditional risk is tightly regulated around α, but not exceeded. This property is crucial to attaining an effective early stopping rule. In principle, instead of determining $\underline{\hat{\eta}}_t$ solely based on empirical risk, we could choose the smallest $\underline{\eta}_t$ whose p-value falls below δ, an approach that is akin to the split fixed sequence testing idea of Angelopoulos et al. (2021). However, we decided to work directly with the empirical risk, as it is arguably straightforward to implement, and we found these two approaches to have similar halt times. In any case, while sensible, the process of finding the vector $\underline{\hat{\eta}}$ is heuristic in the sense that it is not guaranteed to control the conditional risk for future test points. This issue naturally leads us to the next stage.

4.2. Stage 2: Testing

In this testing stage, we build on the candidate vector $\underline{\hat{\eta}}$ to form a statistically valid stopping rule that attains (5). A naive and optimistic approach would be to test for a single null $H_{0,\underline{\lambda}}$ defined in (9) for the choice $\underline{\lambda} = \underline{\hat{\eta}}$. Rejection of this null hypothesis with a significance level of δ implies that $\underline{\hat{\eta}}$ attains (5), achieving a powerful stopping rule due to the design of $\underline{\hat{\eta}}$. However, if we fail to reject this null, our fallback is the trivial configuration $\underline{\hat{\lambda}} = (\infty, \ldots, \infty)$ that results in a conditional accuracy gap of zero. However, in this case, the stopping rule we form is the most conservative one, as the model will never stop early.

To alleviate this, we employ fixed sequence testing, designed to yield an effective stopping rule with FWER-

control, even in cases where the null hypothesis $H_{0,\underline{\lambda}}$ with the "optimistic" configuration $\underline{\lambda} = \underline{\hat{\eta}}$ would not be rejected. Recall the underlying principle of fixed sequence testing: order the hypotheses from the most plausible to the least, without looking at the holdout data $\mathcal{D}_{\text{cal-2}}$. Building on the structure of the ETSC problem, we define the sequence of configurations $\underline{\lambda}^{t_{\max}} = (\infty, \ldots, \infty, \hat{\eta}_{t_{\max}})$, $\underline{\lambda}^{t_{\max}-1} = (\infty, \ldots, \infty, \hat{\eta}_{t_{\max}-1}, \hat{\eta}_{t_{\max}})$, all the way to $\underline{\lambda}^{1} = (\hat{\eta}_1, \hat{\eta}_2, \ldots \hat{\eta}_{t_{\max}})$. That is, the t'-th element in the vector $\underline{\lambda}^t$ is $\underline{\lambda}^t_{t'} = \hat{\eta}_{t'}$ if $t' \geq t$, and $\underline{\lambda}^t_{t'} = \infty$ otherwise. Importantly, the stopping rule $\tau_{\underline{\lambda}^t}$ does not allow stopping the classification process at timesteps smaller than t. With this construction in place, we suggest applying the fixed sequence testing procedure to the hypotheses ordered from the one induced by $\underline{\lambda}^{t_{\max}}$, i.e., $H_{0,\underline{\lambda}^{t_{\max}}}$ down to the one corresponding to $\underline{\lambda}^1$, i.e., $H_{0,\underline{\lambda}^1}$. Note that this ordering is particularly powerful when the accuracy gap of the model tends to decrease with the number of timesteps observed—a sensible characteristic in ETSC. Additionally, the suggested ordering enables us to postpone the testing of hypotheses involving limited sample sizes to later stages of the procedure, which is attractive as it is more likely that we will fail to reject those nulls.

Having defined the ordering of the hypotheses, we turn to describe how to compute a valid p-value for each of the individual hypotheses, using the holdout data $\mathcal{D}_{\text{cal-2}}$. Consider the hypothesis $H_{0,\underline{\lambda}}$ in (9) for the choice $\underline{\lambda} = \underline{\lambda}^t$, and define its finer null hypotheses as follows:

$$H^{t'}_{0,\underline{\lambda}^t} : R^{\leq t'}_{\text{gap}}(\tau_{\underline{\lambda}^t}) > \alpha \;\text{ for }\; t' = t, \ldots, t_{\max}. \tag{10}$$

Observe that $H_{0,\underline{\lambda}^t}$ in (9) is true if and only if there exists $t' \geq t$ such that $H^{t'}_{0,\underline{\lambda}^t}$ is true. Observe also that, by construction, $\tau_{\underline{\lambda}^t}$ cannot stop at timesteps smaller than t, and thus t_0 in (9) satisfies $t_0 \geq t$. Importantly, the formulation of the finer nulls in (10) paves the way to test the individual hypothesis $H_{0,\underline{\lambda}^t}$. Specifically, it implies that we can reject the individual hypothesis $H_{0,\underline{\lambda}^t}$ if all the finer hypotheses $H^{t'}_{0,\underline{\lambda}^t}$, $t' \geq t$ are rejected. This amounts to computing a p-value $\hat{p}^{t'}_{\underline{\lambda}^t}$ for each finer hypothesis $H^{t'}_{0,\underline{\lambda}^t}$ and rejecting $H_{0,\underline{\lambda}^t}$ if $\hat{p}^{t'}_{\underline{\lambda}^t} \leq \delta$ for all $t' \geq t$. Put simply, we reject $H_{0,\underline{\lambda}^t}$ if $\hat{p}_{\underline{\lambda}^t} = \max\{\hat{p}^{t'}_{\underline{\lambda}^t} : t' \geq t\} \leq \delta$.

Algorithm 2 summarizes the proposed testing procedure. The outer loop in this algorithm sequentially iterates over the hypotheses $H_{0,\underline{\lambda}^t}$ from $\underline{\lambda}^{t_{\max}}$ to $\underline{\lambda}^1$. The inner loop tests the null $H_{0,\underline{\lambda}^t}$ under study by breaking it into the finer hypotheses $H^{t'}_{0,\underline{\lambda}^t}$, $t' \geq t$. This algorithm returns the configuration $\hat{\underline{\lambda}} = \underline{\lambda}^t$ corresponding to the smallest t in which $H_{0,\underline{\lambda}^t}$ was rejected. The complexity of Algorithm (2) is $\mathcal{O}(t^2_{\max} \cdot |\mathcal{D}_{\text{cal-2}}|)$, and the validity of the resulting stopping rule $\tau_{\hat{\underline{\lambda}}}$ is as follows.

Proposition 2. *Assuming the calibration and test samples are i.i.d., with $\hat{\underline{\lambda}}$ selected as outlined in Algorithm 2, the stopping rule $\tau_{\hat{\underline{\lambda}}}(X)$ satisfies (5).*

Similarly to Proposition 1, the above result states that Algorithm (2) achieves a finite-sample, distribution-free risk control. But, in contrast with Proposition 1, here we control a stronger notion of error—the conditional accuracy gap.

5. Experiments

In this section, we evaluate the proposed methods both on structured time series datasets that are widely used in the ETSC literature and on the multiple-choice answering task, which was introduced in Section 1.1. The performance metrics include the conditional $R^{\leq t}_{\text{gap}}(\hat{\tau})$ and marginal $R^{\text{marginal}}_{\text{gap}}(\hat{\tau}) = R^{\leq t_{\max}}_{\text{gap}}(\hat{\tau})$ accuracy gap, evaluated on unseen test data $\mathcal{D}_{\text{test}}$. We also report the gain in early stopping, defined as the average normalized halt time: $T_{\text{avg}} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{X_i \in \mathcal{D}_{\text{test}}} \frac{\hat{\tau}(X_i)}{t_{\max}}$. In all experiments, we set the target accuracy gap level to $\alpha = 10\%$, with $\delta = 1\%$ and $\Delta = 0.01$. Throughout this section, the marginal method can be thought of as a baseline, as it closely resembles the calibration procedure suggested by Schuster et al. (2022) to control the accuracy gap for early exit in transformers.

5.1. Application to Structured Data

In this subsection, we test the applicability of our methods on five datasets: `Tiselac` (Ienco, 2017), `ElectricDevices` (Chen et al., 2015), `PenDigits` (Alpaydin & Alimoglu, 1998), `Crop` (Tan et al., 2017), and `WalkingSittingStanding` (Reyes-Ortiz et al., 2012). These datasets are publicly available via the aeon toolkit. We refer to these as structured datasets as $X \in \mathbb{R}^{t_{\max} \times d}$ and $X^t \in \mathbb{R}^d$. See Table C.2 in Appendix C.1 for more details.

To implement and evaluate our methods, we partition each dataset into four distinct sets: 80% of the samples are allocated for model fitting, while the remaining samples are equally divided to form $\mathcal{D}_{\text{cal-1}}$, $\mathcal{D}_{\text{cal-2}}$, and $\mathcal{D}_{\text{test}}$. For the marginal method, we set $\mathcal{D}_{\text{cal}} = \mathcal{D}_{\text{cal-1}} \cup \mathcal{D}_{\text{cal-2}}$. In all experiments, we employ an LSTM model (Hochreiter & Schmidhuber, 1997) as the base sequential classifier. A detailed description of the model architecture and training strategy is provided in Appendix C.2.

The results obtained by the marginal and conditional methods are summarized in Table 1; see Appendix C.3 for more detailed results for each dataset. Following this table, the two methods control the marginal accuracy gap, supporting our theory. However, the marginal method fails to control the conditional risk for sequences with early halt times, in contrast with the conditional approach that attains valid risk control over the accumulated halt times—as guaranteed by our theory. The statistical efficiency of both methods is comparable, as evidenced by the average normalized halt time

Table 1. **Summary of performance metrics for the proposed marginal and conditional methods across all structured datasets**. Results are presented for a nominal accuracy gap of $\alpha = 10\%$ and $\delta = 1\%$. The table provides the accumulated accuracy gap over the 20% and 50% earliest stopping times determined by $\hat{\tau}$ for each method, along with the marginal accuracy gap. The rightmost column presents the average normalized stopping time. All performance metrics are averaged over 100 random calibration/test splits. All standard errors are less than 0.008 and thus omitted.

Dataset	Late Acc.	Method	Early Acc.	Acc. Gap for Earliest $\hat{\tau}(X)$			T_{avg}
				20% earliest	50% earliest	Marginal	
Tiselac	0.816	Marginal	0.757	0.081	0.084	0.093	0.209
		Conditional	0.771	0.064	0.065	0.085	0.215
ElectricDevices	0.873	Marginal	0.809	0.117	0.108	0.079	0.471
		Conditional	0.825	0.030	0.031	0.075	0.552
PenDigits	0.989	Marginal	0.912	0.086	0.090	0.080	0.446
		Conditional	0.940	0.049	0.050	0.051	0.567
Crop	0.673	Marginal	0.608	0.171	0.135	0.086	0.580
		Conditional	0.642	0.057	0.063	0.079	0.450
WalkingSittingStanding	0.962	Marginal	0.884	0.004	0.054	0.079	0.125
		Conditional	0.901	0.033	0.039	0.067	0.061

T_{avg} performance metric. In fact, although the conditional method controls a stronger notion of error, it resulted in a smaller average normalized stopping time T_{avg} in 2 out of 5 datasets. We attribute this gain to our decision to employ a vector of thresholds to form the conditional stopping rule, as opposed to the single threshold used in the baseline marginal approach. Lastly, Figure C.4 in Appendix C.3 illustrates the trade-off between the tolerable accuracy gap α and the average stopping time for the `Tiselac` dataset. There, one can see that the conditional method allows for earlier stopping times when a higher accuracy gap is permitted.

5.2. An NLP Application

We now revisit the reading comprehension task introduced in Section 1.1, where the goal is to select the correct answer from a set of four options based on a given context. To allow the sequential processing of the data, we first divide the context of each question into sentences. These sentences are then grouped into $t_{\max} = 10$ sets. When the total number of sentences cannot be grouped into 10 equally sized sets, we include the remaining sentences in the last set. To formulate the input sequence $X^{\leq t}$, we construct a prompt that includes the context sentences up to timestep t, along with the question and its four options, labeled 'A', 'B', 'C', and 'D'. The prompt concludes with "`The answer is:\n\n`", which is then fed to the Vicuna-13B model (Zheng et al., 2023) to make a prediction; the model is accessible via HuggingFace. We employ the vLLM framework (Kwon et al., 2023) to compute the probability assigned to each of the four options, resulting in $\hat{f}(X^{\leq t}) \in [0,1]^4$. Lastly, we define the function $\hat{\pi}(X^{\leq t}) = \max\{\hat{f}_k(X^{\leq t}) : k = 1, \ldots, 4\}$, which is utilized to formulate the stopping rule $\hat{\tau}$.

The results obtained by the marginal and conditional methods are presented in Figure 2. As portrayed in the left panel, the conditional approach rigorously controls the conditional accuracy gap on the accumulated halt times, in contrast with the marginal method that merely controls the marginal risk. The right panel in Figure 2 shows that the marginal method tends to stop earlier. This is also indicated by its lower average normalized halt time of 0.483 compared to 0.831 for the conditional method. However, this gain is not necessarily desired, as the marginal approach tends to make errors in the early halt times.

Figure E.6 in the Appendix presents an ablation study, underscoring the importance of the testing phase (Stage 2) of the conditional method. As illustrated, the candidate configuration $\hat{\underline{\eta}}$ obtained by the greedy candidate screening algorithm (Stage 1) does not provide rigorous control of the conditional accuracy gap in the sense of (5). This stands in contrast with the conditional method that includes the testing stage. Nevertheless, the candidate $\hat{\underline{\eta}}$ provides a reasonable initial set of configurations for the hyperparameters to be tested, as it yields a stopping rule that roughly centers around the nominal accuracy gap level α.

In Appendix F we provide additional results with the `QuAIL` dataset (Rogers et al., 2020) and a predictive model that is based on Llama 2 70B (Touvron et al., 2023; Lee et al., 2023a), accessible via HuggingFace. This demonstrates the flexibility of our method.

6. Conclusion

In this paper, we presented a novel statistical framework that rigorously controls the accuracy gap conditional on the accumulated halt times. Additionally, we performed a series of numerical experiments that highlight the significance of transitioning from marginal to conditional guarantees, which validates our theory and underscores the practical implications of our proposal.

Our work opens several future research directions. To start with, it would be intriguing to design more effective stopping rules at the cost of increasing the computational complexity of the proposed two-stage calibration procedure. Another direction is to address the limitation of our work—the reliance on the i.i.d. assumption, which may be violated in practice. It would be illuminating to extend the tools we presented and relax this assumption, possibly by relying on the foundations of Barber et al. (2023). It would also be interesting to explore relaxations of the simultaneous accuracy gap guarantee as a way to obtain earlier halt times, e.g., by designing methods with a pointwise accuracy gap guarantee. Another natural progression is to extend the tools we developed to regression problems (Ye et al., 2023). The challenge here is that the loss L_{gap} might not be binary, and thus the exact p-value we used in this paper would not be applicable. More broadly, while this paper focuses on the accuracy and earliness trade-off, our approach can also be used to further control the accuracy and computational complexity trade-off. In this scenario, an early exit mechanism could be applied not only across the time horizon but also across the transformer layers (Schuster et al., 2022). Consequently, determining the threshold parameters would involve considering these two dimensions (time and network depth), calling for the design of an algorithm that optimizes them concurrently, followed by a specialized statistical procedure that tests hypotheses in a strategic order. Pressing this point, given the potential for computational savings along two axes, this novel setup requires assessing, at each test step, which of the two axes yields greater computational benefits. We intend to explore this interplay in future research.

Impact Statement

The methods developed in this paper focus on applications demanding early predictions given sequential data. The proposed methods aim to enhance the reliability of machine learning algorithms in this context, by supporting their predictions with rigorous risk-controlling guarantees. While such guarantees are desired in high-stakes applications, it is crucial to emphasize that the validity of the methods we offer holds under the i.i.d. assumption, which may not be satisfied in practice. Therefore, it is crucial to treat the data and problem at hand with care, especially if it may have social or ethical implications. More generally, the goal of our work is to advance the field of reliable machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

Y. R. and L. R. were supported by the Israel Science Foundation (grant No. 729/21). Y. R. thanks the Career Advancement Fellowship, Technion, for providing research support. Y. R. also thanks Verily for the generous financial support. Y.R. and L.R. thank Anastasios Angelopoulos for his insightful comments and discussions.

References

Alpaydin, E. and Alimoglu, F. Pen-based recognition of handwritten digits. UCI Machine Learning Repository, 1998.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Bauer, P. Multiple testing in clinical trials. *Statistics in medicine*, 10(6):871–890, 1991.

Brown, L. D., Cai, T. T., and DasGupta, A. Interval estimation for a binomial proportion. *Statistical science*, 16(2): 101–133, 2001.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pp. 1–22, 2023.

Chen, H., Zhang, Y., Tian, A., Hou, Y., Ma, C., and Zhou, S. Decoupled early time series classification using varied-length feature augmentation and gradient projection technique. *Entropy*, 24(10):1477, 2022.

Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., and Batista, G. The UCR time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

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*, 2020.

Ebihara, A. F., Miyagawa, T., Sakurai, K., and Imaoka, H. Toward asymptotic optimality: Sequential unsupervised regression of density ratio for early classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Feldman, S., Ringel, L., Bates, S., and Romano, Y. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Fisch, A., Jaakkola, T. S., and Barzilay, R. Calibrated selective classification. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Ghodrati, A., Bejnordi, B. E., and Habibian, A. Frameexit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15608–15618, 2021.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

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

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*, pp. 101–110, 2019.

Hartvigsen, T., Gerych, W., Thadajarassiri, J., Kong, X., and Rundensteiner, E. Stop&hop: Early classification of irregular time series. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 696–705, 2022.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Ienco, D. Tiselac: time series land cover classification challenge, 2017. https://www.timeseriesclassification.com/description.php?Dataset=Tiselac.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Laufer-Goldshtein, B., Fisch, A., Barzilay, R., and Jaakkola, T. S. Efficiently controlling multiple risks with pareto testing. In *International Conference on Learning Representations*, 2022.

Laufer-Goldshtein, B., Fisch, A., Barzilay, R., and Jaakkola, T. Risk-controlling model selection via guided bayesian optimization. *arXiv preprint arXiv:2312.01692*, 2023.

Lee, A. N., Hunter, C. J., and Ruiz, N. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*, 2023a.

Lee, D., Huang, X., Hassani, H., and Dobriban, E. T-cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72, 2023b.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 2014.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction intervals with temporal dependence. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Miller, R. *Simultaneous Statistical Inference*. Springer Series in Statistics. Springer New York, 2012.

Miyagawa, T. and Ebihara, A. F. The power of log-sum-exp: Sequential density ratio matrix estimation for speed-accuracy optimization. In *International Conference on Machine Learning*, pp. 7792–7804. PMLR, 2021.

Pang, R. Y., Parrish, A., Joshi, N., Nangia, N., Phang, J., Chen, A., Padmakumar, V., Ma, J., Thompson, J., He, H., and Bowman, S. R. QuALITY: Question answering with long input texts, yes! In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358, 2022.

Papadopoulos, H. and Haralambous, H. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., and Parra, X. Human activity recognition using smartphones. UCI Machine Learning Repository, 2012.

Rogers, A., Kovaleva, O., Downey, M., and Rumshisky, A. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *AAAI conference on artificial intelligence*, volume 34, pp. 8722–8731, 2020.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Sabet, A., Hare, J., Al-Hashimi, B., and Merrett, G. V. Temporal early exits for efficient video object detection. *arXiv preprint arXiv:2106.11208*, 2021.

Schuster, T., Fisch, A., Jaakkola, T., and Barzilay, R. Consistent accelerated inference via confident adaptive transformers. In *Conference on Empirical Methods in Natural Language Processing*, pp. 4962–4979, 2021.

Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V., Tay, Y., and Metzler, D. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472, 2022.

Shekhar, S., Eswaran, D., Hooi, B., Elmer, J., Faloutsos, C., and Akoglu, L. Benefit-aware early prediction of health outcomes on multivariate EEG time series. *Journal of Biomedical Informatics*, 139:104296, 2023.

Tan, C. W., Webb, G. I., and Petitjean, F. Indexing and classifying gigabytes of time series under time warping. In *SIAM international conference on data mining*, pp. 282–290. SIAM, 2017.

Tang, R., Kumar, K., Xin, J., Vyas, P., Li, W., Yang, G., Mao, Y., Murray, C., and Lin, J. Temporal early exiting for streaming speech commands recognition. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 7567–7571. IEEE, 2022.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Ye, C. T., Han, J., Liu, K., Angelopoulos, A., Griffith, L., Monakhova, K., and You, S. Learned, uncertainty-driven adaptive acquisition for photon-efficient multiphoton microscopy. *arXiv preprint arXiv:2310.16102*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

A. Marginal Risk Control Algorithm

Algorithm A.3 presents the marginal method described in Section 3.

Algorithm A.3 Fixed sequence testing for marginal risk control

Input: Calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$, tolerable accuracy gap α, significance level δ, grid resolution Δ.
$\hat{\lambda} \leftarrow \infty$ // Use ∞ as a fallback if the first null is not rejected.
$\lambda \leftarrow 1$ // Start testing with the largest $\lambda \in \Lambda$.
while $\lambda \geq 0$ **do**
  $\hat{R}_{\text{gap}} \leftarrow \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} L_{\text{gap}}(Y_i, \hat{Y}_i^{\text{full}}, \hat{Y}_i^{\text{early}}(\tau_\lambda))$ // Compute the empircal risk.
  $\hat{p} \leftarrow \text{CDF}_{\text{bin}}\left(\hat{R}_{\text{gap}} \cdot n_{\text{cal}}; n_{\text{cal}}, \alpha\right)$ // Compute a p-value.
  if $\hat{p} > \delta$ **then**
    break // Failed to reject the null, stop testing.
  end if
  $\hat{\lambda} \leftarrow \lambda$ // $H_{0,\lambda}$ was rejected, update the chosen $\hat{\lambda}$.
  $\lambda \leftarrow \lambda - \Delta$ // Next test will test a lower threshold.
end while
Output: $\hat{\lambda}$

Marginal risk control with adaptive thresholds We remark that it is possible to combine the candidate screening algorithm (stage 1) of the conditional method with Algorithm A.3 to obtain better thresholds. One way to implement such a method is to run stage 1 to find $\underline{\hat{\eta}}$ and then construct a sequence of increasingly aggressive stoppers, defined as $\underline{\hat{\eta}} + c_i$, where c_i is a decreasing sequence of constants. This approach can result in more powerful stopping rules and we leave this exploration for future work. However, it is crucial to emphasize that this approach is merely supported by a marginal guarantee, in striking contrast to our conditional testing approach.

B. Proofs

Proof of Proposition 1. The validity of the proposition is a direct consequence of using fixed sequence testing. For completeness, we add a proof that fixed sequence testing controls the FWER at level δ. Denote by $H_{0,j}$ the j-th ordered hypothesis. If all the hypotheses are false, we trivially get that $\mathbb{P}(V \geq 1) = 0$. Next, denote the index of the first true null by j_0, i.e., H_{0,j_0} is true and the preceding $H_{0,j'},\ j' < j_0$ are false. By the construction of the fixed sequence testing procedure, we may encounter this first true null only at step j_0 of the procedure. Now, observe that $\mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0) = 1 - \mathbb{P}(\hat{p}_{\lambda_{j_0}} > \delta) = \mathbb{P}(\hat{p}_{\lambda_{j_0}} \leq \delta) \leq \delta$. Above, the second equality holds since the testing procedure stops the first time that any p-value exceeds δ, and thus we get $V = 0$ if and only if $\hat{p}_{\lambda_{j_0}} > \delta$; under this event, the procedure would terminate without rejecting $H_{0,\lambda_{j_0}}$ and $H_{0,\lambda_{j'}},\ j' > j_0$. The last inequality follows from the validity of the p-value under the null (8). □

Proof of Proposition 2. To prove the result, it suffices to show that Algorithm 2 controls the FWER at level δ. First, observe that the outer loop in Algorithm 2 tests the hypotheses $H_{0,\underline{\lambda}^t}$ sequentially, starting from $\underline{\lambda}^{t_{\max}}$ down to $\underline{\lambda}^1$. As such, it follows the protocol of fixed sequence testing for FWER control. Second, each of the p-values $\hat{p}^{t'}_{\underline{\lambda}^t}$, corresponding to the finer null hypotheses (10), are valid since they are calculated using i.i.d. samples from the distribution $P_{XY|\hat{\tau}(X)\leq t'}$. Third, the max p-value $\hat{p}_{\underline{\lambda}^t} = \max\{\hat{p}^{t'}_{\underline{\lambda}^t} : t' \geq t\}$ used to test each of $H_{0,\underline{\lambda}^t}$ satisfies $\mathbb{P}\left(\hat{p}_{\underline{\lambda}^t} \leq \delta \mid H_{0,\underline{\lambda}^t}\right) \leq \delta$ (Angelopoulos et al., 2021, Proposition 6). Combining these three arguments completes the proof. □

C. Further Details on Experiments with Structured Datasets

In Section 5.1 of the main manuscript, we introduce the structured datasets on which we applied our methods. Here, we provide more details on each dataset, elaborate on the model architecture and training strategy, and present additional results.

C.1. Datasets

Table C.2 provides more details on each dataset.

Table C.2. Summary of structured datasets.

Dataset	#Samples	#Timesteps	#Features	#Classes	Type
`Tiselac`	99687	23	10	9	Image
`ElectricDevices`	16637	96	1	7	Device
`PenDigits`	10992	8	2	10	Motion
`Crop`	24000	46	1	24	Image
`WalkingSittingStanding`	10299	206	3	6	Motion

C.2. Model Architecture and Training Strategy

We used a standard LSTM for feature extraction with one recurrent layer with a hidden size of 32, except for `WalkingSittingStanding` where we used 2 recurrent layers, each with a hidden size of 256. The output of the last recurrent layer is plugged to two fully connected classification heads, one for classifying the label $\hat{f}(X^{\leq t}) \in [0,1]^K$ and the other for estimating the confidence in the classification $\hat{\pi}(X^{\leq t}) \in [0,1]$. The loss $L_{\text{CE}}^{\hat{f}}$ for updating $\hat{f}$ is the cross-entropy, and the loss $L_{\text{BCE}}^{\hat{\pi}}$ for updating $\hat{\pi}$ is the binary cross-entropy. The whole network is trained to minimize $L_{\text{CE}}^{\hat{f}}(\hat{f}(X^{\leq t}), Y) + \gamma \cdot L_{\text{BCE}}^{\hat{\pi}}(\hat{\pi}(X^{\leq t}), B(X^{\leq t}))$, where the function $B \in \{0,1\}$ returns the value 1 if $\hat{f}(X^{\leq t})$ correctly predicts the label Y and zero otherwise. We set the hyperparameter γ to 0.2 in all experiments. We augment the data by fitting the model on all possible prefixes $X^{\leq t}$, $t = 1, \ldots, t_{\max}$. The optimizer used to minimize the objective function is Adam (Kingma & Ba, 2014), with a learning rate of 0.001, and a batch size of 64. We allocate $1/8$ of the training samples to a validation set and optimize the model on the remaining $7/8$ of the samples. Training continues until there is no improvement in the loss on the validation set for 30 epochs. The model with the best validation set loss is then saved.

C.3. Additional Results

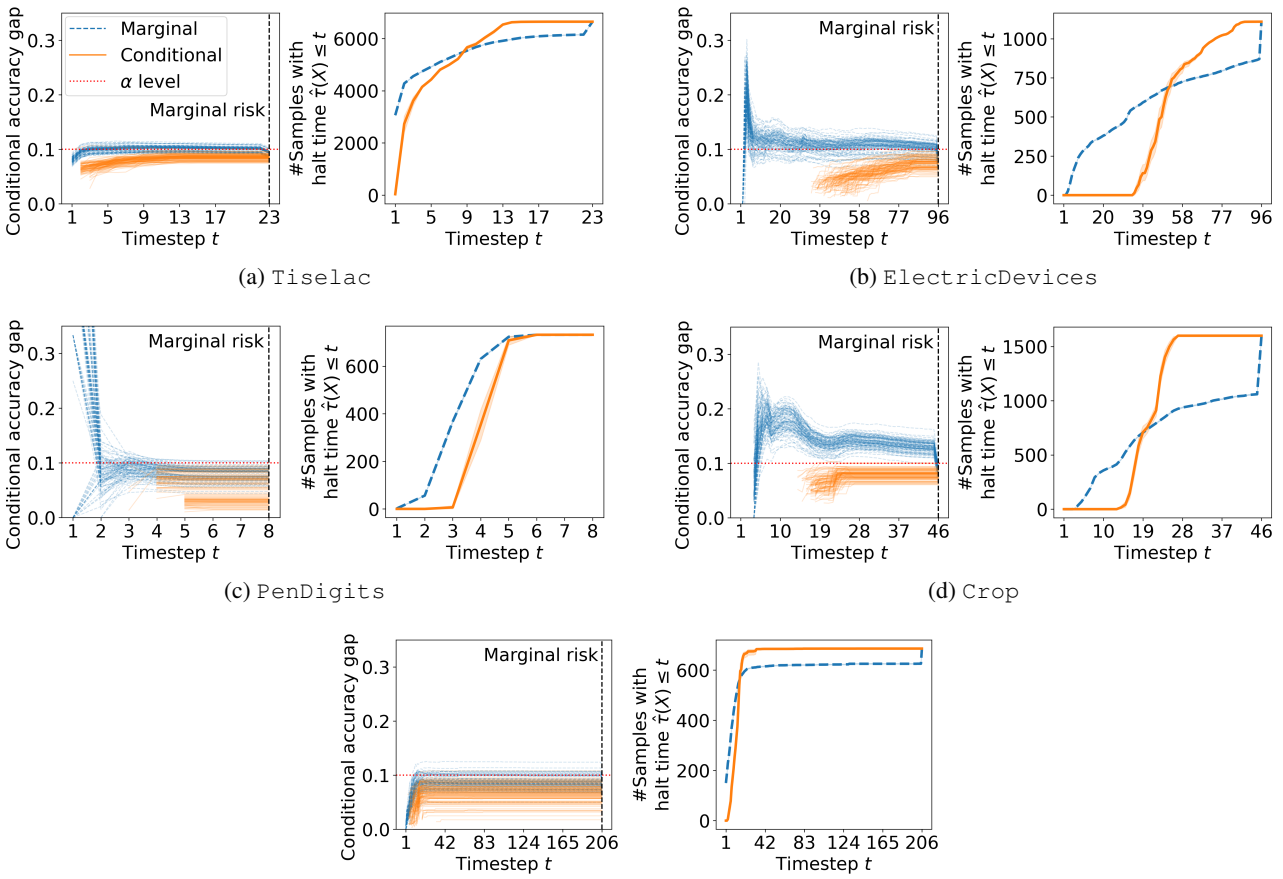
Table 1 from the main manuscript summarizes the performance of the marginal and conditional methods on the structured datasets for $\alpha = 10\%$.

In addition, Figure C.3 presents more detailed results, illustrating the accumulated accuracy gap and accumulated stopping times as a function of t obtained by the marginal and conditional methods.

Figure C.4 shows how different error levels α affect the average halt times with the conditional method. As expected, when allowing for a higher level of risk, the calibration manages to identify thresholds that result in shorter halt times. To further illustrate this trade-off, in Table C.3 we also report the results for $\alpha = 5\%$ of all structured datasets.

Table C.3. **Summary of performance metrics for the proposed marginal and conditional methods across all structured datasets for** $\alpha = 5\%$. Other details are as in Table 1.

Dataset	**Late Acc.**	**Method**	**Early Acc.**	**Acc. Gap for Earliest $\hat{\tau}(X)$**			T_{avg}
				20% earliest	50% earliest	Marginal	
`Tiselac`	0.816	Marginal	0.793	0.041	0.051	0.044	0.390
		Conditional	0.806	0.026	0.031	0.042	0.435
`ElectricDevices`	0.873	Marginal	0.844	0.074	0.066	0.036	0.594
		Conditional	0.862	0.013	0.014	0.031	0.771
`PenDigits`	0.989	Marginal	0.957	0.049	0.044	0.034	0.555
		Conditional	0.968	0.023	0.024	0.024	0.690
`Crop`	0.673	Marginal	0.646	0.123	0.077	0.038	0.708
		Conditional	0.680	0.025	0.030	0.035	0.709
`WalkingSittingStanding`	0.962	Marginal	0.930	0.000	0.029	0.032	0.217
		Conditional	0.939	0.023	0.022	0.028	0.121



(a) `Tiselac` (b) `ElectricDevices`

(c) `PenDigits` (d) `Crop`

(e) `WalkingSittingStanding`

Figure C.3. **Comparison between the marginal and conditional methods for the structured datasets.** The other details are as in Figure 2.

D. The Trend of the Confidence Thresholds Over Time

We found empirically that the confidence thresholds exhibit a decreasing pattern, although it is frequently non-monotonic and varies across different datasets. Figure D.5 presents a series of graphs that illustrate the value of $\hat{\lambda}_t$ as a function of t for all the datasets we studied.

E. Ablation Study on the NLP Application: `QuALITY` Dataset

In this section, we present an ablation study to assess the significance of the second stage in the conditional method: the testing phase. Figure E.6 summarizes the results discussed in Section 5.2 of the main manuscript.

F. Additional NLP Experiments: `QuAIL` Dataset

We present the results of an additional multiple-choice question answering dataset—`QuAIL` (Rogers et al., 2020). We used a different predictive model that is based on Llama 2 70B (Touvron et al., 2023; Lee et al., 2023a), accessible via HuggingFace. Since the contexts in this dataset are relatively short, we split the sentences into 3 timesteps. We use a total of 10000 data points, where 2/3 are used for calibration and the rest for testing. The left panel in Figure F.7 illustrates the empirical conditional accuracy gap as a function of $t = 1, 2, 3$. As depicted in the figure, the accuracy gap of the marginal approach is larger than $\alpha = 10\%$ for the early timesteps, in contrast with the conditional method that controls the accumulated accuracy gap across all three timesteps. Additionally, the right panel of the same figure demonstrates that both the marginal and conditional methods attain meaningful halt times.

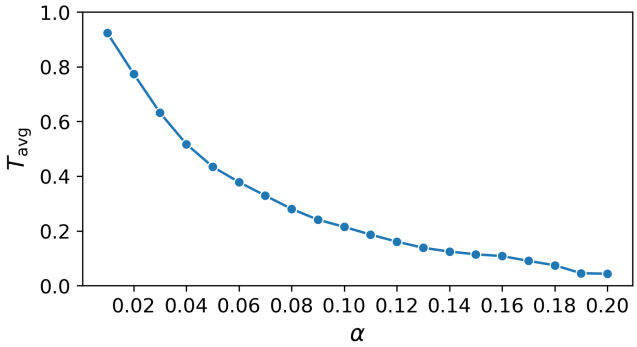
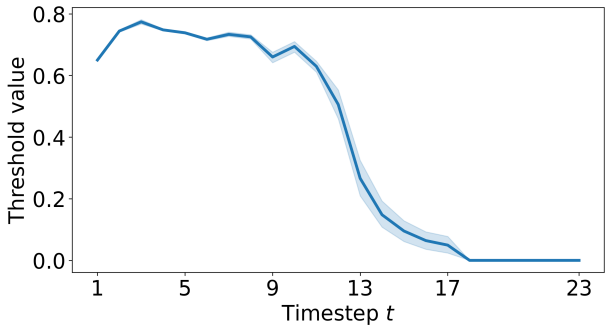
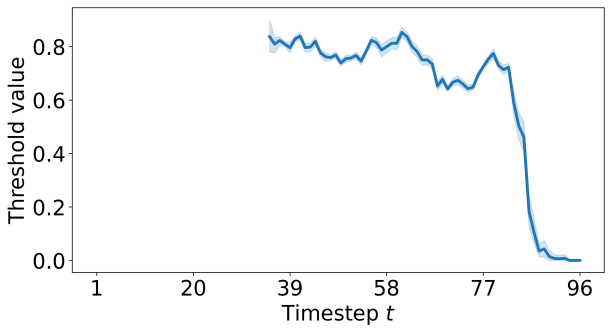


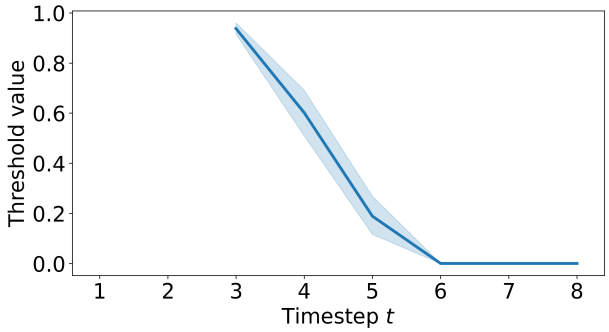
Figure C.4. **Normalized halt time T_{avg} vs. tolerable accuracy gap α.** The results are averaged over 100 random splits of the `Tiselac` dataset, with (tiny) standard error bars.



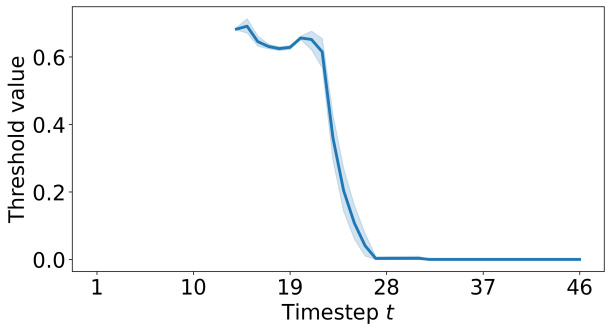
(a) `Tiselac`



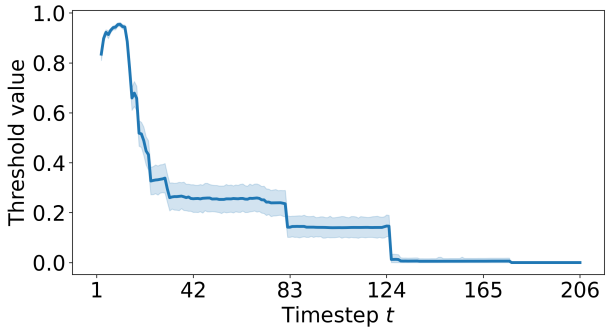
(b) `ElectricDevices`



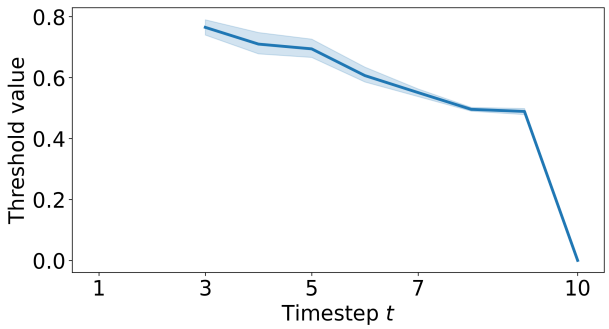
(c) `PenDigits`



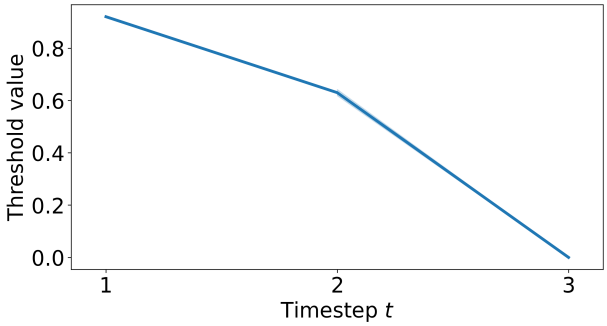
(d) `Crop`



(e) `WalkingSittingStanding`



(f) `QuALITY`



(g) `QuAIL`

Figure D.5. **The trend of the confidence threshold $\hat{\lambda}_t$ as a function of the timestep t.** Each panel in the figure corresponds to a different dataset. The threshold values presented are averaged over 100 random splits, where the shaded area represents a 95% confidence interval. Infinite thresholds are not included in the calculation of the average value.

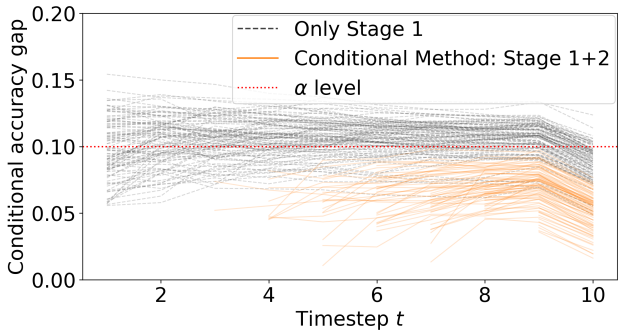


Figure E.6. **The importance of the testing procedure—Stage 2.** Comparison of conditional accuracy gap obtained by candidate screening (Stage 1, black curves) and by the full conditional method (Stage 1+2, orange curves). The results are presented for 100 random calibration/test splits of the `QuALITY` dataset, with each curve corresponding to a different split.

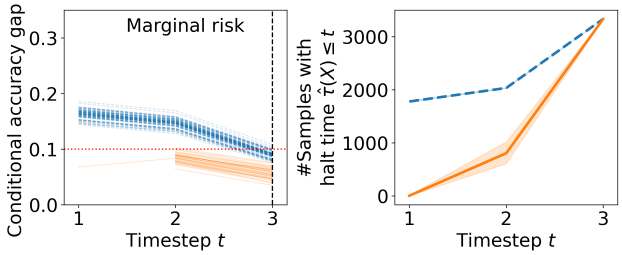


Figure F.7. **Comparison between the marginal and conditional methods for the `QuAIL` dataset.** The other details are as in Figure 2.