Minimax-Optimal Transformer Classification for Functional Data with a Dense–Sparse Phase Transition

Shuoyang Wang
University of Louisville
Louisville, KY 40208, USA
shuoyang.wang@louisville.edu

Yidan Tian
Michigan State University
East Lansing, MI 48824, USA
tianyid1@msu.edu

Guanqun Cao
Michigan State University
East Lansing, MI 48824, USA
caoguanq@msu.edu

Abstract

Despite the remarkable empirical success of transformer models in natural language processing, their theoretical foundations for functional data classification remain largely unexplored. This paper takes a first step toward closing this gap by developing a rigorous statistical framework for transformer-based functional classifiers. We show that a transformer architecture with an expanding attention window in the latent representation attains minimax-optimal excess-risk rates up to logarithmic factors in infinite-dimensional functional classification, without relying on conventional dimension-reduction procedures. Our analysis further reveals a dense-to-sparse phase transition: when the sampling frequency exceeds a critical threshold relative to the sample size, the proposed classifier achieves the optimal dense-observation rate. Otherwise, we precisely characterize the degradation in convergence under sparse sampling. Extensive simulations and real-data experiments support the theory and demonstrate that transformers provide a competitive and theoretically justified approach to functional classification.

1 Introduction

Functional data analysis (FDA) studies data arising from random functions or curves and has become increasingly important in applications such as growth trajectories, wearable-device signals, and brain and medical imaging [Wang et al., 2016]. Let $\{(Y_i, X_i(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S})\}_{i=1}^n$ denote labeled functional observations, where $X_i : \mathcal{S} \to \mathbb{R}$, $Y_i \in \{-1, 1\}$, $\mathcal{S} \subset [0,1]^q$, and $q \in \mathbb{N}^+$. The goal of *functional classification* is to learn a decision rule that accurately predicts the label of a new functional input. This problem is statistically challenging because the predictor is intrinsically infinite-dimensional and, in practice, is often observed only through discrete and possibly sparse measurements.

A large body of work on functional classification is built around functional principal component analysis (FPCA) and related dimension-reduction techniques [Shin, 2008, Delaigle et al., 2012, Delaigle and Hall, 2012, 2013, Galeano et al., 2015, Dai et al., 2017, Berrendero et al., 2018, Park et al., 2020, Adrover et al., 2004]. See Wang et al. [2024] for a comprehensive review. These methods are effective when the classification signal concentrates in a small number of leading functional components. They become fundamentally restrictive, however, when the Bayes decision boundary depends on complex nonlinear structure, localized variation, or diffuse contributions distributed across *infinitely many* functional directions, a situation that arises naturally in high-resolution imaging, dense EEG recordings, and longitudinal biomarker data. Recent deep-learning approaches have

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison with related work. Our FT classifier is the only method in the table with minimax guarantees in both the dense and sparse observation regimes.

Method	Nonparametric setting	Functional classification	Minimax (dense)	Minimax (sparse)
Classic functional classifiers (see details in Wang et al. [2024])	✗	✓	✗	✗
Transformer regression [Takakura and Suzuki, 2023]	✓	✗	✗	✗
FDNN [Wang et al., 2023]	✓	✓	✓	✗
Proposed FT classifier (ours)	✓	✓	✓	✓

broadened the scope of FDA, especially for multi-dimensional functional inputs, and have begun to establish rigorous statistical guarantees [Wang et al., 2023]. In particular, Wang et al. [2023] developed the functional deep neural network (FDNN) and established minimax excess-risk theory for multidimensional functional classification in the densely observed regime. Their analysis, however, rests on a *finite truncation assumption*: the decision boundary is assumed to depend on only finitely many functional projection scores, thereby effectively reducing the problem to a finite-dimensional one and introducing a tuning parameter, namely the truncation level, that must be specified in advance.

In many scientific applications this assumption is too strong. The Bayes decision boundary may aggregate information from infinitely many functional directions, with gradually decaying contributions. This motivates the following open question: *Can one construct a classifier for functional data that operates on an infinite-dimensional latent representation, accommodates Bayes decision boundaries depending on infinitely many latent directions, and still attains minimax-optimal excess-risk guarantees under both dense and sparse regimes?*

We answer this question affirmatively through a **functional transformer (FT) classifier** based on the self-attention mechanism [Vaswani et al., 2017]. The FT architecture is well suited to this problem for three related reasons. First, the self-attention module provides a data-adaptive mechanism for aggregating information across the latent representation. Second, by operating on an infinite-dimensional latent representation, the FT framework can accommodate Bayes decision boundaries that depend on infinitely many latent directions under regular decay. Third, the learned attention and readout weights allow the model to assign heterogeneous importance across latent directions. While recent work has begun to study transformers with infinite-dimensional inputs for sequence-to-sequence *regression* [Takakura and Suzuki, 2023], the statistical theory for transformer-based *functional classification* remains essentially undeveloped.

Related work. The closest work at the intersection of transformers and functional data is SAND [Hong et al., 2024], which introduces a self-attention architecture for functional data *imputation* with approximation-theoretic network-size guarantees. Imputation and classification are fundamentally different statistical tasks, and SAND provides no excess-risk theory. Neural Operator Flows [Shi et al., 2024] extend transformers to infinite-dimensional functional *regression* via operator learning, but do not address classification or phase transitions under discrete sampling. On the theoretical side, Akyürek et al. [2023] established that transformers can implement regression algorithms in context with near-optimal predictive risk. Our paper provides the analogous minimax theory for *classification* with functional inputs, including a phase-transition analysis that is absent from all prior transformer theory. At the architecture level, Heinrichs et al. [2023] proposed shift-invariant functional neural networks for EEG classification, and Boschi et al. [2024] developed a feature-selection algorithm for high-dimensional longitudinal functional data. Neither establishes minimax excess-risk guarantees or addresses infinite-dimensional nonparametric decision boundaries. The FDNN classifier of Wang et al. [2023] is the closest statistical predecessor: it achieves minimax optimality in the dense regime, but under a finite truncation assumption and without theory for sparse observations. Table 1 summarizes the key gap. *The present paper is the first to establish near-sharp minimax excess-risk theory for transformer-based functional classification under both dense and sparse regimes, within a framework with an infinite-dimensional latent representation.*

Contributions. The main contributions of this paper are as follows.

- **Transformer-based framework for functional classification in an infinite-dimensional nonparametric setting.** We introduce a functional transformer classifier for binary classification with functional inputs. Our framework is designed for infinite-dimensional nonparametric classification

problems in which the Bayes decision boundary may depend on infinitely many functional directions. This moves beyond existing theory based on finitely many functional projection scores and provides a flexible mechanism for capturing multiscale dependence across an infinite-dimensional latent representation, with interaction range increasing through the architecture class.

- **Near-sharp minimax excess-risk theory.** We establish architecture-independent minimax lower bounds for functional classification with an infinite-sequence posterior and show that the proposed FT classifier attains these bounds up to logarithmic factors. These results provide, to our knowledge, the first rigorous minimax theory for transformer-based classification with functional inputs, and they show that self-attention can achieve near-optimal statistical performance in infinite-dimensional functional classification problems.
- **Minimax dense-to-sparse phase transition for functional classification.** We characterize how the number of measurements per subject, m, governs the attainable excess-risk rate. When m is sufficiently large relative to the sample size n, the classifier attains the same minimax rate as in the fully observed regime. When m is smaller, we quantify the unavoidable excess risk induced by discrete observation. This yields a unified minimax theory covering both dense and sparse functional observations, as illustrated in Figure 2.

Notation and mathematical conventions

For a matrix $\boldsymbol{A}$, let $\boldsymbol{A}_{j:}$ and $\boldsymbol{A}_{:j}$ denote its jth row and column, respectively. For $N \in \mathbb{N}^+$, write $[N] = \{1, \ldots, N\}$. Let $\mathbb{N}_0^\infty = \bigcup_{d=1}^\infty \mathbb{N}_0^d$ denote the set of infinite sequences with finitely many nonzero terms, where $\mathbb{N}_0^d = \{(l_1, \ldots, l_d, 0, 0, \ldots) : l_j \in \mathbb{N},\, j \le d\}$. For positive sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \lesssim b_n$ if $a_n/b_n \le c$ for some universal constant c and all large n, and $a_n \asymp b_n$ if $a_n \lesssim b_n \lesssim a_n$. We write $\|\cdot\|_\infty$ for the maximum-entry norm of a matrix or vector, $\|\cdot\|_0$ for the number of nonzero entries, $\|\cdot\|$ for the ℓ_2 vector norm, and $\|\cdot\|_2$ for the L^2 function norm. For $f_1, f_2 \in L^2(\mathcal{S})$ observed on a regular grid $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m\} \subset \mathcal{S}$ with mesh tending to zero, the empirical inner product is $\langle f_1, f_2\rangle_m := \frac{1}{m}\sum_{\ell=1}^m f_1(\boldsymbol{s}_\ell) f_2(\boldsymbol{s}_\ell)$, which converges to $\langle f_1, f_2\rangle = \int_{\mathcal{S}} f_1 f_2\, d\boldsymbol{s}$ as $m \to \infty$. Define σ_s by $\sigma_s(x) = 1/(1+e^{-x}), x \in \mathbb{R}$.

2 Functional transformer classifier

We draw i.i.d. samples $\{(Y_i, X_i(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S})\}_{i=1}^n$ from Q, where Q is the joint distribution function. For any generic classifier $\widehat{C}$, the *excess misclassification risk* is assessed by $\mathcal{E}_Q(\widehat{C}) \equiv \mathbb{E}\Big[R_Q(\widehat{C}) - R_Q(C^{Q*})\Big]$, where $R_Q(\widehat{C}) = \mathbb{P}_Q\left(\widehat{C}(X) \neq Y\right)$ and C^{Q*} denotes the Bayes classifier, such that

$$C^{Q*}\big(X(\boldsymbol{s})\big) = \begin{cases} 1, & \mathbb{P}\big(Y^Q = 1 \mid X^Q(\boldsymbol{s})\big) \ge \mathbb{P}\big(Y^Q = -1 \mid X^Q(\boldsymbol{s})\big), \\ -1, & \text{otherwise.} \end{cases}$$

The goal is to construct a classifier that minimizes the worst-case excess risk over a nonparametric class $\mathcal{Q}$ of joint distributions. See Section 3 for the precise definition.

2.1 Architecture

We introduce the FT classifier through three building blocks assembled in sequence: a *functional embedding layer* that converts a discretely observed function into an infinite-dimensional latent token sequence, a *self-attention layer* that refines the tokens by capturing long-range interactions across the sequence, and a *sparse neural network* that maps each refined token to a scalar score. A weighted readout then aggregates the scores into a final classification decision. The full pipeline is illustrated in Figure 1. The proposed architecture is designed to address two primary challenges in functional classification: the infinite-dimensional nature of functional inputs, necessitating a structured embedding into a latent domain, and the presence of complex multiscale interactions, requiring an attention-based aggregation mechanism followed by flexible nonlinear modeling.

Neural network class. First, we introduce the sparse neural network class $\mathcal{NN}(L, P, S, B)$. Let σ denote the rectified linear unit (ReLU) activation function, defined as $\sigma(x) = (x)_+$ for $x \in \mathbb{R}$.

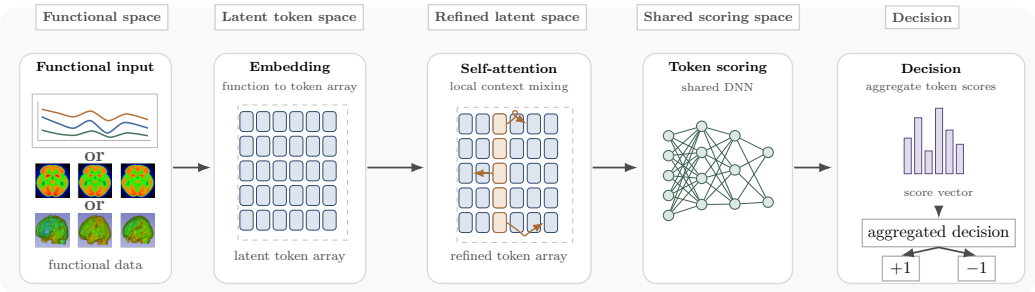


Fig. 1: Schematic illustration of the proposed functional transformer (FT) classifier.

For any two real vectors $\boldsymbol{u} = (u_1, \ldots, u_w)^\top$ and $\boldsymbol{v} = (v_1, \ldots, v_w)^\top$, define the shifted activation function $\sigma_{\boldsymbol{v}}(\boldsymbol{u}) = (\sigma(u_1 - v_1), \ldots, \sigma(u_w - v_w))^\top$. A feedforward neural network with depth $L \in \mathbb{N}^+$ and width $P \in \mathbb{N}^+$ is defined as

$$g(\boldsymbol{x}) = \mathbf{W}_L \sigma_{\boldsymbol{V}_L} \mathbf{W}_{L-1} \sigma_{\boldsymbol{V}_{L-1}} \ldots \mathbf{W}_2 \sigma_{\boldsymbol{V}_2} \mathbf{W}_1 \boldsymbol{x}, \tag{1}$$

where for $l \in [L]$, $\mathbf{W}_l \in \mathbb{R}^{p_{l+1} \times p_l}$ are weight matrices and $\boldsymbol{V}_l \in \mathbb{R}^{p_l}$ are shift vectors, with $\max_{\ell \in [L]} p_\ell \leq P$. The *sparse neural network class* is

$$\mathcal{NN}(L, P, S, B) \quad = \quad \left\{ g \text{ in (1)} : \sum_{l=1}^{L} \|\mathbf{W}_l\|_0 + \|\mathbf{V}_l\|_0 \leq S, \max_{1 \leq l \leq L} (\|\mathbf{W}_l\|_\infty \vee \|\mathbf{V}_l\|_\infty) \leq B \right\},$$

where S controls the total number of nonzero parameters and B bounds their magnitude.

Self-attention layer class. The self-attention layer is designed to model interactions within the latent representation of a functional observation. In contrast to standard feedforward layers, which act coordinatewise across representation indices, self-attention performs an adaptive, data-driven aggregation across indices. This mechanism permits adaptive interaction across latent indices while remaining analytically tractable. Let $\bar{\sigma}_s(\cdot)$ denote the softmax map applied to a finite-dimensional vector. Fix an embedding dimension $d \in \mathbb{N}^+$. For any $\boldsymbol{Z} = (\boldsymbol{Z}_{:1}, \boldsymbol{Z}_{:2}, \ldots) \in \mathbb{R}^{d \times \infty}$ and window size $U \in \mathbb{N}^+$, define the local window around index j by $\boldsymbol{Z}_{[j-U:j+U]} := \big(\boldsymbol{Z}_{:j-U}, \ldots, \boldsymbol{Z}_{:j+U}\big) \in \mathbb{R}^{d \times (2U+1)}$, with the natural modification near boundaries when applicable. The window is defined over latent indices. Under the architecture scaling in Assumptions 4 and 5, U increases with the target approximation accuracy and depends jointly on sample size and sampling frequency in the discretely observed regime. This expanding interaction range yields the near-sharp excess-risk bounds in Section 3. A multi-head self-attention layer $t : \mathbb{R}^{d \times \infty} \to \mathbb{R}^{d \times \infty}$ with $H \in \mathbb{N}^+$ heads is defined columnwise by

$$t(\boldsymbol{Z})_{:j} = \boldsymbol{Z}_{:j} + \sum_{h=1}^{H} V_h \, \boldsymbol{Z}_{[j-U:j+U]} \, \bar{\sigma}_s\Big(\big(K_h \boldsymbol{Z}_{[j-U:j+U]}\big)^\top \big(Q_h \boldsymbol{Z}_{:j}\big) \Big), \qquad j \in \mathbb{N}^+, \tag{2}$$

where, for each head $h \in [H]$, the matrices $Q_h \in \mathbb{R}^{d' \times d}$ and $K_h \in \mathbb{R}^{d' \times d}$ are the *query* and *key* transforms, respectively, and $V_h \in \mathbb{R}^{d \times d}$ is the *value* transform, with $d' \leq d$. The residual connection $\boldsymbol{Z}_{:j}$ promotes stability and facilitates approximation of near-identity mappings. The class of self-attention layers is defined as

$$\mathcal{A}(U, D, H, B) = \Big\{ t(\cdot) \text{ of the form (2)} : \; d \leq D, \; \max_{h \in [H]} \big(\|K_h\|_\infty \vee \|Q_h\|_\infty \vee \|V_h\|_\infty \big) \leq B \Big\}.$$

Functional embedding layer. While the functional observation $X(\boldsymbol{s})$ is defined on the physical domain $\boldsymbol{s} \in \mathcal{S}$, the transformer operates on a sequence indexed by representation order. The embedding layer maps the function from the physical domain into a latent representation domain that encodes global and multi-resolution characteristics of X, while preserving linear structure and enabling controlled complexity.

Given an orthonormal basis $\{\phi_j\}_{j\geq 1}$ of $L^2(\mathcal{S})$, it maps $X \in L^2(\mathcal{S})$ to an infinite token matrix $e(X) \in \mathbb{R}^{d\times\infty}$ in two stages: $e_1(X)_j = \sigma_s\big(\langle X, \phi_j\rangle\big), e_2(\boldsymbol{z})_{:j} = \boldsymbol{a}\, z_j + \boldsymbol{E}_{:j}$, $\boldsymbol{a} \in \mathbb{R}^d$, $\boldsymbol{E} \in \mathbb{R}^{d\times\infty}$, $d \leq D$, and $\|\boldsymbol{a}\|_\infty \vee \|\boldsymbol{E}\|_\infty \leq B$. The logistic activation maps the unbounded projection scores $\langle X, \phi_j\rangle \in \mathbb{R}$ onto $(0, 1)$, providing a canonically bounded infinite-dimensional representation without altering the relative ordering of the scores. The composed map $e = e_2 \circ e_1 : L^2(\mathcal{S}) \to \mathbb{R}^{d\times\infty}$ forms the input to the self-attention layer. In practice, the classifier is trained on the discretized observation $X^{(m)} = \big(X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_m)\big) \in \mathbb{R}^m$, and the implementation of $\mathcal{T}$ is obtained by replacing the population inner product $\langle\cdot,\cdot\rangle$ in the embedding layer with its empirical counterpart $\langle\cdot,\cdot\rangle_m$.

Functional transformer class. The embedding layer, self-attention module, and sparse neural network are integrated to define the functional transformer class, a construction that proceeds in two primary stages. First, a token-wise nonlinear scoring function is applied to the latent tokens refined by the self-attention mechanism. Second, a weighted readout operator aggregates the resulting infinite sequence, mapping the functional representation into a final scalar decision score.

Let $t \in \mathcal{A}(U, D, H, B)$ be a self-attention layer acting on the latent representation. For any $g \in \mathcal{NN}(L, P, S, B)$ with $g : \mathbb{R}^d \to \mathbb{R}$, define the token-wise score map

$$F : L^2(\mathcal{S}) \to \mathbb{R}^\infty, \qquad F\big(X(\boldsymbol{s})\big) := \Big(g\big((t\circ e)(X)_{:1}\big),\, g\big((t\circ e)(X)_{:2}\big),\, \ldots\Big),$$

that is, g is applied *column-wise* to the output of $t \circ e$, where $(t \circ e)(X)_{:j} \in \mathbb{R}^d$ denotes the jth latent token after attention. This step produces an infinite sequence of real-valued scores indexed by the latent representation order.

To obtain a scalar decision score, we apply a weighted readout to $F(X)$. Let $\boldsymbol{w} = (w_j)_{j\geq 1} \in \mathbb{R}^\infty$ and $v \in \mathbb{R}$, and define

$$f : L^2(\mathcal{S}) \to \mathbb{R}, \qquad f\big(X(\boldsymbol{s})\big) := \sum_{j=1}^{\infty} \omega_j\, F_j\big(X(\boldsymbol{s})\big) + \nu, \quad \max_{j\in\mathbb{N}^+} |\omega_j| \vee \nu \leq B,$$

whenever the series is well defined. Here $F_j(X)$ denotes the jth coordinate of $F(X)$. The induced classifier is then given by $C(X) = \operatorname{sign}\{f(X)\}$. We denote the resulting hypothesis class of decision functions as $\mathcal{T}(L, P, S, B, U, D, H)$, hereafter abbreviated as $\mathcal{T}$. The class $\mathcal{T}$ is defined through an infinite-dimensional latent representation at the population level. In practice, the readout is truncated to the first J tokens, for relatively large J. The computational truncation J is a sieve parameter for approximating the infinite-dimensional readout. With J retained tokens, $U \geq J - 1$ recovers full attention.

Classification. Given training data $\{(Y_i, X_i(\boldsymbol{s}))\}_{i=1}^n$, we fit the FT classifier by minimizing the empirical hinge loss, which is convex and well-suited to binary classification under the margin condition in Assumption 1 [Steinwart and Christmann, 2008]: $\widehat{f}_{\text{hinge}} = \arg\min_{f\in\mathcal{T}} \frac{1}{n}\sum_{i=1}^n \max\big(1 - Y_i f(X_i(\boldsymbol{s})),\, 0\big)$. The **FT classifier** is then $\widehat{C}^{\text{FT}}(X(\boldsymbol{s})) = \mathbb{I}\big\{\widehat{f}_{\text{hinge}}(X(\boldsymbol{s})) \geq 0\big\} - \mathbb{I}\big\{\widehat{f}_{\text{hinge}}(X(\boldsymbol{s})) < 0\big\}$. Algorithm 1 in the Appendix summarizes the practical training procedure of the proposed FT classifier. Starting from the functional representation, the algorithm applies the embedding map, local multi-head self-attention, token-wise nonlinear scoring, and a weighted readout to produce the decision score, and then optimizes the empirical hinge loss by backpropagation. To keep the presentation concise in the main text, we only provide the core training pipeline here.

3 Minimax excess-risk theory

The key task is to show that $\widehat{C}^{\text{FT}}$ attains the *minimax excess misclassification risk*, which is defined as $\sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widehat{C}) \asymp \inf_{\widetilde{C}} \sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widetilde{C})$, where $\mathcal{Q}$ is a nonparametric class defined below and the infimum is over all measurable classifiers trained on fully observed $X(\boldsymbol{s})$. We study the minimax excess-risk behavior of the proposed FT classifier over an infinite-dimensional nonparametric classification class. The theory depends on three quantities. The margin parameter α measures classification difficulty near the Bayes boundary, and the effective smoothness index β^* quantifies the complexity of the decision boundary. For discretely observed data, the sampling frequency m determines the reconstruction accuracy of the functional covariate.

Function class. Define the posterior class probability as $\eta(X(\boldsymbol{s})) = \mathbb{P}(Y = 1|X(\boldsymbol{s}))$, and the feature map

$$\mathcal{G} : L^2([0,1]^q) \to [0,1]^\infty, \qquad \mathcal{G}(X) := \left(\sigma_s\left(\langle X, \phi_\ell\rangle\right)\right)_{\ell\in\mathbb{N}^+},$$

which is measurable embedding onto its range. To model infinite-dimensional nonparametric decision boundaries, we assume that the posterior class probability factors through the latent representation $\mathcal{G}(X)$. Specifically, let

$$\widetilde{\mathcal{F}}_\gamma := \left\{g : L^2([0,1]^q) \to [0,1] \,:\, \exists\, f \in \mathcal{F}_\gamma,\ \text{s.t. } g(X) = f\big(\mathcal{G}(X)\big),\ \ \forall\, X \in L^2([0,1]^q)\right\},$$

where $\mathcal{F}_\gamma$ is a γ-smooth function space defined in Definition 2 in the Appendix. Intuitively, any decision rule for functional data induces a decision boundary on the infinite sequence space associated with the latent representation of the underlying functional input. We impose two assumptions on $\eta \in \widetilde{\mathcal{F}}_\gamma$.

Assumption 1. *(Functional margin condition) There exist constants $c_0 > 0$ and $\alpha \geq 0$, such that*

$$\mathbb{P}\left(|\eta(X) - 1/2| \leq x\right) \leq c_0 x^\alpha, \qquad \forall x > 0, \qquad X \in L^2(\mathcal{S}).$$

Assumption 2. *(Regularly varying γ-smoothness) Let $\eta \in \widetilde{\mathcal{F}}_\gamma$, where $\mathcal{F}_\gamma$ is a γ-smooth function space with coordinate-wise smoothness parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)^\intercal$ and $\beta_j \in \mathbb{R}^+$ for all $j \in \mathbb{N}^+$. Let $(\bar{\beta}_j)_{j\geq 1}$ denote the nondecreasing rearrangement of $(\beta_j)_{j\geq 1}$. For $\boldsymbol{w} = (w_j)_{j\in\mathbb{N}^+} \in \mathbb{N}_0^\infty$, assume that there exist constants $0 < c_- < c_+ < \infty$, such that one of the following two regimes holds.*

(i) Mixed smoothness. $\gamma(\boldsymbol{w}) = \sum_{j\geq 1} \beta_j w_j$, $c_-\, j^\theta \ \leq\ \beta_j \ \leq\ c_+\, j^\theta, \forall j \in \mathbb{N}^+, \theta > 0$.

(ii) Anisotropic smoothness. $\gamma(\boldsymbol{w}) = \max_{j\geq 1}\{\beta_j w_j\}$, $c_-\, j^\theta \ \leq\ \bar{\beta}_j \ \leq\ c_+\, j^\theta, \forall j \in \mathbb{N}^+, \theta > 1$.

Assumption 1 extends the traditional margin condition [Tsybakov, 2004] to functional inputs, with larger α indicating an easier problem. Assumption 2 allows the decision boundary to depend on infinitely many latent directions, with complexity governed by coordinate-wise smoothness and its growth across indices. The inverse smoothness index is defined by $\beta^* = \bar{\beta}_1$ for mixed smoothness and $\beta^* = \left(\sum_{j\geq 1} \bar{\beta}_j^{-1}\right)^{-1}$ for anisotropic smoothness.

3.1 Fully observed functional data

We first consider the fully observed regime, in which the functional covariate is available without discretization error ($m = \infty$). Let $\mathcal{Q} = \{Q : \pi \in (0,1), \eta \text{ satisfies both Assumptions 1 and 2}\}$, where $\pi = \mathbb{P}(Y_i = 1)$ is the prior distribution. Define $\underline{r_n^2} = n^{-\frac{(\alpha+1)\beta^*}{(\alpha+2)\beta^*+1}}$ and $r_n^2 = \underline{r_n^2}\,(\log n)^{(1+2/\theta+4/\theta^*)(\alpha+1)/(\alpha+2+1/\beta^*)}$. The following theorem gives the minimax lower bound over $\mathcal{Q}$ and shows that the upper bound can be achieved by the proposed FT classifier up to a logarithmic factor. The architecture parameters are allowed to scale with the target approximation level. See Assumption 4 in the Appendix.

Theorem 3.1 (Dense regime)**.** *Under Assumptions 1 and 2, there exists a universal constant $c_1 > 0$, such that*

$$\inf_{\widetilde{C}} \sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widetilde{C}) \geq c_1 \underline{r_n^2}.$$

Additionally, if the hypothesis class $\mathcal{T}$ satisfies Assumption 4 in the Appendix, there exists a universal constant $c_2 > 0$, such that

$$\sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widehat{C}^{FT}) \leq c_2 r_n^2.$$

Remark 3.2 (Rate interpretation)**.** *The gap between $\underline{r_n^2}$ and r_n^2 is purely logarithmic, so $\widehat{C}^{\mathrm{FT}}$ is minimax optimal up to a $\log n$ factor. The polynomial decay rate is governed by the interaction between the margin parameter α and the effective smoothness index β^*. This result allows the Bayes decision boundary to depend on infinitely many latent directions. The minimax-optimal rate $\underline{r_n^2}$ admits an interpretation analogous to that in Wang et al. [2023]: it is governed by the interaction between an effective smoothness term and an effective dimensionality term. In particular, larger effective smoothness relative to dimension corresponds to lower intrinsic complexity of the decision boundary and hence a faster minimax rate.*

3.2 Discretely observed functional data and phase transition

In practice, functional covariates are observed only on a finite grid. We therefore extend the analysis to discretely observed functional data, where each trajectory is sampled at m design points. The resulting statistical performance depends on how m scales relative to the sample size n. We assume the functional covariate $X_i(\boldsymbol{s})$ is observed at $\boldsymbol{s}_j = j/m$ for $j = 1, \ldots, m$, representing an evenly spaced grid on $[0,1]^q$. Let $\boldsymbol{X}_i^{(m)} = \big(X_i(\boldsymbol{s}_1), \ldots, X_i(\boldsymbol{s}_m)\big)$ denote the discretized observations for $i \in [n]$. Accordingly, in the discretely observed regime, the FT classifier is implemented through the empirical projection scores induced by $\langle \cdot, \cdot \rangle_m$, while the resulting excess-risk bound is analyzed relative to the population classifier defined on the underlying functional input $X(\boldsymbol{s})$. To model the regularity of discretely observed functional covariates and to control the reconstruction error induced by finite sampling, we introduce the Besov-type smoothness class $\mathcal{S}^\tau$.

Definition 1 (Besov-type Smoothness class $\mathcal{S}^\tau$)**.** *Let $q \geq 1$, $\tau > 0$, and $M > 0$. Let $\{\varphi_{j,k}\}_{j\geq 0,\, k\in\mathcal{K}_j}$ be a compactly supported orthonormal wavelet basis of $L^2([0,1]^q)$ with regularity strictly larger than τ. Define the Besov-type ball $\mathcal{S}^\tau \equiv \mathcal{S}^\tau(M) := \left\{ X \in L^2([0,1]^q) : \sup_{j\geq 0} 2^{j\tau} \left(\sum_{k\in\mathcal{K}_j} \big|\langle X, \varphi_{j,k}\rangle\big|^2 \right)^{1/2} \leq M \right\}$.*

Next, over the smoothness class $\mathcal{S}^\tau(M)$, we assume the existence of a reconstruction operator that recovers an $L^2([0,1]^q)$ trajectory from sampled functional data with uniformly controlled approximation error. This trajectory-level requirement is standard and is satisfied by many classical presmoothing or reconstruction procedures [Mallat, 2008, Brenner and Scott, 2008].

Assumption 3 (Reconstruction operator)**.** *There exists a measurable reconstruction operator $\mathcal{O}_m : \mathbb{R}^m \to L^2([0,1]^q)$ and a constant $C > 0$ independent of m such that*

$$\sup_{x\in\mathcal{S}^\tau} \Big\| x - \mathcal{O}_m\big(x(\boldsymbol{s}_1), \ldots, x(\boldsymbol{s}_m)\big) \Big\|_2 \leq C\, m^{-\tau}.$$

The following theorem establishes the minimax convergence rates under finite sampling frequency.

Theorem 3.3 (Sparse-to-dense phase transition)**.** *Suppose $X \in \mathcal{S}^\tau \subset L^2([0,1]^q)$. Under Assumptions 1 and 2, there exists a universal constant $c_3 > 0$, such that*

$$\inf_{\widetilde{C}_m} \sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widetilde{C}_m) \geq c_3 \left(\underline{r}_n^2 + m^{-\tau(\alpha+1)} \right),$$

where $\widetilde{C}_m$ is any generic classifier constructed by $\left\{ \left(Y_i, \boldsymbol{X}_i^{(m)}, X_i \in \mathcal{S}^r \right) \right\}_{i=1}^n$. Additionally, if Assumption 3 holds and the hypothesis class $\mathcal{T}$ satisfies Assumption 5 in the Appendix, there exists a universal constant $c_4 > 0$, such that the classifier $\widehat{C}^{FT} \in \mathcal{T}$ satisfies

$$\sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widehat{C}^{FT}) \leq c_4 \left(r_n^2 + m^{-\tau(\alpha+1)} \right).$$

Remark 3.4 (Phase transition thresholds)**.** *Theorem 3.3 shows that this discretization error is unavoidable and is added to the fully observed minimax rate $\underline{r}_n^2$. Balancing these two terms yields the minimax phase-transition threshold $\underline{m}^* \asymp n^{1/\tau(\alpha+2+1/\beta^*)}$. Thus, when $m \gtrsim \underline{m}^*$, the discretization effect is asymptotically negligible and the problem behaves as in the fully observed setting. When $m \lesssim \underline{m}^*$, the excess risk is dominated by the sampling resolution.*

For the FT classifier, balancing $m^{-\tau(\alpha+1)}$ with r_n^2 gives $m^* \asymp \underline{m}^* (\log n)^{-\delta}$ and $\delta = \frac{1+2/\theta+4/\theta^*}{\tau(\alpha+2+1/\beta^*)}$, so that $m^*/\underline{m}^* \asymp (\log n)^{-\delta}$. Thus, the transition threshold for the FT classifier differs from the minimax benchmark by a logarithmic factor, so the FT classifier is polynomially optimal but not threshold-sharp in the intermediate regime.

The phase-transition behavior is summarized in Figure 2. In the **sparse regime**, where $m \lesssim m^*$, both the minimax lower bound and the FT upper bound are dominated by the discretization term $m^{-\tau(\alpha+1)}$. In the **intermediate regime**, where $m^* \lesssim m \lesssim \underline{m}^*$, the minimax lower bound is still governed by discretization, whereas the FT upper bound is already controlled by r_n^2. In the **dense regime**, where $m \gtrsim \underline{m}^*$, discretization becomes negligible and the problem reduces to the fully observed case, so the FT classifier attains the fully observed rate.

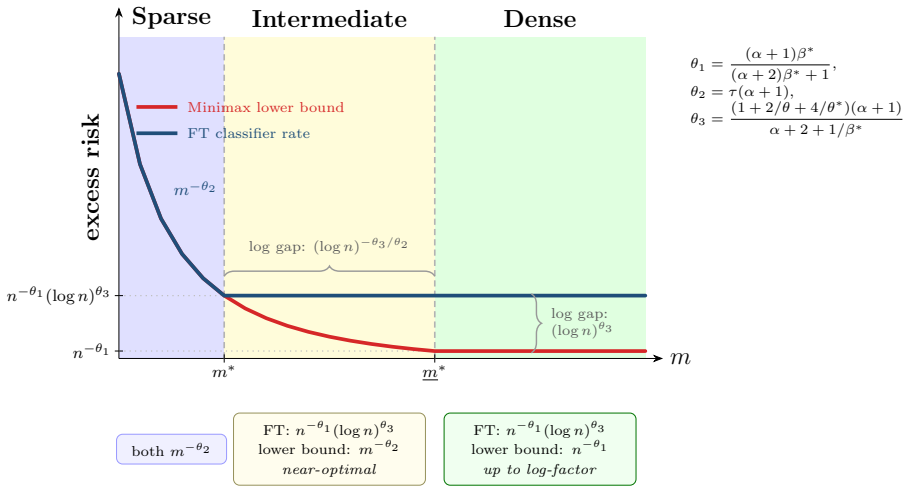


Fig. 2: Schematic illustration of the phase transition in excess risk for **fixed** n as m increases. The two transition thresholds are close, and their discrepancy is due only to a $\log n$ factor.

4 Simulation study

We compare the proposed FT classifier with several competitors, including FDNN [Wang et al., 2023], the vision transformer (ViT) [Dosovitskiy et al., 2020], and the convolutional neural network (CNN) [LeCun et al., 1998]. For 1D functional data, we additionally consider the quadratic discriminant classifier (QD) from Delaigle and Hall [2012, 2013] and the nonparametric Bayes classifier (NB) from Dai et al. [2017]. We study 1D, 2D, and 3D functional classification under six simulation settings. The full data-generating mechanisms are deferred to Section D in the Appendix.

In each setting, we generate n training functions and 500 testing functions for each class, where $n \in \{40, 100, 200, 400\}$ and the training samples are evenly split between the two classes. The reported results are based on 100 Monte Carlo replications. The 1D, 2D, and 3D functional observations are recorded on grids of size 50, 30×30, and $20 \times 20 \times 20$, respectively. We use the Fourier basis for 1D data and tensor-product Fourier bases for 2D and 3D data to extract projection scores. Table 2 reports the misclassification rates, with standard errors in parentheses, for all six settings.

FT attains the lowest misclassification rate in all 1D settings, improving on FDNN, QD and NB. In 2D, FT has the lowest error in three of four sample sizes for DGP3 and two of four for DGP4. In 3D, FT is competitive with ViT and CNN in DGP5 and attains the lowest error at every sample size in DGP6. These results show that FT performs well across the three functional dimensions, with its relative performance varying by setting and sample size.

Table 2: Misclassification rates (%) with standard errors in brackets for DGP1 to DGP6

n	DGP1				DGP2			
	FT	FDNN	QD	NB	FT	FDNN	QD	NB
40	**17.93(0.03)**	**24.68(0.02)**	38.58(0.02)	38.33(0.02)	**19.75(0.09)**	**36.71(0.08)**	39.99(0.01)	39.26(0.03)
100	**16.30(0.02)**	**22.88(0.02)**	37.91(0.02)	41.03(0.02)	**12.74(0.05)**	**32.20(0.07)**	38.42(0.09)	40.27(0.03)
200	**15.87(0.02)**	**22.05(0.02)**	37.35(0.02)	39.92(0.02)	**11.86(0.04)**	**30.39(0.05)**	42.63(0.02)	39.84(0.04)
400	**13.72(0.02)**	**15.07(0.02)**	36.75(0.02)	38.54(0.02)	**9.56(0.01)**	**20.40(0.05)**	43.98(0.09)	38.51(0.04)
n	DGP3				DGP4			
	FT	FDNN	ViT	CNN	FT	FDNN	ViT	CNN
40	**16.43(0.03)**	17.01(0.07)	20.23(0.03)	**16.38(0.03)**	**13.26(0.02)**	13.95(0.06)	**13.61(0.03)**	14.15(0.03)
100	**14.41(0.02)**	14.80(0.06)	15.67(0.02)	**14.50(0.02)**	**12.97(0.01)**	**12.73(0.04)**	13.00(0.02)	12.99(0.01)
200	**13.23(0.02)**	**13.92(0.05)**	14.38(0.02)	14.40(0.03)	**12.68(0.01)**	12.71(0.01)	12.50(0.01)	**12.47(0.01)**
400	**12.31(0.01)**	**12.74(0.04)**	13.40(0.01)	13.16(0.02)	**12.25(0.01)**	12.30(0.01)	12.36(0.01)	**12.28(0.01)**
n	DGP5				DGP6			
	FT	FDNN	ViT	CNN	FT	FDNN	ViT	CNN
40	**15.78(0.04)**	33.85(0.05)	**15.74(0.05)**	16.18(0.04)	**6.24(0.01)**	**6.34(0.01)**	8.05(0.03)	7.63(0.03)
100	**10.57(0.02)**	29.45(0.04)	**10.81(0.03)**	11.49(0.04)	**5.89(0.01)**	**6.02(0.01)**	6.87(0.02)	6.23(0.01)
200	**8.33(0.01)**	27.71(0.04)	8.74(0.02)	**8.14(0.01)**	**5.78(0.01)**	**5.98(0.01)**	6.88(0.02)	5.79(0.01)
400	**8.07(0.01)**	9.11(0.02)	8.13(0.01)	**7.92(0.02)**	**5.60(0.01)**	5.74(0.01)	7.24(0.01)	**5.61(0.01)**

Empirical examination of the phase transition and baseline comparisons. To examine the resolution dependence in Theorem 3.3, we vary the number of design points m over $\{10, 20, 30, 50\}$ in 1D, $\{10^2, 20^2, 30^2\}$ in 2D, and $\{10^3, 15^3, 20^3\}$ in 3D. The underlying process is fixed within each DGP, and each method is re-tuned at every resolution. FT has lower error at the highest resolution than at the lowest in 21 of 24 comparisons over $n \in \{40, 100, 200, 400\}$. The largest improvements occur in DGP1 and DGP2, while DGP6 is nearly insensitive to resolution. These results are qualitatively consistent with the resolution dependence predicted by the theorem. At $n = 400$, FT has the lowest high-resolution error in five of six DGPs, with CNN attaining a lower error in DGP5 (7.92% versus 8.07%). Tables 5 and 6 report the endpoint results, including FPCA+LDA and FPCA+SVM.

We also examine the choice of representation and attention window. Fourier-score FT has lower mean error than FPCA-score FT in all six DGPs at $n = 400$. At each of the 15 examined truncation levels J, the best tested local window has lower mean error than full attention. Tables 7 and 8 give the corresponding results.

5 Real data analysis

We further evaluate the proposed FT classifier on three real-data benchmarks representing 1D, 2D, and 3D functional classification. These include a speech recognition benchmark based on log-periodograms, a chest-radiograph dataset for normal versus abnormal classification, and an ADNI PET imaging task for distinguishing Alzheimer's disease (AD) from early mild cognitive impairment (EMCI). The detailed dataset descriptions and preprocessing steps are deferred to Section E in the Appendix. Table 3 summarizes the predictive and computational results under the original tuning configurations. We then report an expanded comparison with FPCA-based baselines and a broader FT validation grid for ADNI.

The speech benchmark provides the clearest overall advantage for FT. Across all three pairwise phoneme classification tasks, FT attains the lowest misclassification rate. The gain is most visible for `aa` versus `ao`, where FT reduces the error from 20.744% for the next best method (FDNN) to 19.130%. For `aa` versus `iy` and `ao` versus `iy`, the classification problem is nearly separable, and FT again achieves the smallest error, improving on the runner-up by 0.183% and 0.173%, respectively. In all three comparisons, FT clearly outperforms the classical 1D competitors QD and NB.

On the chest-radiograph task, FT achieves the best classification accuracy, with a misclassification rate of 22.135%, compared with 25.788% for FDNN, 25.404% for ViT, and 26.587% for CNN. Thus, FT improves on the strongest alternative by 3.269 percentage points. Computationally, FT remains in a similar range to FDNN, requiring 1.088 minutes and 83.336 MB on average versus 0.458 minutes and 18.503 MB for FDNN, while being substantially more efficient than both ViT and CNN, which require 1.626 minutes and 1898.058 MB, and 5.362 minutes and 3426.210 MB, respectively.

On the ADNI PET task, CNN attains the lowest misclassification rate at 10.314%, while FT is a close second at 11.611%. Still, FT markedly outperforms FDNN and ViT, reducing the error by 7.956 and 4.922 percentage points, respectively. Moreover, FT is far more efficient than CNN and ViT, requiring only 0.264 minutes and 87.055 MB on average, versus 0.850 minutes and 308.327 MB for ViT and 5.267 minutes and 1866.994 MB for CNN. Its efficiency is also comparable to that of FDNN, which requires 0.095 minutes and 19.364 MB, but FT delivers better accuracy. Thus, although CNN is slightly more accurate on this 3D task, FT offers a substantially better tradeoff between accuracy and computational cost.

Taken together, the real-data experiments reinforce the simulation results. FT performs best on the 1D speech benchmark and the 2D chest-radiograph task, and remains highly competitive on the 3D ADNI task while being substantially more computationally efficient than the image-based baselines.

Table 3: Real-data misclassification rates (%, mean with standard error in parentheses). QD and NB are included only for the 1D speech task.

Dataset	Task	FT	FDNN	ViT	CNN	QD	NB
Speech (1D)	aa vs ao	**19.130(0.028)**	20.744(0.016)	24.465(0.037)	23.700(0.025)	25.402(0.026)	25.378(0.021)
	aa vs iy	**0.010(0.001)**	0.193(0.002)	0.215(0.005)	0.330(0.004)	0.288(0.005)	0.273(0.006)
	ao vs iy	**0.010(0.001)**	0.183(0.004)	0.235(0.004)	0.310(0.004)	0.578(0.005)	0.232(0.005)
Chest radiographs (2D)	normal vs abnormal	**22.135(0.019)**	25.788(0.016)	25.404(0.017)	26.587(0.107)	—	—
ADNI PET (3D)	AD vs EMCI	**11.611(0.043)**	19.567(0.078)	16.533(0.074)	**10.314(0.043)**	—	—

Expanded real-data comparison. We extend the real-data comparison by adding FPCA+LDA and FPCA+SVM and broadening the FT validation grid for ADNI. The expanded grid was specified before examining the revised test results, with the same training, validation and test protocol. Within each replication, the Fourier truncation, token count, attention window, embedding dimension, number of heads and Transformer layers, and MLP width were selected using validation data. The selected configuration was then evaluated on the held-out test set. As shown in Table 4, FT attains a mean error of 10.017% on ADNI, a modest improvement of 0.297 percentage points over CNN. For the two nearly separable speech tasks, FPCA+SVM is the strongest alternative, with FT lower by 0.025 and 0.075 percentage points. Table 9 gives the complete comparison. The runtime and memory measurements above apply to the original FT configuration.

Table 4: Additional real-data results. Rates are percentages. Parentheses give standard errors. The ADNI FT entry uses the expanded validation grid.

Task	FT	FPCA+LDA	FPCA+SVM
Speech: aa vs ao	**19.130 (0.028)**	25.410 (0.281)	23.475 (0.260)
Speech: aa vs iy	**0.010 (0.001)**	0.290 (0.031)	0.035 (0.013)
Speech: ao vs iy	**0.010 (0.001)**	0.585 (0.046)	0.085 (0.020)
Chest: normal vs abnormal	**22.135 (0.019)**	30.252 (0.327)	30.412 (0.330)
ADNI PET: AD vs EMCI	**10.017 (0.059)**	20.600 (0.341)	16.050 (0.510)

6 Conclusion and limitations

We establish a first minimax theory for transformer-based functional classification in an infinite-dimensional nonparametric setting. The analysis allows the Bayes decision boundary to depend on infinitely many latent directions and establishes a dense-to-sparse phase transition under discrete observation. This transition identifies when sampling frequency limits the attainable excess-risk rate. Empirically, FT performs strongly across simulations and real-data benchmarks, achieves the best results in most settings, and offers a substantially more favorable computational profile. Limitations include the focus on binary classification, relatively homogeneous functional inputs, and locally structured attention, leaving mixed functional objects and more general observation schemes beyond the scope of this work. Understanding how deeper attention architectures and broader observation settings affect tradeoffs between statistical accuracy and computational cost is an important direction for future study.

Acknowledgments and disclosures

Guanqun Cao's research is partially supported by the National Science Foundation under Grants DMS-2413301, CNS-2319342 and CNS-2319343.

Competing interests: The authors declare no competing interests.

References

Jorge Adrover, Matias Salibian-Barrera, and Ruben Zamar. Globally robust inference for the location and simple linear regression models. *Journal of Statistical Planning and Inference*, 119(2):353–375, 2004.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35:608–633, 2007.

J. R. Berrendero, A. Cuevas, and J. L. Torrecilla. On the use of reproducing kernel hilbert spaces in functional classification. *Journal of the American Statistical Association*, 113(523):1210–1218, 2018.

Tobia Boschi, Francesca Bonin, Rodrigo Ordonez-Hurtado, Alessandra Pascale, and Jonathan Epperlein. A new computationally efficient algorithm to solve feature selection for functional data classification in high-dimensional spaces. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Susanne C. Brenner and Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, 3rd edition, 2008. ISBN 978-0-387-75934-0. URL https://doi.org/10.1007/978-0-387-75934-0.

Xiongtao Dai, Hans-Georg Müller, and Fang Yao. Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560, 2017.

A. Delaigle and P. Hall. Achieving near-perfect classification for functional data. *Journal of the Royal Statistical Society, Series B*, 74:267–286, 2012.

A. Delaigle, P. Hall, and N. Bathia. Componentwise classification and clustering of functional data. *Biometrika*, 99(2):299–313, 2012.

Aurore Delaigle and Peter Hall. Classification using censored functional data. *Journal of the American Statistical Association*, 108(504):1269–1283, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2):161–173, 2003.

Pedro Galeano, Esdras Joseph, and Rosa E. Lillo. The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291, 2015.

Florian Heinrichs, Mavin Heim, and Corinna Weber. Functional neural networks: shift invariant models for functional data with applications to eeg classification. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

Ju-Sheng Hong, Junwen Yao, Jonas W Mueller, and Jane-Ling Wang. Sand: Smooth imputation of sparse and noisy functional data with transformer networks. *Advances in Neural Information Processing Systems*, 37:134024–134042, 2024.

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Stéphane Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3rd edition, 2008.

Juhyun Park, Jeongyoun Ahn, and Yongho Jeon. Sparse functional linear discriminant analysis. *arXiv:2012.06488*, 2020.

Yaozhong Shi, Angela F Gao, Zachary E Ross, and Kamyar Azizzadenesheli. Universal functional regression with neural operator flows. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=rHL329Xa3X.

H. Shin. An extension of fisher's discriminant analysis for stochastic processes. *Journal of Multivariate Analysis*, 99:1191—-1216, 2008.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science and Business Media, 2008.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, pages 33416–33447. PMLR, 2023.

Hans Triebel. Entropy numbers in function spaces with mixed integrability. *Revista matemática complutense*, 24(1):169–188, 2011.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32:135–166, 2004.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

J.L. Wang, J. M. Chiou, and H. G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Shuoyang Wang, Guanqun Cao, and Zuofeng Shang. Deep neural network classifier for multi-dimensional functional data. *Scandinavian Journal of Statistics*, 50:1667–1686, 2023.

Shuoyang Wang, Yuan Huang, and Guanqun Cao. Review on functional data classification. *WIREs Computational Statistics*, 16(1), 2024.

A Additional definitions and assumptions

Define $\mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\}$ and $\mathbb{N}^+ := \{1, 2, \dots\}$, respectively. Throughout this work, all indexing and dimensionality parameters are assumed to belong to $\mathbb{N}^+$ unless otherwise stated. Let $\mathbb{Z}_0^\infty$ denote the set of infinite sequences $(l_j)_{j=1}^\infty$ with finitely many nonzero terms, defined as the union $\mathbb{Z}_0^\infty = \bigcup_{d=1}^\infty \mathbb{Z}_0^d$, where $\mathbb{Z}_0^d = \{(l_1, \dots, l_d, 0, 0, \dots) : l_j \in \mathbb{Z} \text{ for } j = 1, \dots, d\}$. The set represents the space of multi-indices with finite support, which are utilized throughout our analysis to index the functional basis expansions and latent representation coordinates.

A.1 Functional decision boundary and its latent representation

We define the latent space of the decision boundary $\eta(\cdot)$ that quantify directional smoothness in terms of Fourier frequency localization and weighted square norm decay. First, we introduce the orthonormal basis system for $L^2([0,1]^\infty)$. For a multi-index $\boldsymbol{u} = (u_j)_{j\in\mathbb{N}^+} \in \mathbb{Z}_0^\infty$ with finitely many nonzero and an input sequence $\boldsymbol{v} = (v_j)_{j\in\mathbb{N}^+}^\intercal \in [0,1]^\infty$, denote an infinite-dimensional Fourier basis as

$$\psi_{\boldsymbol{u}} : [0,1]^\infty \to \mathbb{R}, \quad \psi_{\boldsymbol{u}}(\boldsymbol{v}) := \prod_{j\in\mathbb{N}^+} \psi_{u_j}(v_j), \quad \psi_{u_j}(v_j) := \begin{cases} \sqrt{2}\cos(2\pi|u_j|v_j), & u_j < 0, \\ 1, & u_j = 0, \\ \sqrt{2}\sin(2\pi|u_j|v_j), & u_j > 0, \end{cases}$$

which yields a complete orthonormal system of $L^2([0,1]^\infty)$ under the product Lebesgue measure, by standard tensor-product Hilbert space arguments. Consequently, any $f \in L^2([0,1]^\infty)$ can be represented by $f = \sum_{\boldsymbol{u}\in\mathbb{Z}_0^\infty} \langle f, \psi_{\boldsymbol{u}}\rangle \psi_{\boldsymbol{u}}$.

Next, we introduce the frequency block function to restricts coefficients of f according to the dyadic frequency levels with base 2. For a nonnegative index function $\boldsymbol{w} = (w_j)_{j\in\mathbb{N}^+} \in \mathbb{N}_0^\infty$, define the frequency block

$$\zeta_{\boldsymbol{w}} : L^2([0,1]^\infty) \to L^2([0,1]^\infty), \quad \zeta_{\boldsymbol{w}}(f) := \sum_{\substack{\boldsymbol{u}\in\mathbb{Z}_0^\infty \\ \lfloor 2^{w_j-1}\rfloor \le |u_j| < 2^{w_j} \ \forall j}} \langle f, \psi_{\boldsymbol{u}}\rangle \psi_{\boldsymbol{u}},$$

which isolates the portion of $f \in L^2([0,1]^\infty)$ at multi-frequency scale $\boldsymbol{u}$.

Definition 2. *For a function $\gamma : \mathbb{N}_0^\infty \to \mathbb{R}$ which is monotonically non-decreasing with respect to each coordinate, define the γ-smooth function space as*

$$\mathcal{F}_\gamma = \left\{ f \in L^2([0,1]^\infty) \mid \left(\sum_{\boldsymbol{w}\in\mathbb{N}_0^\infty} 2^{2\gamma(\boldsymbol{w})} \|\zeta_{\boldsymbol{w}}(f)\|_2^2 \right)^{1/2} < \infty \right\}.$$

Additionally, define the smoothness index of γ as $\beta^* = \left\{ \sup_{\boldsymbol{w}\in\mathbb{N}_0^\infty\setminus\{\boldsymbol{0}\}} \frac{\sum_{j=1}^\infty w_j}{\gamma(\boldsymbol{w})} \right\}^{-1}$.

Remark A.1. *The definition of $\mathcal{F}_\gamma$ provides a unified framework for describing a wide range of smoothness classes on $[0,1]^\infty$ through appropriate choices of the smoothness map γ. Naïvely, this framework includes the classical finite-dimensional Sobolev space as a special case by restricting $\gamma(\boldsymbol{w})$ to depend only on finitely many coordinates of $\boldsymbol{w}$ and taking $\gamma(\boldsymbol{w})$ proportional to the maximum resolution level among those coordinates. In this sense, $\mathcal{F}_\gamma$ extends standard Sobolev regularity to an infinite-dimensional setting. More importantly, the definition naturally captures mixed and anisotropic smoothness classes, which are of primary interest in high- and infinite-dimensional problems. In the mixed smoothness case, γ takes the additive form $\gamma(\boldsymbol{w}) = \sum_{j=1}^\infty \beta_j w_j$, where the sequence $\boldsymbol{\beta} = (\beta_j)_{j\ge1}$ controls the coordinate-wise smoothness. This choice reflects a cumulative contribution of oscillations across all active coordinates and leads to function classes with the complexity of sparse grids. In contrast, the anisotropic smoothness class is obtained by defining $\gamma(\boldsymbol{w}) = \max_{j\ge1}\{\beta_j w_j\}$, which emphasizes the dominant coordinate in each frequency block and models situations where the overall regularity is dictated by the least smooth active direction. For both choices, the inverse smoothness index β^* summarizes the intrinsic complexity of the function class, which further governs the minimax convergence rates. In particular, a larger β^* indicates a smoother class and thus a faster convergence rate. Notably, the ambient dimension does not appear explicitly. Instead, the*

statistical difficulty is determined entirely by the growth behavior of γ through $\boldsymbol{\beta}$, illustrating how mixed and anisotropic structures circumvent the curse of dimensionality in infinite-dimensional settings. Equivalently, $\beta^* = \bar{\beta}_1$ for mixed smoothness, and $\beta^* = \left(\sum_{j\geq 1} \bar{\beta}_j^{-1}\right)^{-1}$ for anisotropic smoothness.

To construct a functional transformer classifier that attains the minimax lower bound, we first assume an optimal structural form designed to trade off approximation accuracy against estimation efficiency.

Assumption 4. *There exist positive universal constants c_U, c_D, c_H, c_L, c_P, c_S, such that the transformer class $\mathcal{T}(L, P, S, B, U, D, H)$ satisfies:*

- $L = \lfloor c_L \log^{2/\theta^*}(\rho^{-1}) \rfloor$,
- $P = \lfloor c_P \rho^{-1/\beta^*} \log^{1/\theta}(\rho^{-1}) \rfloor$,
- $S = \lfloor c_S \rho^{-1/\beta^*} \log^{(2/\theta+2/\theta^*)}(\rho^{-1}) \rfloor$,
- $B = \rho^{-1}$,
- $\log U = c_U \log(\rho^{-1})$,
- $D = \lfloor c_D \log^{1/\theta}(\rho^{-1}) \rfloor$,
- $H = \lfloor c_H \log^{1/\theta}(\rho^{-1}) \rfloor$,
- $\boldsymbol{E}_{:j} = (0, \ldots, 0, \cos(j(2U+1)^{-1}), \sin(j(2U+1)^{-1}))^\intercal$,

where $\theta^* = \min(\theta, 1)$ and $\rho = r_n^2$.

In particular, the attention range U is allowed to increase with the sample size and sampling frequency through the theoretically prescribed architecture class. The next assumption parallels Assumption 4, which imposes structural conditions for optimal transformer neural network classifiers. Unlike the fully observed setting, Assumption 5 is formulated to account jointly for the sample size n and the sampling frequency m.

Assumption 5. *In Assumption 4, we set $\rho = r_{nm}$ with $r_{nm}^2 = r_n^2 \vee m^{-\tau(\alpha+1)}$.*

A.2 Choice of the basis system

In principle, the basis system used in the functional embedding layer can be chosen as any orthonormal basis of $L^2([0,1]^q)$. This flexibility is appealing in practice, as the true structural basis underlying the functional predictor is rarely known. However, an important theoretical issue concerns the potential misspecification of the chosen basis system and its impact on representation and approximation.

Suppose the functional predictor admits the expansion $X(\boldsymbol{s}) = \sum_{\ell=1}^{\infty} \langle X, \phi_\ell \rangle \phi_\ell(\boldsymbol{s})$, where $\{\phi_\ell\}_{\ell\geq 1}$ denotes an oracle orthonormal basis. In contrast, the functional embedding layer may employ a different orthonormal basis $\{\psi_j\}_{j\geq 1}$. In this case, the oracle coordinate $\langle X, \phi_j \rangle$ can be expressed as a linear combination of the coordinates induced by $\{\psi_\ell\}_{\ell\geq 1}$: $\langle X, \phi_j \rangle = \sum_{\ell=1}^{\infty} r_{j\ell} \langle X, \psi_\ell \rangle$, $r_{j\ell} = \langle \phi_j, \psi_\ell \rangle$, $j \in \mathbb{N}^+$. Thus, a misspecified basis system introduces an additional linear transformation between the representation employed by the embedding layer and the oracle representation that would be optimal for the target decision function.

To accommodate this mismatch, we enrich the functional embedding layer by allowing an additional neural network transformation. Specifically, we extend the input of the first-stage embedding map e_1 from $\langle X, \boldsymbol{\psi} \rangle$ to

$$\widetilde{\boldsymbol{W}} \sigma_{\overline{\boldsymbol{V}}} \overline{\boldsymbol{W}} \langle X, \boldsymbol{\psi} \rangle + \boldsymbol{d}, \quad \|\widetilde{\boldsymbol{W}}\|_\infty \vee \|\overline{\boldsymbol{V}}\|_\infty \vee \|\overline{\boldsymbol{W}}\|_\infty \vee |\boldsymbol{d}| \leq B, \tag{3}$$

where $\langle X, \boldsymbol{\psi} \rangle = (\langle X, \psi_1 \rangle, \langle X, \psi_2 \rangle, \ldots)^\intercal$, $\overline{\boldsymbol{W}} \in \mathbb{R}^{\infty \times \infty}$, $\overline{\boldsymbol{V}}, \widetilde{\boldsymbol{W}}, \boldsymbol{d} \in \mathbb{R}^\infty$. This additional layer enables the model to learn linear transformations between different basis-induced representations within the overall architecture.

The validity of this construction follows from a basic property of the ReLU activation: for any $t \in \mathbb{R}$, $t = \mathrm{ReLU}(t) - \mathrm{ReLU}(-t)$. Consequently, linear mappings can be exactly represented by

shallow ReLU networks. Therefore, for an arbitrary orthonormal basis system $\{\psi_j(\boldsymbol{s})\}_{j\geq 1}$, there exist $\{\widetilde{\boldsymbol{W}}\}$, $\{\overline{\boldsymbol{V}}\}$ and $\{\overline{\boldsymbol{W}}\}$ such that $c_j\langle X, \phi_j\rangle = \left(\widetilde{\boldsymbol{W}}\sigma_{\overline{\boldsymbol{V}}}\overline{\boldsymbol{W}}\langle X, \boldsymbol{\psi}\rangle\right)_j$, $j \in \mathbb{N}^+$, where $\{\phi_j\}_{j\in\mathbb{N}^+}$ are the oracle basis system. This shows that the functional transformer architecture is robust to basis misspecification: even when the embedding layer is constructed using a non-oracle basis, the model can recover the oracle representation through learned linear transformations, without affecting the theoretical approximation properties.

The sigmoid maps each basis coefficient into $(0, 1)$ and, being strictly monotone, preserves coordinate-wise information in exact arithmetic. Other bounded strictly monotone links with range in $(0, 1)$ provide the same type of coordinate representation. Our analysis uses a fixed orthonormal basis. Extending it to an estimated FPCA basis requires an additional bound on basis-estimation error. Table 8 compares Fourier and FPCA scores within the FT architecture and supports the use of Fourier scores for the six simulated settings. Comparisons with spline and wavelet representations remain for future work.

B Algorithms and tuning parameter selection

Algorithm 1 provides a detailed version of the FT training algorithm. In particular, it makes explicit the truncated forward pass, the mini-batch Adam updates, and the optimization variables used in practice to fit the classifier. Algorithm 2 then describes the hyperparameter selection procedure adopted in our experiments. Using a fixed training/validation split, the tuning algorithm selects the depth, width, truncation level, attention configuration, batch size, and learning rate by minimizing the validation criterion, and then retrains the model using the selected configuration.

Remark B.1 (Architecture selection)**.** *The architecture scaling in Assumptions 4 and 5 depends on the unknown regularity parameters $(\alpha, \beta^*, \theta, \tau)$. In practice, we select the network hyperparameters by validation, as described in Algorithm 2. Finite architecture size and regularization approximate the theoretical sparsity budget. Establishing adaptive minimax guarantees for this selection procedure remains an open problem.*

In the original numerical experiments, hyperparameters were selected by the validation-based procedure described in Algorithm 2. Section F describes the additional studies of resolution, representation and attention window, together with the expanded ADNI tuning. For each candidate configuration, the model was trained on the training subset and compared on the validation subset. The selected configuration was then refitted on the full training set and evaluated on the test set. For the FT classifier, the candidate set included the directional Fourier basis order, chosen from $\{10, 12\}$ in the 1D experiments, from $\{4, 6, 8\}$ in the 2D experiments, and from $\{2, 3, 4\}$ in the 3D experiments, together with $d \in \{64, 128\}$, $U \in \{8, 16, 32\}$, $H \in \{2, 4\}$, $L \in \{2, 3\}$, $P \in \{128, 192\}$, $M \in \{32, 64\}$, and $\eta \in \{3 \times 10^{-4}, 5 \times 10^{-4}, 7 \times 10^{-4}\}$. We additionally varied the dropout rate in $\{0.1, 0.15\}$, weight decay in $\{10^{-4}, 2 \times 10^{-4}\}$. For FDNN, the candidate set included $(K_x, K_y, K_z) \in \{(3, 3, 2), (3, 3, 3), (4, 4, 2), (4, 4, 3)\}$, hidden width in $\{128, 192, 256\}$, number of hidden layers in $\{2, 3\}$, batch size in $\{32, 64\}$, learning rate in $\{5 \times 10^{-4}, 10^{-3}\}$, dropout rate in $\{0.1, 0.15\}$, and weight decay in $\{10^{-4}, 2 \times 10^{-4}\}$. For ViT, the candidate set included embedding dimension in $\{96, 128\}$, number of attention heads in $\{2, 4\}$, number of transformer layers in $\{2, 3, 4\}$, batch size in $\{32, 64\}$, learning rate in $\{5 \times 10^{-4}, 10^{-3}\}$, dropout rate in $\{0.1, 0.15\}$, and weight decay in $\{10^{-4}, 2 \times 10^{-4}\}$. For CNN, the candidate set included base channel size in $\{16, 32\}$, network depth in $\{3, 4\}$, batch size in $\{32, 64\}$, learning rate in $\{5 \times 10^{-4}, 10^{-3}\}$, dropout rate in $\{0.1, 0.15\}$, and weight decay in $\{10^{-4}, 2 \times 10^{-4}\}$.

C Proofs

Proof roadmap. We first derive a minimax lower bound over the proposed infinite-dimensional classification class. For the upper bound, we combine existing transformer approximation results for structured function classes with a classification argument that translates approximation error into excess risk. For discretely observed data, we then incorporate the reconstruction error into the risk bound to obtain the dense-to-sparse phase transition.

Algorithm 1 Functional Transformer (FT) classifier training via mini-batch Adam

Require: Training data $\{(X_i, Y_i)\}_{i=1}^n$ with $Y_i \in \{-1, 1\}$, embedding dimension d, window size U, number of heads H, truncation level J, functional basis $\{\phi_j\}_{j=1}^J$, mini-batch size M, learning rate η, Adam parameters $\beta_1, \beta_2, \epsilon$, maximum number of epochs T

Initialize

$$\Theta_0 = \Big\{\boldsymbol{a}, \boldsymbol{E}, \{Q_h, K_h, V_h\}_{h=1}^H, \{\mathbf{W}_\ell, \boldsymbol{V}_\ell\}_{\ell=1}^L, \{\omega_j\}_{j=1}^J, \nu\Big\}, \quad \boldsymbol{m}_0 \leftarrow 0,\ \boldsymbol{v}_0 \leftarrow 0,\ s \leftarrow 0.$$

for $t = 1, \ldots, T$ **do**
  Shuffle the training data and partition it into mini-batches $\mathcal{B}_1, \ldots, \mathcal{B}_{K_t}$ of size M
  for $k = 1, \ldots, K_t$ **do**
    $s \leftarrow s + 1$
    for each $(X_i, Y_i) \in \mathcal{B}_k$ **do**
      for $j = 1, \ldots, J$ **do**
        $z_{ij} \leftarrow \sigma_s(\langle X_i, \phi_j \rangle_m), \quad \boldsymbol{Z}_{i,:j}^{(0)} \leftarrow \boldsymbol{a}\, z_{ij} + \boldsymbol{E}_{:j}$
      end for
      for $j = 1, \ldots, J$ **do**

$$\widetilde{\boldsymbol{Z}}_{i,:j} \leftarrow \boldsymbol{Z}_{i,:j}^{(0)} + \sum_{h=1}^{H} V_h\, \boldsymbol{Z}_{i,[j-U:j+U]}^{(0)}\, \bar{\sigma}_s\Big(\big(K_h \boldsymbol{Z}_{i,[j-U:j+U]}^{(0)}\big)^\top \big(Q_h \boldsymbol{Z}_{i,:j}^{(0)}\big)\Big)$$

where boundary indices are handled by natural truncation of the local window

        $F_{ij} \leftarrow g(\widetilde{\boldsymbol{Z}}_{i,:j})$
      end for
      $f_{\Theta_{s-1}}(X_i) \leftarrow \sum_{j=1}^J \omega_j F_{ij} + \nu$
    end for

$$\mathcal{L}_{\mathcal{B}_k}(\Theta_{s-1}) \leftarrow \frac{1}{|\mathcal{B}_k|} \sum_{(X_i, Y_i) \in \mathcal{B}_k} \max\{1 - Y_i f_{\Theta_{s-1}}(X_i), 0\}$$

    $\boldsymbol{G}_s \leftarrow \nabla_\Theta \mathcal{L}_{\mathcal{B}_k}(\Theta_{s-1})$ by backpropagation

$$\boldsymbol{m}_s \leftarrow \beta_1 \boldsymbol{m}_{s-1} + (1 - \beta_1)\boldsymbol{G}_s, \qquad \boldsymbol{v}_s \leftarrow \beta_2 \boldsymbol{v}_{s-1} + (1 - \beta_2)(\boldsymbol{G}_s \odot \boldsymbol{G}_s)$$

$$\widehat{\boldsymbol{m}}_s \leftarrow \boldsymbol{m}_s / (1 - \beta_1^s), \qquad \widehat{\boldsymbol{v}}_s \leftarrow \boldsymbol{v}_s / (1 - \beta_2^s), \qquad \Theta_s \leftarrow \Theta_{s-1} - \eta\, \widehat{\boldsymbol{m}}_s \oslash (\sqrt{\widehat{\boldsymbol{v}}_s} + \epsilon)$$

  end for
  Stop if the maximum epoch is reached or the training loss has converged
end for
return

$$\widehat{C}^{\mathrm{FT}}(X) = \mathbb{I}\{f_{\widehat{\Theta}}(X) \geq 0\} - \mathbb{I}\{f_{\widehat{\Theta}}(X) < 0\}$$

C.1 Proof of the lower bound in Theorem 3.1

Without loss of generality, we consider $\pi = 1/2$. Let $\mathcal{V} = [0, 1]^\infty$ be equipped with the product Lebesgue probability measure μ, and write $L^2 := L^2(\mathcal{V}, \mu)$. Recall that for any $\eta \in \widetilde{\mathcal{F}}_\gamma$ there exists $f_\eta \in \mathcal{F}_\gamma$ such that

$$\eta(X) = f_\eta(\mathcal{G}(X)), \qquad X \in L^2([0, 1]^q).$$

In the lower bound construction we therefore work directly on the latent space $\mathcal{V}$, construct a finite family $\{f_{\eta,\sigma}\} \subset \mathcal{F}_\gamma(M)$, and then induce the corresponding Bayes decision functions $\{\eta_\sigma\}$ by composition with $\mathcal{G}$.

Definition 3 (γ-smooth function space)**.** *Let $\gamma : \mathbb{N}_0^\infty \to \mathbb{R}$ be coordinate-wise nondecreasing. Let $\{\zeta_{\boldsymbol{w}}\}_{\boldsymbol{w} \in \mathbb{N}_0^\infty}$ denote the dyadic frequency-block projections defined in Section A.1. Define the*

Algorithm 2 Hyperparameter selection for the FT classifier by training/validation split

Require: Data $\{(X_i, Y_i)\}_{i=1}^n$, candidate hyperparameter set

$$\Lambda = \{(J, d, U, H, L, P, M, \eta)\},$$

training proportion $\rho \in (0, 1)$

Randomly split the data into a training set $\mathcal{D}_{\mathrm{tr}}$ of size $\lfloor \rho n \rfloor$ and a validation set $\mathcal{D}_{\mathrm{val}}$
Initialize the best validation score $S^* \leftarrow \infty$ and best hyperparameter $\lambda^* \leftarrow \varnothing$
for each $\lambda = (J, d, U, H, L, P, M, \eta) \in \Lambda$ **do**
    Train the FT classifier on $\mathcal{D}_{\mathrm{tr}}$ using Algorithm 1 with hyperparameter λ, yielding fitted score $\widehat{f}_\lambda$
    Compute the validation criterion

$$S(\lambda) \leftarrow \frac{1}{|\mathcal{D}_{\mathrm{val}}|} \sum_{(X_i, Y_i) \in \mathcal{D}_{\mathrm{val}}} \max\{1 - Y_i \widehat{f}_\lambda(X_i), 0\}$$

    if $S(\lambda) < S^*$ **then**
        $S^* \leftarrow S(\lambda)$ and $\lambda^* \leftarrow \lambda$
    end if
end for
Retrain the FT classifier on the full dataset using λ^*
return selected hyperparameter λ^* and final classifier $\widehat{C}^{\mathrm{FT}}_{\lambda^*}$

γ-smooth norm

$$\|f\|_{\mathcal{F}_\gamma} := \left(\sum_{\boldsymbol{w} \in \mathbb{N}_0^\infty} 2^{2\gamma(\boldsymbol{w})} \|\zeta_{\boldsymbol{w}}(f)\|_2^2 \right)^{1/2}, \qquad f \in L^2(\mathcal{V}),$$

and for $M > 0$ define the associated ball

$$\mathcal{F}_\gamma(M) := \{f \in \mathcal{F}_\gamma : \ \|f\|_{\mathcal{F}_\gamma} \le M\}.$$

Entropy numbers and covering numbers. Let A and B be Banach spaces and let $T : A \to B$ be a bounded linear operator. The kth *entropy number* of T is defined as

$$e_k(T) := \inf \Big\{ \varepsilon > 0 : \ T(U_A) \subset \bigcup_{j=1}^{2^{k-1}} \big(b_j + \varepsilon U_B\big) \text{ for some } b_1, \dots, b_{2^{k-1}} \in B \Big\},$$

where U_A and U_B denote the unit balls of A and B, respectively. Equivalently, $e_k(T)$ is the smallest radius ε such that the image $T(U_A)$ can be covered by at most 2^{k-1} balls of radius ε in B.

In our setting, $\mathcal{F}_\gamma(M)$ is viewed as a bounded subset of $L^2(\mu)$, and its metric complexity is governed by the compact embedding into $L^2(\mu)$. By definition of entropy numbers,

$$N\big(\varepsilon, \mathcal{F}_\gamma(M), \|\cdot\|_2\big) \le 2^{k-1} \quad \Longleftrightarrow \quad e_k\big(\mathcal{F}_\gamma(M); L^2(\mu)\big) \le \varepsilon.$$

Consequently, an entropy-number decay

$$e_k\big(\mathcal{F}_\gamma(M); L^2(\mu)\big) \asymp k^{-\beta^*}$$

is equivalent to the metric entropy growth

$$\log N\big(\varepsilon, \mathcal{F}_\gamma(M), \|\cdot\|_2\big) \asymp \varepsilon^{-1/\beta^*}, \qquad \varepsilon \downarrow 0. \tag{4}$$

The regularly varying smoothness condition in Assumption 2 places the mixed and anisotropic γ-smooth classes in the same regularly varying sparsity and locality regime considered in the approximation literature. Motivated by the Besov-type interpretation of these classes, we adopt the entropy scaling in (4), which is consistent with the classical behavior of anisotropic and mixed Besov

embeddings. See, for example, Triebel [2011]. This is the entropy regime used in the lower-bound construction below.

Lemma 1 (Localized blocks from metric entropy)**.** *Assume* (4). *Then there exist constants* $\widetilde{c}_0, \widetilde{c}_1, \widetilde{c}_2, \widetilde{c}_3 > 0$ *such that for every* $a \in (0, \widetilde{c}_0)$ *there exist an integer* $\widetilde{m} = \widetilde{m}(a)$*, pairwise disjoint measurable sets* $A_1, \dots, A_{\widetilde{m}} \subset \mathcal{V}$ *and functions* $\varphi_1, \dots, \varphi_{\widetilde{m}} \in \mathcal{F}_\gamma(M)$ *satisfying*

1. $\widetilde{m} \geq \widetilde{c}_1 \, a^{-1/\beta^*}$.
2. $\mu(A_k) = 1/\widetilde{m}$ *and* $\operatorname{supp}(\varphi_k) \subset A_k$ *for all* k.
3. $0 \leq \varphi_k \leq 1$ *for all* k.
4. $\displaystyle\int_{A_k} \varphi_k \, d\mu \geq \widetilde{c}_2 \, \widetilde{m}^{-1}$ *and* $\|\varphi_k\|_2^2 \geq \widetilde{c}_3 \, \widetilde{m}^{-1}$ *for all* k.

Lemma 2 (Hypercube in $\mathcal{F}_\gamma(M)$)**.** *Let* $\{\varphi_k, A_k\}$ *be as in Lemma 1. For* $\sigma \in \{-1, +1\}^{\widetilde{m}}$ *define*

$$f_{\eta,\sigma}(z) := \frac{1}{2} + \frac{a}{2} \sum_{k=1}^{\widetilde{m}} \sigma_k \varphi_k(z), \qquad z \in \mathcal{V}.$$

Then for a *sufficiently small,* $f_{\eta,\sigma}(z) \in [3/8, 5/8]$ *for all* $z \in \mathcal{V}$ *and* $f_{\eta,\sigma} - \frac{1}{2} \in \mathcal{F}_\gamma(M)$*. Moreover, for Hamming neighbors* $\sigma \sim \sigma'$*,*

$$\|f_{\eta,\sigma} - f_{\eta,\sigma'}\|_2^2 \geq \widetilde{c}_4 \, \frac{a^2}{\widetilde{m}}$$

for some universal constant $\widetilde{c}_4 > 0$*.*

Proof of the lower bound in Theorem 3.1. Fix $a \in (0, \widetilde{c}_0)$ and take $\widetilde{m}$, $\{A_k\}$ and $\{\varphi_k\}$ from Lemma 1, and $\{f_{\eta,\sigma}\}$ from Lemma 2. Let $A_0 \subset \mathcal{V} \setminus \cup_{k=1}^{\widetilde{m}} A_k$ with $\mu(A_0) > 0$ and define a density p with respect to μ by

$$p(z) = \sum_{k=1}^{\widetilde{m}} \frac{w}{\mu(A_k)} \mathbf{1}\{z \in A_k\} + \frac{1 - \widetilde{m} w}{\mu(A_0)} \mathbf{1}\{z \in A_0\}, \qquad 0 < w \leq 1/\widetilde{m}.$$

Let $Z \sim p \, d\mu$, and, given $Z = z$, let

$$Y \mid Z = z \sim \text{Bernoulli}(f_{\eta,\sigma}(z)).$$

Denote the resulting joint distribution by Q_σ and define

$$\eta_\sigma(X) := f_{\eta,\sigma}(\mathcal{G}(X)).$$

Then $\eta_\sigma \in \widetilde{\mathcal{F}}_\gamma$ by construction.

We verify the margin condition in Assumption 1. Since $f_{\eta,\sigma} \equiv 1/2$ on A_0 and $|f_{\eta,\sigma}(z) - 1/2| \leq a/2$ everywhere, for any $t \in (0, a/2]$,

$$Q_\sigma\big(|\eta_\sigma(X) - \tfrac{1}{2}| \leq t\big) \leq Q_\sigma\Big(Z \in \bigcup_{k=1}^{\widetilde{m}} A_k\Big) = \widetilde{m} w.$$

Choosing $w = w(a)$ such that $\widetilde{m} w \leq c_0 \, a^\alpha$ yields

$$Q_\sigma\big(|\eta_\sigma(X) - \tfrac{1}{2}| \leq t\big) \leq c_0 \, t^\alpha, \qquad \forall \, t \in (0, a/2],$$

and hence Assumption 1 holds.

For Hamming neighbors $\sigma \sim \sigma'$, the two distributions Q_σ and $Q_{\sigma'}$ differ only on one cell A_k. Since $f_{\eta,\sigma}(z), f_{\eta,\sigma'}(z) \in [3/8, 5/8]$, a standard bound for Bernoulli models yields

$$\mathrm{KL}(Q_\sigma^{\otimes n}, Q_{\sigma'}^{\otimes n}) \leq C \, n \, \mathbb{E}\big[\big(f_{\eta,\sigma}(Z) - f_{\eta,\sigma'}(Z)\big)^2\big]$$

for a universal constant $C > 0$. Moreover,

$$\mathbb{E}\big[\big(f_{\eta,\sigma}(Z) - f_{\eta,\sigma'}(Z)\big)^2\big] = \frac{a^2}{4}\int \varphi_k(z)^2 p(z)\, d\mu(z) = \frac{a^2}{4}\cdot\frac{w}{\mu(A_k)}\int_{A_k}\varphi_k(z)^2\, d\mu(z) \le \frac{a^2}{4}\, w,$$

since $\varphi_k \le 1$ and $\mu(A_k) = 1/\widetilde{m}$. Hence

$$\mathrm{KL}(Q_\sigma^{\otimes n}, Q_{\sigma'}^{\otimes n}) \le C'\, n\, a^2 w$$

for some universal constant $C' > 0$. Choose $w \asymp (na^2)^{-1}$ so that $\mathrm{KL}(Q_\sigma^{\otimes n}, Q_{\sigma'}^{\otimes n})$ is bounded by a small absolute constant for all Hamming neighbors.

Applying Assouad's lemma (Theorem 4.1 of Audibert and Tsybakov, 2007) then gives

$$\inf_{\widetilde{C}}\sup_{\sigma}\big(R_{Q_\sigma}(\widetilde{C}) - R_{Q_\sigma}(C_\sigma^*)\big) \gtrsim \widetilde{m}\, w\, a,$$

where the implicit constant is universal. With $w \asymp (na^2)^{-1}$ this yields

$$\inf_{\widetilde{C}}\sup_{\sigma}\big(R_{Q_\sigma}(\widetilde{C}) - R_{Q_\sigma}(C_\sigma^*)\big) \gtrsim \frac{\widetilde{m}}{na}.$$

Using Lemma 1, $\widetilde{m} \ge c_1 a^{-1/\beta^*}$ for some $c_1 > 0$. The margin constraint $\widetilde{m}w \le c_0 a^\alpha$ together with $w \asymp (na^2)^{-1}$ implies the feasibility condition

$$\frac{a^{-1/\beta^*}}{na^2} \lesssim a^\alpha, \qquad \text{equivalently} \qquad a^{\alpha+2+1/\beta^*} \gtrsim n^{-1}.$$

Moreover, the lower bound obtained from Assouad's lemma satisfies

$$\inf_{\widetilde{C}}\sup_{\sigma}\big(R_{Q_\sigma}(\widetilde{C}) - R_{Q_\sigma}(C_\sigma^*)\big) \gtrsim \frac{\widetilde{m}}{na} \gtrsim n^{-1}a^{-(1+1/\beta^*)}.$$

Since $a \mapsto n^{-1}a^{-(1+1/\beta^*)}$ is decreasing, the lower bound is maximized up to universal constants by choosing a of the smallest order compatible with the feasibility condition, namely

$$a \asymp n^{-1/(\alpha+2+1/\beta^*)}.$$

Substituting this choice into $n^{-1}a^{-(1+1/\beta^*)}$ yields

$$\inf_{\widetilde{C}}\sup_{Q\in\mathcal{Q}}\mathcal{E}_Q(\widetilde{C}) \ge c_1\, n^{-\frac{(\alpha+1)\beta^*}{(\alpha+2)\beta^*+1}},$$

for some constant $c_1 > 0$. This completes the proof. □

C.2 Proof of the upper bound in Theorem 3.1

The main novelty of the upper-bound argument is not a new approximation theorem itself, but rather the way an existing transformer approximation result is incorporated into the functional classification framework to obtain minimax excess-risk control. We first introduce the approximation error for estimating η.

Lemma 3. *For any $\eta \in \widetilde{\mathcal{F}}_\gamma$ and $\epsilon > 0$, if Assumption 2 holds, there exists a transformer network $\widehat{\eta} \in \mathcal{T}(L, P, S, B, U, D, H)$ with*

- $\log U = c_U \log\big(\epsilon^{-2}\big)$,
- $D = \Big\lfloor c_D \log^{1/\theta}\big(\epsilon^{-2}\big)\Big\rfloor$,
- $H = \Big\lfloor c_H \log^{1/\theta}\big(\epsilon^{-2}\big)\Big\rfloor$,
- $L = \Big\lfloor c_L \log^{2/\theta^*}\big(\epsilon^{-2}\big)\Big\rfloor$,

- $P = \left\lfloor c_P \epsilon^{-2/((\alpha+1)\beta^*)} \log^{1/\theta}\left(\epsilon^{-2}\right) \right\rfloor$,
- $B = \epsilon^{-2/(\alpha+1)}$,
- $S = \left\lfloor c_S \epsilon^{-2/((\alpha+1)\beta^*)} \log^{(2/\theta+2/\theta^*)}\left(\epsilon^{-1}\right) \right\rfloor$,
- $\boldsymbol{E}_{:j} = \left(0, \ldots, 0, \cos\left(j(2U_1+1)^{-1}\right), \sin\left(j(2U_1+1)^{-1}\right)\right)^\intercal$,

where c_U, c_D, c_H, c_L, c_P, c_S are positive universal constants, such that $\|\eta - \widehat{\eta}\|_\infty \leq \epsilon^{2/(\alpha+1)}$.

Proof. By definition, for any $\eta \in \widetilde{\mathcal{F}}_\gamma$, there exists $f_\eta \in \mathcal{F}_\gamma$, where f_η is a γ-smooth function defined on $[0,1]^\infty$, such that $\eta = f_\eta(\mathcal{G}(X))$. Similarly, $\widehat{\eta}$ admits $\widehat{\eta} = f_{\widehat{\eta}}(\mathcal{G}(X))$, and $\|\eta - \widehat{\eta}\|_\infty = \|f_\eta - f_{\widehat{\eta}}\|_\infty$. Our scalar target can be embedded into the sequence-to-sequence framework by the constant extension. This is a degenerate special case of the shift-equivariant regression model considered in Takakura and Suzuki [2023]. Therefore, the consistency of Theorem 4.2 in Takakura and Suzuki [2023] guarantees for $\widehat{F}$ apply directly, and a scalar estimate is obtained via averaged pooling. Specifically, the weights $\omega_j = J^{-1}$ and $\nu = 0$ for any $J \in \mathbb{N}^+$, such that $f\big(X(\boldsymbol{s})\big) = \sum_{j=1}^J \omega_j F_j\big(X(\boldsymbol{s})\big) + \nu$. By letting $T = 2/(\alpha+1)\log\epsilon^{-1}$, since $\log^{1/\theta^*}(\epsilon^{-1}) \lesssim \log(\epsilon^{-1})$, the proof can be directly obtained by Theorem 4.2 in Takakura and Suzuki [2023], and the conditions in Lemma 3 holds. □

Next, we introduce the covering number of an arbitrary function class $\mathcal{F}$ given any error $\kappa > 0$.

Definition 4. *For some $\kappa > 0$, a subset $\{a_k \in \mathcal{F}\}_{k\geq 1}$ is called a κ-covering set of $\mathcal{F}$ with respect to $\|\cdot\|_\infty$, if for all $a \in \mathcal{F}$, there exists an a_k such that $\|a_k - a\|_\infty \leq \kappa$. The κ-covering number of $\mathcal{G}$ with respect to $\|\cdot\|_\infty$ is defined by*

$$\mathcal{N}(\kappa, \mathcal{F}, \|\cdot\|_\infty) = \inf\left\{ N \in \mathbb{N} : \exists a_1, \ldots, a_N, s.t. \mathcal{F} \subset \bigcup_{k=1}^N \{a \in \mathcal{F} : \|a_k - a\|_\infty \leq \kappa\} \right\}.$$

The following lemma shows the covering number of the considered transformer network class.

Lemma 4. *For any $\epsilon > 0$, if $\mathcal{T}(L, P, S, B, U, D, H)$ satisfies the structure defined in Lemma 3, then*

$$\log \mathcal{N}(\epsilon, \mathcal{T}(L, P, S, B, U, D, H), \|\cdot\|_\infty) \lesssim \left(\epsilon^{-2}\right)^{1/(\alpha+1)\beta^*} \log^{(1+2/\theta+4/\theta^*)}\left(\epsilon^{-1}\right).$$

Proof. By the conditions in Lemma 3 and Theorem 5.3 in Takakura and Suzuki [2023], it satisfies that

$$\begin{aligned} & \log \mathcal{N}(\epsilon, \mathcal{T}(L, P, S, B, U, D, H), \|\cdot\|_\infty) \\ \lesssim\ & L(S + HD^2) \log\left(DHLPB\epsilon^{-2}\right) \\ \asymp\ & \left(\epsilon^{-2}\right)^{1/(\alpha+1)\beta^*} \log^{(1+2/\theta+4/\theta^*)}\left(\epsilon^{-1}\right), \end{aligned}$$

thus the proof is complete. □

Proof of the upper bound in Theorem 3.1. We use the network $\widehat{\eta}$ obtained in Lemma 3 and construct a network

$$\widehat{f} = 2\left(\sigma\left(\epsilon^{-2/(\alpha+1)}\left(\widehat{\eta} - \frac{1}{2}\right)\right) - \sigma\left(\epsilon^{-2/(\alpha+1)}\left(\widehat{\eta} - \frac{1}{2}\right) - 1\right)\right) - 1.$$

This step requires two layer, which can be obtained by

$$\widehat{\eta} \mapsto \sigma\left(\epsilon^{-2/(\alpha+1)}\left(\widehat{\eta} - \frac{1}{2}\right)\right),$$

$$\widehat{\eta} \mapsto \sigma\left(\epsilon^{-2/(\alpha+1)}\left(\widehat{\eta} - \frac{1}{2}\right) - 1\right)$$

with the maximal value of weights is bounded above by $\epsilon^{-2/(\alpha+1)}$. Since the subtraction is multiplied by two, we need double the width, and the last layer of additive structure with bias term -1. The construction doesn't change the asymptotic orders of L and P. Define $\boldsymbol{A} = \left\{X : |\eta(X) - \frac{1}{2}| > 2\epsilon^{2/(\alpha+1)}\right\}$.

First, when $X \in \boldsymbol{A}$, it is suffice to show that $\widehat{f}(X) = C^{Q*}$. This is because when $\eta(X) - \frac{1}{2} > 2\epsilon^{2/(\alpha+1)}$, we have

$$\widehat{\eta}(X) - \frac{1}{2} = \widehat{\eta}(X) - \eta(X) + \eta(X) - \frac{1}{2} > -\epsilon^{2/(\alpha+1)} + 2\epsilon^{2/(\alpha+1)} = \epsilon^{2/(\alpha+1)}$$

and when $\eta(X) - \frac{1}{2} < -2\epsilon^{2/(\alpha+1)}$, we have

$$\widehat{\eta}(X) - \frac{1}{2} = \widehat{\eta}(X) - \eta(X) + \eta(X) - \frac{1}{2} < \epsilon^{2/(\alpha+1)} - 2\epsilon^{2/(\alpha+1)} = -\epsilon^{2/(\alpha+1)}.$$

Next, we show the behavior of $\widehat{f}$ when $X \notin \boldsymbol{A}$. According to Theorem 2.31 of Steinwart and Christmann [2008], there exists a universal constant c', such that

$$\sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widehat{C}^{FT}) \leq c' \sup_{Q\in\mathcal{Q}} E\left[\phi\left(Y\widehat{f}\right) - \phi\left(YC^{Q*}\right)\right].$$

Therefore, since

$$\begin{aligned}
& E\left[\phi\left(Y\widehat{f}(X)\right) - \phi\left(YC^{Q*}(X)\right)\right] \\
= & \int \left|\widehat{f}(X) - C^{Q*}(X)\right| |2\eta(X) - 1| \, dP_Q(X) \\
= & 2\int_{\boldsymbol{A}^c} \left|\widehat{f}(X) - C^{Q*}(X)\right| \left|\eta(X) - \frac{1}{2}\right| dP_Q(X) \\
\leq & 4\int_{\boldsymbol{A}^c} \left|\eta(X) - \frac{1}{2}\right| dP_Q(X) \\
\leq & 8\epsilon^{2/(\alpha+1)} \mathrm{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leq 2\epsilon^{2/(\alpha+1)}\right) \leq 2^{\alpha+3} c_0 \epsilon^2,
\end{aligned}$$

it suffices to show that

$$\sup_{Q\in\mathcal{Q}} \mathcal{E}_Q(\widehat{C}^{FT}) \leq 2^{\alpha+3} c_0 c' \epsilon^2.$$

According to Lemma 4 and Theorem 3.3 in Kim et al. [2021], by choosing

$$n\left(\epsilon^2\right)^{(\alpha+2)/(\alpha+1)} \gtrsim \left(\epsilon^{-2}\right)^{1/(\alpha+1)\beta^*} \log^{(1+2/\theta+4/\theta^*)}\left(\epsilon^{-1}\right),$$

we obtain the best possible convergence rate

$$\epsilon^2 = n^{-\frac{(\alpha+1)\beta^*}{(\alpha+2)\beta^*+1}} (\log n)^{(1+2/\theta+4/\theta^*)(\alpha+1)/(\alpha+2+1/\beta^*)}.$$

Thus, by letting $r_n^2 = \epsilon^2$ and $c_2 = 2^{\alpha+3} c_0 c'$, the proof is complete.

□

C.3 Proof of the lower bound in Theorem 3.3

Proof. Let $\boldsymbol{X}_i^{(m)} = (X_i(\boldsymbol{s}_1), \ldots, X_i(\boldsymbol{s}_m))$ denote the discretely observed trajectory of the ith sample, and let $\widetilde{C}_m$ be any classifier constructed from $\{(Y_i, \boldsymbol{X}_i^{(m)}, X_i \in \mathcal{S}^r)\}_{i=1}^n$. Since $\boldsymbol{X}_i^{(m)}$ is a measurable function of the fully observed trajectory X_i, any classifier based on $\{\boldsymbol{X}_i^{(m)}\}_{i=1}^n$ has access to no more information than one based on $\{X_i\}_{i=1}^n$. Consequently,

$$\inf_{\widetilde{C}_m} \sup_{Q\in\mathcal{Q}} \mathcal{E}_h(\widetilde{C}_m) \geq \inf_{\widetilde{C}} \sup_{Q\in\mathcal{Q}} \mathcal{E}_h(\widetilde{C}),$$

where the infimum on the right-hand side is taken over all classifiers based on fully observed trajectories. By Theorem 3.1, this yields the statistical lower bound $c_1\, \underline{r}_n^2$.

To establish the discretization-induced lower bound, consider the restriction of $\mathcal{Q}$ to distributions for which the conditional class probability $\eta(X)$ depends on oscillatory components of X that are not identifiable from $\{X(\boldsymbol{s}_j)\}_{j=1}^m$. Because $X \in \mathcal{S}^\tau$, there exist functions $x_+, x_- \in \mathcal{S}^\tau$ such that $x_+(\boldsymbol{s}_j) = x_-(\boldsymbol{s}_j)$ for all $j = 1, \ldots, m$, while $\|x_+ - x_-\|_2 \asymp m^{-\tau}$. Such pairs can be constructed, for instance, by adding localized high-frequency oscillations supported between grid points in Sobolev, Hölder, or Besov balls of smoothness τ.

Define a subfamily of distributions $Q \subset \mathcal{Q}$ for which X takes values in $\{x_+, x_-\}$ with equal probability and

$$\eta(x_\pm) = \frac{1}{2} \pm c\, m^{-\tau},$$

for a sufficiently small constant $c > 0$. Under this construction, the discretized observations $\boldsymbol{X}_i^{(m)}$ have identical distributions for x_+ and x_-, and hence no classifier $\widetilde{C}_m$ based on $\{\boldsymbol{X}_i^{(m)}\}_{i=1}^n$ can distinguish between the two hypotheses. As a result, any such classifier incurs an excess risk bounded from below by $\mathbb{E}|\eta(X) - 1/2| \asymp m^{-\tau}$.

Assumption 1 further implies that the margin behavior of $\eta(X)$ converts this perturbation into a classification excess risk of order $m^{-\tau(\alpha+1)}$. More precisely, the functional margin condition ensures that when $|\eta(X) - 1/2|$ is of order $m^{-\tau}$, there exists a constant c_1', such that the corresponding excess risk is at least of $c_1' m^{-\tau(\alpha+1)}$ uniformly over $Q \in \mathcal{Q}$.

Combining the statistical lower bound $\underline{r}_n^2$ with the discretization-induced lower bound $m^{-\tau(\alpha+1)}$, by letting $c_3 = c_1 \wedge c_1'$, it yields

$$\inf_{\widetilde{C}_m} \sup_{Q \in \mathcal{Q}} \mathcal{E}_h(\widetilde{C}_m) \;\geq\; c_3\big(\underline{r}_n^2 + m^{-\tau(\alpha+1)}\big),$$

which completes the proof. □

C.4 Proof of the upper bound in Theorem 3.3

Proof. Let $\boldsymbol{X}_i^{(m)} = (X_i(\boldsymbol{s}_1), \ldots, X_i(\boldsymbol{s}_m))$ denote the discretely observed trajectory of the ith sample, and let

$$\widetilde{X}_i := \mathcal{O}_m(\boldsymbol{X}_i^{(m)}) \in L^2([0,1])$$

be its reconstruction according to Assumption 3. The classifier $\widehat{C}$ is constructed by applying the functional transformer architecture in Assumption 5 to the reconstructed sample $\{(\widetilde{X}_i, Y_i)\}_{i=1}^n$.

Denote by C_m^{Q*} the Bayes classifier based on the reconstructed trajectory $\widetilde{X}$, and by C^{Q*} the Bayes classifier based on the fully observed trajectory X. By definition of excess risk,

$$\mathcal{E}_Q(\widehat{C}^{FT}) = \mathbb{E}\Big[R_Q(\widehat{C}^{FT}) - R_Q(C_m^{Q*})\Big] + \mathbb{E}\big[R_Q(C_m^{Q*}) - R_Q(C^{Q*})\big]\,.$$

The first term corresponds to the estimation error when learning from the reconstructed functional data. Since $\widetilde{X}_i \in \mathcal{S}^\tau$ and the decision function η satisfies Assumptions 1 and 2, the learning problem based on $\{(\widetilde{X}_i, Y_i)\}_{i=1}^n$ falls within the fully observed functional framework analyzed earlier. Applying the upper bound in Theorem 3.1 for fully observed functional data yields

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}\Big[R_Q(\widehat{C}^{FT}) - R_Q(C_m^{Q*})\Big] \;\leq\; c_2\, r_n^2.$$

It remains to control the discretization-induced approximation error $R_Q(C_m^{Q*}) - R_Q(C^{Q*})$. By Assumption 3, the reconstruction operator satisfies

$$\|X - \widetilde{X}\|_2 \leq C\, m^{-\tau} \quad \text{uniformly over } X \in \mathcal{S}^\tau.$$

Since $\eta \in \widetilde{\mathcal{F}}_\gamma$ and $\mathcal{G}$ is an isometry, η is Lipschitz on $\mathcal{S}^\tau$ with respect to the L^2 norm. Consequently, there exists a constant $C' > 0$ such that

$$|\eta(X) - \eta(\widetilde{X})| \leq C'\, m^{-\tau}.$$

If the Bayes decisions based on X and $\widetilde{X}$ differ, then necessarily $|\eta(X) - 1/2| \leq C' m^{-\tau}$. Therefore,

$$R_Q(C_m^{Q*}) - R_Q(C^{Q*}) \leq \mathbb{E}\big[|2\eta(X) - 1|\, \mathbf{1}\big\{|\eta(X) - 1/2| \leq C' m^{-\tau}\big\}\big]\,.$$

Applying the functional margin condition in Assumption 1 yields

$$R_Q(C_m^{Q*}) - R_Q(C^{Q*}) \;\leq\; c_2'\, m^{-\tau(\alpha+1)}.$$

Combining the bounds for the estimation and discretization terms, by letting $c_4 = c_2 \vee c_2'$, it holds that

$$\sup_{Q \in \mathcal{Q}} \mathcal{E}_Q(\widehat{C}^{FT}) \;\leq\; c_4\big(r_n^2 + m^{-\tau(\alpha+1)}\big),$$

which establishes the desired upper bound and completes the proof. □

D Detailed simulation settings

This section provides the full data-generating mechanisms and additional discussion for the simulation study in Section 4. For the 1D CNN, each function is treated as a one-dimensional sequence, and convolution and pooling are applied along the domain to extract local features. We consider six settings covering 1D, 2D, and 3D functional classification, including both Gaussian and non-Gaussian scenarios.

DGP1 (1D) Generate

$$X(s) = \sum_{j=1}^{3} \xi_j \psi_j(s), \qquad s \in [0,1],$$

where

$$\psi_1(s) = \log(s+2), \qquad \psi_2(s) = s, \qquad \psi_3(s) = s^3.$$

Under class k, generate independently

$$(\xi_1, \xi_2, \xi_3)^\top \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where

$$\boldsymbol{\mu}_1 = (-1, 2, -3)^\top, \qquad \boldsymbol{\Sigma}_1^{1/2} = \operatorname{diag}\left(\frac{3}{5}, \frac{2}{5}, \frac{1}{5}\right),$$

and

$$\boldsymbol{\mu}_{-1} = \left(-\frac{1}{2}, \frac{5}{2}, -\frac{5}{2}\right)^\top, \qquad \boldsymbol{\Sigma}_{-1}^{1/2} = \operatorname{diag}\left(\frac{9}{10}, \frac{1}{2}, \frac{3}{10}\right).$$

DGP2 (1D) Generate

$$X(s) = \sum_{j=1}^{3} \xi_j \psi_j(s), \qquad s \in [0,1],$$

with the same basis functions as in DGP1. Under class 1,

$$(\xi_1, \xi_2, \xi_3)^\top \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),$$

where

$$\boldsymbol{\mu}_1 = (-1, 2, -3)^\top, \qquad \boldsymbol{\Sigma}_1^{1/2} = \operatorname{diag}(3, 2, 1).$$

Under class -1, generate ξ_j independently from t_{7-2j} for $j = 1, 2, 3$.

DGP3 (2D) Generate

$$X(s_1, s_2) = \sum_{j=1}^{4} \xi_j \psi_j(s_1, s_2), \qquad 0 \leq s_1, s_2 \leq 1,$$

where

$$\psi_1(s_1, s_2) = s_1 s_2, \qquad \psi_2(s_1, s_2) = s_1 s_2^2,$$
$$\psi_3(s_1, s_2) = s_1^2 s_2, \qquad \psi_4(s_1, s_2) = s_1^2 s_2^2.$$

Under class k, generate independently

$$(\xi_1, \xi_2, \xi_3, \xi_4)^\top \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where

$$\boldsymbol{\mu}_1 = (8, -6, 4, -2)^\top, \qquad \boldsymbol{\Sigma}_1^{1/2} = \mathrm{diag}(8, 6, 4, 2),$$

and

$$\boldsymbol{\mu}_{-1} = \left(-\frac{7}{2}, -\frac{5}{2}, \frac{3}{2}, -\frac{1}{2}\right)^\top, \qquad \boldsymbol{\Sigma}_{-1}^{1/2} = \mathrm{diag}\left(\frac{9}{2}, \frac{7}{2}, \frac{5}{2}, \frac{3}{2}\right).$$

DGP4 (2D) Generate

$$X(s_1, s_2) = \sum_{j=1}^{4} \xi_j \psi_j(s_1, s_2), \qquad 0 \le s_1, s_2 \le 1,$$

with the same basis functions as in DGP3. For $j = 1, 2, 3, 4$, under class 1, generate ξ_j independently from $t_{2j}(0)$. Under class -1, generate ξ_j independently from the non-central t distribution $t_{2j+1}(\mu_j)$, where

$$\mu_1 = 2, \qquad \mu_2 = \frac{3}{2}, \qquad \mu_3 = 1, \qquad \mu_4 = \frac{1}{2}.$$

DGP5 (3D) Generate

$$X(s_1, s_2, s_3) = \sum_{j=1}^{8} \xi_j \psi_j(s_1, s_2, s_3), \qquad 0 \le s_1, s_2, s_3 \le 1,$$

where

$$\psi_1(s_1, s_2, s_3) = s_1 s_2 s_3, \qquad \psi_2(s_1, s_2, s_3) = s_1^2 s_2 s_3, \qquad \psi_3(s_1, s_2, s_3) = s_1 s_2^2 s_3,$$
$$\psi_4(s_1, s_2, s_3) = s_1 s_2 s_3^2, \qquad \psi_5(s_1, s_2, s_3) = s_1^2 s_2^2 s_3, \qquad \psi_6(s_1, s_2, s_3) = s_1^2 s_2 s_3^2,$$
$$\psi_7(s_1, s_2, s_3) = s_1 s_2^2 s_3^2, \qquad \psi_8(s_1, s_2, s_3) = s_1^2 s_2^2 s_3^2.$$

Under class 1, generate

$$(\xi_1, \ldots, \xi_8)^\top \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),$$

and under class -1, generate

$$(\xi_1, \ldots, \xi_8)^\top \sim N(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1}),$$

where

$$\boldsymbol{\mu}_1 = (1.0, -0.7, 0.7, -0.5, 0.5, -0.35, 0.25, -0.15)^\top, \qquad \boldsymbol{\mu}_{-1} = -\boldsymbol{\mu}_1,$$

and

$$\boldsymbol{\Sigma}_1 = \boldsymbol{D} + \boldsymbol{A}_1, \qquad \boldsymbol{\Sigma}_{-1} = \boldsymbol{D} + \boldsymbol{A}_{-1},$$

where

$$\boldsymbol{D} = \mathrm{diag}(1.2, 1.1, 1.0, 0.9, 0.9, 0.8, 0.7, 0.6).$$

Let

$$L = \{1, 2, 3, 4\}, \qquad H = \{5, 6, 7, 8\},$$

and, for $k \in \{1, -1\}$, let $\boldsymbol{A}_k = (a_{ij}^{(k)})_{1 \le i,j \le 8}$ be symmetric with

$$a_{ii}^{(k)} = 0,$$

and, for $i \neq j$,

$$a_{ij}^{(k)} = \begin{cases} 0.18, & i,j \in L \text{ or } i,j \in H, \\ 0.22, & (i,j) \in \mathcal{E}_k \text{ or } (j,i) \in \mathcal{E}_k, \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\mathcal{E}_1 = \{(1,5),(2,6),(3,7),(4,8)\}, \qquad \mathcal{E}_{-1} = \{(1,6),(2,7),(3,8),(4,5)\}.$$

DGP6 (3D) Generate

$$X(s_1,s_2,s_3) = \sum_{j=1}^{8} \xi_j \psi_j(s_1,s_2,s_3), \qquad 0 \leq s_1,s_2,s_3 \leq 1,$$

with the same basis functions as in DGP5. For $j = 1,\ldots,8$, under class 1, generate ξ_j independently from $t_{\nu_j}(0)$ with

$$(\nu_1,\ldots,\nu_8) = (4,6,8,10,12,14,16,18).$$

Under class -1, generate ξ_j independently from the non-central t distribution $t_{\nu_j+1}(\delta_j)$, where

$$(\delta_1,\ldots,\delta_8) = (2,1.75,1.5,1.25,1,0.75,0.5,0.25).$$

For each DGP, we generate n training functions and 500 testing functions for each class, where $n \in \{40,100,200,400\}$ and the training samples are evenly split between the two classes. The reported misclassification rates are based on 100 Monte Carlo replications. The 1D, 2D, and 3D functional observations are recorded on grids of size 50, 30×30, and $20 \times 20 \times 20$, respectively. We use the Fourier basis for 1D data and tensor-product Fourier bases for 2D and 3D data to extract projection scores.

Table 2 reports the results summarized in Section 4. FT attains the lowest error at every sample size in DGP1, DGP2 and DGP6. For DGP3 and DGP4, it has the lowest error at three and two of the four sample sizes, respectively. In DGP5, FT attains the lowest error at $n = 100$, while ViT has the lowest error at $n = 40$ and CNN at $n = 200$ and 400. Thus, the comparisons support FT across functional dimensions while also showing settings in which ViT or CNN performs better.

E Additional details for the real-data experiments

This section provides the dataset descriptions and preprocessing details for the real-data experiments reported in Section 5.

Speech benchmark (1D). The first dataset is a standard speech-recognition benchmark derived from the TIMIT Acoustic-Phonetic Continuous Speech Corpus, which has been widely used in functional data analysis and speech classification studies [Ferraty and Vieu, 2003]. We retain the three phoneme classes that appear in Table 3: `aa`, `ao`, and `iy`. Each speech frame is represented through its log-periodogram, and the first 150 frequency components are used as the functional input. For each pairwise classification task, we use 300 samples from each class for training and the remaining 100 samples from each class for testing. The reported results are averages over repeated random splits.

Chest radiographs (2D). The second dataset consists of chest radiographs collected from U.S. coal workers and maintained by the National Institute for Occupational Safety and Health (NIOSH). We consider the binary task of classifying normal versus abnormal radiographs. A curated set of $1,042$ posterior-anterior chest radiographs is used for model development, and an additional test set of 131 radiographs is used for evaluation. The images were preprocessed by the MIDI lab at Michigan State University, including resizing each DICOM image to 512×512 pixels and extracting radiomic representations. Consensus readings from certified B-readers are treated as the reference labels.

ADNI PET images (3D). The third dataset is drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We focus on the binary classification problem of distinguishing Alzheimer's disease (AD) from early mild cognitive impairment (EMCI), which is substantially more challenging

than comparisons involving cognitively normal controls. The analysis uses PET scans from 79 AD subjects and 45 EMCI subjects. Following the study design, we use the third-visit scan for AD subjects and the second-visit scan for EMCI subjects. All images were spatially normalized and reoriented to $79 \times 95 \times 68$ voxels. Since the most discriminative signal is concentrated away from the final slices, we retain the first 60 slices for analysis. For evaluation, we repeatedly split the data into balanced training and testing sets using a 7:3 ratio.

Overall, these three datasets provide complementary empirical evidence for the proposed method across one-, two-, and three-dimensional functional inputs. The speech data assess low-dimensional functional discrimination, the chest-radiograph study evaluates 2D image classification in a clinically relevant setting, and the ADNI experiment examines a more demanding 3D neuroimaging task involving two closely related disease groups.

F Additional numerical results

This section examines the effects of sampling resolution, truncation level, attention window and basis representation, and reports the expanded real-data comparison. These experiments supplement the original results in Tables 2 and 3. Parentheses denote standard deviations in the attention and representation sensitivity tables, and standard errors in the real-data tables.

F.1 Resolution dependence and added classical baselines

For all six DGPs and $n \in \{40, 100, 200, 400\}$, the underlying process is fixed within each DGP and every method is re-tuned at each grid resolution. The 1D resolutions are $m \in \{10, 20, 30, 50\}$, the 2D resolutions are $m \in \{10^2, 20^2, 30^2\}$, and the 3D resolutions are $m \in \{10^3, 15^3, 20^3\}$. Table 5 reports the low- and high-resolution endpoints of FT's path. Table 6 compares the endpoints at $n = 400$ with FDNN, FPCA+LDA, FPCA+SVM and the strongest remaining comparator. The two added FPCA classifiers use FPCA scores in every functional dimension, complementing the original functional and image-based baselines.

Table 5: FT misclassification rates (%) from the lowest to highest resolution. Each entry reports low $\rightarrow$ high resolution.

DGP	Resolution	$n = 40$	$n = 100$	$n = 200$	$n = 400$
DGP1	$10 \rightarrow 50$	$30.01 \rightarrow 17.93$	$25.94 \rightarrow 16.30$	$24.71 \rightarrow 15.87$	$21.65 \rightarrow 13.72$
DGP2	$10 \rightarrow 50$	$43.38 \rightarrow 19.75$	$39.65 \rightarrow 12.74$	$38.41 \rightarrow 11.86$	$35.15 \rightarrow 9.56$
DGP3	$10^2 \rightarrow 30^2$	$22.88 \rightarrow 16.30$	$21.18 \rightarrow 14.41$	$16.17 \rightarrow 13.23$	$14.06 \rightarrow 12.31$
DGP4	$10^2 \rightarrow 30^2$	$13.44 \rightarrow 13.26$	$13.38 \rightarrow 12.97$	$12.92 \rightarrow 12.45$	$12.72 \rightarrow 12.25$
DGP5	$10^3 \rightarrow 20^3$	$15.73 \rightarrow 15.72$	$11.40 \rightarrow 10.57$	$10.21 \rightarrow 8.33$	$8.38 \rightarrow 8.07$
DGP6	$10^3 \rightarrow 20^3$	$6.40 \rightarrow 6.24$	$5.75 \rightarrow 5.89$	$5.68 \rightarrow 5.78$	$5.59 \rightarrow 5.60$

Table 6: Low- to high-resolution comparison at $n = 400$. Entries are misclassification rates (%). "Best other" excludes FT, FDNN, FPCA+LDA and FPCA+SVM. Bold marks the lowest mean at each endpoint.

DGP	FT	FDNN	FPCA+LDA	FPCA+SVM	Best other
DGP1	$\mathbf{21.65} \rightarrow \mathbf{13.72}$	$23.10 \rightarrow 15.07$	$38.03 \rightarrow 38.50$	$49.22 \rightarrow 49.11$	36.98 (QD) $\rightarrow 36.75$ (QD)
DGP2	$\mathbf{35.15} \rightarrow \mathbf{9.56}$	$35.17 \rightarrow 20.40$	$49.83 \rightarrow 49.86$	$43.74 \rightarrow 43.97$	38.56 (NB) $\rightarrow 38.51$ (NB)
DGP3	$14.06 \rightarrow \mathbf{12.31}$	$15.26 \rightarrow 12.74$	$42.27 \rightarrow 41.61$	$15.80 \rightarrow 15.50$	$\mathbf{13.14}$ (ViT) $\rightarrow 13.16$ (CNN)
DGP4	$12.72 \rightarrow \mathbf{12.25}$	$12.41 \rightarrow 12.30$	$28.72 \rightarrow 29.02$	$14.37 \rightarrow 14.11$	$\mathbf{12.23}$ (ViT) $\rightarrow 12.28$ (CNN)
DGP5	$8.38 \rightarrow 8.07$	$8.89 \rightarrow 9.11$	$23.63 \rightarrow 23.16$	$11.34 \rightarrow 10.26$	$\mathbf{6.78}$ (CNN) $\rightarrow \mathbf{7.92}$ (CNN)
DGP6	$5.59 \rightarrow \mathbf{5.60}$	$5.63 \rightarrow 5.74$	$27.67 \rightarrow 28.05$	$6.39 \rightarrow 6.45$	$\mathbf{5.57}$ (CNN) $\rightarrow 5.61$ (CNN)

FT has lower error at the highest resolution than at the lowest in 21 of 24 comparisons. The three exceptions occur in DGP6, where error increases by 0.14, 0.10 and 0.01 percentage points for $n = 100, 200, 400$, respectively. DGP1 and DGP2 show the largest resolution effects, DGP3 a moderate effect, and DGP4 to DGP6 are closer to a plateau. At $n = 400$, FT is best in five of six DGPs at the highest resolution but only two at the lowest. CNN remains better in DGP5 by

0.15 percentage points at the highest resolution. These patterns are qualitatively consistent with the resolution dependence in Theorem 3.3, which gives asymptotic rate bounds up to constants and logarithmic factors.

F.2 Joint sensitivity to truncation and attention window

Table 7 covers representative 1D, 2D and 3D settings. The finite truncation J is selected jointly with U by validation, subject to the observed grid resolution. Full attention contains the local class, and $U \geq J - 1$ recovers full attention for J retained tokens. The table uses $U = J$ as the full-attention convention. The theoretical bounds use an increasing local window to attain the near-sharp rate. The resulting finite-dimensional softmax operations allow us to control the covering number of the attention class.

For all 15 examined values of J, the best tested local window has lower mean error than full attention at the same J, by 0.12 to 2.14 percentage points. From the smallest to the largest examined J, the best observed local error falls by 3.28, 6.40 and 5.14 points in DGP1, DGP3 and DGP5, respectively. These results support using a sufficiently large candidate range for J and selecting J jointly with U by validation. Establishing adaptive minimax guarantees for this selection remains an open problem.

Table 7: Sensitivity of FT to J and U. Rates are percentages, parentheses give standard deviations, and the difference is full attention minus the best local rate. Bold marks the lowest local mean within each setting.

Setting	J	Best local U	Best local rate	Full ($U = J$)	Difference
DGP1, $n = 400, m = 30$	9	7	21.50 (0.05)	23.11 (0.07)	1.61
	13	11	20.10 (0.05)	21.55 (0.05)	1.45
	17	5	20.05 (0.04)	21.47 (0.08)	1.42
	21	9	18.86 (0.05)	20.36 (0.07)	1.50
	25	11	**18.22 (0.03)**	20.31 (0.06)	2.09
DGP3, $n = 100, m = 20^2$	25	5	21.70 (0.04)	21.94 (0.04)	0.24
	81	15	19.57 (0.04)	19.69 (0.05)	0.12
	169	15	17.21 (0.05)	17.79 (0.04)	0.58
	225	25	15.87 (0.03)	16.04 (0.04)	0.17
	289	81	**15.30 (0.02)**	15.86 (0.04)	0.56
DGP5, $n = 100, m = 15^3$	45	15	16.16 (0.08)	18.26 (0.09)	2.10
	125	15	12.95 (0.05)	15.02 (0.06)	2.07
	245	15	11.07 (0.02)	13.21 (0.04)	2.14
	405	30	11.26 (0.02)	13.09 (0.04)	1.83
	605	15	**11.02 (0.02)**	11.15 (0.03)	0.13

F.3 Fourier-score and FPCA-score representations

Table 8 compares Fourier and FPCA scores within the same FT architecture at $n = 400$. The Fourier-score version has lower mean error in all six DGPs, by 0.20 to 11.87 percentage points. This comparison supports the use of Fourier scores in the six simulated settings. Our theoretical analysis uses a fixed orthonormal basis. An extension to an estimated FPCA basis would require controlling the additional basis-estimation error.

Table 8: Fourier-score versus FPCA-score FT at $n = 400$. Rates are percentages. Parentheses give standard deviations. Difference is FPCA minus Fourier, in percentage points.

DGP	Fourier-score FT	FPCA-score FT	Difference
DGP1	**13.72 (0.02)**	15.24 (0.31)	1.52
DGP2	**9.56 (0.01)**	21.43 (0.79)	11.87
DGP3	**12.31 (0.01)**	12.69 (0.17)	0.38
DGP4	**12.25 (0.01)**	13.26 (0.51)	1.01
DGP5	**8.07 (0.01)**	9.63 (0.38)	1.56
DGP6	**5.60 (0.01)**	5.80 (0.15)	0.20

F.4 Complete revised real-data comparison

Table 9 compares all methods after expanding the ADNI validation grid and adding FPCA+LDA and FPCA+SVM. The expanded ADNI candidate grid was specified before examining the revised test results. Each replication selected the Fourier truncation, number of tokens, local window, embedding dimension, heads, Transformer layers and MLP width using validation data only, then evaluated the held-out test data after fixing the configuration. The revised ADNI FT result, 10.017%, improves on CNN's 10.314% by 0.297 percentage points. The improvement in mean error is modest. The ADNI runtime and memory measurements in Section 5 apply to the original FT configuration.

Table 9: Complete revised real-data misclassification rates (%). Parentheses give standard errors. Bold marks the lowest mean for each task across both panels. QD and NB apply only to speech tasks.

Task	FT	FDNN	ViT	CNN
Speech: aa vs ao	**19.130 (0.028)**	20.744 (0.016)	24.465 (0.037)	23.700 (0.025)
Speech: aa vs iy	**0.010 (0.001)**	0.193 (0.002)	0.215 (0.005)	0.330 (0.004)
Speech: ao vs iy	**0.010 (0.001)**	0.183 (0.004)	0.235 (0.004)	0.310 (0.004)
Chest: normal/abnormal	**22.135 (0.019)**	25.788 (0.016)	25.404 (0.017)	26.587 (0.107)
ADNI: AD vs EMCI	**10.017 (0.059)**	19.567 (0.078)	16.533 (0.074)	10.314 (0.043)
Task	QD	NB	FPCA+LDA	FPCA+SVM
Speech: aa vs ao	25.402 (0.026)	25.378 (0.021)	25.410 (0.281)	23.475 (0.260)
Speech: aa vs iy	0.288 (0.005)	0.273 (0.006)	0.290 (0.031)	0.035 (0.013)
Speech: ao vs iy	0.578 (0.005)	0.232 (0.005)	0.585 (0.046)	0.085 (0.020)
Chest: normal/abnormal	—	—	30.252 (0.327)	30.412 (0.330)
ADNI: AD vs EMCI	—	—	20.600 (0.341)	16.050 (0.510)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: Section 1 addresses the limitations of using a deep neural network based classifier. Section 2 introduces functional transformer (FT) as a solution. Theoretical properties of FT are rigorously discussed in Section 3, and numerical study results are presented in Section 4 and 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:We discuss the limitations of this work in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The assumptions for the theoretical results are stated in Section 3, and additional assumptions are saved in Appendix Section A. The complete proofs are provided in Appendix Section C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Section 4 specifies the simulation design, including the sample sizes, grid resolutions, number of Monte Carlo replications, competing methods, and basis systems used for feature extraction. Section 5 summarizes the real-data tasks, and the full simulation mechanisms and real-data preprocessing details are deferred to Appendix Sections D and E, respectively.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized code package with the experiment scripts and run instructions is provided in the supplementary material. The simulation experiments are fully reproducible from the released scripts, and the real-data experiments rely on benchmark datasets whose access and preprocessing procedures are documented in Appendix Section E. The 1D speech experiment uses the phoneme log-periodogram dataset from the archived R package `ElemStatLearn`. The 2D chest radiograph data and the 3D ADNI PET data are not publicly accessible due to data-use restrictions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation and tuning details are provided in Appendix B. The full simulation mechanisms are provided in Appendix D, and real-data preprocessing details are provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Tables 2 and 3 report misclassification rates with standard errors across repeated Monte Carlo replications or repeated random splits. Tables 4 and 9 report standard errors. Tables 7 and 8 report standard deviations. Tables 5 and 6 report endpoint means.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The runtime and GPU-memory comparisons for the original configurations are reported in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The authors cite all references to the datasets used and include the licenses for the code employed to run some of the baseline methods.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper focuses on developing a rigorous statistical framework for transformer-based functional classifiers and showing the proposed classifier attains minimax-optimal excess-risk. Given the scope and focus of the paper, the discussion of negative societal impacts is deemed irrelevant and therefore not included.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not involve text datasets and image generators, nor does it deal with high-risk models. The focus is on functional data classification, which does not present significant misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly credits the creators and original owners of all assets used. All usage complies with the respective licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce any new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: This paper does not describe the usage of LLMs. We focus on functional data classification problem.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.