BEYOND LIKELIHOOD: ROBUST MULTICLASS CLASSIFICATION WITH TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a robust multiclass classification approach based on the rho-estimation framework. An oracle inequality for the resulting estimator controls estimation error when the model is well specified and provides robustness guarantees under model misspecification and data contamination. For i.i.d. data, we further derive an upper bound on the excess misclassification error of the corresponding plug-in classifier. We then explore its implementation with Transformers. Our analysis establishes a constructive approximation result for Hölder functions near low-dimensional manifolds. The construction implements a stable projection through standard self-attention and separates network approximation error from the effect of neighborhood thickness. Building on this approximation result, we derive explicit convergence rates for softmax Transformer logit models, with exponents governed by the projection dimension rather than the ambient dimension. Experiments on synthetic and real-world datasets support the theoretical findings.

1 INTRODUCTION

Classification is a central problem in modern machine learning, with applications ranging from image recognition and language understanding to medical diagnosis (Krizhevsky et al., 2012; Devlin et al., 2019; Esteva et al., 2017). In multiclass problems, probabilistic classifiers estimate conditional class probabilities and use them to make predictions. These models are commonly trained by maximizing conditional likelihood, or equivalently by minimizing cross-entropy, as in multinomial logistic regression and many deep neural classifiers (Hastie et al., 2009; Krizhevsky et al., 2012). In practice, however, both model specification and data quality pose challenges. First, the candidate model class may not contain the true conditional distribution, even when the data are uncontaminated. Meaningful statistical guarantees must therefore account for the discrepancy between the truth and the chosen model class. Second, large-scale data collection makes data quality difficult to control, and labeling errors have been documented even in widely used benchmarks (Northcutt et al., 2021). Such errors can undermine cross-entropy training (Ghosh et al., 2017). For example, a mislabeled observation may receive an extremely small probability $Q_f(y \mid x)$ under a model that describes the clean data well. Its negative log-likelihood contribution $-\log Q_f(y \mid x)$ can then be arbitrarily large and contribute disproportionately to the empirical loss. These challenges motivate robust methods for multiclass conditional-distribution estimation, with risk guarantees that explicitly accommodate both model misspecification and data contamination.

A classical approach to robust estimation modifies the loss to reduce sensitivity to extreme observations, as exemplified by Huber's M-estimation principle (Huber, 1964). Subsequent developments include high-dimensional robust M-estimation and adaptive calibration of Huber regression (Loh, 2017; 2021). Median-of-means (MOM) ideas provide another approach, with applications to multivariate mean estimation (Lugosi & Mendelson, 2019) and robust binary classification (Lecué et al., 2020). Estimation through robust tests underlies T-estimators (Birgé, 2006) and the subsequent ρ-estimation framework (Baraud et al., 2017; Baraud & Birgé, 2018). The latter provides Hellinger risk guarantees that explicitly account for model misspecification and contamination. Building on this framework, Baraud & Chen (2024) study regression in one-parameter exponential families, including binary logistic regression. Their theoretical analysis leverages the specific density structure of these families to transfer VC-subgraph dimension bounds from the scalar-valued parameter class to the induced density class. This argument, unfortunately, does not directly extend to multiclass softmax models, where the normalization couples all class probabilities. The extension to the multiclass

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

case requires a dedicated analysis linking the complexity of vector-valued logit classes to that of the induced comparison functions. This is one of the problems we address in this paper and establish explicit risk guarantees for conditional-distribution estimation and the resulting plug-in classifier.

A complementary question concerns the models used to represent these conditional probabilities. Transformers are widely used in machine translation, language understanding, and image recognition (Vaswani et al., 2017; Devlin et al., 2019; Dosovitskiy et al., 2021). Self-attention supports parallel computation and efficient training (Vaswani et al., 2017). Large-scale pretraining also enables effective transfer across tasks with limited architectural changes (Devlin et al., 2019; Dosovitskiy et al., 2021). These empirical successes motivate quantitative studies of Transformer approximation and learning (Yun et al., 2020; Jiao et al., 2025; Lai & Sun, 2026). Data geometry is particularly relevant to image classification, where estimated intrinsic dimensions can be far below pixel dimensions (Pope et al., 2021). Measurement noise further motivates approximate manifold support, under which Jiao et al. (2023) establish regression guarantees for deep ReLU networks. For Transformers, existing results cover ReLU attention on exact manifolds (Havrilla & Liao, 2024; Shi & Labate, 2026) and manifold neighborhoods with targets that factor through the nearest-point projection (Shen et al., 2025). Approximation and learning guarantees have also been established for softmax Transformers on exact manifolds (Shi & Liao, 2026). We characterize the approximation capabilities of standard softmax Transformers for general Hölder targets on manifold neighborhoods of positive thickness. Combining these bounds with the oracle inequality for multiclass ρ-estimation yields convergence rates for excess classification risk.

Our main contributions are as follows.

- **Robust multiclass conditional-distribution estimation.** Building on the ρ-general framework, we develop a ρ-estimation method for multiclass conditional distributions that requires neither knowledge of the covariate marginal distributions nor additional robustness tuning parameters. We establish a nonasymptotic oracle inequality for independent, potentially non-identically distributed data, providing robustness guarantees under model misspecification and contamination.
- **Approximation on approximate manifolds.** We show that standard softmax Transformers can exploit approximate low-dimensional structure when inputs lie within distance ζ of a low-dimensional manifold. Our construction implements a stable projection through self-attention and approximates β-Hölder targets with error $\varepsilon + O(\zeta^\beta)$ and parameter complexity $O(\varepsilon^{-q/\beta})$, where q is the effective projection dimension.
- **Explicit estimation and classification rates.** We establish convergence guarantees for multiclass learning on approximate low-dimensional manifolds. In the well-specified i.i.d. setting, the expected squared Hellinger and excess classification risks have upper bounds of orders

$$\varepsilon_n^2 + \zeta_n^{2\beta} \quad \text{and} \quad \varepsilon_n + \zeta_n^\beta, \qquad \varepsilon_n = \left(\frac{(\log n)^2}{n}\right)^{\beta/(2\beta+q)}.$$

  Here, ζ_n measures the manifold neighborhood thickness. When $\zeta_n \lesssim [(\log n)^2/n]^{1/(2\beta+q)}$, the estimator attains the typical nonparametric rates up to logarithmic factors, with exponents governed by the effective projection dimension q rather than the ambient dimension. Thus, low-dimensional learning rates remain achievable beyond exact manifold support, while our oracle inequality additionally guarantees robustness to misspecification and contamination.

The remainder of the paper is organized as follows. Section 2 introduces the statistical setting. Section 3 presents the estimation procedure and its theoretical guarantees. Section 4 develops the Transformer approximation theory and establishes convergence rates. Section 5 describes the numerical experiments. Proofs and additional experimental details are provided in the appendix.

Notation. Let $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, $\mathbb{N} = \mathbb{N}_0 \setminus \{0\}$, and $[m] = \{1, \ldots, m\}$ for $m \in \mathbb{N}$. For a vector $\boldsymbol{u}$, $\|\boldsymbol{u}\|_r$ denotes its ℓ_r-norm. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_{\mathrm{F}}$ and $\|\mathbf{A}\|_{\mathrm{op}}$ denote its Frobenius and spectral norms. For $x > 0$, write $\log_+(x) = \max\{1, \log x\}$, and let $\mathbf{1}_D = (1, \ldots, 1)^\top \in \mathbb{R}^D$.

2 Statistical Setting

Let $K \geq 2$ be the number of classes and let $\mathcal{X} = [0,1]^{d_{\mathrm{in}} \times D}$ be the covariate space. We observe independent, not necessarily identically distributed, random pairs $W_i = (X_i, Y_i)$, $i \in [n]$, taking

values in $\mathcal{W} = \mathcal{X} \times [K]$. For each i, let P_{X_i} stand for the marginal distribution of X_i and let $Q_i^\star(\cdot \mid x) \in \Delta_K$ be a version of the conditional distribution of Y_i given $X_i = x$, where

$$\Delta_K = \left\{ p = (p_1, \ldots, p_K) \in [0,1]^K : \sum_{k=1}^K p_k = 1 \right\}.$$

Thus the joint distribution $P_i^\star$ of W_i satisfies $P_i^\star(dx, \{y\}) = Q_i^\star(y \mid x) P_{X_i}(dx)$, and the full sample has distribution $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$. The goal is to estimate the conditional-distribution tuple $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$, rather than only the associated Bayes classifier.

Let $\mathcal{F}$ be a class of vector-valued logit functions $f = (f_1, \ldots, f_K)^\top : \mathcal{X} \to \mathbb{R}^K$. Each $f \in \mathcal{F}$ induces the softmax conditional distribution

$$Q_f(k \mid x) = \frac{\exp(f_k(x))}{\sum_{\ell=1}^K \exp(f_\ell(x))}, \qquad k \in [K].$$

The softmax map is invariant under adding the same scalar function to all coordinates. An identifiability constraint such as $f_K = 0$ may therefore be imposed, but none of the results below requires a particular normalization.

For each $f \in \mathcal{F}$, let $P_{i,f}(dx, \{y\}) = Q_f(y \mid x) P_{X_i}(dx)$ and $\mathbf{P}_f = \bigotimes_{i=1}^n P_{i,f}$. For probability vectors $p, q \in \Delta_K$, their squared Hellinger distance is given by

$$h^2(p,q) = \frac{1}{2} \sum_{k=1}^K \left(\sqrt{p_k} - \sqrt{q_k} \right)^2.$$

For two tuples of conditional distributions $\mathbf{Q} = (Q_i)_{i=1}^n$ and $\mathbf{Q}' = (Q_i')_{i=1}^n$, we define the pseudo-Hellinger distance

$$\mathbf{h}^2(\mathbf{Q}, \mathbf{Q}') = \sum_{i=1}^n \int_{\mathcal{X}} h^2 \left(Q_i(\cdot \mid x), Q_i'(\cdot \mid x) \right) dP_{X_i}(x).$$

Equivalently, $\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f) = \sum_{i=1}^n h^2(P_i^\star, P_{i,f})$. Since Hellinger distance is independent of the choice of dominating measure, this loss compares conditional laws without relying on a particular density representation. In particular, in the i.i.d. case, where $P_{X_i} = P_X$ and $Q_i^\star = Q^\star$ for all i, we use the normalized risk

$$\bar{h}^2(Q^\star, Q_f) = \int_{\mathcal{X}} h^2 \left(Q^\star(\cdot \mid x), Q_f(\cdot \mid x) \right) dP_X(x) = \frac{1}{n} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f).$$

3 Methodology and Theoretical Performance

The estimation approach is based on ρ-estimation; see Baraud et al. (2017) and Baraud & Birgé (2018) for the theoretical foundations of a general estimation framework.

Let $\mathcal{F}$ be a finite or countable class of vector-valued logit functions $f : \mathcal{X} \to \mathbb{R}^K$. The countability requirement is imposed only to avoid measurability issues. In neural-network applications, one may first restrict the parameters to rational values and then pass to the full parameter space by a standard approximation argument.

Let $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ denote the observed sample. Define the comparison function

$$\psi(t) = \frac{t-1}{t+1}, \qquad t \in [0, +\infty],$$

with the convention $\psi(+\infty) = 1$. For $f, g \in \mathcal{F}$, define the empirical ρ-contrast

$$\rho(\mathcal{D}_n; f, g) = \sum_{i=1}^n \psi \left(\sqrt{\frac{Q_g(Y_i \mid X_i)}{Q_f(Y_i \mid X_i)}} \right).$$

For each $f \in \mathcal{F}$, define the ρ-criterion $\Upsilon(\mathcal{D}_n; f) = \sup_{g \in \mathcal{F}} \rho(\mathcal{D}_n; f, g)$. A ρ-estimator is any measurable element $\widehat{f}_\rho \in \mathcal{F}$ satisfying

$$\Upsilon(\mathcal{D}_n; \widehat{f}_\rho) \le \inf_{f \in \mathcal{F}} \Upsilon(\mathcal{D}_n; f) + 1. \tag{1}$$

The resulting estimator of $\mathbf{Q}^\star$ is given by $\mathbf{Q}_{\widehat{f}_\rho} = \left(Q_{\widehat{f}_\rho}, \ldots, Q_{\widehat{f}_\rho}\right)$. The above estimation procedure does not involve the covariate distributions P_{X_i}, which therefore can be unknown. The particular choice of the tolerance 1 in (1) is inessential and is made only for simplicity of presentation.

For $f, g : \mathcal{X} \to \mathbb{R}^K$, write

$$d_\infty(f, g) = \|f - g\|_{L^\infty(\mathcal{X}; \ell_\infty^K)} = \sup_{x \in \mathcal{X}} \max_{k \in [K]} |f_k(x) - g_k(x)|.$$

The covering number $\mathcal{N}(\eta, \mathcal{F}, d_\infty)$ is the minimum cardinality of a subset $\mathcal{C} \subseteq \mathcal{F}$ such that every $f \in \mathcal{F}$ satisfies $d_\infty(f, g) \le \eta$ for some $g \in \mathcal{C}$; it is infinite if no finite cover exists. The covering number measures the complexity of the logit class. It determines the variance term in the risk bound for the resulting ρ-estimator. In the following result, we establish a large-deviation inequality for the risk of an estimator based on a general class of vector-valued functions.

Theorem 1. *Let $\mathcal{F}$ be a finite or countable class of K-class logit functions, and let $\widehat{f}_\rho$ be a ρ-estimator based on $\{Q_f : f \in \mathcal{F}\}$. Suppose that there exist $V, A \ge 1$ such that, for every $0 < \eta \le 1$, $\log \mathcal{N}(\eta, \mathcal{F}, d_\infty) \le V \log(A/\eta)$. Assume further that $c_\mathcal{F} = \inf_{f \in \mathcal{F}} \inf_{(x,y) \in \mathcal{W}} Q_f(y \mid x) > 0$. There exists a universal constant $C > 0$ such that, for every $\xi > 0$,*

$$\mathbb{P}\left\{\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{f}_\rho}) > C\left[\inf_{f \in \mathcal{F}} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f) + V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_\mathcal{F}}}\right) + \xi\right]\right\} \le e^{-\xi}.$$

Integrating the above tail bound with respect to ξ yields

$$\mathbb{E}\left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{f}_\rho})\right] \le C\left[\inf_{f \in \mathcal{F}} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f) + V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_\mathcal{F}}}\right)\right]. \tag{2}$$

The bound comprises an approximation (bias) term and a complexity (variance) term. The former quantifies model misspecification and vanishes under correct specification, when $\mathbf{Q}^\star = \mathbf{Q}_{f^\star}$ for some $f^\star \in \mathcal{F}$. The robustness to contamination follows from the boundedness of Hellinger loss. Suppose that, for some $f^\star \in \mathcal{F}$ and $\mathcal{I} \subseteq [n]$, $Q_i^\star = Q_{f^\star}$, P_{X_i}-almost surely, for every $i \notin \mathcal{I}$. Then

$$\frac{1}{n}\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{Q}_{f^\star}\right) = \frac{1}{n}\sum_{i \in \mathcal{I}} \int_\mathcal{X} h^2\left(Q_i^\star(\cdot \mid x), Q_{f^\star}(\cdot \mid x)\right) dP_{X_i}(x) \le \frac{|\mathcal{I}|}{n}.$$

Hence, arbitrary contamination at a fraction $|\mathcal{I}|/n$ contributes at most the same order to the normalized approximation error. This stability contrasts with likelihood-based analysis through the Kullback–Leibler divergence, which admits no distribution-free upper bound.

For the remainder of this section, we specialize to the i.i.d. setting. Let (X, Y) be an independent test observation with $X \sim P_X$ and $Y \mid X = x \sim Q^\star(\cdot \mid x)$. For a measurable classifier $g : \mathcal{X} \to [K]$, define $\mathcal{R}(g) = \mathbb{P}\left(g(X) \ne Y\right)$. The Bayes risk is

$$\mathcal{R}^\star = \inf_{g : \mathcal{X} \to [K]} \mathcal{R}(g) = \int_\mathcal{X} \left[1 - \max_{k \in [K]} Q^\star(k \mid x)\right] dP_X(x),$$

where the infimum ranges over measurable classifiers. The estimator $Q_{\widehat{f}_{\rho,n}}$ induces plug-in classifier

$$\widehat{g}_{\rho,n}(x) = \min \arg\max_{k \in [K]} Q_{\widehat{f}_{\rho,n}}(k \mid x).$$

with ties broken by the smallest index. The following result bounds the excess risk $\mathcal{R}(\widehat{g}_{\rho,n}) - \mathcal{R}^\star$ in terms of the estimator's Hellinger loss. Its proof is deferred to the appendix.

Corollary 2. *Under the preceding i.i.d. setting, the plug-in classifier $\widehat{g}_{\rho,n}$ satisfies, almost surely,*

$$0 \le \mathcal{R}(\widehat{g}_{\rho,n}) - \mathcal{R}^\star \le 2\sqrt{2}\left\{\bar{h}^2\left(Q^\star, Q_{\widehat{f}_{\rho,n}}\right)\right\}^{1/2}.$$

4 Implementation using Transformers

In this section, we consider implementing the proposed method using transformer models. We first specify the adopted architecture, following Jiao et al. (2025) and Lai & Sun (2026), which retains the

embedding, multihead softmax self-attention, and tokenwise feedforward components of the original Transformer (Vaswani et al., 2017).

The embedding layer $F_{\mathrm{EB}} : \mathbb{R}^{d_{\mathrm{in}} \times D} \to \mathbb{R}^{d_0 \times D}$ is the affine map

$$F_{\mathrm{EB}}(\boldsymbol{x}) = \mathbf{W}_{\mathrm{EB}}\boldsymbol{x} + \mathbf{B}_{\mathrm{EB}}, \quad \text{where} \quad \mathbf{W}_{\mathrm{EB}} \in \mathbb{R}^{d_0 \times d_{\mathrm{in}}}, \quad \mathbf{B}_{\mathrm{EB}} \in \mathbb{R}^{d_0 \times D}.$$

A feedforward block $F_{\mathrm{FF}} : \mathbb{R}^{d_{\mathrm{FF}}^{(\mathrm{in})} \times D} \to \mathbb{R}^{d_{\mathrm{FF}}^{(\mathrm{out})} \times D}$ with depth L and width p acts columnwise. Starting from $\mathbf{F}^{(0)} = \boldsymbol{x}$, its hidden representations are

$$\mathbf{F}^{(\ell)} = \sigma_{\mathrm{ReLU}}\left(\mathbf{W}_\ell \mathbf{F}^{(\ell-1)} + \mathbf{b}_\ell \mathbf{1}_{1\times D}\right), \qquad \ell \in [L-1],$$

and its output is $F_{\mathrm{FF}}(\boldsymbol{x}) = \mathbf{W}_L \mathbf{F}^{(L-1)} + \mathbf{b}_L \mathbf{1}_{1\times D}$. Here $\sigma_{\mathrm{ReLU}}(x) = \max\{x, 0\}$ is applied entrywise. The weight matrices satisfy $\mathbf{W}_1 \in \mathbb{R}^{p \times d_{\mathrm{FF}}^{(\mathrm{in})}}$, $\mathbf{W}_\ell \in \mathbb{R}^{p\times p}$, for $\ell = 2, \ldots, L-1$, and $\mathbf{W}_L \in \mathbb{R}^{d_{\mathrm{FF}}^{(\mathrm{out})} \times p}$. The bias vectors satisfy $\mathbf{b}_\ell \in \mathbb{R}^p$, for $\ell \in [L-1]$, and $\mathbf{b}_L \in \mathbb{R}^{d_{\mathrm{FF}}^{(\mathrm{out})}}$. A self-attention layer $F_{\mathrm{SA}} : \mathbb{R}^{d_{\mathrm{SA}} \times D} \to \mathbb{R}^{d_{\mathrm{SA}} \times D}$ with H attention heads of dimension S is

$$F_{\mathrm{SA}}(\boldsymbol{x}) = \boldsymbol{x} + \sum_{h=1}^{H} \mathbf{W}_O^{(h)} \mathbf{W}_V^{(h)} \boldsymbol{x}\, \sigma_{\mathrm{S}}\left(\boldsymbol{x}^\top \mathbf{W}_K^{(h)\top} \mathbf{W}_Q^{(h)} \boldsymbol{x}\right),$$

where $\mathbf{W}_O^{(h)} \in \mathbb{R}^{d_{\mathrm{SA}} \times S}$, and $\mathbf{W}_V^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_Q^{(h)} \in \mathbb{R}^{S \times d_{\mathrm{SA}}}$. Here σ_{S} applies softmax to each column. In particular, $[\sigma_{\mathrm{S}}(x)]_i = \exp(x_i)/\sum_{j=1}^{D} \exp(x_j)$ for $x \in \mathbb{R}^D$ and $i \in [D]$.

A Transformer $T : \mathbb{R}^{d_{\mathrm{in}} \times D} \to \mathbb{R}^{d_{\mathrm{out}} \times D}$ of length R is obtained by first applying an embedding layer $F_{\mathrm{EB}} : \mathbb{R}^{d_{\mathrm{in}} \times D} \to \mathbb{R}^{d_0 \times D}$, followed by $R+1$ feedforward blocks $F_{\mathrm{FF}}^{(r)} : \mathbb{R}^{d_r \times D} \to \mathbb{R}^{d_{r+1} \times D}$, $r = 0, \ldots, R$, and R self-attention layers $F_{\mathrm{SA}}^{(r)} : \mathbb{R}^{d_r \times D} \to \mathbb{R}^{d_r \times D}$, $r \in [R]$, where $d_{R+1} = d_{\mathrm{out}}$. More precisely,

$$T(\boldsymbol{x}) = F_{\mathrm{FF}}^{(R)} \circ F_{\mathrm{SA}}^{(R)} \circ F_{\mathrm{FF}}^{(R-1)} \circ \cdots \circ F_{\mathrm{FF}}^{(1)} \circ F_{\mathrm{SA}}^{(1)} \circ F_{\mathrm{FF}}^{(0)} \circ F_{\mathrm{EB}}(\boldsymbol{x}). \tag{3}$$

The architecture of T is summarized by the ordered size tuple

$$\big((L_0, p_0), (H_1, S_1), (L_1, p_1), \ldots, (H_R, S_R), (L_R, p_R)\big)$$

and the dimension vector $\boldsymbol{d} = (d_{\mathrm{in}}, d_0, d_1, \ldots, d_R, d_{\mathrm{out}})$. Here, (L_r, p_r) specifies the depth and width of the r-th feedforward block, while (H_r, S_r) specifies the number of heads and the head dimension of the rth self-attention layer.

Let B_{EB}, B_{FF}, and B_{SA} denote upper bounds on the absolute values of the parameters in the embedding layer, feedforward blocks, and self-attention layers, respectively. Similarly, let M_{EB}, M_{FF}, and M_{SA} denote the corresponding total numbers of parameters.

4.1 Approximation on an approximate manifold

We examine whether Transformers can exploit latent low-dimensional covariate structure when the data lie near, rather than exactly on, a manifold. The neighborhood thickness contributes an explicit approximation term (Jiao et al., 2023).

Let $t = d_{\mathrm{in}} D$. We identify $\mathbb{R}^{d_{\mathrm{in}} \times D}$ with $\mathbb{R}^t$ through the column-wise vectorization map $\mathrm{vec} : \mathbb{R}^{d_{\mathrm{in}} \times D} \longrightarrow \mathbb{R}^t$. Since vectorization only rearranges the entries of a matrix,

$$\|\,\mathrm{vec}(\boldsymbol{x}) - \mathrm{vec}(\boldsymbol{x}')\|_2 = \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathrm{F}}, \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_{\mathrm{in}} \times D}. \tag{4}$$

Throughout this section, let $0 < \beta \le 1$. For a function $f : \mathcal{X} \to \mathbb{R}$, define

$$[f]_{\mathcal{H}^\beta(\mathcal{X})} = \sup_{\substack{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} \\ \boldsymbol{x} \neq \boldsymbol{x}'}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{x}')|}{\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathrm{F}}^\beta}, \quad \text{and} \quad \|f\|_{\mathcal{H}^\beta(\mathcal{X})} = \max\left\{\|f\|_{L^\infty(\mathcal{X})}, [f]_{\mathcal{H}^\beta(\mathcal{X})}\right\}.$$

For $B_0 > 0$, let $\mathcal{H}^\beta(\mathcal{X}, B_0) = \left\{f : \|f\|_{\mathcal{H}^\beta(\mathcal{X})} \le B_0\right\}$ denote the β-Hölder ball of radius B_0.

Assumption 1. *Let $\mathcal{M} \subset [0,1]^{d_{\text{in}} \times D}$ be a compact $d_{\mathcal{M}}$-dimensional Riemannian submanifold satisfying the geometric regularity conditions in Theorem 3.1 of Baraniuk & Wakin (2009). For $\zeta \in [0,1)$, define its closed ζ-neighborhood in $\mathcal{X}$ by $\mathcal{M}_\zeta = \{\boldsymbol{x} \in \mathcal{X} : \inf_{\boldsymbol{x}' \in \mathcal{M}} \|\boldsymbol{x} - \boldsymbol{x}'\|_{\text{F}} \le \zeta\}$ and assume that $P_{X_i}(\mathcal{M}_\zeta) = 1$, $i \in [n]$.*

Let $\mathfrak{M} = \text{vec}(\mathcal{M}) \subset [0,1]^t$. We denote by $V_{\mathcal{M}}$, $R_{\mathcal{M}}$, and $1/\tau$, respectively, the volume, geodesic covering regularity, and condition number appearing in the stable manifold projection theorem (Theorem 3.1 in Baraniuk & Wakin (2009)). The following result establishes a quantitative approximation guarantee for Transformers on approximate manifolds. The proof is provided in the appendix.

Theorem 3. *Suppose that Assumption 1 holds. Let $u_0 \in \mathcal{H}^\beta(\mathcal{X}, B_0)$ for some $0 < \beta \le 1$, $B_0 > 0$, and let $C_{\text{BW}} > 0$ be a sufficiently large universal constant. Fix $\delta \in (0,1)$ and define*

$$q = \min\left\{t,\ \max\left\{d_{\mathcal{M}},\ \left\lceil C_{\text{BW}} \frac{d_{\mathcal{M}}}{\delta^2} \log_+\left(\frac{t V_{\mathcal{M}} R_{\mathcal{M}}}{\tau \delta}\right)\right\rceil\right\}\right\}.$$

For any $\varepsilon \in (0,1)$, there exists a Transformer T_ε of the form defined in (3), with architecture specified in the proof, such that its scalar output $t_\varepsilon(\boldsymbol{x}) = [T_\varepsilon(\boldsymbol{x})]_{1,1}$ satisfies

$$\sup_{\boldsymbol{x} \in \mathcal{M}_\zeta} |t_\varepsilon(\boldsymbol{x}) - u_0(\boldsymbol{x})| \le \varepsilon + B_0 \left[\frac{1}{(1-\delta)^\beta}\left(\frac{t}{q}\right)^{\beta/2} + 1\right] \zeta^\beta.$$

The bound separates the network approximation error ε from the geometric error $O(\zeta^\beta)$, with the former dominating whenever $\zeta \lesssim \varepsilon^{1/\beta}$. The construction in the proof yields a Transformer with two self-attention layers and a total parameter count satisfying

$$M_{\text{EB}} + M_{\text{FF}} + M_{\text{SA}} \lesssim \left(\frac{\mathfrak{B}_{q,\delta}}{\varepsilon}\right)^{q/\beta},$$

where $\mathfrak{B}_{q,\delta}$ is independent of ζ and ε. Thus, the dependence of the parameter count on ε is governed by the effective projection dimension q rather than the ambient dimension $t = d_{\text{in}} D$. The choice of δ introduces a tradeoff: a smaller δ reduces the projection distortion but may increase q.

4.2 Convergence Rate

Building on the approximation result above, we derive a convergence rate in the well-specified i.i.d. setting, where the target logit function is Hölder-smooth and the covariates are supported in a neighborhood of a Riemannian manifold.

Assumption 2. *There exists a logit function $f_0 : \mathcal{X} \to \mathbb{R}^K$ with $f_{0,k} \in \mathcal{H}^\beta(\mathcal{X}, B_0)$ for all $k \in [K]$ such that $Q^\star(\cdot \mid x) = Q_{f_0}(\cdot \mid x)$ for P_X-almost every x, where $0 < \beta \le 1$ and $B_0 > 0$.*

For $\mathcal{B} \ge B_0$, let $\Pi_{\mathcal{B}}(u) = (-\mathcal{B}) \vee (u \wedge \mathcal{B})$ act coordinatewise. For $\varepsilon \in (0,1)$, let $\mathcal{T}_\varepsilon$ denote the Transformer class with the architecture and parameter bounds specified in the proof of Theorem 3. Define the clipped multiclass logit class by

$$\mathcal{F}_{\varepsilon,\mathcal{B}} = \left\{\boldsymbol{x} \mapsto \Pi_{\mathcal{B}}\left(\left([T_1(\boldsymbol{x})]_{1,1}, \ldots, [T_K(\boldsymbol{x})]_{1,1}\right)^\top\right) : T_k \in \mathcal{T}_\varepsilon,\ k \in [K]\right\}.$$

Corollary 4. *Suppose that Assumptions 1 and 2 hold in the i.i.d. setting, with $P_X(\mathcal{M}_{\zeta_n}) = 1$ for some $\zeta_n \in [0,1)$. Fix $\delta \in (0,1)$, and let q be defined as in Theorem 3. For $n \ge 3$, set $\varepsilon_n = [(\log n)^2/n]^{\beta/(2\beta+q)}$. Let $\mathcal{F}_n$ be a fixed countable d_∞-dense subclass of $\mathcal{F}_{\varepsilon_n,\mathcal{B}}$, and let $\widehat{f}_{\rho,n}$ be a ρ-estimator based on $\{Q_f : f \in \mathcal{F}_n\}$. Write*

$$\Gamma_{\mathcal{M}} = B_0 \left[\frac{1}{(1-\delta)^\beta}\left(\frac{t}{q}\right)^{\beta/2} + 1\right].$$

Then there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\bar{h}^2\left(Q^\star, Q_{\widehat{f}_{\rho,n}}\right)\right] \le C\left[\varepsilon_n^2 + \Gamma_{\mathcal{M}}^2 \zeta_n^{2\beta}\right].$$

In particular, if $\Gamma_{\mathcal{M}}\zeta_n^{\beta} \le \varepsilon_n$, *then*

$$\mathbb{E}\left[\bar{h}^2\left(Q^\star, Q_{\widehat{f}_{\rho,n}}\right)\right] \le C\left(\frac{(\log n)^2}{n}\right)^{2\beta/(2\beta+q)}.$$

The constant C *may depend on the fixed geometric, smoothness, and architectural quantities, including* K, $\mathcal{B}$, *and* δ, *but not on* n *or* ζ_n.

Up to logarithmic factors, the rate $n^{-2\beta/(2\beta+q)}$ matches the classical nonparametric rate for estimating β-Hölder functions in q dimensions. The effective projection dimension replaces the ambient dimension in the convergence exponent, mitigating the curse of dimensionality when q is substantially smaller than t. Together with Theorem 1, this result indicates that the ρ-estimator achieves the standard convergence rate when the model is well specified while remaining robust to small departures from the ideal setting. We further examine these properties through simulation studies.

5 Numerical Experiments

In this section, we conduct a small-scale study using synthetic and real-world data to support our theoretical findings and evaluate the practical performance of the method proposed in Section 3.

5.1 Searching for ρ-estimator

Finding an estimator satisfying (1) requires approximately solving a min–max problem, which can be challenging for neural network classes. This formulation resembles the adversarial optimization used in GANs (Goodfellow et al., 2014; Mescheder et al., 2018). However, the principle of estimation through testing that underlies ρ-estimation motivates the iterative search procedure described below.

For simplicity, suppose that the data are i.i.d. with distribution $P^\star$. As shown in Baraud (2021, Proposition 14), $\rho(\mathcal{D}_n; f, g)$ provides a robust test of H_0: P_f is closer to $P^\star$ in Hellinger distance than P_g, against H_1: P_g is closer to $P^\star$ than P_f. Specifically, for fixed $f, g \in \mathcal{F}$, $\rho(\mathcal{D}_n; f, g) > 0$ with high probability when $h^2(P_g, P^\star) \ll h^2(P_f, P^\star)$ and $1/n \ll h^2(P_f, P^\star)$, whereas $\rho(\mathcal{D}_n; f, g) < 0$ with high probability when $h^2(P_f, P^\star) \ll h^2(P_g, P^\star)$ and $1/n \ll h^2(P_g, P^\star)$. These comparison properties motivate an iterative search for a ρ-estimator guided by $\rho(\mathcal{D}_n; f, g)$. Moreover, since

$$\Upsilon(\mathcal{D}_n; f) = \sup_{g\in\mathcal{F}} \rho(\mathcal{D}_n; f, g) \ge \rho(\mathcal{D}_n; f, f) = 0, \qquad f \in \mathcal{F},$$

it suffices to find a candidate $\widehat{f}_\rho \in \mathcal{F}$ with $\Upsilon(\mathcal{D}_n; \widehat{f}_\rho) \le 1$, because then

$$\Upsilon(\mathcal{D}_n; \widehat{f}_\rho) \le 1 \le \inf_{f\in\mathcal{F}} \Upsilon(\mathcal{D}_n; f) + 1. \tag{5}$$

Thus, $\widehat{f}_\rho$ satisfies (1) and is a ρ-estimator. We therefore conduct this search by repeatedly solving a challenger problem using J initializations, as described in Algorithm 1. The operator $\mathrm{LocalMax}(g_\ell, \Theta, \boldsymbol{s}_{\ell,j}; M)$ performs at most M local ascent steps from that point with $\boldsymbol{\theta}$ held fixed and returns the best point found. After updating the current candidate to the best returned challenger, the algorithm stops when the largest attained contrast falls below 1 or the outer iteration limit is reached. Related training strategies based on best responses have been explored for GANs, including Fictitious GAN (Ge et al., 2018) and DO-GAN (Aung et al., 2022).

5.2 Synthetic data

To assess the theoretical predictions, we evaluate the proposed procedure in the well-specified setting and under varying proportions of high-leverage contamination. Since the true conditional distribution is known for the synthetic data, we assess both distribution estimation and classification performance. We present five-class results here and three-class results in the appendix.

In the experimental setting, each input contains five scalar tokens combining a shared mixture-distributed level, a positional trend, and token-specific Gaussian noise. A two-head attention model defines the conditional class probabilities. Its scores depend on token content and position. The logits use the mean and contrast of the two attention summaries. The five classes represent low/high levels,

Algorithm 1 Searching for a ρ-estimator with J initializations

Require: Training data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$; initial parameter $\boldsymbol{\theta}_0$; outer iteration limit L; local optimizer iteration limit M; number of initializers $J \geq 1$; perturbation scales $\sigma_1, \ldots, \sigma_{J-1}$.

Ensure: $\widehat{f}_\rho$

$\boldsymbol{\theta} \leftarrow \Pi_\Theta(\boldsymbol{\theta}_0)$
for $\ell = 1, \ldots, L$ **do**
  With $\boldsymbol{\theta}$ fixed, define $g_\ell(\boldsymbol{\eta}) = \rho(\mathcal{D}_n; f_{\boldsymbol{\theta}}, f_{\boldsymbol{\eta}})$.
  $\boldsymbol{s}_{\ell,1} \leftarrow \boldsymbol{\theta}$
  for $j = 2, \ldots, J$ **do**
    Draw independently $\boldsymbol{\xi}_{\ell,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_\theta})$.
    $\boldsymbol{s}_{\ell,j} \leftarrow \Pi_\Theta(\boldsymbol{\theta} + \sigma_{j-1}\boldsymbol{\xi}_{\ell,j})$
  end for
  for $j = 1, \ldots, J$ **do**
    $\widetilde{\boldsymbol{\theta}}_{\ell,j} \leftarrow \mathrm{LocalMax}(g_\ell, \Theta, \boldsymbol{s}_{\ell,j}; M)$
  end for
  $j_\ell \in \arg\max_{j \in \{1,\ldots,J\}} g_\ell(\widetilde{\boldsymbol{\theta}}_{\ell,j})$
  $V_\ell \leftarrow g_\ell(\widetilde{\boldsymbol{\theta}}_{\ell,j_\ell})$
  if $V_\ell \leq 1$ **then**
    break
  end if
  $\boldsymbol{\theta} \leftarrow \widetilde{\boldsymbol{\theta}}_{\ell,j_\ell}$
end for
$\widehat{f}_\rho \leftarrow f_{\boldsymbol{\theta}}$
return $\widehat{f}_\rho$

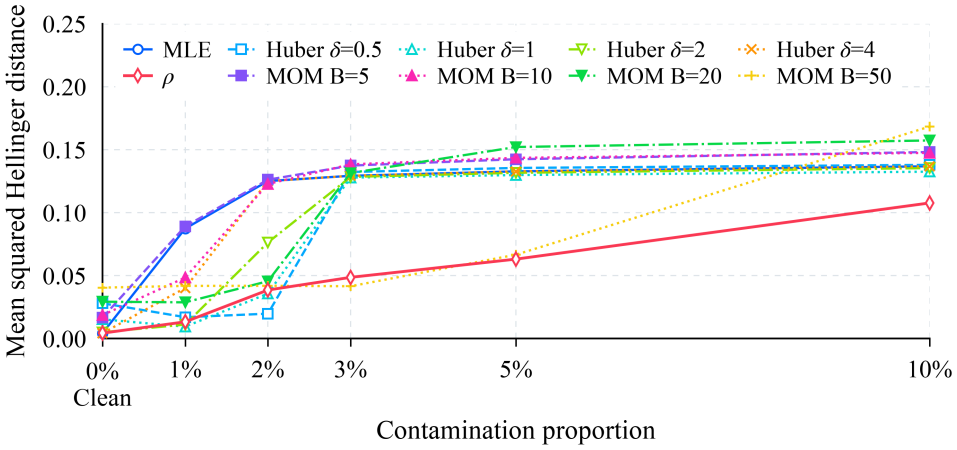

Figure 1: Squared Hellinger distance on clean test data under varying proportions of high-leverage contamination. Lines connect means over 100 replications.

small/large contrasts, and an intermediate state. All estimators fit the same correctly specified family, jointly learning four attention parameters and twelve classification coefficients.

Each replication uses $n = 500$ training observations and $N = 10{,}000$ independent clean test observations. For $\varepsilon \in \{0, 0.01, 0.02, 0.03, 0.05, 0.10\}$, we replace exactly $n\varepsilon$ training observations with extreme positive token sequences. We compare the ρ-estimator with maximum likelihood (MLE) and two robust alternatives: median-of-means (MOM) objectives with $B \in \{5, 10, 20, 50\}$ blocks and Huber-based objectives with $\delta \in \{0.5, 1, 2, 4\}$. Each tuning setting is reported separately. Results are presented in Figure 1 and Figure 2. The appendix details the experimental setup, data-generating distribution, replacement rule, objectives, random seeds, and competitors.

In the experiment, the ρ-estimator matches MLE's mean prediction error on clean data and achieves nearly identical squared Hellinger distance. Its advantage becomes clear as contamination increases: it attains the lowest mean prediction error among all methods at 3%, 5%, and 10% contamination. At 5%, its prediction error is 36.39%, compared with 48.34% for MLE, 46.40% for the best Huber setting, and 39.17% for the best MOM setting. It also achieves the lowest mean squared Hellinger distance at 5% and 10% contamination. Although some Huber settings slightly perform better under lighter contamination, these results highlight the protection offered by ρ-estimation as contamination increases, while preserving clean-data accuracy. This is consistent with the theoretical guarantees of accuracy under correct specification and robustness to contamination. Moreover, it is worth emphasizing that, unlike Huber and MOM, ρ-estimation requires no additional robustness tuning parameter, such as the Huber threshold δ or the MOM block count B.

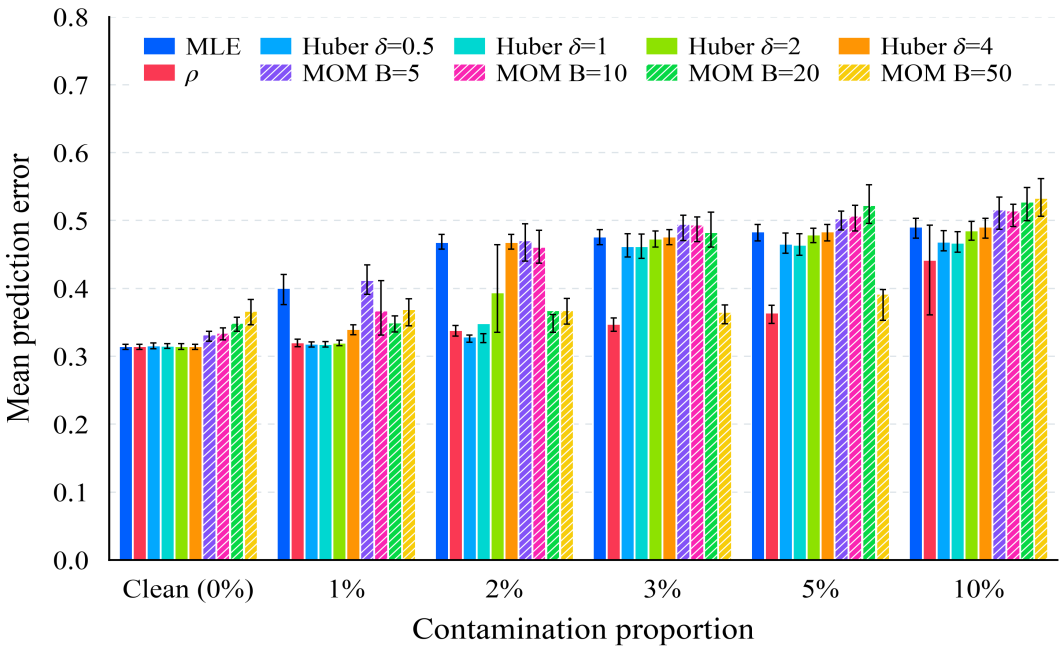

Figure 2: Prediction error on clean test data under varying proportions of high-leverage contamination. Bars show means over 100 replications, with error bars spanning the 25th–75th percentiles.

5.3 Real-world datasets

We compare ρ-estimation with MLE, Huber ($c = 1$), and MOM (10 blocks) on Vowel (11 classes, 10 features) and Letter (26 classes, 16 features), using two-layer Transformers. We contaminate 0%, 5%, 10%, 20%, or 30% of training observations by replacing their features with random ± 50 values and assigning uniformly sampled incorrect labels, while keeping test data clean. The results shown in Figure 3 are averaged over 100 runs, each with a different random seed. The proposed method remains competitive on clean data and achieves the lowest mean test misclassification error at every positive contamination level on both datasets, albeit with greater computational cost than MLE. On real datasets, "clean" data do not guarantee correct model specification. The slight advantage of ρ over MLE on Letter may therefore reflect mild model misspecification. Further details of the experimental settings, numerical results, and running times are provided in the appendix.

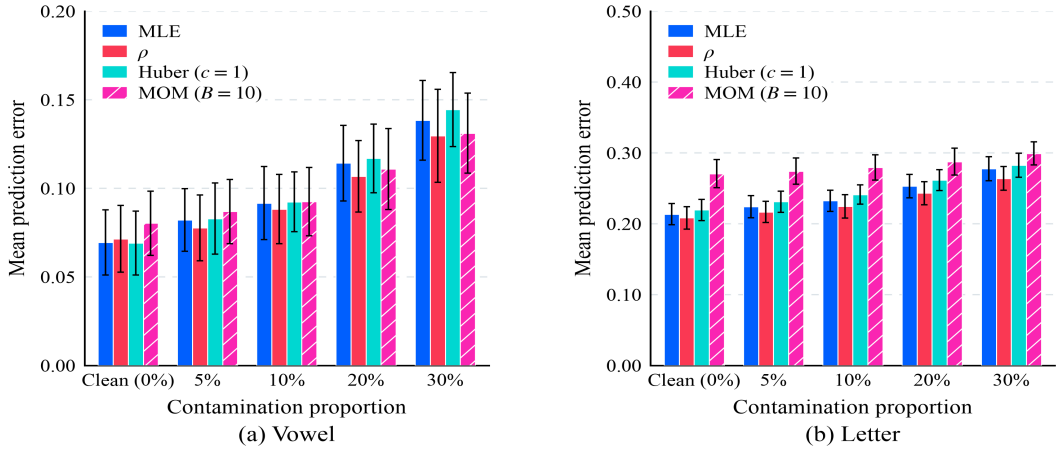

Figure 3: Mean test misclassification error on Vowel and Letter. Error bars indicate standard deviations over 100 runs.

AI USE STATEMENT

In this work, we used generative AI tools for brainstorming, sourcing and searching for information, and editing the manuscript to improve readability. We did not use generative AI tools for any tasks requiring mandatory disclosure. We have reviewed all AI-assisted work to ensure its accuracy and appropriateness and to verify that language revisions preserve the intended scientific meaning. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or any practices that raise concerns regarding privacy, security, or legal compliance. The authors declare no conflicts of interest and no sponsorship that could influence the results. This research fully complies with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide complete theoretical proofs and derivations in the appendix, along with detailed experimental settings, including hyperparameters, model architectures, training procedures, and evaluation protocols. The source code for all experiments is publicly available at https://anonymous.4open.science/status/DHBAJHOIDAJS27F2. The two datasets, Vowel and Letter, used in our experiments are fully public and are available via the links https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#vowel and https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#letter, respectively. These materials provide the information and resources needed to reproduce our results.

REFERENCES

Aye Phyu Phyu Aung, Xinrun Wang, Runsheng Yu, Bo An, Senthilnath Jayavelu, and Xiaoli Li. DO-GAN: A double oracle framework for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11265–11274, 2022.

Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

Yannick Baraud. Tests and estimation strategies associated to some loss functions. *Probability Theory and Related Fields*, 180:799–846, 2021.

Yannick Baraud and Lucien Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018.

Yannick Baraud and Juntong Chen. Robust estimation of a regression function in exponential families. *Journal of Statistical Planning and Inference*, 233:106167, 2024.

Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones Mathematicae*, 207(2):425–517, 2017.

Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 42(3):273–325, 2006.

Juntong Chen. Robust classification with convolutional neural networks. *Communications in Information and Systems*, 25(4):787–812, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

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. In *International Conference on Learning Representations*, 2021.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

Hao Ge, Yin Xia, Xu Chen, Randall Berry, and Ying Wu. Fictitious GAN: Training GANs with historical models. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pp. 122–137. Springer, 2018.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1919–1925, 2017.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

Alex Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. In *Advances in Neural Information Processing Systems*, pp. 42162–42210, 2024.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

Yuling Jiao, Yanming Lai, Defeng Sun, Yang Wang, and Bokai Yan. Approximation bounds for transformer networks with application to regression. *arXiv preprint arXiv:2504.12175*, 2025.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Yanming Lai and Defeng Sun. Standard transformers achieve the minimax rate in nonparametric regression with $C^{s,\lambda}$ targets. *arXiv preprint arXiv:2602.20555 preprint arXiv:2602.20555*, 2026.

Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via MOM minimization. *Machine Learning*, 109:1635–1665, 2020.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

Po-Ling Loh. Scale calibration for high-dimensional robust regression. *Electronic Journal of Statistics*, 15(2):5933–5994, 2021.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

Edward J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

Curtis Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Zhaiming Shen, Alex Havrilla, Rongjie Lai, Alexander Cloninger, and Wenjing Liao. Transformers for learning on noisy and task-level manifolds: Approximation and generalization insights. *arXiv preprint arXiv:2505.03205*, 2025.

Ji Shi and Demetrio Labate. Dimension-independent approximations on low-dimensional manifolds using transformers. *Mathematics*, 14(9):1559, 2026.

Zhongjie Shi and Wenjing Liao. Learning theory of transformers: Local-to-global approximation via softmax partition of unity. *arXiv preprint arXiv:2605.08811*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

A Appendix

We provide detailed proofs of the results established in the main context, experimental settings and additional numerical experiments in this section.

A.1 Proofs of the statistical results

We begin with a dimension-free inequality that transfers uniform logit bounds to Hellinger bounds for the induced conditional distributions.

Lemma 1. *For all $z, z' \in \mathbb{R}^K$,*

$$h^2(\mathrm{softmax}(z), \mathrm{softmax}(z')) \le \frac{1}{8}\|z - z'\|_\infty^2.$$

Consequently, for any two logit functions $f, g : \mathcal{X} \to \mathbb{R}^K$ and any $x \in \mathcal{X}$,

$$h^2\big(Q_f(\cdot \mid x), Q_g(\cdot \mid x)\big) \le \frac{1}{8}\|f(x) - g(x)\|_\infty^2.$$

Proof. Define $\Phi = (\Phi_1, \dots, \Phi_K)^\top : \mathbb{R}^K \to \mathbb{R}^K$ by $\Phi(z) = \sqrt{\mathrm{softmax}(z)}$, where the square root is taken coordinatewise, and let $p = \mathrm{softmax}(z)$. Since

$$\frac{\partial p_k}{\partial z_\ell} = p_k\big(\mathbf{1}_{\{k=\ell\}} - p_\ell\big),$$

the Jacobian $D\Phi(z)$ has entries

$$\frac{\partial \Phi_k(z)}{\partial z_\ell} = \frac{1}{2}\sqrt{p_k}\big(\mathbf{1}_{\{k=\ell\}} - p_\ell\big).$$

A direct calculation gives

$$D\Phi(z)^\top D\Phi(z) = \frac{1}{4}\left(\mathrm{diag}(p) - pp^\top\right).$$

Hence, for every $v \in \mathbb{R}^K$,

$$\|D\Phi(z)v\|_2^2 = \frac{1}{4}\,\mathrm{Var}_{J\sim p}(v_J).$$

Since the variance of a random variable supported on an interval of length r is at most $r^2/4$,

$$\mathrm{Var}_{J\sim p}(v_J) \le \frac{1}{4}\left(\max_{k\in[K]} v_k - \min_{k\in[K]} v_k\right)^2 \le \|v\|_\infty^2.$$

It follows that

$$\|D\Phi(z)v\|_2 \le \frac{1}{2}\|v\|_\infty.$$

Applying the fundamental theorem of calculus along the line segment joining z' and z yields

$$\|\Phi(z) - \Phi(z')\|_2 \le \frac{1}{2}\|z - z'\|_\infty.$$

Writing $p' = \mathrm{softmax}(z')$ and using

$$h^2(p, p') = \frac{1}{2}\|\sqrt{p} - \sqrt{p'}\|_2^2,$$

we conclude that

$$h^2(\mathrm{softmax}(z), \mathrm{softmax}(z')) \le \frac{1}{8}\|z - z'\|_\infty^2.$$

Taking $z = f(x)$ and $z' = g(x)$ proves the second assertion. □

Lemma 2. *Let $\mathcal{F}$ be a class of K-class logit functions and let $p : \mathcal{W} \to (0,\infty)$ be measurable with $p(x,y) \ge c_p > 0$ on $\mathcal{W}$. Define*

$$\mathcal{G}(\mathcal{F}, p) = \left\{(x,y) \mapsto \psi\left(\sqrt{\frac{Q_f(y \mid x)}{p(x,y)}}\right) : f \in \mathcal{F}\right\}.$$

Then, for every $\delta > 0$,

$$\log \mathcal{N}\big(\delta, \mathcal{G}(\mathcal{F}, p), \|\cdot\|_{L^\infty(\mathcal{X}\times[K])}\big) \le \log \mathcal{N}\big(\delta\sqrt{c_p}, \mathcal{F}, d_\infty\big).$$

Proof. For any $a, b \ge 0$,

$$|\psi(a) - \psi(b)| = \frac{2|a-b|}{(1+a)(1+b)} \le 2|a-b|.$$

Thus, ψ is 2-Lipschitz on $[0,\infty)$.

Fix $f, g \in \mathcal{F}$. For every $(x,y) \in \mathcal{X}\times[K]$, the Lipschitz continuity of ψ and the lower bound $p(x,y) \ge c_p$ give

$$\begin{aligned}\left|\psi\left(\sqrt{\frac{Q_f(y \mid x)}{p(x,y)}}\right) - \psi\left(\sqrt{\frac{Q_g(y \mid x)}{p(x,y)}}\right)\right| &\le \frac{2}{\sqrt{p(x,y)}}\left|\sqrt{Q_f(y \mid x)} - \sqrt{Q_g(y \mid x)}\right| \\ &\le \frac{2}{\sqrt{c_p}}\left\|\sqrt{Q_f(\cdot \mid x)} - \sqrt{Q_g(\cdot \mid x)}\right\|_2,\end{aligned}$$

where the square root of a probability vector is taken coordinatewise.

By Lemma 1,

$$\begin{aligned}\left\|\sqrt{Q_f(\cdot \mid x)} - \sqrt{Q_g(\cdot \mid x)}\right\|_2^2 &= 2h^2(Q_f(\cdot \mid x), Q_g(\cdot \mid x)) \\ &\le \frac{1}{4}\|f(x) - g(x)\|_\infty^2.\end{aligned}$$

Consequently,

$$\left\|\sqrt{Q_f(\cdot \mid x)} - \sqrt{Q_g(\cdot \mid x)}\right\|_2 \le \frac{1}{2}\|f(x) - g(x)\|_\infty.$$

Combining the preceding inequalities yields

$$\left|\psi\left(\sqrt{\frac{Q_f(y \mid x)}{p(x,y)}}\right) - \psi\left(\sqrt{\frac{Q_g(y \mid x)}{p(x,y)}}\right)\right| \le \frac{1}{\sqrt{c_p}}\|f(x) - g(x)\|_\infty.$$

Taking the supremum over $(x, y) \in \mathcal{X} \times [K]$, we obtain

$$\left\| \psi\left(\sqrt{\frac{Q_f}{p}}\right) - \psi\left(\sqrt{\frac{Q_g}{p}}\right) \right\|_{L^\infty(\mathcal{X}\times[K])} \leq \frac{1}{\sqrt{c_p}} \|f - g\|_{L^\infty(\mathcal{X};\ell_\infty^K)}.$$

Therefore, the image of every $\delta\sqrt{c_p}$-cover of $\mathcal{F}$ under the map

$$f \longmapsto \left[(x, y) \mapsto \psi\left(\sqrt{\frac{Q_f(y \mid x)}{p(x, y)}}\right)\right]$$

is a δ-cover of $\mathcal{G}(\mathcal{F}, p)$. Hence,

$$\mathcal{N}\big(\delta, \mathcal{G}(\mathcal{F}, p), \|\cdot\|_{L^\infty(\mathcal{X}\times[K])}\big) \leq \mathcal{N}\big(\delta\sqrt{c_p}, \mathcal{F}, \|\cdot\|_{L^\infty(\mathcal{X};\ell_\infty^K)}\big).$$

Taking logarithms completes the proof. □

We next bound the ρ-dimension of Baraud & Birgé (2018). The argument uses the following empirical-process inequality, stated as Theorem A.1 in Chen (2025).

Lemma 3. *Let $W_1, \ldots, W_n$ be independent random variables taking values in a measurable space $(\mathcal{W}, \mathscr{W})$, and let $\mathcal{F}$ be a class of functions from $\mathcal{W}$ to $[-1, 1]$. Suppose that there exist constants $M_1, M_2 \geq 1$ such that, for every $0 < \delta \leq 1$,*

$$\log \mathcal{N}\left(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\mathcal{W})}\right) \leq M_1 \log\left(\frac{M_2}{\delta}\right).$$

Assume further that, for some $\sigma \in (0, 1]$,

$$\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[f^2(W_i)\right] \leq \sigma^2 \leq 1.$$

Then

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (f(W_i) - \mathbb{E}[f(W_i)])\right|\right] \leq 2304\, \mathscr{H}(M_1, M_2, \sigma) + 24\sqrt{n\sigma^2 \mathscr{H}(M_1, M_2, \sigma)},$$

where

$$\mathscr{H}(M_1, M_2, \sigma) = M_1 \log\left(\frac{2M_2}{\sigma}\right).$$

Proposition 1. *Let $\mathcal{F}$ be a finite or countable class of K-class logit functions, and let $\mathscr{P}_\mathcal{F} = \{\mathbf{P}_f : f \in \mathcal{F}\}$ be the corresponding product model. Suppose that there exist $V, A \geq 1$ such that, for every $0 < \eta \leq 1$,*

$$\log \mathcal{N}\big(\eta, \mathcal{F}, \|\cdot\|_{L^\infty(\mathcal{X};\ell_\infty^K)}\big) \leq V \log\left(\frac{A}{\eta}\right).$$

Fix $\bar{f} \in \mathcal{F}$ and suppose that

$$c_{\bar{f}} = \inf_{(x,y)\in\mathcal{X}\times[K]} Q_{\bar{f}}(y \mid x) > 0.$$

Then, for every product probability distribution $\mathbf{P}^\star$ on $\mathcal{W}^n$,

$$D^{\mathscr{P}_\mathcal{F}}\left(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}\right) \leq 2800\, V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_{\bar{f}}}}\right).$$

Proof. For product probability distributions $\mathbf{R} = \bigotimes_{i=1}^n R_i$ and $\mathbf{R}' = \bigotimes_{i=1}^n R_i'$ on $\mathcal{W}^n$, recall that

$$\mathbf{h}^2(\mathbf{R}, \mathbf{R}') = \sum_{i=1}^n h^2(R_i, R_i').$$

Fix $\bar{f} \in \mathcal{F}$. For every $r > 0$, consider the localized class

$$\mathcal{H}_r = \left\{\varphi_f : f \in \mathcal{F},\ \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_f) + \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) < r^2\right\},$$

where

$$\varphi_f(x, y) = \psi\left(\sqrt{\frac{Q_f(y \mid x)}{Q_{\bar{f}}(y \mid x)}}\right).$$

Write $W_i = (X_i, Y_i)$, so that W_i has distribution $P_i^\star$ under

$$\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star.$$

By Proposition 3 of Baraud & Birgé (2018), the function ψ satisfies the required variance inequality, and hence

$$\sup_{\varphi \in \mathcal{H}_r} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_i^\star}\left[\varphi^2(W_i)\right] \le \sigma^2(r) = \left(\frac{a_2^2 r^2}{n}\right) \wedge 1, \qquad a_2^2 = 3\sqrt{2}.$$

Consider also the nonlocalized class

$$\mathcal{H} = \{\varphi_f : f \in \mathcal{F}\}.$$

Clearly, $\mathcal{H}_r \subseteq \mathcal{H}$. Since $Q_{\bar{f}}(y \mid x) \ge c_{\bar{f}}$ for every $(x, y) \in \mathcal{X} \times [K]$, Lemma 2, applied with $p(x, y) = Q_{\bar{f}}(y \mid x)$, implies that, for every relevant $0 < \delta \le 1$,

$$\begin{aligned}
\log \mathcal{N}\big(\delta, \mathcal{H}_r, \|\cdot\|_{L^\infty(\mathcal{W})}\big) &\le \log \mathcal{N}\big(\delta, \mathcal{H}, \|\cdot\|_{L^\infty(\mathcal{W})}\big) \\
&\le \log \mathcal{N}\big(\delta\sqrt{c_{\bar{f}}}, \mathcal{F}, \|\cdot\|_{L^\infty(\mathcal{X}; \ell_\infty^K)}\big) \\
&\le V \log\left(\frac{A}{\delta\sqrt{c_{\bar{f}}}}\right).
\end{aligned}$$

Thus, the entropy condition of Lemma 3 holds with

$$M_1 = V \qquad \text{and} \qquad M_2 = \frac{A}{\sqrt{c_{\bar{f}}}}.$$

Define

$$w^{\mathscr{P}_\mathcal{F}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}, r) = \mathbb{E}_{\mathbf{P}^\star}\left[\sup_{\varphi \in \mathcal{H}_r} \left|\sum_{i=1}^n \left(\varphi(W_i) - \mathbb{E}_{P_i^\star}[\varphi(W_i)]\right)\right|\right].$$

Applying Lemma 3 gives

$$w^{\mathscr{P}_\mathcal{F}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}, r) \le 24\sqrt{n\sigma^2(r) V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right)} + 2304 V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right).$$

Since $n\sigma^2(r) \le a_2^2 r^2$, we obtain

$$w^{\mathscr{P}_\mathcal{F}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}, r) \le 24 a_2 r\sqrt{V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right)} + 2304 V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right).$$

Let $a_1 = 3/8$ and $b_\rho = a_1/(4a_2)$, and set

$$\overline{H} = V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_{\bar{f}}}}\right) \qquad \text{and} \qquad D_0 = 2800\,\overline{H}.$$

We next show that

$$V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right) \le \overline{H} \qquad \text{whenever} \qquad r \ge b_\rho^{-1}\sqrt{D_0}.$$

If $a_2^2 r^2/n \le 1$, then

$$\sigma(r) = \frac{a_2 r}{\sqrt{n}},$$

and therefore

$$\begin{aligned} V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right) &= V \log\left(\frac{2A\sqrt{n}}{a_2 r\sqrt{c_{\bar{f}}}}\right) \\ &\le V \log\left(\frac{2b_\rho A\sqrt{n}}{a_2\sqrt{D_0}\sqrt{c_{\bar{f}}}}\right) \\ &\le V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_{\bar{f}}}}\right) = \overline{H}. \end{aligned}$$

The last inequality follows from $D_0 \ge a_1^2/(16a_2^4)$. This condition holds because $V \ge 1$, $A \ge 1$, $c_{\bar{f}} \le 1$, and $n \ge 1$.

If $a_2^2 r^2/n > 1$, then $\sigma(r) = 1$, and hence

$$V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}\,\sigma(r)}\right) = V \log\left(\frac{2A}{\sqrt{c_{\bar{f}}}}\right) \le V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_{\bar{f}}}}\right) = \overline{H}.$$

Thus, in both cases,

$$w^{\mathscr{P}_{\mathcal{F}}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}, r) \le 24 a_2 r\sqrt{\overline{H}} + 2304\overline{H}.$$

For $r \ge b_\rho^{-1}\sqrt{D_0}$, we have

$$\frac{1}{r} \le \frac{b_\rho}{\sqrt{D_0}} \qquad \text{and} \qquad \frac{1}{r^2} \le \frac{b_\rho^2}{D_0}.$$

Consequently,

$$\begin{aligned} w^{\mathscr{P}_{\mathcal{F}}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}, r) &\le \frac{a_1 r^2}{8}\left[\frac{8 \times 24 a_2\sqrt{\overline{H}}}{a_1 r} + \frac{8 \times 2304\overline{H}}{a_1 r^2}\right] \\ &\le \frac{a_1 r^2}{8}\left[\frac{48\sqrt{\overline{H}}}{\sqrt{D_0}} + \frac{1152 a_1 \overline{H}}{a_2^2 D_0}\right] \\ &\le \frac{a_1 r^2}{8}\left[\frac{48\sqrt{\overline{H}}}{\sqrt{D_0}} + \frac{102\overline{H}}{D_0}\right], \end{aligned}$$

where we used

$$\frac{1152 a_1}{a_2^2} = 72\sqrt{2} < 102.$$

Since $D_0 = 2800\,\overline{H}$,

$$\frac{48\sqrt{\overline{H}}}{\sqrt{D_0}} + \frac{102\overline{H}}{D_0} = \frac{48}{\sqrt{2800}} + \frac{102}{2800} < 1.$$

It follows that

$$w^{\mathscr{P}_{\mathcal{F}}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}, r) \le \frac{a_1 r^2}{8}$$

for every $r \ge b_\rho^{-1}\sqrt{D_0}$. By the definition of the ρ-dimension,

$$D^{\mathscr{P}_{\mathcal{F}}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) \le D_0.$$

Finally,

$$D^{\mathscr{P}_{\mathcal{F}}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) \le 2800\, V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_{\bar{f}}}}\right),$$

which completes the proof. □

Proof of Theorem 1. For each i, let $P_i^\star$ be the joint distribution determined by $Q_i^\star$ and P_{X_i}, and, for each $f \in \mathcal{F}$, let $P_{i,f}$ be the joint distribution determined by Q_f and P_{X_i}. More precisely,

$$P_i^\star(dx, \{y\}) = Q_i^\star(y \mid x) P_{X_i}(dx), \qquad P_{i,f}(dx, \{y\}) = Q_f(y \mid x) P_{X_i}(dx).$$

Set $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$ and $\mathbf{P}_f = \bigotimes_{i=1}^n P_{i,f}$. By construction,

$$\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_f) = \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f).$$

Let $\widehat{\mathbf{P}} = \mathbf{P}_{\widehat{f}_\rho} = \bigotimes_{i=1}^n P_{i,\widehat{f}_\rho}$ and $\widehat{\mathbf{Q}} = \mathbf{Q}_{\widehat{f}_\rho}$. The general oracle inequality for ρ-estimators (Baraud & Birgé, 2018, Theorem 1) implies that, for every $\bar{f} \in \mathcal{F}$ and every $\xi > 0$, with probability at least $1 - e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}}) \le C_0 \left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) + D^{\mathscr{P}_\mathcal{F}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) + 1 + \xi \right],$$

where $C_0 > 0$ is a universal constant.

For a fixed $\bar{f} \in \mathcal{F}$, write

$$c_{\bar{f}} = \inf_{(x,y) \in \mathcal{X} \times [K]} Q_{\bar{f}}(y \mid x).$$

Since $c_{\bar{f}} \ge c_\mathcal{F}$, Proposition 1 gives

$$D^{\mathscr{P}_\mathcal{F}}(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) \le 2800\, V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_{\bar{f}}}}\right) \le 2800\, V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_\mathcal{F}}}\right).$$

Substituting this bound into the oracle inequality yields, with probability at least $1 - e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}}) \le C_1 \left[\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) + V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_\mathcal{F}}}\right) + \xi \right],$$

where $C_1 > 0$ is universal. Here, the additive constant has been absorbed into the entropy term, since $V, A, n \ge 1$ and $c_\mathcal{F} \le 1$.

Using

$$\mathbf{h}^2(\mathbf{P}^\star, \widehat{\mathbf{P}}) = \mathbf{h}^2(\mathbf{Q}^\star, \widehat{\mathbf{Q}}), \qquad \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\bar{f}}) = \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\bar{f}}),$$

we obtain

$$\mathbf{h}^2(\mathbf{Q}^\star, \widehat{\mathbf{Q}}) \le C_1 \left[\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\bar{f}}) + V \log\left(\frac{2A\sqrt{n}}{\sqrt{c_\mathcal{F}}}\right) + \xi \right].$$

This bound holds for every fixed comparator $\bar{f} \in \mathcal{F}$. For each $\varepsilon > 0$, choose a deterministic $f_\varepsilon \in \mathcal{F}$ such that

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{f_\varepsilon}) \le \inf_{f \in \mathcal{F}} \mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f) + \varepsilon.$$

Applying the preceding inequality with $\bar{f} = f_\varepsilon$ and then letting ε decrease to zero proves the claimed large-deviation bound. □

Proof of Corollary 2. Fix a realization of the training sample for which $\widehat{f}_{\rho,n}$ is defined, and write $\widehat{Q} = Q_{\widehat{f}_{\rho,n}}$. Since K is finite, a measurable Bayes classifier can be chosen as

$$g^\star(x) = \min\left\{ k \in [K] : Q^\star(k \mid x) = \max_{\ell \in [K]} Q^\star(\ell \mid x) \right\}.$$

For P_X-almost every x, the definition of $\widehat{g}_{\rho,n}$ gives

$$\widehat{Q}(\widehat{g}_{\rho,n}(x) \mid x) = \max_{k \in [K]} \widehat{Q}(k \mid x) \ge \widehat{Q}(g^\star(x) \mid x).$$

Consequently,

$$\begin{aligned}
&Q^\star(g^\star(x) \mid x) - Q^\star(\widehat{g}_{\rho,n}(x) \mid x) \\
&\quad = Q^\star(g^\star(x) \mid x) - \widehat{Q}(g^\star(x) \mid x) \\
&\qquad + \widehat{Q}(g^\star(x) \mid x) - \widehat{Q}(\widehat{g}_{\rho,n}(x) \mid x) \\
&\qquad + \widehat{Q}(\widehat{g}_{\rho,n}(x) \mid x) - Q^\star(\widehat{g}_{\rho,n}(x) \mid x) \\
&\quad \le \left| Q^\star(g^\star(x) \mid x) - \widehat{Q}(g^\star(x) \mid x) \right| \\
&\qquad + \left| Q^\star(\widehat{g}_{\rho,n}(x) \mid x) - \widehat{Q}(\widehat{g}_{\rho,n}(x) \mid x) \right| \\
&\quad \le \left\| Q^\star(\cdot \mid x) - \widehat{Q}(\cdot \mid x) \right\|_1 .
\end{aligned}$$

For any probability vectors $p, q \in \Delta_K$, the Cauchy–Schwarz inequality yields

$$\begin{aligned}
\|p - q\|_1 &= \sum_{k=1}^{K} |\sqrt{p_k} - \sqrt{q_k}| \, (\sqrt{p_k} + \sqrt{q_k}) \\
&\le \left\{ \sum_{k=1}^{K} (\sqrt{p_k} - \sqrt{q_k})^2 \right\}^{1/2} \left\{ \sum_{k=1}^{K} (\sqrt{p_k} + \sqrt{q_k})^2 \right\}^{1/2} .
\end{aligned}$$

Moreover,

$$\sum_{k=1}^{K} (\sqrt{p_k} - \sqrt{q_k})^2 = 2h^2(p, q)$$

and

$$\sum_{k=1}^{K} (\sqrt{p_k} + \sqrt{q_k})^2 = 2 + 2\sum_{k=1}^{K} \sqrt{p_k q_k} \le 4.$$

It follows that

$$\|p - q\|_1 \le 2\sqrt{2}\, h(p, q).$$

Therefore,

$$\begin{aligned}
\mathcal{R}(\widehat{g}_{\rho,n}) - \mathcal{R}^\star &= \int_{\mathcal{X}} [Q^\star(g^\star(x) \mid x) - Q^\star(\widehat{g}_{\rho,n}(x) \mid x)] \, dP_X(x) \\
&\le 2\sqrt{2} \int_{\mathcal{X}} h\left(Q^\star(\cdot \mid x), Q_{\widehat{f}_{\rho,n}}(\cdot \mid x) \right) dP_X(x) \\
&\le 2\sqrt{2} \left\{ \int_{\mathcal{X}} h^2\left(Q^\star(\cdot \mid x), Q_{\widehat{f}_{\rho,n}}(\cdot \mid x) \right) dP_X(x) \right\}^{1/2} \\
&= 2\sqrt{2} \left\{ \bar{h}^2\left(Q^\star, Q_{\widehat{f}_{\rho,n}} \right) \right\}^{1/2} .
\end{aligned}$$

The last inequality follows from the Cauchy–Schwarz inequality and $P_X(\mathcal{X}) = 1$. Nonnegativity follows from the optimality of the Bayes classifier. Since the argument applies to every realization for which $\widehat{f}_{\rho,n}$ is defined, the result holds almost surely. □

Proof of Corollary 4. Applying Theorem 3 to each coordinate of f_0, and using the fact that clipping fixes f_0 and does not increase distances, we obtain $f_{\varepsilon_n} \in \mathcal{F}_{\varepsilon_n, \mathcal{B}}$ satisfying

$$\sup_{\boldsymbol{x} \in \mathcal{M}_{\zeta_n}} |f_{\varepsilon_n}(\boldsymbol{x}) - f_0(\boldsymbol{x})|_\infty \le \varepsilon_n + \Gamma_{\mathcal{M}} \zeta_n^\beta .$$

Since $Q^\star = Q_{f_0}$, P_X-almost surely, and $P_X(\mathcal{M}_{\zeta_n}) = 1$, Lemma 1 gives

$$\bar{h}^2(Q^\star, Q_{f_{\varepsilon_n}}) \le \frac{1}{8} \left(\varepsilon_n + \Gamma_{\mathcal{M}} \zeta_n^\beta \right)^2 .$$

The same lemma implies that $f \mapsto \bar{h}^2(Q^\star, Q_f)$ is continuous under d_∞. Therefore, by the density of $\mathcal{F}_n$,

$$\inf_{f\in\mathcal{F}_n} \bar{h}^2(Q^\star, Q_f) = \inf_{f\in\mathcal{F}_{\varepsilon_n,\mathcal{B}}} \bar{h}^2(Q^\star, Q_f) \le \frac{1}{8}\left(\varepsilon_n + \Gamma_{\mathcal{M}}\zeta_n^\beta\right)^2.$$

We next control the entropy of the logit class. For $\varepsilon \in (0,1)$, write $\varepsilon_{q,\delta} = \varepsilon/\mathfrak{B}_{q,\delta}$. Since $R = 2$ and the remaining token, head, and hidden dimensions are fixed, the architectural and parameter bounds in the proof of Theorem 3 give

$$\log \mathfrak{L} \le C\left(1 + \log\frac{1}{\varepsilon_{q,\delta}}\right)^2.$$

Moreover, the total parameter count is bounded by $C\varepsilon_{q,\delta}^{-q/\beta}$, and the logarithm of each parameter magnitude bound is at most $C\log(1/\varepsilon_{q,\delta})$. Applying Proposition 2 with $\mathcal{T} = \mathcal{T}_\varepsilon$ therefore yields

$$\begin{aligned}\log \mathcal{N}\left(\eta, \mathcal{F}_{\varepsilon,\mathcal{B}}, d_\infty\right) &\le K \sum_{a\in\{\mathrm{EB,FF,SA}\}} M_a \log\left(\frac{2B_a\mathfrak{L}}{\eta}\right) \\ &\le C\varepsilon^{-q/\beta}\left[\left(1+\log\frac{1}{\varepsilon}\right)^2 + \log\frac{1}{\eta}\right], \qquad 0 < \eta \le 1,\end{aligned}$$

where the fixed quantities K and $\mathfrak{B}_{q,\delta}$ have been absorbed into C. In particular, $\mathcal{F}_{\varepsilon,\mathcal{B}}$ is totally bounded and admits a countable d_∞-dense subclass.

By the density of $\mathcal{F}_n$, each center of an $\eta/2$-cover of $\mathcal{F}_{\varepsilon_n,\mathcal{B}}$ can be approximated within $\eta/2$ by an element of $\mathcal{F}_n$. Hence,

$$\mathcal{N}(\eta, \mathcal{F}_n, d_\infty) \le \mathcal{N}\left(\eta/2, \mathcal{F}_{\varepsilon_n,\mathcal{B}}, d_\infty\right).$$

Consequently, for a sufficiently large constant $C_1 \ge 1$, the entropy condition in Theorem 1 holds with

$$V_n = C_1\varepsilon_n^{-q/\beta}, \qquad A_n = \exp\left[\left(1+\log\frac{1}{\varepsilon_n}\right)^2\right].$$

Since every $f \in \mathcal{F}_n$ takes values in $[-\mathcal{B}, \mathcal{B}]^K$,

$$c_{\mathcal{F}_n} \ge \frac{1}{1+(K-1)e^{2\mathcal{B}}} > 0.$$

Using $\log(1/\varepsilon_n) \le C\log n$ for $n \ge 3$, we obtain

$$\begin{aligned}\frac{V_n}{n}\log\left(\frac{2A_n\sqrt{n}}{\sqrt{c_{\mathcal{F}_n}}}\right) &\le C\frac{\varepsilon_n^{-q/\beta}}{n}\left[\left(1+\log\frac{1}{\varepsilon_n}\right)^2 + \log n + 1\right] \\ &\le C\frac{\varepsilon_n^{-q/\beta}(\log n)^2}{n}.\end{aligned}$$

Combining the preceding approximation and complexity bounds with the expectation bound in Theorem 1, and using $\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_f) = n\bar{h}^2(Q^\star, Q_f)$, gives

$$\mathbb{E}\left[\bar{h}^2\left(Q^\star, Q_{\widehat{f}_{\rho,n}}\right)\right] \le C\left[\left(\varepsilon_n + \Gamma_{\mathcal{M}}\zeta_n^\beta\right)^2 + \frac{\varepsilon_n^{-q/\beta}(\log n)^2}{n}\right].$$

By the choice of ε_n,

$$\frac{\varepsilon_n^{-q/\beta}(\log n)^2}{n} = \varepsilon_n^2.$$

Together with $(a+b)^2 \le 2a^2 + 2b^2$, this proves the expectation bound. The final assertion follows from $\Gamma_{\mathcal{M}}^2\zeta_n^{2\beta} \le \varepsilon_n^2$. □

A.2 PROOFS OF THE APPROXIMATION RESULTS

In this section, we prove Theorem 3. We first establish the stable dimensionality-reduction result required for the proof.

Lemma 4. *Suppose that Assumption 1 holds. Fix $\delta \in (0,1)$, and define*

$$\overline{q}_\delta = \max\left\{ d_{\mathcal{M}}, \left\lceil C_{\mathrm{BW}} \frac{d_{\mathcal{M}}}{\delta^2} \log_+\left(\frac{t V_{\mathcal{M}} R_{\mathcal{M}}}{\tau \delta} \right) \right\rceil \right\},$$

where $C_{\mathrm{BW}} > 0$ is a sufficiently large numerical constant. Set $q = \min\{t, \overline{q}_\delta\}$. Then there exists a linear map $\Psi : \mathbb{R}^t \to \mathbb{R}^q$ such that

$$\Psi\Psi^\top = \frac{t}{q}\mathbf{I}_q$$

and for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathfrak{M}$,

$$(1-\delta)\|\boldsymbol{x}-\boldsymbol{x}'\|_2 \le \|\Psi\boldsymbol{x}-\Psi\boldsymbol{x}'\|_2 \le (1+\delta)\|\boldsymbol{x}-\boldsymbol{x}'\|_2.$$

Proof. Suppose first that $\overline{q}_\delta < t$. By Theorem 3.1 of Baraniuk & Wakin (2009), with C_{BW} sufficiently large, there exists a q-dimensional orthoprojector

$$\Phi : \mathbb{R}^t \to \mathbb{R}^q, \qquad \Phi\Phi^\top = \mathbf{I}_q,$$

satisfying

$$(1-\delta)\sqrt{\frac{q}{t}}\,\|\boldsymbol{x}-\boldsymbol{x}'\|_2 \le \|\Phi\boldsymbol{x}-\Phi\boldsymbol{x}'\|_2 \le (1+\delta)\sqrt{\frac{q}{t}}\,\|\boldsymbol{x}-\boldsymbol{x}'\|_2 \tag{6}$$

simultaneously for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathfrak{M}$. Define $\Psi = \sqrt{t/q}\,\Phi$ hence $\Psi\Psi^\top = t/q\mathbf{I}_q$. Multiplying (6) by $\sqrt{t/q}$ yields the desired conclusion.

If $\overline{q}_\delta \ge t$, set $q = t$ and $\Psi = \mathbf{I}_t$. Thus, $\Psi\Psi^\top = \mathbf{I}_t = t/q\mathbf{I}_q$, and the conclusion holds trivially. □

The following result constructs a low-dimensional Hölder representation of the target function together with its extension to the projected space.

Lemma 5. *Suppose that Assumption 1 holds, and let Ψ be the linear map provided by Lemma 4. If $u_0 \in \mathcal{H}^\beta(\mathcal{X}, B_0)$ for some $0 < \beta \le 1$ and $B_0 > 0$, then there exists a function $G_\Psi : \mathbb{R}^q \to \mathbb{R}$ such that*

$$G_\Psi\big(\Psi\,\mathrm{vec}(\boldsymbol{x})\big) = u_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{M}, \tag{7}$$

and

$$\|G_\Psi\|_{L^\infty(\mathbb{R}^q)} \le B_0.$$

Moreover,

$$|G_\Psi(\boldsymbol{z}) - G_\Psi(\boldsymbol{z}')| \le \frac{B_0}{(1-\delta)^\beta}\|\boldsymbol{z}-\boldsymbol{z}'\|_2^\beta, \qquad \boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^q. \tag{8}$$

Consequently,

$$\|G_\Psi\|_{\mathcal{H}^\beta(\mathbb{R}^q)} \le \frac{B_0}{(1-\delta)^\beta}.$$

Proof. Define the projected manifold $\mathcal{M}_\Psi = \Psi(\mathfrak{M}) \subset \mathbb{R}^q$. The lower bound in Lemma 4 implies that the restriction $\Psi|_{\mathfrak{M}} : \mathfrak{M} \to \mathcal{M}_\Psi$ is injective. Indeed, if $\Psi\boldsymbol{x} = \Psi\boldsymbol{x}'$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathfrak{M}$, then

$$(1-\delta)\|\boldsymbol{x}-\boldsymbol{x}'\|_2 \le \|\Psi\boldsymbol{x}-\Psi\boldsymbol{x}'\|_2 = 0,$$

and therefore $\boldsymbol{x} = \boldsymbol{x}'$. Hence, for every $\boldsymbol{z} \in \mathcal{M}_\Psi$, there exists a unique $\boldsymbol{x}_{\boldsymbol{z}} \in \mathfrak{M}$ satisfying $\Psi\boldsymbol{x}_{\boldsymbol{z}} = \boldsymbol{z}$. Define

$$g_\Psi(\boldsymbol{z}) = u_0\big(\mathrm{vec}^{-1}(\boldsymbol{x}_{\boldsymbol{z}})\big), \qquad \boldsymbol{z} \in \mathcal{M}_\Psi.$$

Let $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}_\Psi$. By the Hölder continuity of u_0, (4), and the lower bound in Lemma 4,

$$\begin{aligned} |g_\Psi(\boldsymbol{z}_1) - g_\Psi(\boldsymbol{z}_2)| &= \big|u_0\big(\mathrm{vec}^{-1}(\boldsymbol{x}_{\boldsymbol{z}_1})\big) - u_0\big(\mathrm{vec}^{-1}(\boldsymbol{x}_{\boldsymbol{z}_2})\big)\big| \\ &\le B_0\|\boldsymbol{x}_{\boldsymbol{z}_1} - \boldsymbol{x}_{\boldsymbol{z}_2}\|_2^\beta \\ &\le \frac{B_0}{(1-\delta)^\beta}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2^\beta. \end{aligned}$$

Thus, g_Ψ is β-Hölder on $\mathcal{M}_\Psi$, with Hölder constant $L_\delta = B_0/(1-\delta)^\beta$. For $0 < \beta \le 1$, the function

$$d_\beta(\boldsymbol{z}, \boldsymbol{z}') = \|\boldsymbol{z} - \boldsymbol{z}'\|_2^\beta$$

defines a metric on $\mathbb{R}^q$. Indeed, for $a, b \ge 0$, $(a+b)^\beta \le a^\beta + b^\beta$, and hence the usual Euclidean triangle inequality implies

$$d_\beta(\boldsymbol{z}_1, \boldsymbol{z}_3) \le d_\beta(\boldsymbol{z}_1, \boldsymbol{z}_2) + d_\beta(\boldsymbol{z}_2, \boldsymbol{z}_3).$$

Therefore, g_Ψ is an L_δ-Lipschitz function on the metric subspace $(\mathcal{M}_\Psi, d_\beta)$. By the McShane extension theorem (McShane, 1934), define

$$\widetilde{G}_\Psi(\boldsymbol{z}) = \inf_{\boldsymbol{u} \in \mathcal{M}_\Psi} \left\{ g_\Psi(\boldsymbol{u}) + L_\delta \|\boldsymbol{z} - \boldsymbol{u}\|_2^\beta \right\}, \qquad \boldsymbol{z} \in \mathbb{R}^q. \tag{9}$$

First, by definition,

$$\widetilde{G}_\Psi(\boldsymbol{z}) = g_\Psi(\boldsymbol{z}), \qquad \boldsymbol{z} \in \mathcal{M}_\Psi.$$

Next, for arbitrary $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^q$, the triangle inequality for d_β gives

$$\|\boldsymbol{z} - \boldsymbol{u}\|_2^\beta \le \|\boldsymbol{z} - \boldsymbol{z}'\|_2^\beta + \|\boldsymbol{z}' - \boldsymbol{u}\|_2^\beta$$

for every $\boldsymbol{u} \in \mathcal{M}_\Psi$. Consequently,

$$\widetilde{G}_\Psi(\boldsymbol{z}) \le \widetilde{G}_\Psi(\boldsymbol{z}') + L_\delta \|\boldsymbol{z} - \boldsymbol{z}'\|_2^\beta.$$

Interchanging $\boldsymbol{z}$ and $\boldsymbol{z}'$ gives

$$|\widetilde{G}_\Psi(\boldsymbol{z}) - \widetilde{G}_\Psi(\boldsymbol{z}')| \le L_\delta \|\boldsymbol{z} - \boldsymbol{z}'\|_2^\beta.$$

Note that the McShane extension is not necessarily uniformly bounded on all of $\mathbb{R}^q$. Define the truncation operator

$$\mathcal{T}_{B_0}(a) = \max\{-B_0, \min\{a, B_0\}\},$$

and set

$$G_\Psi = \mathcal{T}_{B_0} \circ \widetilde{G}_\Psi.$$

Since $|g_\Psi(\boldsymbol{z})| \le B_0$ for every $\boldsymbol{z} \in \mathcal{M}_\Psi$, the truncation leaves g_Ψ unchanged on $\mathcal{M}_\Psi$. Hence $G_\Psi|_{\mathcal{M}_\Psi} = g_\Psi$, which gives (7).

Moreover, $\mathcal{T}_{B_0}$ is 1-Lipschitz, and therefore

$$|G_\Psi(\boldsymbol{z}) - G_\Psi(\boldsymbol{z}')| \le |\widetilde{G}_\Psi(\boldsymbol{z}) - \widetilde{G}_\Psi(\boldsymbol{z}')| \le L_\delta \|\boldsymbol{z} - \boldsymbol{z}'\|_2^\beta.$$

Thus,

$$[G_\Psi]_{\mathcal{H}^\beta(\mathbb{R}^q)} \le L_\delta.$$

By construction,

$$\|G_\Psi\|_{L^\infty(\mathbb{R}^q)} \le B_0.$$

Since $L_\delta = B_0/(1-\delta)^\beta \ge B_0$, it follows that

$$\|G_\Psi\|_{\mathcal{H}^\beta(\mathbb{R}^q)} = \max\left\{ \|G_\Psi\|_{L^\infty(\mathbb{R}^q)}, [G_\Psi]_{\mathcal{H}^\beta(\mathbb{R}^q)} \right\} \le \frac{B_0}{(1-\delta)^\beta}.$$

□

We next show that the preceding linear projection can be implemented within a standard softmax self-attention layer.

Lemma 6. *Let $\Psi \in \mathbb{R}^{q \times (d_{\text{in}} D)}$ be fixed. There exists a standard Transformer prefix which maps any $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_D] \in \mathbb{R}^{d_{\text{in}} \times D}$ to a representation whose final q coordinates at every token are exactly $\Psi \operatorname{vec}(\boldsymbol{x})$. The projection self-attention layer can be chosen with $H_{\text{proj}} = D$, $S_{\text{proj}} = \min\{d_{\text{in}}, q\}$, and embedding dimension $d_{\text{proj}} = d_{\text{in}} + D + 1 + q$. Its number of self-attention parameters is*

$$M_{\text{SA}}^{(\text{proj})} = 4D\, d_{\text{proj}} S_{\text{proj}}.$$

Proof. Write $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D]$, $\boldsymbol{x}_j \in \mathbb{R}^{d_{\text{in}}}$, and, consistently with column-wise vectorization, partition $\Psi = [\Psi_1, \ldots, \Psi_D]$, $\Psi_j \in \mathbb{R}^{q \times d_{\text{in}}}$. Then

$$\Psi \operatorname{vec}(\boldsymbol{x}) = \sum_{j=1}^{D} \Psi_j \boldsymbol{x}_j.$$

Define

$$P = \frac{1}{2}\mathbf{I}_D + \frac{1}{2D}\mathbf{1}_D\mathbf{1}_D^\top = (P_{hj})_{h,j=1}^{D}.$$

The matrix P is strictly positive, row stochastic, and invertible, with

$$P^{-1} = 2\mathbf{I}_D - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^\top.$$

For $h = 1, \ldots, D$, define

$$C_h = \sum_{k=1}^{D} \Psi_k (P^{-1})_{kh}.$$

Then, for every j,

$$\sum_{h=1}^{D} P_{hj} C_h = \sum_{k=1}^{D} \Psi_k \sum_{h=1}^{D} (P^{-1})_{kh} P_{hj} = \Psi_j. \tag{10}$$

We first consider the embedding layer. Set $d_{\text{proj}} = d_{\text{in}} + D + 1 + q$. Choose

$$\mathbf{W}_{\text{EB}} = \begin{pmatrix} \mathbf{I}_{d_{\text{in}}} \\ 0_{D \times d_{\text{in}}} \\ 0_{1 \times d_{\text{in}}} \\ 0_{q \times d_{\text{in}}} \end{pmatrix}, \qquad \mathbf{B}_{\text{EB}} = \begin{pmatrix} 0_{d_{\text{in}} \times D} \\ \mathbf{I}_D \\ \mathbf{1}_D^\top \\ 0_{q \times D} \end{pmatrix}.$$

Then, we get

$$F_{\text{EB}}(\boldsymbol{x}) = \mathbf{W}_{\text{EB}}\boldsymbol{x} + \mathbf{B}_{\text{EB}} = \begin{pmatrix} \boldsymbol{x} \\ \mathbf{I}_D \\ \mathbf{1}_D^\top \\ 0_{q \times D} \end{pmatrix},$$

so that its j-th token is

$$\boldsymbol{y}_j = (\boldsymbol{x}_j^\top, \boldsymbol{e}_j^\top, 1, 0_q^\top)^\top \in \mathbb{R}^{d_{\text{proj}}},$$

where $\boldsymbol{e}_j \in \mathbb{R}^D$ is the j-th canonical basis vector.

According to the Transformer architecture specified in Section 4, the embedding layer is followed by a feedforward block before the initial self-attention layer. We choose $F_{\text{FF}}^{(0)}$ to implement the identity map exactly. Let $\mathbf{I}_{\text{proj}} = \mathbf{I}_{d_{\text{proj}}}$. For every $\boldsymbol{z} \in \mathbb{R}^{d_{\text{proj}}}$,

$$\boldsymbol{z} = (\mathbf{I}_{\text{proj}} \quad -\mathbf{I}_{\text{proj}})\, \sigma_{\text{ReLU}} \left(\begin{pmatrix} \mathbf{I}_{\text{proj}} \\ -\mathbf{I}_{\text{proj}} \end{pmatrix} \boldsymbol{z} \right).$$

Hence, $F_{\text{FF}}^{(0)}$ can be chosen as an exact identity ReLU block of depth two and width $2d_{\text{proj}}$.

We next construct the projection self-attention layer $F_{\text{SA}}^{(1)}$. Let $S_{\text{proj}} = \min\{d_{\text{in}}, q\}$. The h-th attention head is designed to produce

$$C_h \sum_{j=1}^{D} P_{hj} \boldsymbol{x}_j$$

in the final q coordinates of every token.

For each $h \in [D]$, define

$$\widetilde{C}_h \in \mathbb{R}^{d_{\text{proj}} \times d_{\text{proj}}}$$

by

$$\widetilde{C}_h = \begin{pmatrix} 0_{d_{\text{in}} \times d_{\text{in}}} & 0_{d_{\text{in}} \times D} & 0_{d_{\text{in}} \times 1} & 0_{d_{\text{in}} \times q} \\ 0_{D \times d_{\text{in}}} & 0_{D \times D} & 0_{D \times 1} & 0_{D \times q} \\ 0_{1 \times d_{\text{in}}} & 0_{1 \times D} & 0 & 0_{1 \times q} \\ C_h & 0_{q \times D} & 0_{q \times 1} & 0_{q \times q} \end{pmatrix}.$$

Hence

$$\widetilde{C}_h \boldsymbol{y}_j = \begin{pmatrix} 0_{d_{\text{in}}} \\ 0_D \\ 0 \\ C_h \boldsymbol{x}_j \end{pmatrix}.$$

Thus, $\widetilde{C}_h$ ignores the positional and constant coordinates, acts on the original feature vector $\boldsymbol{x}_j$ through C_h, and writes the resulting vector $C_h \boldsymbol{x}_j$ into the final q coordinates.

Since adding zero rows and columns does not change the rank of the nonzero block,

$$\operatorname{rank}(\widetilde{C}_h) = \operatorname{rank}(C_h) \le \min\{d_{\text{in}}, q\} = S_{\text{proj}}.$$

Therefore, there exist matrices

$$\mathbf{W}_O^{(h)} \in \mathbb{R}^{d_{\text{proj}} \times S_{\text{proj}}}, \qquad \mathbf{W}_V^{(h)} \in \mathbb{R}^{S_{\text{proj}} \times d_{\text{proj}}}$$

such that

$$\mathbf{W}_O^{(h)} \mathbf{W}_V^{(h)} = \widetilde{C}_h.$$

Next, let $\boldsymbol{e}_1^{(S_{\text{proj}})}$ be the first canonical basis vector of $\mathbb{R}^{S_{\text{proj}}}$. For head h, choose the query map such that

$$W_Q^{(h)} \boldsymbol{y}_\ell = \boldsymbol{e}_1^{(S_{\text{proj}})}$$

for every $\ell \in [D]$. This is possible because every embedded token $\boldsymbol{y}_\ell$ contains the same constant coordinate 1. Similarly, choose the key map such that

$$\mathbf{W}_K^{(h)} \boldsymbol{y}_j = (\log P_{hj}) \boldsymbol{e}_1^{(S_{\text{proj}})}, \qquad j \in [D].$$

This is possible because the positional component of $\boldsymbol{y}_j$ is the canonical vector $\boldsymbol{e}_j$, which allows the key map to assign the prescribed value $\log P_{hj}$ to token j. As a consequence, for every source token j and target token ℓ,

$$\boldsymbol{y}_j^\top \mathbf{W}_K^{(h)\top} \mathbf{W}_Q^{(h)} \boldsymbol{y}_\ell = \log P_{hj}.$$

In particular, this score does not depend on ℓ, so every column of the score matrix is $(\log P_{h1}, \dots, \log P_{hD})^\top$. Since the entries in the h-th row of P are strictly positive and sum to one,

$$\operatorname{softmax}\big(\log P_{h1}, \dots, \log P_{hD}\big) = (P_{h1}, \dots, P_{hD})^\top.$$

Hence, head h uses the same attention weights $(P_{h1}, \dots, P_{hD})$ for every target token.

For every target token, the attention weights of head h are $(P_{h1}, \dots, P_{hD})^\top$. Hence, the corresponding value aggregation is

$$\sum_{j=1}^{D} P_{hj} \boldsymbol{y}_j.$$

Since $\mathbf{W}_O^{(h)} \mathbf{W}_V^{(h)} = \widetilde{C}_h$, and $\widetilde{C}_h$ reads only the first d_{in} coordinates of each embedded token, we obtain

$$\widetilde{C}_h \left(\sum_{j=1}^{D} P_{hj} \boldsymbol{y}_j \right) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ C_h \sum_{j=1}^{D} P_{hj} \boldsymbol{x}_j \end{pmatrix}.$$

Thus, the contribution of head h to the final q coordinates of every target token is exactly $C_h \sum_{j=1}^{D} P_{hj}\boldsymbol{x}_j$. Summing over all D heads and using (10), we obtain

$$\begin{aligned}\sum_{h=1}^{D} C_h \sum_{j=1}^{D} P_{hj}\boldsymbol{x}_j &= \sum_{j=1}^{D}\left(\sum_{h=1}^{D} P_{hj}C_h\right)\boldsymbol{x}_j \\ &= \sum_{j=1}^{D} \Psi_j \boldsymbol{x}_j \\ &= \Psi\,\mathrm{vec}(\boldsymbol{x}).\end{aligned}$$

Thus, the same vector $\Psi\,\mathrm{vec}(\boldsymbol{x})$ is written into the final q coordinates of every token. By construction, these q coordinates were initialized to zero by the embedding layer. Hence, the residual connection does not alter the projected representation.

Finally, each head contains the four matrices $\mathbf{W}_O^{(h)}$, $\mathbf{W}_V^{(h)}$, $\mathbf{W}_K^{(h)}$, and $\mathbf{W}_Q^{(h)}$, with $4d_{\mathrm{proj}}S_{\mathrm{proj}}$ scalar parameters in total. Since $H_{\mathrm{proj}} = D$, the projection self-attention layer contains

$$M_{\mathrm{SA}}^{(\mathrm{proj})} = 4D\, d_{\mathrm{proj}} S_{\mathrm{proj}}$$

parameters. □

Proof of Theorem 3. We denote that

$$\mathfrak{B}_{q,\delta} = \max\left\{1,\, B_0,\, B_0\left(\frac{2t}{(1-\delta)\sqrt{q}}\right)^{\beta}\right\}, \qquad \varepsilon_{q,\delta} = \frac{\varepsilon}{\mathfrak{B}_{q,\delta}}.$$

The proof consists of four steps.

Step 1: Construction of a low-dimensional Hölder target. By Lemma 5, there exists $G_\Psi \in \mathcal{H}^\beta(\mathbb{R}^q)$ such that

$$G_\Psi\big(\Psi\,\mathrm{vec}(\boldsymbol{x})\big) = u_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{M},$$

and

$$\|G_\Psi\|_{L^\infty(\mathbb{R}^q)} \le B_0, \qquad [G_\Psi]_{\mathcal{H}^\beta(\mathbb{R}^q)} \le \frac{B_0}{(1-\delta)^\beta}.$$

We also know that, from Lemma 4, $\|\Psi\|_{\mathrm{op}} = \sqrt{t/q}$. Indeed, all nonzero singular values of Ψ are $\sqrt{t/q}$. For every $\boldsymbol{x} \in [0,1]^t$, $\|\boldsymbol{x}\|_2 \le \sqrt{t}$, and hence

$$\|\Psi\boldsymbol{x}\|_\infty \le \|\Psi\boldsymbol{x}\|_2 \le \sqrt{\frac{t}{q}}\|\boldsymbol{x}\|_2 \le \frac{t}{\sqrt{q}}.$$

Define $R_\delta = t/\sqrt{q}$, $E_\delta = [-R_\delta, R_\delta]^q$. Then $\Psi([0,1]^t) \subset E_\delta$. Define the affine rescaling

$$\Psi_\delta(\boldsymbol{z}) = \frac{\boldsymbol{z} + R_\delta \mathbf{1}_q}{2R_\delta}, \qquad \boldsymbol{z} \in E_\delta, \tag{11}$$

and let

$$h_\Psi(\boldsymbol{u}) = G_\Psi\left(2R_\delta \boldsymbol{u} - R_\delta \mathbf{1}_q\right), \qquad \boldsymbol{u} \in [0,1]^q.$$

Then $\|h_\Psi\|_{L^\infty([0,1]^q)} \le B_0$. Moreover, for every $\boldsymbol{u}, \boldsymbol{v} \in [0,1]^q$,

$$\begin{aligned}|h_\Psi(\boldsymbol{u}) - h_\Psi(\boldsymbol{v})| &\le \frac{B_0}{(1-\delta)^\beta}\|2R_\delta(\boldsymbol{u}-\boldsymbol{v})\|_2^\beta \\ &= \frac{B_0(2R_\delta)^\beta}{(1-\delta)^\beta}\|\boldsymbol{u}-\boldsymbol{v}\|_2^\beta.\end{aligned}$$

Therefore,

$$\|h_\Psi\|_{\mathcal{H}^\beta([0,1]^q)} \le \mathfrak{B}_{q,\delta}.$$

Step 2: Approximation of the projected target by a Transformer. Define $\mathbf{h}_\Psi(\boldsymbol{u}) =$

$(h_\Psi(\boldsymbol{u}), 0, \ldots, 0)^\top \in \mathbb{R}^q$, and regard $\boldsymbol{u} \in [0,1]^q$ as a one-token input in $[0,1]^{q\times 1}$. Set $\widetilde{\mathbf{h}}_{q,\delta} = \mathbf{h}_\Psi/\mathfrak{B}_{q,\delta}$, $\varepsilon_{q,\delta} = \varepsilon/\mathfrak{B}_{q,\delta}$. Since $\mathfrak{B}_{q,\delta} \geq 1$, we have $0 < \varepsilon_\Psi < 1$. Moreover, every component of $\widetilde{\mathbf{h}}_\Psi$ has $\mathcal{H}^\beta$-norm at most one. Apply Theorem 2 of Lai & Sun (2026) with

$$s = 0, \qquad \lambda = \beta, \qquad d = q, \qquad n = 1.$$

There exists a Transformer $\widetilde{T}_{\mathrm{low},\varepsilon} : \mathbb{R}^{q\times 1} \to \mathbb{R}^{q\times 1}$ such that

$$\sup_{\boldsymbol{u}\in[0,1]^q} \left| [\widetilde{T}_{\mathrm{low},\varepsilon}(\boldsymbol{u})]_1 - \frac{h_\Psi(\boldsymbol{u})}{\mathfrak{B}_{q,\delta}} \right| \leq \varepsilon_{q,\delta}.$$

Moreover, the architecture of $\widetilde{T}_{\mathrm{low},\varepsilon}$ is given by the size tuple

$$\big((\widetilde{L}_0, \widetilde{p}_0), (\widetilde{H}_1, \widetilde{S}_1), (\widetilde{L}_1, \widetilde{p}_1)\big),$$

where

$$\widetilde{L}_0 = 4, \quad \widetilde{p}_0 = O\left(\varepsilon_{q,\delta}^{-1/\beta}\right), \quad \widetilde{H}_1 = (3q+1)3^q, \quad \widetilde{S}_1 = \max\{q, 3\},$$

and

$$\widetilde{L}_1 = O\left(\log \frac{1}{\varepsilon_{q,\delta}}\right), \quad \widetilde{p}_1 = O\left(\varepsilon_{q,\delta}^{-q/\beta}\right).$$

Its dimension vector is

$$\boldsymbol{d}_{\mathrm{low}} = \big(q,\ q+5,\ 7q3^q,\ q\big).$$

Moreover,

$$\widetilde{M}_{\mathrm{EB}} = O(1), \qquad \widetilde{M}_{\mathrm{FF}} = O\left(\varepsilon_{q,\delta}^{-q/\beta}\right), \qquad \widetilde{M}_{\mathrm{SA}} = O(1),$$

and

$$\widetilde{B}_{\mathrm{EB}} = O(1), \qquad \widetilde{B}_{\mathrm{FF}} = O\left(\varepsilon_{q,\delta}^{-(6q+2)/\beta}\right), \qquad \widetilde{B}_{\mathrm{SA}} = O\left(\log \frac{1}{\varepsilon_{q,\delta}}\right).$$

Multiplying the final affine output transformation by $\mathfrak{B}_{q,\delta}$ yields another standard Transformer $T_{\mathrm{low},\varepsilon}$ satisfying

$$\sup_{\boldsymbol{u}\in[0,1]^q} |[T_{\mathrm{low},\varepsilon}(\boldsymbol{u})]_1 - h_\Psi(\boldsymbol{u})| \leq \varepsilon. \tag{12}$$

This rescaling changes only parameter magnitudes and not the number of parameters.

Step 3: Incorporating the projection into the Transformer architecture. By Lemma 6, after applying F_{EB}, $F_{\mathrm{FF}}^{(0)}$, and $F_{\mathrm{SA}}^{(1)}$, the designated q coordinates of each of the D tokens are exactly $\Psi\,\mathrm{vec}(\boldsymbol{x})$.

After $F_{\mathrm{SA}}^{(1)}$, the designated q coordinates of every token are equal to $\Psi\,\mathrm{vec}(\boldsymbol{x})$, while the remaining coordinates may still depend on the token index. Let Proj_q denote the linear map that extracts these q coordinates, and define

$$\boldsymbol{u}(\boldsymbol{x}) = \Psi_\delta(\Psi\,\mathrm{vec}(\boldsymbol{x})) \in [0,1]^q,$$

where Ψ_δ is defined in (11). Conceptually, $\boldsymbol{u}(\boldsymbol{x})$ serves as the input to the low-dimensional Transformer $T_{\mathrm{low},\varepsilon}$, and would therefore first pass through its embedding layer $F_{\mathrm{EB}}^{\mathrm{low}}$ and then its first feedforward block $F_{\mathrm{FF},0}^{\mathrm{low}}$.

To preserve the standard Transformer architecture, we do not insert $F_{\mathrm{EB}}^{\mathrm{low}}$ as an additional embedding layer after $F_{\mathrm{SA}}^{(1)}$. Instead, since Proj_q, Ψ_δ, and $F_{\mathrm{EB}}^{\mathrm{low}}$ are all affine maps, their composition

$$F_{\mathrm{EB}}^{\mathrm{low}} \circ \Psi_\delta \circ \mathrm{Proj}_q$$

can be absorbed exactly into the first affine transformation of $F_{\mathrm{FF},0}^{\mathrm{low}}$. We denote the resulting feedforward block by $F_{\mathrm{FF}}^{(1)}$. Thus, if $\boldsymbol{z}_j$ denotes the j-th token at the output of $F_{\mathrm{SA}}^{(1)}$, then

$$F_{\mathrm{FF}}^{(1)}(\boldsymbol{z}_j) = F_{\mathrm{FF},0}^{\mathrm{low}}\left(F_{\mathrm{EB}}^{\mathrm{low}}\left(\Psi_\delta\big(\mathrm{Proj}_q(\boldsymbol{z}_j)\big)\right)\right).$$

Since $\mathrm{Proj}_q(\boldsymbol{z}_j) = \Psi\,\mathrm{vec}(\boldsymbol{x})$ for every $j \in [D]$, the output of $F_{\mathrm{FF}}^{(1)}$ is identical across all D tokens. In particular, each token contains the same hidden representation

$$F_{\mathrm{FF},0}^{\mathrm{low}}\left(F_{\mathrm{EB}}^{\mathrm{low}}\left(\boldsymbol{u}(\boldsymbol{x})\right)\right).$$

After the merged feedforward block $F_{\mathrm{FF}}^{(1)}$, all D tokens are identical. We next show that the remaining one-token computation of $T_{\mathrm{low},\varepsilon}$ can be replicated exactly across these identical tokens.

Suppose that, at some stage of the network, the token matrix has the form

$$\boldsymbol{\Gamma} = \boldsymbol{\gamma}\mathbf{1}_D^\top$$

for some vector $\boldsymbol{\gamma}$. Consider an arbitrary self-attention head. We have

$$\boldsymbol{\Gamma}^\top \mathbf{W}_K^\top \mathbf{W}_Q \boldsymbol{\Gamma} = \mathbf{1}_D\big(\boldsymbol{\gamma}^\top \mathbf{W}_K^\top \mathbf{W}_Q \boldsymbol{\gamma}\big)\mathbf{1}_D^\top = c\mathbf{1}_D\mathbf{1}_D^\top,$$

where $c = \boldsymbol{\gamma}^\top \mathbf{W}_K^\top \mathbf{W}_Q \boldsymbol{\gamma} \in \mathbb{R}$. Thus all entries of the attention-score matrix are identical. Its column-wise softmax is therefore

$$\frac{1}{D}\mathbf{1}_D\mathbf{1}_D^\top.$$

Consequently,

$$\mathbf{W}_V\boldsymbol{\Gamma}\left(\frac{1}{D}\mathbf{1}_D\mathbf{1}_D^\top\right) = \mathbf{W}_V\boldsymbol{\gamma}\mathbf{1}_D^\top.$$

Thus each attention head produces the same update at every token, and this update coincides exactly with the corresponding one-token computation. The same conclusion therefore holds for the complete self-attention layer, including its residual connection. Moreover, since each feedforward block acts independently and identically on the token columns, it also preserves this replication property. Induction over the remaining layers shows that every subsequent hidden state is exactly the corresponding one-token hidden state replicated across all D tokens. Hence the remaining computation of $T_{\mathrm{low},\varepsilon}$ is realized exactly while preserving the specified alternating feedforward/self-attention architecture.

It remains to account for the architectural quantities and the number of parameters introduced by the construction. The resulting Transformer T_ε has $R = 2$ self-attention layers and size tuple

$$\big((L_0, p_0), (H_1, S_1), (L_1, p_1), (H_2, S_2), (L_2, p_2)\big),$$

where

$$L_0 = 2, \qquad p_0 = 2d_{\mathrm{proj}}, \qquad H_1 = D, \qquad S_1 = S_{\mathrm{proj}} = \min\{d_{\mathrm{in}}, q\},$$
$$L_1 = 4, \qquad p_1 = O\left(\varepsilon_{q,\delta}^{-1/\beta}\right), \qquad H_2 = (3q+1)3^q, \qquad S_2 = \max\{q, 3\},$$

and

$$L_2 = O\left(\log\frac{1}{\varepsilon_{q,\delta}}\right), \qquad p_2 = O\left(\varepsilon_{q,\delta}^{-q/\beta}\right).$$

Here (L_0, p_0) corresponds to the identity feedforward block in the projection module, (H_1, S_1) to the projection self-attention layer, and the remaining quantities are inherited from $T_{\mathrm{low},\varepsilon}$. The dimension vector of T_ε is

$$\boldsymbol{d} = \big(d_{\mathrm{in}}, d_{\mathrm{proj}}, d_{\mathrm{proj}}, 7q3^q, q\big), \qquad d_{\mathrm{proj}} = d_{\mathrm{in}} + D + 1 + q.$$

The embedding layer contains $M_{\mathrm{EB}} = d_{\mathrm{proj}}(d_{\mathrm{in}} + D)$ parameters. The two self-attention layers contain

$$M_{\mathrm{SA}} = 4D\,d_{\mathrm{proj}}S_{\mathrm{proj}} + 4(3q+1)3^q(7q3^q)\max\{q,3\} \le C_0$$

parameters, where C_0 is independent of ε and ζ.

The projection feedforward block $F_{\mathrm{FF}}^{(0)}$ contains only a fixed number of parameters independent of ε and ζ. From Step 2, the feedforward blocks of $T_{\mathrm{low},\varepsilon}$ contain at most $C_1\varepsilon_{q,\delta}^{-q/\beta}$ parameters. It remains to account for the merging operation. Before merging, the first weight matrix of $F_{\mathrm{FF},0}^{\mathrm{low}}$ has input dimension $q+5$ and output dimension $p_1 = O(\varepsilon_{q,\delta}^{-1/\beta})$. After absorbing Proj_q, Ψ_δ, and $F_{\mathrm{EB}}^{\mathrm{low}}$, this matrix is replaced by a matrix with input dimension d_{proj} and the same output dimension p_1.

Hence, the number of parameters contributed by the modified affine transformation is bounded, up to a constant independent of ε and ζ, by $d_{\text{proj}}\varepsilon_{q,\delta}^{-1/\beta}$. Consequently,

$$M_{\text{FF}}(T_\varepsilon) \le C_0 + C_1\varepsilon_{q,\delta}^{-q/\beta} + C_2 d_{\text{proj}}\varepsilon_{q,\delta}^{-1/\beta}.$$

Since $q \ge 1$, $0 < \varepsilon_{q,\delta} < 1$, and d_{proj} is independent of ε and ζ,

$$d_{\text{proj}}\varepsilon_{q,\delta}^{-1/\beta} \le d_{\text{proj}}\varepsilon_{q,\delta}^{-q/\beta}.$$

After adjusting the constants and using $\varepsilon_{q,\delta} = \varepsilon/\mathfrak{B}_{q,\delta}$, we obtain

$$M_{\text{FF}}(T_\varepsilon) \lesssim \left(\frac{\mathfrak{B}_{q,\delta}}{\varepsilon}\right)^{q/\beta}.$$

Finally, the parameter magnitudes satisfy

$$B_{\text{EB}} \le C, \qquad B_{\text{FF}} \le C\varepsilon_{q,\delta}^{-(6q+2)/\beta}, \qquad B_{\text{SA}} \le C\log\frac{1}{\varepsilon_{q,\delta}}.$$

The constants in these bounds may depend on the fixed geometric, smoothness, and architectural quantities, but not on ε or ζ.

Step 4: Error caused by the thickness of the approximate manifold. Fix $\boldsymbol{x} \in \mathcal{M}_\zeta$. Since $\mathcal{M}$ is compact, there exists $\widetilde{\boldsymbol{x}} \in \mathcal{M}$ attaining the distance from $\boldsymbol{x}$ to $\mathcal{M}$, so that

$$\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\text{F}} = \text{dist}_{\text{F}}(\boldsymbol{x}, \mathcal{M}) \le \zeta.$$

Set $\boldsymbol{v} = \text{vec}(\boldsymbol{x})$, $\widetilde{\boldsymbol{v}} = \text{vec}(\widetilde{\boldsymbol{x}})$. By Lemma 5, $G_\Psi(\Psi\widetilde{\boldsymbol{v}}) = u_0(\widetilde{\boldsymbol{x}})$. Therefore,

$$\begin{aligned}|t_\varepsilon(\boldsymbol{x}) - u_0(\boldsymbol{x})| \le |t_\varepsilon(\boldsymbol{x}) - G_\Psi(\Psi\boldsymbol{v})| + |G_\Psi(\Psi\boldsymbol{v}) - G_\Psi(\Psi\widetilde{\boldsymbol{v}})| \\ + |u_0(\widetilde{\boldsymbol{x}}) - u_0(\boldsymbol{x})|. \end{aligned} \tag{13}$$

The first term in the right hand side of (13) is bounded by ε. For the second term, Lemmas 4 and 5 yield

$$\begin{aligned}|G_\Psi(\Psi\boldsymbol{v}) - G_\Psi(\Psi\widetilde{\boldsymbol{v}})| &\le \frac{B_0}{(1-\delta)^\beta}\|\Psi(\boldsymbol{v} - \widetilde{\boldsymbol{v}})\|_2^\beta \\ &\le \frac{B_0}{(1-\delta)^\beta}\|\Psi\|_{\text{op}}^\beta\|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|_2^\beta \\ &= \frac{B_0}{(1-\delta)^\beta}\left(\frac{t}{q}\right)^{\beta/2}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\text{F}}^\beta \\ &\le \frac{B_0}{(1-\delta)^\beta}\left(\frac{t}{q}\right)^{\beta/2}\zeta^\beta.\end{aligned}$$

For the third term, the Hölder continuity of u_0 gives

$$|u_0(\widetilde{\boldsymbol{x}}) - u_0(\boldsymbol{x})| \le B_0\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|_{\text{F}}^\beta \le B_0\zeta^\beta.$$

Substituting these estimates into (13) gives

$$|t_\varepsilon(\boldsymbol{x}) - u_0(\boldsymbol{x})| \le \varepsilon + B_0\left[\frac{1}{(1-\delta)^\beta}\left(\frac{t}{q}\right)^{\beta/2} + 1\right]\zeta^\beta.$$

Since $\boldsymbol{x} \in \mathcal{M}_\zeta$ was arbitrary, taking the supremum proves the assertion. □

Proposition 2. *Let $\mathcal{T}$ be the scalar function class obtained from a fixed output coordinate of Transformers on $\mathcal{X} \subseteq [0,1]^{d_{\text{in}} \times D}$. The Transformers share a fixed standard architecture, including the trainable parameter locations, with $R \ge 1$ self-attention layers, maximal feedforward depth L, maximal feedforward width p, maximal number of attention heads H, maximal head dimension S, and dimension vector $\boldsymbol{d} = (d_{\text{in}}, d_0, \ldots, d_R, d_{\text{out}})$. Let M_{EB}, M_{FF}, and M_{SA} bound the respective*

numbers of embedding, feedforward, and self-attention parameters. These parameters range independently over the corresponding intervals $[-B_{\mathrm{EB}}, B_{\mathrm{EB}}]$, $[-B_{\mathrm{FF}}, B_{\mathrm{FF}}]$, *and* $[-B_{\mathrm{SA}}, B_{\mathrm{SA}}]$, *where* $B_{\mathrm{EB}}, B_{\mathrm{FF}}, B_{\mathrm{SA}} \geq 1$. *For* $K \geq 2$ *and* $\mathcal{B} > 0$, *define*

$$\mathcal{F}_{\mathcal{T},\mathcal{B}} = \left\{ \Pi_{\mathcal{B}} \circ (t_1, \ldots, t_K)^\top : t_k \in \mathcal{T},\ k \in [K] \right\}.$$

Then, for every $0 < \varsigma \leq 1$,

$$\log \mathcal{N}\left(\varsigma, \mathcal{F}_{\mathcal{T},\mathcal{B}}, d_\infty\right) \leq K \left[M_{\mathrm{EB}} \log\left(\frac{2B_{\mathrm{EB}}\mathfrak{L}}{\varsigma}\right) + M_{\mathrm{FF}} \log\left(\frac{2B_{\mathrm{FF}}\mathfrak{L}}{\varsigma}\right) + M_{\mathrm{SA}} \log\left(\frac{2B_{\mathrm{SA}}\mathfrak{L}}{\varsigma}\right) \right],$$

where

$$\mathfrak{L} = (2R+2)6^R 4^{R^2+R+4} D^{R^2+5R/2+3} d_{\mathrm{in}}^{R+1/2} d_0^{2R+1} \left(\prod_{r=1}^{R} d_r^{4(R-r)+6}\right) d_{\mathrm{out}}^{1/2} \times H^{R^2+R-1} S^{R^2+2R+1} L^{R^2+2R+3} p^{(L-1)(R^2+3R+3)} B_{\mathrm{EB}}^{2R+1} B_{\mathrm{FF}}^{L(R^2+3R+3)} B_{\mathrm{SA}}^{2(R^2+2R+1)}.$$

Proof. The pointwise parameter perturbation estimate in the proof of Lai & Sun (2026, Lemma 4), followed by discretization within the admissible parameter boxes, gives

$$\mathcal{N}\left(\varsigma, \mathcal{T}, \|\cdot\|_{L^\infty(\mathcal{X})}\right) \leq \prod_{a \in \{\mathrm{EB},\mathrm{FF},\mathrm{SA}\}} \left(\frac{2B_a\mathfrak{L}}{\varsigma}\right)^{M_a}.$$

The Cartesian product of K scalar ς-covers is a ς-cover of $\mathcal{T}^K$ under d_∞. Since coordinatewise clipping is 1-Lipschitz under this metric,

$$\mathcal{N}\left(\varsigma, \mathcal{F}_{\mathcal{T},\mathcal{B}}, d_\infty\right) \leq \mathcal{N}\left(\varsigma, \mathcal{T}, \|\cdot\|_{L^\infty(\mathcal{X})}\right)^K.$$

Taking logarithms proves the result. □

A.3 Comparison with Existing Transformer Approximation Results

Table 1 compares our approximation result with existing Transformer results in terms of attention mechanisms, geometric assumptions, target function classes, and error bounds.

B Experimental Details and Additional Results

B.1 The ρ-estimator and competing methods

We use the notation of Section 3: $f = (f_1, \ldots, f_K)^\top$ is a vector-valued logit function, Q_f is its induced conditional distribution, and K denotes the number of classes. Let $\mathcal{F}_\Theta = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be the fitted family, with parameter dimension d_θ. For an observed training sample $\widetilde{\mathcal{D}}_n = \{(\widetilde{\boldsymbol{X}}_i, \widetilde{Y}_i)\}_{i=1}^n$, write

$$\ell_i(\boldsymbol{\theta}) = -\log Q_{f_{\boldsymbol{\theta}}}(\widetilde{Y}_i \mid \widetilde{\boldsymbol{X}}_i).$$

All methods optimize the same model parameters over the same set Θ. We denote B for the number of MOM blocks.

B.2 Numerical approximation to the ρ-estimator

The implemented procedure is the local search in Algorithm 1. For a current parameter $\boldsymbol{\theta}$ and a challenger $\boldsymbol{\eta}$, recall that

$$\rho(\widetilde{\mathcal{D}}_n; f_{\boldsymbol{\theta}}, f_{\boldsymbol{\eta}}) = \sum_{i=1}^n \psi\left(\sqrt{\frac{Q_{f_{\boldsymbol{\eta}}}(\widetilde{Y}_i \mid \widetilde{\boldsymbol{X}}_i)}{Q_{f_{\boldsymbol{\theta}}}(\widetilde{Y}_i \mid \widetilde{\boldsymbol{X}}_i)}}\right).$$

Table 1: Architectures and approximation settings of selected Transformer results.

Work	Attention	Uses manifold geometry	Domain	Target class	Uniform error bound
Havrilla & Liao (2024)	Multi-head ReLU	✓	Exact manifold	Hölder	ε
Shen et al. (2025)	Multi-head ReLU	✓	Manifold tube	$f = g \circ \pi_{\mathcal{M}}$; Hölder g	ε
Shi & Labate (2026)	Single-head ReLU	✓	Exact manifold	Hölder	ε
Jiao et al. (2025)	Multi-head softmax	×	Euclidean cube	Hölder	ε
Lai & Sun (2026)	Multi-head softmax	×	Euclidean cube	$C^{s,\lambda}$	ε
Shi & Liao (2026)	Softmax; multi-head then single-head	✓	Exact manifold	Hölder	ε
This work	Multi-head softmax	✓	Manifold neighborhood	General ambient Hölder	$\varepsilon + \Gamma\zeta^{\beta}$

Note. ✓/× indicate whether the compared theorem exploits manifold geometry, not architectural capability; ζ denotes neighborhood thickness. Bounds are shown without their differing complexity requirements.

To illustrate, consider $J = 3$. At iteration ℓ, the current parameter $\boldsymbol{\theta}^{(\ell)}$ is held fixed while three local challenger problems are initialized at

$$\boldsymbol{\theta}^{(\ell)}, \qquad \Pi_{\Theta}\left(\boldsymbol{\theta}^{(\ell)} + \sigma_1 \boldsymbol{\xi}_{\ell,1}\right), \qquad \Pi_{\Theta}\left(\boldsymbol{\theta}^{(\ell)} + \sigma_2 \boldsymbol{\xi}_{\ell,2}\right),$$

where $\boldsymbol{\xi}_{\ell,1}$ and $\boldsymbol{\xi}_{\ell,2}$ are independent draws from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_\theta})$, and Π_{Θ} clips each coordinate to its feasible interval. The perturbations affect all fitted parameters. Let $\boldsymbol{\eta}^{(\ell)}$ be the returned challenger with the largest attained contrast, and set

$$V_\ell = \rho(\widetilde{\mathcal{D}}_n; f_{\boldsymbol{\theta}^{(\ell)}}, f_{\boldsymbol{\eta}^{(\ell)}}), \qquad \boldsymbol{\theta}^{(\ell+1)} = \boldsymbol{\eta}^{(\ell)}.$$

After updating the parameter, the procedure stops if $V_\ell < 1$ or the outer iteration limit is reached. In the synthetic experiment, $\sigma_1 = 0.10$, $\sigma_2 = 0.35$, and the outer iteration limit is ten. Thus the code stops when its minimized statistic is strictly greater than -1. The returned parameter is not challenged again after this stopping test. The experimental ρ fit is a numerical approximation to the theoretical estimator.

B.2.1 Competing objectives

Maximum likelihood. The MLE objective is the total negative log-likelihood,

$$L_{\mathrm{MLE}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}).$$

Median of means. For a block count B, a random permutation of the observation indices is divided into blocks $\mathcal{I}_1, \ldots, \mathcal{I}_B$. We minimize

$$L_{\mathrm{MOM},B}(\boldsymbol{\theta}) = \mathrm{median}_{b\in[B]} \left\{ \frac{1}{|\mathcal{I}_b|} \sum_{i\in\mathcal{I}_b} \ell_i(\boldsymbol{\theta}) \right\}.$$

For even B, the median is the arithmetic mean of the two central ordered block losses. The blocks remain fixed throughout optimization and across the two initializations. All blocks share a single parameter vector. We report $B \in \{5, 10, 20, 50\}$ separately; with $n = 500$, the corresponding block sizes are $100, 50, 25, 10$.

Huber-based likelihood objective. For $\delta > 0$, define

$$H_\delta(r) = \begin{cases} r^2/2, & 0 \le r \le \delta, \\ \delta r - \delta^2/2, & r > \delta, \end{cases} \qquad L_{\mathrm{Huber},\delta}(\boldsymbol{\theta}) = \sum_{i=1}^{n} H_\delta\Big(\sqrt{2\ell_i(\boldsymbol{\theta})}\Big).$$

Each contribution equals $\ell_i(\boldsymbol{\theta})$ when $\ell_i(\boldsymbol{\theta}) \le \delta^2/2$ and equals $\delta\sqrt{2\ell_i(\boldsymbol{\theta})} - \delta^2/2$ otherwise. Thus the Huber transformation acts on the square root of twice the negative log-likelihood. We report $\delta \in \{0.5, 1, 2, 4\}$ separately. The terms MOM and Huber refer to these specific objectives; no tuning value is selected using the test set.

B.3 Synthetic experimental setting

This subsection specifies the experiment summarized in Section 5.2. We use $K = 5$ classes, $D = 5$ scalar tokens, and $H = 2$ attention heads. With $d_{\mathrm{in}} = 1$, we identify each token array with a vector $\boldsymbol{x}$ in the experimental input space $\mathcal{X}_{\mathrm{syn}} = \mathbb{R}^D$.

B.3.1 Model and clean-data distribution

Let $r_j = -1 + 2(j-1)/(D-1)$, $j \in [D]$, be the fixed token positions. Write $\boldsymbol{\vartheta} = (a_1, b_1, a_2, b_2)^\top$ for the attention parameter vector and define

$$\alpha_{hj}(\boldsymbol{x}; \boldsymbol{\vartheta}) = \frac{\exp(a_h x_j + b_h r_j)}{\sum_{v=1}^{D} \exp(a_h x_v + b_h r_v)},$$
$$z_h(\boldsymbol{x}; \boldsymbol{\vartheta}) = \sum_{j=1}^{D} \alpha_{hj}(\boldsymbol{x}; \boldsymbol{\vartheta}) x_j, \qquad h \in [H].$$

The feature vector and coefficient matrix are

$$\boldsymbol{\phi}_{\boldsymbol{\vartheta}}(\boldsymbol{x}) = \big(1, z_1(\boldsymbol{x}; \boldsymbol{\vartheta}), z_2(\boldsymbol{x}; \boldsymbol{\vartheta})\big)^\top, \qquad \mathbf{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{K-1})^\top \in \mathbb{R}^{(K-1)\times 3},$$

where $\boldsymbol{\beta}_k \in \mathbb{R}^3$. The full parameter vector is

$$\boldsymbol{\theta} = (\boldsymbol{\vartheta}^\top, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_{K-1}^\top)^\top \in \mathbb{R}^{16}.$$

We impose the reference-logit constraint $f_{\boldsymbol{\theta},K} = 0$ and set

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \boldsymbol{\beta}_k^\top \boldsymbol{\phi}_{\boldsymbol{\vartheta}}(\boldsymbol{x}), \quad k \in [K-1], \qquad Q_{f_{\boldsymbol{\theta}}}(k \mid \boldsymbol{x}) = \frac{\exp(f_{\boldsymbol{\theta},k}(\boldsymbol{x}))}{\sum_{v=1}^{K} \exp(f_{\boldsymbol{\theta},v}(\boldsymbol{x}))}.$$

This restricted attention classifier has four attention parameters and twelve classification coefficients. The common parameter domain Θ is specified by

$$a_1 \in [0, 2.5], \qquad a_2 \in [-2.5, 0], \qquad b_1, b_2 \in [-2.5, 2.5],$$
$$\beta_{kq} \in [-30, 30] \qquad (k \in [K-1],\ q \in [3]).$$

For each clean observation, independently draw

$$G \sim \mathcal{N}(0, 0.9^2), \qquad \mathbb{P}(M = -1.5) = \mathbb{P}(M = 1.5) = \tfrac{1}{4}, \qquad \mathbb{P}(M = 0) = \tfrac{1}{2},$$

and generate the coordinates by

$$X_j = G + M + 0.65 r_j + \xi_j, \qquad \xi_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.65^2), \quad j \in [D].$$

The variables G, M, and $\xi_1, \dots, \xi_D$ are mutually independent. The shared mixture shift controls the sequence level, while the positional trend and coordinate noises induce variation within a sequence.

The true attention and classification parameters are

$$\boldsymbol{\vartheta}^\star = (1.6, 0.7, -1.3, -0.6)^\top, \qquad \mathbf{B}^\star = 1.5 \begin{pmatrix} -1.8 & -0.8 & -0.8 \\ -1.8 & 0.8 & 0.8 \\ 3.0 & -3.0 & 3.0 \\ -5.4 & 3.0 & -3.0 \end{pmatrix}.$$

Let $f^\star = f_{\boldsymbol{\theta}^\star}$ and $Q^\star = Q_{f^\star}$. With $z_h^\star(\boldsymbol{x}) = z_h(\boldsymbol{x}; \boldsymbol{\vartheta}^\star)$, define

$$m(\boldsymbol{x}) = \frac{z_1^\star(\boldsymbol{x}) + z_2^\star(\boldsymbol{x})}{2}, \qquad u(\boldsymbol{x}) = z_1^\star(\boldsymbol{x}) - z_2^\star(\boldsymbol{x}).$$

The true logit function can equivalently be written as

$$f^\star(\boldsymbol{x}) = \begin{pmatrix} -2.7 - 2.4m(\boldsymbol{x}) \\ -2.7 + 2.4m(\boldsymbol{x}) \\ 4.5 - 4.5u(\boldsymbol{x}) \\ -8.1 + 4.5u(\boldsymbol{x}) \\ 0 \end{pmatrix}.$$

These logits favor low and high levels, small and large attention-summary contrasts, and an intermediate reference class. The contrast $u(\boldsymbol{x})$ is the difference between the two attention summaries, rather than the raw range or variance of the tokens. Labels are sampled according to

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{Categorical}(Q^\star(\cdot \mid \boldsymbol{x})).$$

Thus the clean model is correctly specified, since $\boldsymbol{\theta}^\star \in \Theta$, while the labels remain stochastic. The true parameters are used for data generation and evaluation, not for fitting.

B.3.2 Replacement contamination

Each replication begins with $n = 500$ clean training observations and an independent clean test sample of size $N = 10{,}000$. For

$$\varepsilon \in \{0, 0.01, 0.02, 0.03, 0.05, 0.10\}, \qquad c_\varepsilon = \mathrm{round}(n\varepsilon),$$

the replacement counts are $0, 5, 10, 15, 25, 50$. Retain the first $n - c_\varepsilon$ clean observations and replace the final c_ε observations by the following pairs, indexed by $t = 0, \ldots, c_\varepsilon - 1$:

$$\boldsymbol{x}_t^{\mathrm{out}} = 20\mathbf{1}_D + (2 \times 10^{-6}) t\, \boldsymbol{e}_{1+(t \bmod D)},$$
$$y_t^{\mathrm{out}} = \min \arg\min_{k \in [K]} Q^\star(k \mid \boldsymbol{x}_t^{\mathrm{out}}).$$

Here $\boldsymbol{e}_j$ is the j-th coordinate vector, and $\mathbf{1}_D$ is the vector of ones. The small deterministic perturbation reproduces the implementation; the label rule assigns the least likely class and breaks ties by the smallest index. The outlier magnitude is 20. No additional random outlier sampling or randomization of the replacement indices is performed.

The training size remains 500. All methods receive the same contaminated sample at a given level, and all levels share the same underlying clean sample and test set within a replication. At $\varepsilon = 0$, the clean fit is reused. Evaluation always uses uncontaminated test observations.

B.3.3 Replication, evaluation, and reporting

We use $N_{\mathrm{rep}} = 100$ replications, each with a distinct base seed to generate the clean training inputs and labels, followed by the clean test inputs and labels.

Let $\widehat{f}$ be a fitted logit function and let $\widehat{g}(\boldsymbol{x}) = \min \arg\max_{k \in [K]} Q_{\widehat{f}}(k \mid \boldsymbol{x})$ be its plug-in classifier. For the clean test sample $\{(\boldsymbol{X}_j^{\mathrm{test}}, Y_j^{\mathrm{test}})\}_{j=1}^N$, estimate the normalized Hellinger loss defined in the main text by

$$\begin{aligned} \widehat{\bar{h}}_N^2(Q^\star, Q_{\widehat{f}}) &= \frac{1}{N} \sum_{j=1}^N h^2\Big(Q^\star(\cdot \mid \boldsymbol{X}_j^{\mathrm{test}}), Q_{\widehat{f}}(\cdot \mid \boldsymbol{X}_j^{\mathrm{test}})\Big) \\ &= \frac{1}{N} \sum_{j=1}^N \left[1 - \sum_{k=1}^K \sqrt{Q^\star(k \mid \boldsymbol{X}_j^{\mathrm{test}}) Q_{\widehat{f}}(k \mid \boldsymbol{X}_j^{\mathrm{test}})}\right]. \end{aligned}$$

This is a Monte Carlo estimate of $\bar{h}^2(Q^\star, Q_{\widehat{f}})$.

Prediction error estimates the main-text classification risk $\mathcal{R}(\widehat{g})$:

$$\widehat{\mathcal{R}}_N(\widehat{g}) = \frac{1}{N} \sum_{j=1}^N \mathbf{1}_{\{\widehat{g}(\boldsymbol{X}_j^{\mathrm{test}}) \neq Y_j^{\mathrm{test}}\}}.$$

This quantity compares predictions with sampled labels and includes the irreducible classification error. Prediction error is displayed as a percentage. Detailed results are shown in Table 2.

Table 2: Squared Hellinger distance and prediction error in the five-class synthetic experiment over 100 repetitions. Squared Hellinger distances are reported on their original scale, and prediction errors as percentages. Within each contamination level, the lowest value in each column is shown in bold.

Contamination	Method	Squared Hellinger distance			Prediction error (%)		
		Mean	$Q_{0.25}$	$Q_{0.75}$	Mean	$Q_{0.25}$	$Q_{0.75}$
Clean (0%)	MLE	**0.0043**	**0.0033**	**0.0051**	**31.44**	31.02	**31.76**
	ρ	**0.0043**	**0.0033**	**0.0051**	**31.44**	**31.01**	31.78
	Huber ($\delta = 0.5$)	0.0280	0.0247	0.0323	31.55	31.13	31.97
	Huber ($\delta = 1$)	0.0153	0.0127	0.0180	31.52	31.14	31.88
	Huber ($\delta = 2$)	0.0050	0.0039	0.0063	31.47	31.03	31.88
	Huber ($\delta = 4$)	**0.0043**	**0.0033**	**0.0051**	**31.44**	31.02	**31.76**
	MOM ($B = 5$)	0.0166	0.0118	0.0202	33.16	32.23	33.70
	MOM ($B = 10$)	0.0186	0.0134	0.0213	33.41	32.43	34.17
	MOM ($B = 20$)	0.0293	0.0232	0.0348	34.88	33.69	35.74
	MOM ($B = 50$)	0.0404	0.0319	0.0494	36.65	34.63	38.37
1%	MLE	0.0873	0.0840	0.0911	40.05	37.60	42.05
	ρ	0.0133	**0.0043**	0.0223	32.03	31.44	32.54
	Huber ($\delta = 0.5$)	0.0170	0.0139	0.0200	31.81	**31.38**	**32.13**
	Huber ($\delta = 1$)	**0.0095**	0.0074	**0.0111**	**31.77**	**31.38**	32.16
	Huber ($\delta = 2$)	0.0111	0.0089	0.0126	31.98	31.55	32.36
	Huber ($\delta = 4$)	0.0401	0.0346	0.0437	34.00	33.18	34.66
	MOM ($B = 5$)	0.0891	0.0831	0.1010	41.22	39.11	43.48
	MOM ($B = 10$)	0.0484	0.0160	0.0911	36.69	33.12	41.16
	MOM ($B = 20$)	0.0288	0.0203	0.0336	34.95	33.59	35.95
	MOM ($B = 50$)	0.0418	0.0283	0.0518	36.92	34.48	38.49
2%	MLE	0.1251	0.1229	0.1271	46.82	45.77	47.94
	ρ	0.0385	0.0325	0.0439	33.85	33.00	34.55
	Huber ($\delta = 0.5$)	**0.0198**	0.0140	**0.0211**	**32.88**	32.07	**33.15**
	Huber ($\delta = 1$)	0.0357	**0.0123**	0.0216	34.87	**32.02**	33.33
	Huber ($\delta = 2$)	0.0763	0.0367	0.1225	39.37	33.55	46.42
	Huber ($\delta = 4$)	0.1251	0.1229	0.1271	46.82	45.77	47.94
	MOM ($B = 5$)	0.1261	0.1181	0.1363	47.07	44.01	49.52
	MOM ($B = 10$)	0.1227	0.1151	0.1331	46.10	43.74	48.54
	MOM ($B = 20$)	0.0455	0.0218	0.0389	36.74	33.55	36.14
	MOM ($B = 50$)	0.0419	0.0318	0.0512	36.77	34.74	38.51
3%	MLE	0.1292	0.1264	0.1317	47.60	46.43	48.66
	ρ	0.0486	0.0411	0.0548	**34.75**	**33.69**	**35.63**
	Huber ($\delta = 0.5$)	0.1323	0.1305	0.1361	46.17	44.61	48.04
	Huber ($\delta = 1$)	0.1278	0.1250	0.1303	46.19	44.41	48.02
	Huber ($\delta = 2$)	0.1284	0.1253	0.1309	47.32	46.10	48.44
	Huber ($\delta = 4$)	0.1292	0.1264	0.1317	47.60	46.43	48.66
	MOM ($B = 5$)	0.1374	0.1317	0.1433	49.49	47.04	50.76
	MOM ($B = 10$)	0.1384	0.1318	0.1445	49.39	46.92	50.52
	MOM ($B = 20$)	0.1315	0.1331	0.1538	48.24	46.08	51.25
	MOM ($B = 50$)	**0.0415**	**0.0331**	**0.0467**	36.50	34.81	37.57
5%	MLE	0.1328	0.1294	0.1358	48.34	47.02	49.43
	ρ	**0.0631**	0.0520	0.0713	**36.39**	**34.83**	**37.53**
	Huber ($\delta = 0.5$)	0.1355	0.1327	0.1383	46.57	45.17	48.15
	Huber ($\delta = 1$)	0.1298	0.1270	0.1321	46.40	44.90	48.05
	Huber ($\delta = 2$)	0.1316	0.1280	0.1344	47.90	46.74	48.85
	Huber ($\delta = 4$)	0.1328	0.1294	0.1358	48.34	47.02	49.43
	MOM ($B = 5$)	0.1424	0.1360	0.1483	50.29	48.61	51.37
	MOM ($B = 10$)	0.1435	0.1367	0.1484	50.66	48.47	52.27
	MOM ($B = 20$)	0.1521	0.1441	0.1576	52.27	49.58	55.26
	MOM ($B = 50$)	0.0666	**0.0402**	**0.0655**	39.17	35.29	39.81
10%	MLE	0.1368	0.1333	0.1399	49.09	47.39	50.31
	ρ	**0.1077**	**0.0749**	0.1372	**44.16**	**36.10**	49.33
	Huber ($\delta = 0.5$)	0.1379	0.1346	0.1410	46.85	45.56	48.51
	Huber ($\delta = 1$)	0.1325	0.1297	**0.1355**	46.71	45.35	**48.39**
	Huber ($\delta = 2$)	0.1354	0.1319	0.1384	48.52	47.08	49.85
	Huber ($\delta = 4$)	0.1368	0.1333	0.1399	49.09	47.39	50.31
	MOM ($B = 5$)	0.1481	0.1403	0.1545	51.56	48.74	53.46
	MOM ($B = 10$)	0.1475	0.1409	0.1520	51.44	49.11	52.41
	MOM ($B = 20$)	0.1573	0.1494	0.1639	52.74	49.99	54.86
	MOM ($B = 50$)	0.1683	0.1607	0.1758	53.33	50.61	56.19

B.4 Additional simulation results

We consider an additional setting with $K = 3$ classes, retaining the input distribution, $D = 5$ scalar tokens, and $H = 2$ attention heads from Section B.3.1. The fitted classifier has $\mathbf{B} \in \mathbb{R}^{2\times 3}$ and

$d_\theta = 10$, with reference logit $f_{\boldsymbol{\theta},3} = 0$ and the same coordinatewise parameter bounds. The true parameters are

$$\boldsymbol{\vartheta}^\star = (0.9, 0.6, -0.8, -0.5)^\top, \qquad \mathbf{B}^\star = \begin{pmatrix} -1.4 & 2.2 & 0.5 \\ -1.4 & -0.5 & -2.2 \end{pmatrix}.$$

Labels are sampled from the resulting three-class softmax distribution $Q^\star$. The training and test sample sizes remain $n = 500$ and $N = 10{,}000$. Competing objectives and tuning grids follow Section B.1, while contamination levels and the replacement rule follow Section B.3.2, with outlier labels determined by the least likely class under the three-class $Q^\star$. We run 100 replications with different random seeds and report means and empirical 25th–75th percentile intervals for results evaluated on clean test data. The results are shown as below.

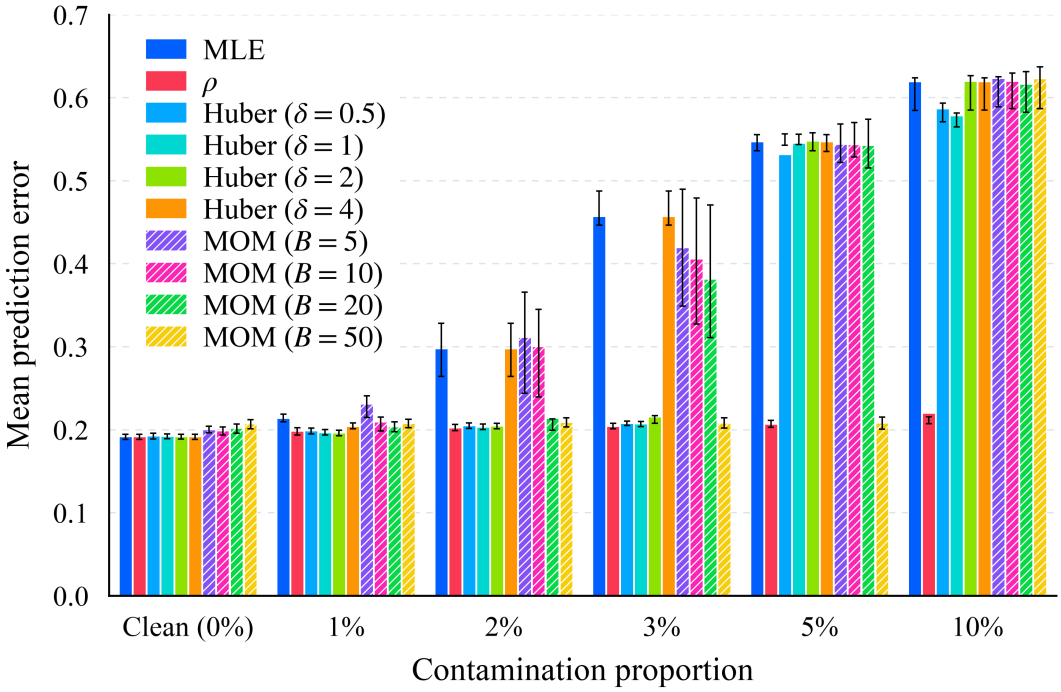

Figure 4: Prediction error on clean test data under varying proportions of high-leverage contamination. Bars show means over 100 replications, with error bars spanning the 25th–75th percentiles.

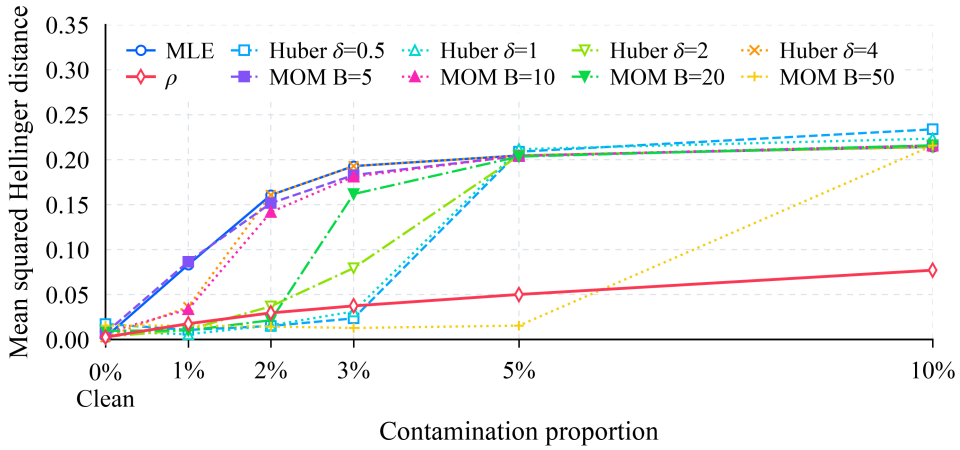

Figure 5: Squared Hellinger distance on clean test data under varying proportions of high-leverage contamination. Lines connect means over 100 replications.

Table 3: Squared Hellinger distance and prediction error in the three-class synthetic experiment over 100 repetitions. Squared Hellinger distances are reported on their original scale, and prediction errors as percentages. Within each contamination level, the lowest value in each column is shown in bold.

Contamination	Method	Squared Hellinger distance			Prediction error (%)		
		Mean	$Q_{0.25}$	$Q_{0.75}$	Mean	$Q_{0.25}$	$Q_{0.75}$
Clean (0%)	MLE	**0.0027**	**0.0017**	**0.0034**	**19.18**	**18.83**	19.47
	ρ	**0.0027**	**0.0017**	**0.0034**	**19.18**	18.85	19.47
	Huber ($\delta = 0.5$)	0.0174	0.0146	0.0202	19.28	18.90	19.57
	Huber ($\delta = 1$)	0.0092	0.0071	0.0110	19.23	18.93	19.52
	Huber ($\delta = 2$)	0.0031	0.0020	0.0038	19.19	18.90	**19.45**
	Huber ($\delta = 4$)	**0.0027**	**0.0017**	**0.0034**	**19.18**	**18.83**	19.47
	MOM ($B = 5$)	0.0073	0.0051	0.0091	20.01	19.64	20.42
	MOM ($B = 10$)	0.0076	0.0053	0.0096	19.91	19.36	20.33
	MOM ($B = 20$)	0.0098	0.0065	0.0117	20.19	19.60	20.67
	MOM ($B = 50$)	0.0156	0.0093	0.0193	20.74	20.12	21.24
1%	MLE	0.0832	0.0813	0.0845	21.39	20.97	21.88
	ρ	0.0177	0.0134	0.0229	19.86	19.39	20.27
	Huber ($\delta = 0.5$)	0.0107	0.0086	0.0125	19.91	19.49	20.22
	Huber ($\delta = 1$)	**0.0059**	**0.0044**	**0.0073**	19.70	19.35	20.04
	Huber ($\delta = 2$)	0.0110	0.0091	0.0127	**19.62**	**19.27**	**19.96**
	Huber ($\delta = 4$)	0.0363	0.0338	0.0386	20.44	20.11	20.82
	MOM ($B = 5$)	0.0866	0.0829	0.0919	23.12	21.47	24.11
	MOM ($B = 10$)	0.0335	0.0066	0.0810	20.93	19.87	21.54
	MOM ($B = 20$)	0.0101	0.0072	0.0130	20.39	19.76	20.97
	MOM ($B = 50$)	0.0156	0.0117	0.0183	20.80	20.23	21.28
2%	MLE	0.1607	0.1574	0.1635	29.79	26.41	32.85
	ρ	0.0297	0.0215	0.0369	**20.24**	**19.88**	**20.65**
	Huber ($\delta = 0.5$)	0.0151	0.0128	0.0172	20.49	20.18	20.81
	Huber ($\delta = 1$)	0.0158	0.0132	0.0181	20.35	20.02	20.69
	Huber ($\delta = 2$)	0.0370	0.0344	0.0391	20.45	20.10	20.79
	Huber ($\delta = 4$)	0.1607	0.1574	0.1635	29.79	26.41	32.85
	MOM ($B = 5$)	0.1516	0.1394	0.1704	31.14	24.39	36.58
	MOM ($B = 10$)	0.1417	0.1227	0.1628	30.02	23.98	34.52
	MOM ($B = 20$)	0.0215	**0.0074**	**0.0157**	21.38	19.95	21.30
	MOM ($B = 50$)	**0.0145**	0.0096	0.0183	20.90	20.33	21.44
3%	MLE	0.1929	0.1919	0.1941	45.70	44.65	48.74
	ρ	0.0374	0.0285	0.0472	**20.43**	**20.10**	**20.77**
	Huber ($\delta = 0.5$)	0.0236	0.0208	0.0259	20.76	20.51	21.03
	Huber ($\delta = 1$)	0.0313	0.0284	0.0336	20.75	20.40	21.00
	Huber ($\delta = 2$)	0.0795	0.0751	0.0818	21.58	20.80	21.71
	Huber ($\delta = 4$)	0.1929	0.1919	0.1941	45.70	44.65	48.74
	MOM ($B = 5$)	0.1829	0.1729	0.1955	41.96	34.91	48.97
	MOM ($B = 10$)	0.1810	0.1673	0.1949	40.62	32.76	47.91
	MOM ($B = 20$)	0.1618	0.1629	0.1906	38.15	31.09	47.10
	MOM ($B = 50$)	**0.0129**	**0.0093**	**0.0156**	20.81	20.20	21.45
5%	MLE	0.2046	0.2034	0.2057	54.73	53.60	55.58
	ρ	0.0502	0.0370	0.0602	**20.69**	20.30	**21.12**
	Huber ($\delta = 0.5$)	0.2088	0.2143	0.2219	53.15	54.27	55.64
	Huber ($\delta = 1$)	0.2116	0.2122	0.2164	54.54	54.36	55.58
	Huber ($\delta = 2$)	0.2048	0.2036	0.2060	54.80	53.60	55.77
	Huber ($\delta = 4$)	0.2046	0.2034	0.2057	54.71	53.55	55.56
	MOM ($B = 5$)	0.2040	0.2010	0.2088	54.41	52.20	56.85
	MOM ($B = 10$)	0.2044	0.2013	0.2080	54.38	52.87	57.01
	MOM ($B = 20$)	0.2034	0.2002	0.2092	54.26	51.53	57.40
	MOM ($B = 50$)	**0.0155**	**0.0090**	**0.0187**	20.83	**20.06**	21.51
10%	MLE	0.2142	0.2130	0.2155	61.94	58.49	62.39
	ρ	**0.0771**	**0.0571**	**0.0926**	**22.03**	**20.75**	**21.57**
	Huber ($\delta = 0.5$)	0.2337	0.2297	0.2371	58.68	57.10	59.34
	Huber ($\delta = 1$)	0.2234	0.2208	0.2256	57.84	56.47	58.16
	Huber ($\delta = 2$)	0.2144	0.2132	0.2157	62.00	58.49	62.68
	Huber ($\delta = 4$)	0.2142	0.2130	0.2155	61.94	58.51	62.39
	MOM ($B = 5$)	0.2157	0.2135	0.2178	62.36	58.90	62.55
	MOM ($B = 10$)	0.2157	0.2131	0.2176	61.98	58.69	62.99
	MOM ($B = 20$)	0.2159	0.2128	0.2192	61.64	58.26	63.14
	MOM ($B = 50$)	0.2156	0.2133	0.2186	62.29	58.70	63.72

As we can see in Figure 4, Figure 5 and Table 3, on clean data, the ρ-estimator closely matches MLE in both metrics. Its Hellinger error grows more gradually under contamination, reaching 0.077 at 10% contamination, compared with 0.214-0.234 for all competing methods and tuning settings. The advantage is especially clear for prediction error. The ρ-estimator achieves the lowest mean

error at every tested contamination level from 2% to 10%. Its error rises only from 19.2% on clean data to 22.0% at 10% contamination, whereas competitors reach 57.8%–62.4%. Even MOM with $B = 50$, which remains competitive at 5% contamination, deteriorates sharply at 10%. These results demonstrate strong predictive performance across the tested contamination levels, consistent with the theoretical guarantees of accuracy under correct specification and robustness to contamination.

B.5 Real-world datasets

The Vowel dataset contains 990 observations with 10 acoustic features for classifying 11 vowel sounds (https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#vowel). The Letter dataset contains 20,000 observations with 16 numerical features extracted from character images, covering the 26 uppercase English letters (https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#letter). We use the scaled versions available from the LIBSVM repository.

We compare ρ-estimation with MLE, Huber, and MOM using the same Transformer architecture, with two self-attention layers, four attention heads, an embedding dimension of 48, and a feedforward width of 96. Preliminary experiments on these two datasets showed limited sensitivity to the Huber and MOM hyperparameters within the ranges considered; we therefore use $c = 1$ and 10 blocks, respectively, as representative settings. For Vowel, we use 623 training observations and 297 test observations. The data split and contamination are regenerated for each of the 100 repetitions. For Letter, each repetition independently samples 1,040 training observations and 1,040 test observations, also over 100 repetitions. For both datasets, distinct random seeds are used across repetitions. At contamination levels of 0%, 5%, 10%, 20%, and 30%, each feature of a selected training observation is independently replaced by -50 or 50 with equal probability, and its label is reassigned uniformly among the incorrect classes. No clipping or rescaling is applied after contamination, and the test data remain clean. Within each repetition, all methods use the same data and contamination. We report the means and empirical quartiles of test misclassification error and fitting time over 100 repetitions. Fitting time excludes final evaluation and file saving. The results are presented in Table 4, Table 5 and Figure 6.

On both Vowel and Letter, ρ-estimation performs comparably to MLE on clean data and achieves the lowest mean test misclassification error among the four methods at every positive contamination level considered. These results support its robustness to joint feature and label contamination. This improvement comes with additional computational cost: its recorded mean fitting time is approximately 3.5-4.9 times that of MLE on Vowel and 4.4-5.2 times on Letter under our experimental settings.

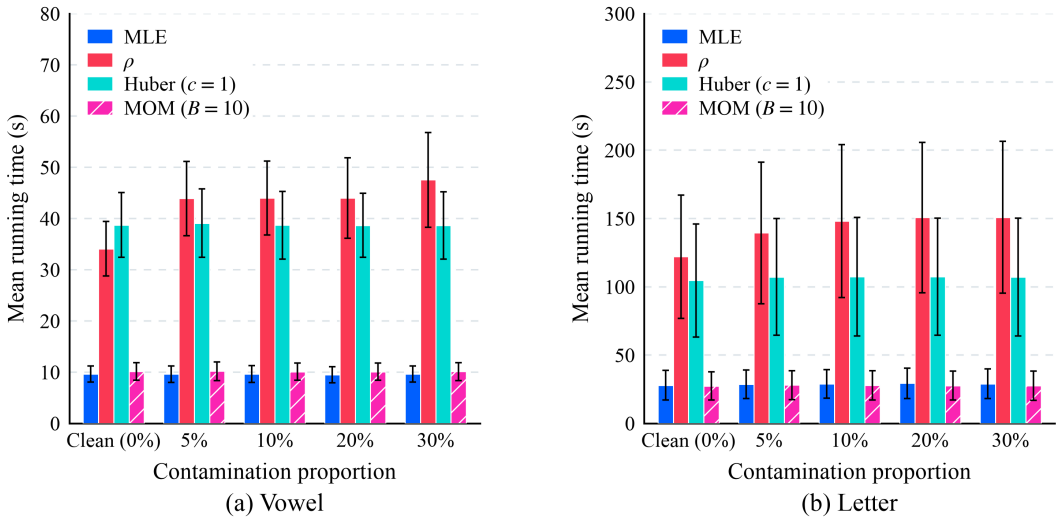

Figure 6: Mean running time on Vowel and Letter at different training-data contamination levels. Error bars indicate standard deviations over 100 runs.

Table 4: Prediction error and running time on the Vowel dataset over 100 runs. Within each contamination level, the lowest value in each prediction-error column is shown in bold.

Contamination	Method	Prediction error (%)			Running time (s)		
		Mean	$Q_{0.25}$	$Q_{0.75}$	Mean	$Q_{0.25}$	$Q_{0.75}$
Clean (0%)	MLE	6.94	5.72	**8.08**	9.63	8.63	10.24
	ρ	7.15	5.98	**8.08**	34.10	31.05	36.03
	Huber ($c = 1$)	**6.91**	**5.39**	8.16	38.72	34.94	41.43
	MOM ($B = 10$)	8.03	6.73	9.09	10.12	9.16	10.90
5%	MLE	8.21	7.07	9.43	9.62	8.69	10.24
	ρ	**7.76**	**6.40**	**9.09**	43.91	39.49	46.72
	Huber ($c = 1$)	8.30	6.99	9.76	39.11	35.39	42.22
	MOM ($B = 10$)	8.69	7.41	9.85	10.19	9.14	10.85
10%	MLE	9.16	7.74	10.77	9.65	8.78	10.34
	ρ	**8.83**	**7.41**	**10.10**	43.98	40.23	46.50
	Huber ($c = 1$)	9.24	8.08	10.19	38.70	34.67	41.22
	MOM ($B = 10$)	9.25	7.74	10.52	10.08	9.07	10.79
20%	MLE	11.43	10.10	12.54	9.51	8.59	10.14
	ρ	**10.68**	**9.09**	**12.12**	44.03	39.40	47.57
	Huber ($c = 1$)	11.69	10.10	13.13	38.65	35.14	40.85
	MOM ($B = 10$)	11.09	9.43	12.79	10.10	9.13	10.81
30%	MLE	13.85	12.12	15.24	9.62	8.77	10.31
	ρ	**12.96**	**11.45**	**14.14**	47.56	41.02	52.78
	Huber ($c = 1$)	14.45	13.05	15.82	38.66	34.41	41.60
	MOM ($B = 10$)	13.11	**11.45**	14.56	10.13	9.04	11.01

Table 5: Prediction error and running time on the 26-class Letter dataset over 100 runs. Within each contamination level, the lowest value in each prediction-error column is shown in bold.

Contamination	Method	Prediction error (%)			Running time (s)		
		Mean	$Q_{0.25}$	$Q_{0.75}$	Mean	$Q_{0.25}$	$Q_{0.75}$
Clean (0%)	MLE	21.36	20.34	22.50	27.97	18.16	38.47
	ρ	**20.83**	**19.69**	**22.12**	122.11	79.55	168.66
	Huber ($c = 1$)	21.95	20.87	22.98	104.64	64.55	146.39
	MOM ($B = 10$)	27.07	25.70	28.39	27.36	17.44	37.64
5%	MLE	22.41	21.42	23.58	28.58	18.38	39.23
	ρ	**21.67**	**20.48**	**22.69**	139.46	95.11	187.71
	Huber ($c = 1$)	23.11	21.83	23.94	107.28	65.56	148.98
	MOM ($B = 10$)	27.42	26.25	28.68	28.03	17.57	37.55
10%	MLE	23.25	22.31	24.33	28.98	18.75	38.62
	ρ	**22.45**	**21.42**	**23.58**	148.19	96.93	198.67
	Huber ($c = 1$)	24.13	23.08	25.00	107.48	64.80	149.18
	MOM ($B = 10$)	27.95	26.71	29.06	27.94	17.35	38.11
20%	MLE	25.33	24.18	26.63	29.35	18.40	39.08
	ρ	**24.32**	**23.15**	**25.41**	150.71	96.41	199.31
	Huber ($c = 1$)	26.16	25.10	27.12	107.34	64.69	148.93
	MOM ($B = 10$)	28.77	27.38	29.93	27.74	17.21	37.31
30%	MLE	27.77	26.73	28.85	29.07	18.39	38.27
	ρ	**26.39**	**25.10**	**27.50**	150.96	96.62	202.21
	Huber ($c = 1$)	28.27	26.83	29.52	107.12	64.16	147.42
	MOM ($B = 10$)	29.93	28.75	31.15	27.53	17.09	38.10