DIMENSION-FREE MINIMAX RATES FOR LEARNING PAIRWISE INTERACTIONS IN ATTENTION-STYLE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the convergence rate of learning pairwise interactions in single-layer attention-style models, where tokens interact through a weight matrix and a non-linear activation function. We prove that the minimax rate is $M^{-\frac{2\beta}{2\beta+1}}$ with M being the sample size, depending only on the smoothness β of the activation, and crucially independent of token count, ambient dimension, or rank of the weight matrix. These results highlight a fundamental dimension-free statistical efficiency of attention-style nonlocal models, even when the weight matrix and activation are not separately identifiable and provide a theoretical understanding of the attention mechanism and its training.

1 INTRODUCTION

The transformer architecture (Vaswani et al., 2017) has achieved remarkable success in natural language processing, computer vision, and other AI domains, with its impact most visible in large language models (LLMs) such as GPT (OpenAI, 2024), LLaMA (Touvron et al., 2023), and BERT (Devlin et al., 2019). At its core, attention mechanisms model nonlocal dependencies between input tokens through pairwise interactions, creating a function class capable of representing intricate contextual relationships.

Despite the empirical success, our theoretical understanding remains incomplete. The attention mechanism computes weighted averages of token representations using pairwise similarities, but we observe only the aggregated outputs and not the underlying interaction structure that generates them. This creates a fundamental inverse problem with critical *sample complexity* questions: can we recover the interaction function from these aggregated observations, how many samples are needed to learn token-to-token interactions for a given accuracy level, and how does the convergence rate depend on embedding dimension, number of tokens, and smoothness of the activation function? Recent phenomena like extreme attention weights on certain tokens (Sun et al., 2024; Guo et al., 2024b; Xiao et al., 2024; Wang et al., 2021) further highlight gaps in our understanding of how transformers process token interactions.

In this paper, we tackle these questions by analyzing an Interacting Particle System (IPS) model for attention-style mechanisms. Tokens are viewed as "particles," and the self-attention aggregates pairwise interactions between them. The interaction is a composite of an unknown embedding matrix and an unknown nonlinear activation function, both are learned from data. This makes the problem challenging as it is fundamentally *nonconvex*. Our IPS approach provides a natural framework for understanding how transformers process inputs with a large number of correlated tokens, moving beyond the restrictive assumption of independent, isotropic token distributions.

We summarize our main contribution below:

- We establish a connection between transformers and IPS models, enabling us to address the challenging inverse problem of inferring nonlinear interactions learned by attention mechanisms. Our analysis extends beyond the standard assumption of independent, isotropic token distributions to allow for dependent and anisotropic data.

- Inferring the interaction function is an inverse problem. We prove that under a *coercivity condition* (Lemma 3.4), this problem is well-posed in the large sample limit. This condition holds for a large class of input distributions.

- We prove that the rate of $M^{-\frac{2\beta}{2\beta+1}}$ is the optimal (up to logarithmic factors) minimax convergence rate in estimating the $2d$-dimensional pairwise interaction function where M is the sample size and β is the Hölder exponent of the function. Importantly, this rate is independent of the embedding dimension d and the weight matrix rank. This dimension-free rate stems from the model's intrinsic structure of a scalar activation on a bilinear form, which reduces the sample complexity of the problem from learning a $2d$ dimensional interaction function to a scalar function applied to a 1D bilinear form. The rate $M^{-\frac{2\beta}{2\beta+1}}$ is precisely the optimal rate for this 1D estimation problem, confirming that the attention-style model evades the curse of dimensionality.

1.1 Related works

Neural networks and IPS. Modeling neural networks as dynamical systems through depth was introduced in Chen et al. (2018), which framed updates in ResNet architectures as the dynamics of a state vector. This perspective has been generalized to various architectures, typically treating skip connections as the evolving state across layers. Following this approach, in Geshkovski et al. (2023; 2025) they view tokens as interacting particles, analyze the attention as an IPS, and study clustering phenomena in continuous time (in depth). Similarly, Dutta et al. (2021) leverages a similar framework to compute attention outputs directly from an initial state evolved over depth, thereby reducing computational costs. While these works provide valuable insights, they focus exclusively on the dynamics of tokens through the layers. To our knowledge, no existing work addresses the learning theory for estimating the pairwise interactions in such particle systems.

Inference in attention models. Many theoretical works have studied the learnability of attention, focusing on specific regimes. Some consider simplified variants, such as linear or random feature target attention models (Wang et al., 2020; Lu et al., 2025; Marion et al., 2025; Hron et al., 2020; Fu et al., 2023), which explore the capability of this model under simple regression tasks. Deora et al. (2024) analyze logistic-loss optimization and prove a generalization rate under a "good" initialization. Others consider a more specific architecture, Li et al. (2023) study the training of shallow vision transformers (ViT) and show that, with suitable initialization and enough stochastic gradient steps, a transformer with additional ReLU layer can achieve zero error. Several works study softmax attention layers with trainable key and query matrices in the limit of high embedding dimension quadratically proportionate to samples with i.i.d. tokens Troiani et al. (2025); Cui et al. (2024); Cui (2025); Boncoraglio et al. (2025), which is further expanded in Troiani et al. (2025) for $\mathrm{softmax}$ attention (without the value matrix) with multiple layers. These works mainly focused on the linear/softmax attention model and do not consider a general interaction function. In addition, most studies assume the tokens are independent and do not draw the connection to the IPS system.

Inference for systems of interacting particles. There is a large body of work on the inference of systems of interacting particles; we state a few here. Parametric inference has been studied in Amorino et al. (2023); Chen (2021); Della Maestra & Hoffmann (2023); Kasonga (1990); Liu & Qiao (2022); Sharrock et al. (2021) for the operator (drift term) and in Huang et al. (2019) for the noise variance (diffusion term). Nonparametric inference on estimating the entire operator R_g, but not the kernel g, has been studied in Della Maestra & Hoffmann (2022); Yao et al. (2022). The closest to this study are Lu et al. (2021a; 2022; 2019); Wang et al. (2025). A key difference from these studies is that their goal is to estimate the radial interaction kernel, whereas our $2d$-dimensional pairwise interaction function is not shift-invariant due to the weight matrix. In addition, all these studies focus on IPS in general, without a clear connection to attention models.

Activation function in transformer layer. Recent work has shown that attention models suffer from the "extreme-token phenomenon", where certain tokens receive disproportionately high weights, creating challenges for downstream tasks (Sun et al., 2024; Guo et al., 2024b; Xiao et al., 2024; Wang et al., 2021). To address this, it was proposed to replace softmax with alternatives, such as ReLU (Guo et al., 2024a; Zhang et al., 2021), which can "turn off" irrelevant tokens, a capability that softmax lacks. While linear attention can outperform softmax in regression tasks by avoiding additional error offsets (Von Oswald et al., 2023; Katharopoulos et al., 2020; Yu et al., 2024;

Han et al., 2024), it may be inferior for classification (Oymak et al., 2023). These findings suggest no universally optimal activation function exists, making the theoretical analysis of a general interaction function in transformer-type models crucial. As for vision tasks, several Vision Transformer (ViT) variants remove the softmax activation while remaining competitive. For example, Lu et al. (2021b) consider an attention mechanism based on a Gaussian kernel, and Koohpayegani & Pirsiavash (2024) apply linear attention after normalizing the Key-Query columns. Furthermore, Ramapuram et al. (2025) examine a sigmoid function as the attention activation, showing it acts as a universal function approximator and benefits from improved regularity compared to softmax attention.

Nonparametric and Semiparametric Estimation for Neural Networks Classical nonparametric estimation provides optimal minimax rates for simple structures. Gaıffas & Lecué (2007) provide bounds for the single index model $f(w^\top x)$ of order $M^{\frac{-2\beta}{2\beta+1}}$. For the more general projection pursuit model $f(x) = \sum_{j=1}^{K} f_j(\langle x, \beta_j \rangle)$, Györfi et al. (2006) shows that the minimax rate is the standard rate up to a log factor. These results directly apply to small single-layer neural networks.

Closer to deep learning, Horowitz & Mammen (2007) analyze generalized additive models with nested k-times differentiable compositions, showing the rate is $M^{-\frac{2k}{2k+1}}$. Schmidt-Hieber (2020) proves that connected deep ReLU networks achieve a near-optimal minimax rate (up to log factors) over a class of composed functions. In Bhattacharya et al. (2024) they study a nonparametric interaction model in high dimension settings and show sparsity assumptions and associated regularization are required in order to obtain optimal rates of convergence.

Notation. Throughout the paper, we use C to denote universal constants independent of the sample size M, particles N and the embedding dimension d, r. The notations C_β or $C_{\beta,L}$ denote constants depending on the subscripts. We introduce the L^2_ρ inner product as $\langle f, g\rangle_{L^2_\rho} = \int f(r)g(r)\rho(dr)$ and denote the L^p_ρ norm by $\|f\|^p_{L^p_\rho} = \int |f(r)|^p \rho(dr)$ for all $p \geqslant 1$. For vectors $a, b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ we write $\langle a, b\rangle_A := a^\top A b$.

2 Problem Formulation

In this section, we describe our statistical task and connect it to the attention model.

Model setup and learning task. We consider a model of N interacting particles,

$$Y_i = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \phi_\star\big(\langle X_i, X_j\rangle_{A_\star}\big) + \eta_i \tag{2.1}$$

where $\eta \in \mathbb{R}^N$ is noise as specified in Assumption 2.2, $\phi_\star : \mathbb{R} \to \mathbb{R}$ is an unknown interaction kernel, and $A_\star \in \mathbb{R}^{d\times d}$ is an unknown interaction matrix. Here, we write $\langle x, y\rangle_A := x^\top A y$ for $x, y \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$. The input $X = (X_1, \ldots, X_N)^\top \in \mathcal{C}^N_d := ([0,1]^d/\sqrt{d})^N \subset \mathbb{R}^{N\times d}$ denotes the particle positions (or token values), and the output $Y = (Y_1, \ldots, Y_N) \in \mathbb{R}^{N\times 1}$ represents the average interactions between the particles.

We observe M i.i.d. samples

$$\mathcal{D}_M = \{(X^m, Y^m)\}_{m=1}^M, \qquad X^m \in \mathcal{C}^N_d := ([0,1]^d/\sqrt{d})^N, Y^m \in \mathbb{R}^N,$$

allowing the N particles and their entries to be dependent. The task is to learn the pairwise interaction function $g_\star : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$g_\star(x, y) := \phi_\star\big(\langle x, y\rangle_{A_\star}\big), \qquad (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \tag{2.2}$$

from the dataset of observations $\mathcal{D}_M$. We introduce the vectorized view of the model via the forward operator R_g for any candidate interaction function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $R_g[X]_i := \frac{1}{N-1}\sum_{j=1, j\neq i}^{N} g(X_i, X_j)$. Accordingly, our model in equation 2.1 becomes $Y_i = R_{g_\star}[X]_i + \eta_i$.

Connection to self-attention layer. We view self-attention through the lens of an IPS: tokens are "particles," and attention aggregates pairwise interactions between them. A typical self-attention layer is composed of an attention block with learnable query, key, and value matrices, $W_Q, W_K \in \mathbb{R}^{d\times d_k}$ with $d_k \leqslant d$ and $W_V \in \mathbb{R}^{d\times d_v}$ that compute

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \qquad Q = XW_Q,\ K = XW_K,\ V = XW_V. \tag{2.3}$$

The attention operation is then often followed by an application of a multilayer perceptron (MLP), which maps the above into some other nonlinear function. The pairwise structure of attention motivates modeling token interactions via a scalar kernel function applied to a bilinear form of some score interaction matrix $A_\star$ that can be viewed as the learned projections through $\frac{1}{\sqrt{d_k}}W_Q W_K^\top$, i.e.,

$$\mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big) = \mathrm{softmax}\Big(XA_\star X^\top\Big), \quad XA_\star X^\top = \big(X_i^\top \tfrac{W_Q W_K^\top}{\sqrt{d_k}} X_j\big)_{1\leqslant i,j\leqslant N}.$$

The interaction function in equation 2.2 can be interpreted as either a function induced by the MLP and softmax function, or as a general activation function with a constant value matrix, see mode details in Appendix A. As stated in the related work, such a setup for a general activation function is often desirable due to the extreme-token phenomenon (Sun et al., 2024; Guo et al., 2024b; Xiao et al., 2024; Wang et al., 2021)

Consequently, the problem of estimating $g_\star$ from the samples described in equation 2.1 is analogous to the joint estimation of the activation function and weight matrix governing nonlocal token–token interactions in a single-layer self-attention mechanism.

Goal of this study. Our goal is to characterize the optimal (minimax) convergence rate of estimators of $g_\star$ as the sample size M grows.

To assess the estimation error for the interaction function, we introduce empirical measures over pairs of particles/tokens (x, y). Termed *exploration measures*, they quantify the extent to which the data explores the argument space relevant to the function.

Definition 2.1 (Exploration measure) *Let $\{X^m \in \mathcal{C}_d\}_{m=1}^M$ be sampled sequence. Define the* empirical exploration measure *of off-diagonal pairs of particles*

$$\rho_M(B) := \frac{1}{MN(N-1)} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \mathbf{1}_{\{(X_i^m, X_j^m)\in B\}}$$

and the population exploration measure *as $\rho(B) := \lim_{M\to\infty} \rho_M(B) = \mathbb{E}[\rho_M(B)]$, for any Lebesgue measurable set $B \subset \mathbb{R}^d \times \mathbb{R}^d$.*

We aim to provide matching upper and lower bound rates for the L^2_ρ error of the estimator $\widehat{g}$, so as to obtain a minimax convergence rate:

$$\mathbb{E}\Big[\|\widehat{g} - g_\star\|^2_{L^2_\rho}\Big] \asymp M^{-\frac{2\beta}{2\beta+1}}, \quad \text{as } M \to \infty, \tag{2.4}$$

where β is the Hölder exponent of $g_\star$ (which is determined by the smoothness of $\phi_\star$). This then demonstrates that the attention model is not susceptible to the curse of dimensionality. In particular, we aim to characterize the dependence of the rate on the embedding dimension d, the rank r of the interaction matrix, and the number of tokens N.

2.1 Assumptions on the data distribution

We now state the assumptions on the distributions of the input and the noise used throughout this work. We do not assume that the N tokens are independent of each other.

Assumption 2.1 (Data Distribution) *We assume the entries of the $\mathcal{C}_d^N = ([0,1]^d/\sqrt{d})^N$-valued random variable $X = (X_1, \ldots, X_N)$ satisfy the following conditions:*

(A1) *The components of the random vector $X = (X_1, \ldots, X_N)$ are exchangeable.*

(A2) *The joint distribution of (X_i, X_j) has a continuous density function for each pair.*

These assumptions simplify the inverse problem and may be replaced by weaker constraints; see Wang et al. (2025) for a discussion and references therein. The exchangeability in (A1) simplifies the exploration measure in Lemma B.1. It enables the coercivity condition for the inverse problem to be well-posed, as detailed in Lemma 3.4, and is only used in the upper bound in Theorem 3.1. The continuity in Assumption (A2) ensures that the exploration measure has a continuous density, which is used in proving the lower minimax rate Theorem 4.4.

We next specify the noise setting. Assumption 2.2 details the constraints we assume for the noise:

Assumption 2.2 (Noise Distribution) *The noise $\eta \in \mathbb{R}^N$ is centered and independent of the random array X. Moreover, we assume the following conditions:*

(B1) *The entries of the noise vector $\eta = (\eta_1, \dots, \eta_N)$ are sub-Gaussian in the sense that for all i, $\mathbb{E}[e^{c\eta_i^2}] < \infty$ for some $c > 0$.*

(B2) *There exists a constant $c_\eta > 0$ such that The density p_η of η satisfies the following:*

$$\int_{\mathbb{R}^N} p_\eta(u) \log \frac{p_\eta(u)}{p_\eta(u+v)} du \leqslant c_\eta \|v\|^2, \quad \forall v \in \mathbb{R}^N. \tag{2.5}$$

We note that assumptions (B1) and (B2) hold for instance for Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I_N)$ with $c_\eta = 1/(2\sigma_\eta^2)$.

2.2 Function classes and model/estimator assumptions

We introduce the functional classes where $g_\star$ lies. Our goal is to consider as large a class of functions as possible while also tracking the properties of the models $\phi_\star$ that control the rate. For that purpose, we introduce the Hölder class and assume that $\phi_\star$ satisfies some smoothness order of β.

Definition 2.2 (Hölder classes) *For $\beta, L, \bar{a} > 0$, the Hölder class $\mathcal{C}^\beta(L, \bar{a})$ on $[-\bar{a}, \bar{a}]$ is given by*

$$\mathcal{C}^\beta(L, \bar{a}) = \Big\{ f : [-\bar{a}, \bar{a}] \to \mathbb{R} : |f^{(l)}(x) - f^{(l)}(y)| \leqslant L|x-y|^{\beta - l}, \forall x, y \in [-\bar{a}, \bar{a}] \Big\}, \tag{2.6}$$

where $f^{(j)}$ denotes the j-th order derivative of functions f and $l = \lfloor \beta \rfloor$.

Low-rank Key and Query matrices often play an important role in the attention model. To keep track of the effects of the rank on the minimax rate, we introduce the following matrix class for the interaction matrix $A_\star$, which is the product of the Key and Query matrices.

Definition 2.3 (Interaction matrix class) *For $\bar{a} > 0$, the d-dimensional matrix class $\mathcal{A}_d(r, \bar{a})$ with rank $r \in \mathbb{N}$ and $2 \leqslant r \leqslant d$ is given by*

$$\mathcal{A}_d(r, \bar{a}) = \{ A \in \mathbb{R}^{d \times d} : 2 \leqslant \operatorname{rank}(A) \leqslant r\ ,\ \|A\|_{\text{op}} \leqslant \bar{a} \}\,. \tag{2.7}$$

Combining both classes, we consider the following function class $\mathcal{G}_r^\beta$ for all the possible pair-wise interaction functions.

Definition 2.4 (Target function class) *Given $L, B_\phi, \bar{a} > 0$ and rank $r \geqslant 2$, $\beta > 0$ define*

$$\mathcal{G}_r^\beta(L, B_\phi, \bar{a}) = \Big\{ g_{\phi,A}(x,y) := \phi(x^\top A y)\ : \phi \in \mathcal{C}^\beta(L, \bar{a}),\ \|\phi\|_\infty \leqslant B_\phi, A \in \mathcal{A}_d(r, \bar{a}) \Big\}. \tag{2.8}$$

For any $g \in \mathcal{G}_r^\beta = \mathcal{G}_r^\beta(L, B_\phi, \bar{a})$, moreover it follows $|R_g[X]_i| \leqslant B_\phi$. For technical reasons we requires $L \leqslant B_\phi (2\bar{a})^\beta$. This holds without loss of generality for any $\bar{a} \geqslant 1$ and $L \leqslant B_\phi$.

We provide both lower and upper bounds for the possible error rate by the number of samples for the interaction $g(\cdot, \cdot) \in \mathcal{G}_r^\beta$. We consider the following functional class for our estimator:

Definition 2.5 (Estimator function class) *Let $s := \max(\lfloor \beta \rfloor, 1)$ and $K_M \in \mathbb{N}$. Let $\Phi_{K_M}^s$ denote the class of piecewise polynomials of degree s, defined on K_M equal sub-intervals of $[-\bar{a}, \bar{a}]$. The corresponding* estimator *model class is*

$$\mathcal{G}_{r,K_M}^s := \Big\{ g_{\phi,A} :\ \phi \in \Phi_{K_M}^s,\ \|\phi\|_\infty + \|\phi'\|_\infty \leqslant B_\phi,\ A \in \mathcal{A}_d(r, \bar{a}) \Big\}\ \subseteq\ \mathcal{G}_r^\beta. \tag{2.9}$$

3 Upper bound

In this section, we provide an upper bound on estimating the token-token interaction. We propose the following estimator $\widehat{g}_M(x,y) = \hat{\phi}(\langle x,y\rangle_{\hat{A}})$ as the empirical risk minimizer over the functional class 2.5

$$\begin{cases} \widehat{g}_M = \underset{g_{\phi,A}\in\mathcal{G}^s_{r,K_M}}{\arg\min}\ \mathcal{E}_M(g_{\phi,A}) := \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}^{(i)}_M(g_{\phi,A}) \quad \text{with} \\ \mathcal{E}^{(i)}_M(g_{\phi,A}) := \frac{1}{M}\sum_{m=1}^{M}\|Y^m_i - R_{g_{\phi,A}}[X^m]_i\|^2. \end{cases} \tag{3.1}$$

Here, $R_{g_{\phi,A}}[X]_i = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} g_{\phi,A}(X_i,X_j)$ the forward operator with interaction function $g_{\phi,A}$. Our goal is to prove that the estimator $\widehat{g}_M$ achieves the optimal upper bound. The large sample limit of $\mathcal{E}_M(g_{\phi,A})$ is then

$$\mathcal{E}_\infty(g_{\phi,A}) := \lim_{M\to\infty}\mathcal{E}_M(g_{\phi,A}) = \frac{1}{N}\mathbb{E}\left[\|Y - R_{g_{\phi,A}}[X]\|^2_2\right].$$

The i-th error $\mathcal{E}^{(i)}_\infty(g_{\phi,A})$ for any $1 \leqslant i \leqslant N$ is defined in the same manner.

The next theorem states that this estimator achieves the nearly optimal rate in estimating the interaction function. This rate matches the lower bound in Theorem 4.4 up to a logarithmic factor. Its proof is deferred to Appendix B.1.

Theorem 3.1 *Suppose $rd \leqslant (M/\log M)^{\frac{1}{2\beta+1}}$. Consider the estimator $\widehat{g}_M$ defined in equation 3.1 computed on data M i.i.d. observation satisfying Assumptions 2.1 and* (B1). *Then, for $\widehat{g}_M$ defined in equation 3.1 it holds that*

$$\limsup_{M\to\infty}\ \sup_{g_\star\in\mathcal{G}^\beta_r(L,B_\phi,\bar{a})}\mathbb{E}\Big[M^{\frac{2\beta}{2\beta+1}}\|\widehat{g}_M - g_\star\|^2_{L^2_\rho}\Big] \lesssim C_{N,L,\bar{a},\beta,s}, \tag{3.2}$$

where $C_{N,L,\bar{a},\beta,s} = N[C_1^\beta\frac{L^2(s\bar{a})^{2\beta}}{(s!)^2} + C_2]$ for some universal positive constants C_1, C_2.

Remark 3.2 *The symbol $\lesssim$ indicates that the upper bound holds up to a logarithmic factor of $(\log M)^{\frac{2\beta}{2\beta+1}+4\max(2\beta,1)}$. We believe this factor can be improved, as it currently creates a gap between our upper and lower bounds, representing a limitation of our methods. It is worth noting that by working with uniformly bounded noise, this factor can be simplified (e.g., see Theorem 22.2 in Györfi et al. (2006)). In simpler settings, such as standard regression or when the interaction matrix A is constant (e.g., for Euclidean distances), this logarithmic factor can be removed using more advanced techniques. This topic is discussed in several works, including Wang et al. (2025); Györfi et al. (2006); Van der Vaart (2000) and the references therein. However, in our model, the optimization depends on both the interaction matrix A and the function ϕ, which makes the problem non-convex. This difficulty makes the aforementioned techniques harder to implement. We therefore leave this for future work.*

Remark 3.3 *This theorem demonstrates that the attention-style model is free from the curse of dimensionality. In particular, the embedding dimension d can be very large, satisfying the bound $rd \leqslant (M/\log M)^{\frac{1}{2\beta+1}}$. This condition becomes looser as β decreases, corresponding to rougher activation functions. When this condition is not satisfied, the error coming from the estimation of $A_\star$ dominates. See Corollary B.2.*

The proof extends the technique in Györfi et al. (2006, Theorem 22.2) originally developed for the projection pursuit algorithm for multi-index models. Our setup differs from the multi-index setup in which one estimates $Y = \sum_{i=1}^{K} f_i(b_i^\top X) + \eta$ with $\{f_i : \mathbb{R}\to\mathbb{R} \text{ and } b_i \in \mathbb{R}^d\}_{i=1}^K$ from sample data $\{(X^m, Y^m)\}_{m=1}^M$, where the data Y depends locally on a projected values of single particle X. Here, the attention-style model involves averaging multiple values of the pairwise interaction function, which is a composition of the unknown $\phi_\star$ and A. This nonlocal dependence, combined with the mixture of parametric and nonparametric estimations, presents a significant challenge.

We list below the main challenges we address in the proof of Theorem 3.1.

1. *Nonlocal dependency.* The nonlocal dependence presents a challenge in estimating the interaction function. The forward operator $R_g[X]$ depends on the g non-locally through the weighted sum of multiple values of g of pairwise interaction. Thus, this is a type of inverse problem that raises significant hurdles in both well-posedness and the construction of estimators to achieve the minimax rate. To address these challenges, we show first that the inverse problem in the large sample limit is well-posed for a large class of distributions of X satisfying Assumption 2.1. A crucial condition for well-posedness of this inverse problem is the *coercivity condition* studied in Li & Lu (2023); Li et al. (2021); Lu et al. (2019); Wang et al. (2025):

$$\frac{1}{N-1}\mathbb{E}\left[\|\widehat{g}_M - g_\star\|^2_{L^2_\rho}\right] \leqslant \mathcal{E}_\infty(\widehat{g}_M) - \mathcal{E}_\infty(g_\star).$$

   We prove this condition holds for a general function in our class in Lemma 3.4. Importantly, differing from these studies where the goal is to estimate the radial interaction kernel, our interaction is not shift-invariant due to the matrix and it is a $2d$-dimensional pairwise interaction function.

2. *Tail decay noise distribution.* The proof in Györfi et al. (2006) is limited to bounded noise. We provide a more general statement for any sub-Gaussian noise. This is done by decomposing the error bound now into three parts.

$$\mathcal{E}_\infty(\widehat{g}_M) - \mathcal{E}_\infty(g_\star) \leqslant \mathbb{E}[T_{1,M}] + \mathbb{E}[T_{2,M}] + \mathbb{E}[T_{3,M}]. \tag{3.3}$$

   The first two terms are a clever form of a bias-variance decomposition applied to a truncated version of the target. To bound these terms, we use a similar technique as in Györfi et al. (2006). To control the last term $T_{3,M}$ due to the truncation, we apply a lemma proved in (Kohler & Mehnert, 2011, Lemma 2).

3. *Covering numbers estimates.* Since our interaction is of the form $\frac{1}{N-1}\sum_{j=1}^N \phi(X_i^\top A X_j)$ instead of working in the space of vectors, we provide a covering estimate for the class of matrices with rank less than or equal to r. This is done in Lemma B.3.

The next lemma proves the crucial condition for the well-posedness of the inverse problem of estimating the interaction function. This Lemma assumes exchangeability and allows us to extract the error of the mean interaction and obtain a dimension-free rate for that error. Its proof is based on the exchangeability of the particle distribution and is postponed to Appendix B.1.

Lemma 3.4 (Coercivity) *Let $g, g_\star \in \mathcal{G}_r^\beta(L, B_\phi, \bar{a})$. Under exchangeability of $(X_i)_{i=1}^N$ in Assumption (A1) , we have*

$$\frac{1}{N-1}\|g - g_\star\|^2_{L^2_\rho} \leqslant \mathcal{E}_\infty(g) - \mathcal{E}_\infty(g_\star).$$

4 Lower bound

This section establishes a lower bound for estimating $g_\star(x, y) := \phi_\star(x^\top A_\star y)$ that matches the upper bound in Theorem 3.1; together, these results determine the minimax rate.

The main challenge lies in the nonlocal dependence of the output Y_i on $g_\star$, which is determined through averaging over all particles, as we don't directly observe any value of $g_\star$. Thus, the estimation of $g_\star$ is a deconvolution-type inverse problem, which is harder than estimating the single index model $Y = f(b^\top X) + \eta$ in Gaïffas & Lecué (2007). Importantly, the nonlinear joint dependence of $g_\star$ on the unknown $\phi_\star$ and $A_\star$ further complicates the problem.

We address the challenge by first reducing the supremum over all $g_\star$ to the supremum over all $\phi_\star$ with a fixed $A_\star \in \mathcal{A}_d(r, \bar{a})$, building on a technical result in Lemma 4.1. This reduces the problem to the minimax lower bound of estimating the interaction kernel $\phi_\star$ only. We derive this lower bound using the scheme in Wang et al. (2025), a variant of the Fano-Tsybakov method in Tsybakov (2008).

Let $A_\star \in \mathcal{A}_d(r, \bar{a})$ and let

$$U_{ij} := X_i^\top A_\star X_j, \quad U \sim p_U(u) := \frac{1}{N(N-1)}\sum_{i=1}^N \sum_{j=1, j\neq i}^N p_{U_{ij}}(u), \tag{4.1}$$

where $p_{U_{ij}}$ denotes the probability density of U_{ij}. Here, the density $p_{U_{ij}}$ exists and is continuous because $\mathrm{rank}(A_\star) \geqslant 2$ and the joint density of (X_i, X_j) exists by Assumption (A2), see Lemma C.1. Hence, the density p_U is continuous. Furthermore, since $\|A_\star\|_{\text{op}} \leqslant \bar{a}$ and $X_i \in \mathcal{C}_d$, we have $|U_{ij}| \leqslant \bar{a}$ and $\mathrm{supp}(p_U) \subset [-\bar{a}, \bar{a}]$. In particular, when the distribution X is exchangeable, we have $p_{U_{ij}}(u) = p_{U_{12}}(u) = p_U(u)$ for all (i,j), $u \in [-\bar{a}, \bar{a}]$. However, our proof below works for non-exchangeable distributions.

The next lemma allows us to reduce the supremum over all $g_\star(x,y) = \phi_\star(x^\top A_\star y)$ to all $\phi_\star$ by bounding $\|\widehat{g} - g_\star\|^2_{L^2_\rho}$ from below by $\|\widehat{\psi} - \phi_\star\|^2_{L^2_{p_U}}$ for a function $\widehat{\psi}$ determined by $\widehat{g}$ and $A_\star$. Its proof can be found in Section C.

Lemma 4.1 *Suppose Assumption* (A2) *holds. Let $A_\star, \hat{A} \in \mathcal{A}_d(r, \bar{a})$. Recall the definitions of U_{ij} and $U \sim p_U$ (defined according to $A_\star$) in equation 4.1. Let $\phi_\star, \hat{\phi} \in L^2_{p_U}$, and define a function $\hat{\psi}$ that is determined by $(\hat{\phi}, \hat{A}, A_\star)$ and the distribution of X as*

$$\widehat{\psi}_{ij}(u) := \mathbb{E}\big[\widehat{\phi}(X_i^\top \hat{A} X_j) \big| U_{ij} = u\big], \quad \widehat{\psi}(u) := \sum_{i=1}^N \sum_{j=1, j\neq i}^N \frac{p_{U_{ij}(u)}}{N(N-1)p_U(u)} \hat{\psi}_{ij(u)}. \tag{4.2}$$

Then, the following inequality holds:

$$\|\widehat{g} - g_\star\|^2_{L^2_\rho} \geqslant \int_{-\bar{a}}^{\bar{a}} \big|\hat{\psi}(u) - \phi_\star(u)\big|^2 p_U(u) \mathrm{d}u.$$

The next lemma constructs a finite family of hypothesis functions that are well-separated in $L^2_{p_U}$, while their induced distributions remain close with a slowly increasing total Kullback-Leibler divergence, enabling the application of Fano's method to derive the minimax lower bound. Its proof follows the scheme in Wang et al. (2025) and is postponed to Section C.

Lemma 4.2 *For each data set $\{(X^m, Y^m)\}_{m=1}^M$ sampled from the model $Y = R_{\phi_\star, A_\star}[X] + \eta$, where $A_\star \in \mathcal{A}_d(r, \bar{a})$ satisfying assumptions* (B2) *and* (A2)*, there exists a set of hypothesis functions $\{\phi_{0,M} \equiv 0, \phi_{1,M}, \cdots, \phi_{K,M}\}$ and positive constants $\{C_0, C_1\}$ independent of M, N, d, r, where*

$$K \geqslant 2^{\bar{K}/8}, \quad \text{with } \bar{K} = \lceil c_{0,N} M^{\frac{1}{2\beta+1}} \rceil, \quad c_{0,N} = C_0 N^{\frac{1}{2\beta+1}}, \tag{4.3}$$

such that the following conditions hold:

(D1) *Holder continuity: $\phi_{k,M} \in \mathcal{C}^\beta(L, \bar{a})$ and $\|\phi_{k,M}\|_\infty \leqslant B_\phi$ for each $k = 1, \cdots, K$;*

(D2) *$2s_{N,M}$-separated: $\|\phi_{k,M} - \phi_{k',M}\|_{L^2_{p_U}} \geqslant 2s_{N,M}$ with $s_{N,M} = C_1 c_{0,N}^{-\beta} M^{-\frac{\beta}{2\beta+1}}$;*

(D3) *Kullback-Leibler divergence estimate: $\frac{1}{K}\sum_{k=1}^K D_{\mathrm{KL}}(\bar{\mathbb{P}}_k, \bar{\mathbb{P}}_0) \leqslant \alpha \log(K)$ with $\alpha < 1/8$,*

where $\bar{\mathbb{P}}_k(\cdot) = \mathbb{P}_{\phi_{k,M}}(\cdot \mid X^1, \ldots, X^M)$ and p_U is the density of U defined in equation 4.1.

The following theorem provides a lower minimax rate for estimating $\phi_\star$ when $A_\star$ is given. Its proof is available in Section C.

Theorem 4.3 *Suppose Assumptions* (A2) *and* (B2) *hold. Let p_U be the density of U defined in equation 4.1. Then, for any $\beta > 0$, there exists a constant $c_0 > 0$ independent of M, d, r and N such that*

$$\liminf_{M\to\infty} \inf_{\hat{\psi}_M \in L^2_{p_U}} \sup_{\substack{\phi_\star \in \mathcal{C}^\beta(L,\bar{a}) \\ \|\phi_\star\|_\infty \leqslant B_\phi}} \mathbb{E}_{\phi_\star}\left[M^{\frac{2\beta}{2\beta+1}} \|\hat{\psi}_M - \phi_\star\|^2_{L^2_{p_U}}\right] \geqslant c_0 N^{-\frac{2\beta}{2\beta+1}} \tag{4.4}$$

where $\hat{\psi}_M$ is estimated based on the observation model with M i.i.d. samples.

Following the above results, we can now provide a lower bound for the convergence rate when estimating $g_\star$ over all possible estimators in the worst-case scenario.

Theorem 4.4 (Minimax lower bound) *Suppose Assumptions* (A2) *and* (B2) *hold. Then, for any* $\beta > 0$ *there exists a constant* $c_0 > 0$ *independent of* M, d, r *and* N*, such that the following inequality holds:*

$$\liminf_{M\to\infty} \inf_{\widehat{g}} \sup_{g_\star \in \mathcal{G}_r^\beta(L, B_\phi, \bar{a})} M^{\frac{2\beta}{2\beta+1}} \mathbb{E}\Big[\|\widehat{g} - g_\star\|_{L^2_\rho}^2\Big] \geqslant c_0 N^{-\frac{2\beta}{(2\beta+1)}} \tag{4.5}$$

where the infimum $\inf_{\widehat{g}}$ *is taken over all* $\widehat{g}(x, y) = \widehat{\phi}(x^\top \widehat{A} y)$ *with* $\widehat{A} \in \mathcal{A}_d(r, \bar{a})$ *and* $\widehat{\phi}$ *such that* $\widehat{g} \in L^2_\rho$.

5 Numerical simulations

In this section, we empirically verify the convergence rates predicted by our theory, emphasizing their independence from the ambient dimension d and their dependence on the activation function's smoothness.

For all experiments, we use B-splines to represent the ground-truth activation $\phi_\star$: a degree-p B-spline is C^{p-1}, so the degree directly controls the smoothness (Lyche et al., 2017). B-splines are linear in their basis coefficients, allowing us to efficiently compute an optimal coefficient estimator by least squares. Our estimator for the interaction function $\hat{g}$ exploits this structure: we first fit $\phi_\star$ in the B-spline basis by least squares, then approximate the fitted activation with a multi-layer perceptron to enable backpropagation when estimating $A_\star$. This design enables us to control both the smoothness and the approximation accuracy of $\hat{g}$, ensuring that it achieves the minimax rate. Full simulation and parameter details appear in Appendix D.

Our experiments confirm the theoretical minimax rates.

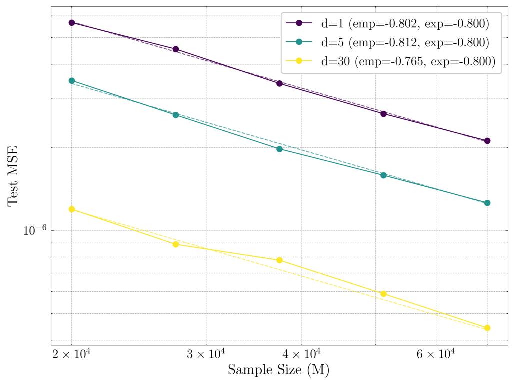
- *Independence from the ambient dimension d.* Figure 1(a) compares convergence across embedding dimensions $d \in \{1, 5, 30\}$. In the log-log plots, the slopes (which encode the rates) are nearly parallel and close to the theoretical exponent $-2\beta/(2\beta + 1)$ for all three dimensions, indicating that the convergence rate is independent of d.
- *Dependence on the activation function's smoothness.* Figure 1(b) reports rates for varying smoothness exponents β, controlled by the B-spline degree used to represent $\phi_\star$. As the spline degree (and hence β) increases, the log-log slope steepens as predicted by theory: for example, the empirical slopes are ≈ -0.81 for degree $P = 3$ and ≈ -0.899 for $P = 8$, closely matching the theoretical values -0.80 and -0.933.

The two plots illustrate that the minimax rate is fully determined by the smoothness β and it is dimension-free.

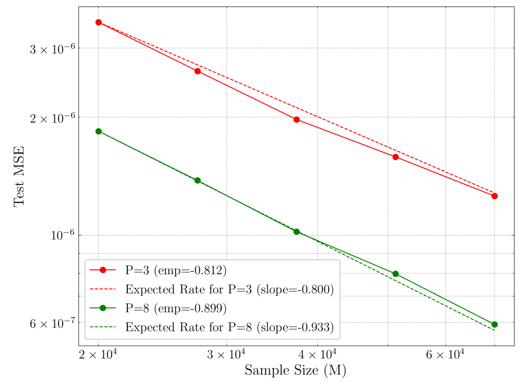
6 Conclusions

We have established *dimension-free* minimax convergence rates in sample size for estimating the pairwise interaction functions in self-attention style models. Using a direct connection to interacting particle systems (IPS), we have proved that under a coercivity condition, one can learn the interaction function at an optimal rate $M^{-2\beta/(2\beta+1)}$ with β being the smoothness of the function. Notably, this rate is independent of both the embedding dimension and the number of tokens. Our analysis extends beyond the standard assumption of independent, isotropic token distributions to allow for correlated and anisotropic token distributions.

These dimension-free rates illuminate how attention can avoid the curse of dimensionality in high-dimensional regimes. Viewing attention through the IPS lens suggests a broad research agenda for understanding the attention models. Promising next steps include extending the theory to multi-head attention, residual connections and self-attention interactions induced by the value matrix. Advances in these directions will improve our understanding of learning mechanisms and generalization in transformers.



(a)



(b)

Figure 1: **(a)** Convergence rates with $d \in \{1, 5, 30\}$. Composed test Mean Squared Error (MSE) vs. sample size M in log scale; dashed lines show the expected rate $M^{-2\beta/(2\beta+1)}$; and the markers represent the median across seeds. The convergence rates are nearly the same for different values of d. **(b)** Convergence rates with varying smoothness exponents, which are controlled by the spline degree of $\phi_\star$ and the estimator, with $P_{\text{true}} = P_{\text{est}} \in \{3, 8\}$, corresponding to $\beta \in \{2, 7\}$ and expected slopes -0.800 and -0.933. The parameters in each simulation are described in Appendix D.

REFERENCES

Chiara Amorino, Akram Heidari, Vytautė Pilipauskaitė, and Mark Podolskij. Parameter estimation of discretely observed interacting particle systems. *Stochastic Processes and their Applications*, 163:350–386, 2023. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2023.06.011. URL `https://www.sciencedirect.com/science/article/pii/S0304414923001321`.

Sohom Bhattacharya, Jianqing Fan, and Debarghya Mukherjee. Deep neural networks for nonparametric interaction models with diverging dimension. *The Annals of Statistics*, 52(6):2738 – 2766, 2024. doi: 10.1214/24-AOS2442. URL `https://doi.org/10.1214/24-AOS2442`.

Fabrizio Boncoraglio, Emanuele Troiani, Vittorio Erba, and Lenka Zdeborová. Bayes optimal learning of attention-indexed models, 2025. URL `https://arxiv.org/abs/2506.01582`.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=eBS3dQQ8GV`.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6572–6583, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html`.

Xiaohui Chen. Maximum likelihood estimation of potential energy in interacting particle systems from single-trajectory data. *Electron. Commun. Probab.*, pp. 1–13, 2021.

Hugo Cui. High-dimensional learning of narrow neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(2):023402, 2025.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. *Advances in Neural Information Processing Systems*, 37:36342–36389, 2024.

Laetitia Della Maestra and Marc Hoffmann. Nonparametric estimation for interacting particle systems: McKean-Vlasov models. *Probability Theory and Related Fields*, pp. 1–63, 2022.

Laetitia Della Maestra and Marc Hoffmann. The lan property for mckean–vlasov models in a mean-field regime. *Stochastic Processes and their Applications*, 155:109–146, 2023.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention, 2024. URL https://arxiv.org/abs/2310.12680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems, 2021. URL https://arxiv.org/abs/2109.15142.

Lawrence Craig Evans. *Measure theory and fine properties of functions*. Routledge, 2018.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. *Advances in Neural Information Processing Systems*, 36:11912–11951, 2023.

Stéphane Gaıffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1:538–573, 2007.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aMjaEkkXJx.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms, 2024a. URL https://arxiv.org/abs/2410.13835.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in kv cache reduction: Value also matters, 2024b. URL https://arxiv.org/abs/2406.12335.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. *Advances in Neural Information Processing Systems*, 37:79221–79245, 2024.

Joel L. Horowitz and Enno Mammen. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *The Annals of Statistics*, 35(6):2589 – 2619, 2007. doi: 10.1214/009053607000000415. URL https://doi.org/10.1214/009053607000000415.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pp. 4376–4386. PMLR, 2020.

Hui Huang, Jian-Guo Liu, and Jianfeng Lu. Learning interacting particle systems: Diffusion parameter estimation for aggregation equations. *Mathematical Models and Methods in Applied Sciences*, 29(01):1–29, 2019.

Raphael A. Kasonga. Maximum likelihood theory for large interacting systems. *SIAM J. Appl. Math.*, 50(3):865–875, 1990. ISSN 0036-1399, 1095-712X. doi: 10.1137/0150050.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Michael Kohler and Jens Mehnert. Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors. *Neural Networks*, 24(3):273–279, 2011. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2010.11.003. URL `https://www.sciencedirect.com/science/article/pii/S0893608010002157`.

Soroush Abbasi Koohpayegani and Hamed Pirsiavash. Sima: Simple softmax-free attention for vision transformers. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2595–2605, 2024. doi: 10.1109/WACV57701.2024.00259.

Hongkang Li, Meng Wang, Sijia Liu, and Pin yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity, 2023. URL `https://arxiv.org/abs/2302.06015`.

Zhongyang Li and Fei Lu. On the coercivity condition in the learning of interacting particle systems. *Stochastics and Dynamics*, pp. 2340003, 2023.

Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, and Cheng Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135–163, 2021.

Meiqi Liu and Huijie Qiao. Parameter estimation of path-dependent McKean-Vlasov stochastic differential equations. *Acta Mathematica Scientia*, 42(3):876–886, 2022.

Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proc. Natl. Acad. Sci. USA*, 116(29):14424–14433, 2019.

Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *Journal of Machine Learning Research*, 22(32):1–67, 2021a.

Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *Foundations of Computational Mathematics*, 22: 1013–1067, 2022.

Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: Softmax-free transformer with linear complexity. In *NeurIPS*, 2021b.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2502599122`.

Tom Lyche, Carla Manni, and Hendrik Speleers. B-splines and spline approximation. 2017. URL `https://api.semanticscholar.org/CorpusID:195737484`.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=DVlPp7Jd7P`.

OpenAI. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning, 2023. URL `https://arxiv.org/abs/2306.03435`.

Jason Ramapuram, Federico Danieli, Eeshan Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russ Webb. Theory, analysis, and best practices for sigmoid self-attention. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=Zhdhg6n2OG`.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530. PMLR, 28–30 Mar 2022.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

Louis Sharrock, Nikolas Kantas, Panos Parpas, and Grigorios A. Pavliotis. Parameter stimation for the McKean-Vlasov stochastic differential equation. *ArXiv210613751 Math Stat*, 2021.

Christopher D Sogge. *Fourier integrals in classical analysis*, volume 210. Cambridge University Press, 2017.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models, 2024. URL https://arxiv.org/abs/2402.17762.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: High-dimensional asymptotics and sharp thresholds. *arXiv preprint arXiv:2502.00901*, 2025.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 1st edition, 2008. ISBN 978-0-387-79051-0.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Shulun Wang, Feng Liu, and Bin Liu. Escaping the gradient vanishing: Periodic alternatives of softmax in attention mechanism. *IEEE Access*, 9:168749–168759, 2021. doi: 10.1109/ACCESS.2021.3138201.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. URL https://arxiv.org/abs/2006.04768.

Xiong Wang, Inbar Seroussi, and Fei Lu. Optimal minimax rate of learning nonlocal interaction kernels, 2025. URL https://arxiv.org/abs/2311.16852.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Rentian Yao, Xiaohui Chen, and Yun Yang. Mean-field nonparametric estimation of interacting particle systems. In *Conference on Learning Theory*, pp. 2242–2275. PMLR, 2022.

Yue Yu, Ning Liu, Fei Lu, Tian Gao, Siavash Jafarzadeh, and Stewart A Silling. Nonlocal attention operator: Materializing hidden knowledge towards interpretable physics discovery. *Advances in Neural Information Processing Systems*, 37:113797–113822, 2024.

Biao Zhang, Ivan Titov, and Rico Sennrich. Sparse attention with linear units. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6507–6520. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.523. URL https://doi.org/10.18653/v1/2021.emnlp-main.523.

A Appendix: Reduction from attention to IPS attention model

In this section, we provide a direct connection between the IPS attention model and the softmax self-attention layer, which typically includes an additional normalization step. Consider a sequence of tokens $\{X_i\}_{i=1}^N$. The output of the softmax self-attention layer is typically composed of an attention block with learnable query, key, and value matrices, $W_Q, W_K \in \mathbb{R}^{d\times d_k}$ with $d_k \leqslant d$ and $W_V \in \mathbb{R}^{d\times d_v}$ that compute

$$Y = \mathrm{Att}(Q,K,V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \qquad Q = XW_Q,\ K = XW_K,\ V = XW_V. \tag{A.1}$$

As explained in the main text, we denote by $A = \frac{1}{\sqrt{d_k}} W_Q W_K^\top$ the score interaction matrix. Using the definition of the softmax function, the output of the softmax self-attention layer for each particle can then be written

$$Y_i = \sum_{j=1}^N \frac{e^{\beta X_i^\top A X_j}}{Z_i[X]} V_j, \quad Z_i[X] = \sum_{\ell=1}^N e^{\beta X_i^\top A X_\ell}$$

with $\beta > 0$ being the inverse temperature parameter. When the number of particles is large, the partition function $Z_i[X]$ concentrates around its mean-field value with respect to the empirical distribution of the particles. If we denote by μ the continuum limit of the empirical measure, then $Z_i[X] \approx N\mathcal{Z}_i = N\int e^{\beta X_i^\top A y}\mathrm{d}\mu(y)$ conditioned on the i-th particle.

For the IPS surrogate we consider in this paper, we adopt two standard simplifications (Sander et al., 2022; Geshkovski et al., 2025; Bruno et al., 2025): we set $d_v = 1$, and treat $\mathcal{Z}_i$ as a constant (independent of X) that can be absorbed into the nonlinearity, and focus only on the self-interaction for $i \neq j$, and setting V_j to be a constant, we get our IPS Attention Model:

$$Y_i = \frac{1}{N}\sum_{j\neq i}\phi(X_i^\top A X_j).$$

This reduction is similar in spirit to the surrogate model (USA) presented in Geshkovski et al. (2025). We note that a possible extension of our model to account for the softmax normalization would be to learn a function for each particle, ϕ_i. We suspect it will not change the overall rate. In fact, as stated in Geshkovski et al. (2025), this reduction seems to capture the essence of the dynamics of the self-attention (SA) model. Therefore, to simplify the setting, we focus on estimating a single function.

B Appendix: upper bound proofs

We begin by reducing the distribution of the pair-wise particles to the distribution of one pair by exchangeability. The exchangeability not only simplifies the proof of the upper bound, but also provides a sufficient condition for the coercivity, which makes the inverse problem well-posed.

Lemma B.1 (Exploration measure under exchangeability) *Under Assumption 2.1, the measure ρ is the distribution of $(X_1, X_2) \in \mathbb{R}^d \times \mathbb{R}^d$ and has a continuous density.*

Proof. The exchangeability in Assumption 2.1 implies that the distributions of (X_i, X_j) and (X_1, X_2) are the same for any $i \neq j$. Hence, by definition, the exploration measure is the distribution of the random variables (X_1, X_2):

$$\rho(B) = \mathbb{P}((X_1, X_2) \in B)$$

which has a continuous density by Assumption 2.1. □

B.1 Proof of the upper bound in theorem 3.1

In this section, we provide the proof of the upper bound.

We begin with the proof of the key coercivity lemma, which is crucial in bounding the error of the interaction function and making the inverse problem well-posed in the large sample limit.

Proof of Lemma 3.4 Recall $R_g(X)_i = \frac{1}{N-1}\sum_{j\neq i} g(X_i, X_j)$. By definition

$$\mathcal{E}_\infty(g) - \mathcal{E}_\infty(g_\star) = \frac{1}{N}\mathbb{E}\langle R_{g-g_\star}[X], R_{g-g_\star}[X]\rangle = \frac{1}{N(N-1)^2}\sum_{i=1}^{N}\sum_{j\neq i}\sum_{j'\neq i}\mathbb{E}\langle \Delta_{ij}, \Delta_{ij'}\rangle,$$

where $\Delta_{ij} = (g - g_\star)(X_i, X_j)$ and $\sum_{j\neq i} = \sum_{j=1, j\neq i}^{N}$. By exchangeability,

$$\begin{aligned}\frac{1}{N(N-1)^2}\sum_{i=1}^{N}\sum_{j\neq i}\sum_{j'\neq i}\mathbb{E}\langle \Delta_{ij}, \Delta_{ij'}\rangle &= \frac{1}{N-1}\mathbb{E}\|\Delta_{12}\|^2 + \frac{N-2}{N-1}\mathbb{E}\langle \Delta_{12}, \mathbb{E}[\Delta_{13} \mid X_1]\rangle \\ &\geqslant \frac{1}{N-1}\mathbb{E}\|\Delta_{12}\|^2,\end{aligned}$$

since $\mathbb{E}\|\mathbb{E}[\Delta_{13} \mid X_1]\|^2 \geqslant 0$. The statement of the Lemma follows. □

Proof of Theorem 3.1 The proof is divided into five steps.

Step 1: Error decomposition. In this step, we decompose the mean squared error $\mathbb{E}[\|\widehat{g}_M - g_\star\|^2_{L^2_\rho}]$ to two terms. Using Lemma 3.4, i.e., the coercivity condition and the definition of $\mathcal{E}_\infty(g) = \frac{1}{N}\mathbb{E}\left[\|Y - R_g[X]\|^2\right]$, we have for $c_{\mathcal{H}} = \frac{1}{N-1}$

$$\begin{aligned} &c_{\mathcal{H}}\mathbb{E}\left[\int |\widehat{g}_M(x,y) - g_\star(x,y)|^2 \mathrm{d}\rho(x,y)\right] \\ &\leqslant \mathcal{E}_\infty(\widehat{g}_M) - \mathcal{E}_\infty(g_\star) \\ &= \frac{1}{N}\mathbb{E}[\|Y - R_{\widehat{g}_M}[X]\|^2] - \frac{1}{N}\mathbb{E}\left[\|Y - R_{g_\star}[X]\|^2\right] \\ &= \frac{1}{N}\mathbb{E}\big[\mathbb{E}\left[\|Y - R_{\widehat{g}_M}[X]\|^2 \mid \mathcal{D}_M\right]\big] - \frac{1}{N}\mathbb{E}\left[\|Y - R_{g_\star}[X]\|^2\right]. \end{aligned} \tag{B.1}$$

Let $B_M := c_1 \log(M)$ with some constant $c_1 > 0$ and $Y_M := \min(B_M, \max(-B_M, Y))$. Let us denote

$$T_{1,M} := 2[\mathcal{E}_M(\widehat{g}_M) - \mathcal{E}_M(g_\star)] \tag{B.2}$$

and

$$T_{2,M} := \frac{1}{N}\mathbb{E}\big[\|Y_M - R_{\widehat{g}_M}[X]\|^2 \mid \mathcal{D}_M\big] - \frac{1}{N}\mathbb{E}\big[\|Y_M - R_{g_\star}[X]\|^2\big] - T_{1,M}, \tag{B.3}$$

$$\begin{aligned} T_{3,M} := &\frac{1}{N}\mathbb{E}\left[\|Y - R_{\widehat{g}_M}[X]\|^2 \mid \mathcal{D}_M\right] - \frac{1}{N}\mathbb{E}\left[\|Y - R_{g_\star}[X]\|^2\right] \\ &- \frac{1}{N}\mathbb{E}\big[\|Y_M - R_{\widehat{g}_M}[X]\|^2 \mid \mathcal{D}_M\big] + \frac{1}{N}\mathbb{E}\big[\|Y_M - R_{g_\star}[X]\|^2\big]. \end{aligned} \tag{B.4}$$

By equation B.1, we can decompose the upper bound of the mean squared error as

$$\begin{aligned} \mathbb{E}\left[\|\widehat{g}_M - g_\star\|^2_{L^2_\rho}\right] &= \mathbb{E}\left[\int |\widehat{g}_M(x,y) - g_\star(x,y)|^2 \mathrm{d}\rho(x,y)\right] \\ &\leqslant c_{\mathcal{H}}^{-1}\Big(\mathbb{E}[T_{1,M}] + \mathbb{E}[T_{2,M}] + \mathbb{E}[T_{3,M}]\Big). \end{aligned} \tag{B.5}$$

We shall proceed with our proof by bounding $\{\mathbb{E}[T_{i,M}]\}_{i=1}^{3}$ in the following Steps 2-4 via approximation error estimate, covering number estimate, and sub-Gaussian property, respectively.

Step 2: Bounding $\mathbb{E}[T_{1,M}]$ via polynomial approximation. Recall that $\widehat{g}_M$ is the minimizer of the empirical error functional $\mathcal{E}_M(g)$ over the estimator space $\mathcal{G}^s_{r,K_M}$. Thus, we have

$$\mathcal{E}_M(\widehat{g}_M) - \mathcal{E}_M(g_\star) \leqslant \mathcal{E}_M(g_{\star,\mathcal{G}_r}) - \mathcal{E}_M(g_\star),$$

where $g_{\star,\mathcal{G}^s_{r,K_M}}$ is a minimizer in $\mathcal{G}^s_{r,K_M}$ attaining $\inf_{g\in\mathcal{G}^s_{r,K_M}} [\mathcal{E}_\infty(g) - \mathcal{E}_\infty(g_\star)]$ (see, (Györfi et al., 2006, Lemma 11.1)). Therefore,

$$\begin{aligned}
\frac{1}{2}\mathbb{E}[T_{1,M}] &= \mathbb{E}\left[\mathcal{E}_M(\widehat{g}_M) - \mathcal{E}_M(g_\star)\right] \\
&\leqslant \mathbb{E}\left[\mathcal{E}_M(g_{\star,\mathcal{G}^s_{r,K_M}}) - \mathcal{E}_M(g_\star)\right] \\
&= \mathcal{E}_\infty(g_{\star,\mathcal{G}^s_{r,K_M}}) - \mathcal{E}_\infty(g_\star) = \inf_{g\in\mathcal{G}^s_{r,K_M}} \left[\mathcal{E}_\infty(g) - \mathcal{E}_\infty(g_\star)\right] . \qquad \text{(B.6)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathcal{E}_\infty(g) - \mathcal{E}_\infty(g_\star) &= \frac{1}{N}\mathbb{E}\left[\|R_{g_\star - g}[X] + \eta\|^2 - \|\eta\|^2\right] \\
&= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left|\frac{1}{N-1}\sum_{j=1,j\neq i}^N [g-g_\star](X_i,X_j)\right|^2\right] . \qquad \text{(B.7)}
\end{aligned}$$

Applying Jensen's inequality to get

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left|\frac{1}{N-1}\sum_{j=1,j\neq i}^N [g-g_\star](X_i,X_j)\right|^2\right] \leqslant \frac{1}{N(N-1)}\sum_{i=1}^N\sum_{j=1,j\neq i}^N \mathbb{E}\left[|[g-g_\star](X_i,X_j)|^2\right] \qquad \text{(B.8)}$$

and by the exchangeability assumption (A1), we have that the expectations are equal and thus

$$\begin{aligned}
\frac{1}{2}\mathbb{E}[T_{1,M}] &\leqslant \inf_{g\in\mathcal{G}^s_{r,K_M}} \|g - g_\star\|^2_{L^2_\rho} \\
&= \inf_{\phi\in\Phi^s_{K_M}, \|A\|_{\text{op}}\leqslant\bar{a}} \int |\phi(\langle x,y\rangle_A) - \phi_\star(\langle x,y\rangle_{A_\star})|^2 \mathrm{d}\rho(x,y) . \qquad \text{(B.9)}
\end{aligned}$$

Next, setting $A = A_\star$ in equation B.9, it is clear that

$$\begin{aligned}
\frac{1}{2}\mathbb{E}[T_{1,M}] &\leqslant \inf_{\phi\in\Phi^s_{K_M}} \int |\phi(\langle x,y\rangle_{A_\star}) - \phi_\star(\langle x,y\rangle_{A_\star})|^2 \mathrm{d}\rho(x,y) \\
&\leqslant \inf_{\phi\in\Phi^s_{K_M}} \left\{\sup_{u\in[-\bar{a},\bar{a}]} |\phi(u) - \phi_\star(u)|^2\right\} .
\end{aligned}$$

Then, one can choose ϕ following the construction in (Györfi et al., 2006, Lemma 11.1) and that $\phi_\star \in C^\beta(L,\bar{a})$ which shows that there exists a piecewise polynomial function f of degree β or less with respect to an equidistant partition of $[-\bar{a},\bar{a}]$ consisting of K_M intervals of length $1/K_M$. For any $x, y \sim \rho$ and any matrix $A \in \mathbb{R}^{d\times d}$ such that $u = \langle x,y\rangle_A \in [-\bar{a},\bar{a}]$, we will choose the dimension K_M (to be specified later) so that

$$\sup_{u\in[-\bar{a},\bar{a}]} |\phi(u) - \phi_\star(u)| \leqslant \frac{L\bar{a}^\beta}{\lfloor\beta\rfloor! K_M^\beta} .$$

We thus conclude that

$$\mathbb{E}[T_{1,M}] \leqslant \frac{2L^2\bar{a}^{2\beta}}{(\lfloor\beta\rfloor!)^2 K_M^{2\beta}} . \qquad \text{(B.10)}$$

Step 3: Bounding $\mathbb{E}[T_{2,M}]$ via covering number estimates. We introduce the following notations to simplify the presentation. Define $\Delta\mathcal{E}_M^{(i)}(g) := \mathcal{E}_M^{(i)}(g) - \mathcal{E}_M^{(i)}(g_\star)$. Also, we denote

$$\Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(\widehat{g}_M) := \mathbb{E}[|Y_i - R_{\widehat{g}_M}[X]_i|^2 \mid \mathcal{D}_M] - \mathbb{E}[|Y_i - R_{g_\star}[X]_i|^2],$$

where $\widehat{g}_M \in \mathcal{G}_{r,K_M}^s$ depends on the samples $\mathcal{D}_M = \{(X^m, Y^m)\}_{m=1}^M$ and write similarly

$$\Delta\mathcal{E}_\infty^{(i)}(g) := \mathbb{E}[|Y_i - R_g[X]_i|^2] - \mathbb{E}[|Y_i - R_{g_\star}[X]_i|^2],$$

for any $g \in \mathcal{G}_{r,K_M}^s$. Note that $\Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(g) = \Delta\mathcal{E}_\infty^{(i)}(g)$ for any (deterministic) $g \in \mathcal{G}_{r,K_M}^s$.

It is straightforward to observe that $\mathbb{E}[T_{2,M}]$ can be expressed as the average error per particle, that is, $\mathbb{E}[T_{2,M}] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[T_{2,M}^{(i)}]$ where $T_{2,M}^{(i)} := \Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(\widehat{g}_M) - T_{1,M}^{(i)}$ with $T_{1,M}^{(i)} = 2\Delta\mathcal{E}_M^{(i)}(\widehat{g}_M)$. To estimate $\mathbb{E}[T_{2,M}^{(i)}]$, it suffices to bound the following probability tail for the i-th particle

$$\begin{aligned}\mathbb{P}\left\{T_{2,M}^{(i)} > t\right\} &= \mathbb{P}\left\{\Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(\widehat{g}_M) - \Delta\mathcal{E}_M^{(i)}(\widehat{g}_M) > \frac{1}{2}[t + \Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(\widehat{g}_M)]\right\} \\ &\leqslant \mathbb{P}\left\{\exists f \in \mathcal{G}_{r,K_M}^s : \Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(f) - \Delta\mathcal{E}_M^{(i)}(f) > \frac{1}{2}[t + \Delta\mathcal{E}_{\mathcal{D}_M}^{(i)}(f)]\right\} \\ &= \mathbb{P}\left\{\exists f \in \mathcal{G}_{r,K_M}^s : \Delta\mathcal{E}_\infty^{(i)}(f) - \Delta\mathcal{E}_M^{(i)}(f) > \frac{1}{2}[t + \Delta\mathcal{E}_\infty^{(i)}(f)]\right\}. \end{aligned} \tag{B.11}$$

We first observe that the probability tail above depends on the joint distribution of all particles since the term $\Delta\mathcal{E}_M^{(i)}(f)$ in equation B.11 involves all particles. To bound the tail probability of $T_{2,M}^{(i)}$, we invoke Györfi et al. (2006, Theorem 11.4), which is applicable to classes of uniformly bounded functions. In our setting, this condition translates to the boundedness of the operator R_g. Specifically, if $\|g\|_\infty \leqslant B_\phi$, then for all $i \in [N]$, we have $|R_g[X]_i| \leqslant B_\phi$. Recall that $B_M := c_1 \log(M)$ and $\mathcal{C}_d := ([0,1]/\sqrt{d})^d$. Applying Theorem 11.4 in Györfi et al. (2006) to equation B.11 (with $\alpha = \beta = t/2$ and $\epsilon = 1/2$), we get for arbitrary $t \geqslant 1/M$

$$\begin{aligned}\mathbb{P}\left\{T_{2,M}^{(i)} > t\right\} &\leqslant 14 \sup_{\{X^m \in \mathcal{C}_d^N\}_{m=1}^M} \mathcal{N}_1\left(\frac{t}{80B_M}, \mathcal{G}_{r,K_M}^s, \rho_M\right) e^{-\frac{tM}{24\cdot 214 B_M^4}} \\ &\leqslant 14 \sup_{\{X^m \in \mathcal{C}_d^N\}_{m=1}^M} \mathcal{N}_1\left(\frac{1}{80B_M M}, \mathcal{G}_{r,K_M}^s, \rho_M\right) e^{-\frac{tM}{24\cdot 214 B_M^4}} \end{aligned} \tag{B.12}$$

where $\mathcal{N}_1(\varepsilon, \mathcal{G}_{r,K_M}^s, \rho_M)$ is the empirical covering number with respect to the L_ρ^1 radius smaller than ε over the function class $\mathcal{G}_{r,K_M}^s$.

Employing the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\mathrm{d}t$ and the standard integral decomposition $\int_0^\infty = \int_0^\varepsilon + \int_\varepsilon^\infty$ with ε to be determined, we get for $\varepsilon \geqslant 1/M$

$$\mathbb{E}[T_{2,M}^{(i)}] = \int_0^\infty \mathbb{P}(T_{2,M}^{(i)} > t)\mathrm{d}t \leqslant \varepsilon + \int_\varepsilon^\infty \mathbb{P}(T_{2,M}^{(i)} > t)\mathrm{d}t. \tag{B.13}$$

Then, substituting equation B.12 in equation B.13 leads to

$$\mathbb{E}[T_{2,M}^{(i)}] \leqslant \varepsilon + \int_\varepsilon^\infty 14 \sup_{\{X^m \in \mathcal{C}_d^N\}_{m=1}^M} \mathcal{N}_1\left(\frac{t}{80B_M}, \mathcal{G}_{r,K_M}^s, \rho_M\right) e^{-\frac{tM}{24\cdot 214 B_M^4}}\mathrm{d}t. \tag{B.14}$$

Notice that we can bound the covering number by its value at $1/M$ since $\varepsilon \geqslant 1/M$ inside the integral in equation B.14 when $t \geqslant \varepsilon$. It then follows that

$$\mathbb{E}[T_{2,M}^{(i)}] \leqslant \varepsilon + 14 \sup_{\{X^m \in \mathcal{C}_d^N\}_{m=1}^M} \mathcal{N}_1\left(\frac{1}{80B_M M}, \mathcal{G}_{r,K_M}^s, \rho_M\right) \int_\varepsilon^\infty e^{-\frac{tM}{24\cdot 214\cdot B^4}}\mathrm{d}t. \tag{B.15}$$

We can now apply the estimate of covering number in equation B.22:

$$\begin{aligned}\mathbb{E}[T_{2,M}] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}[T_{2,M}^{(i)}] \\ &\leqslant \varepsilon + 42\cdot(L_{1,M}M)^{2rd}(L_{2,M}M)^{2K_M(s+1)+2}\cdot\frac{L_{3,M}}{M}e^{-\frac{\varepsilon M}{L_{3,M}}} \end{aligned} \tag{B.16}$$

with $L_{1,M} := 12r\bar{a}B_\phi \cdot 80B_M$, $L_{2,M} := 24eB_\phi \cdot 80B_M$ and $L_{3,M} := 24 \cdot 214 \cdot B_M^4$. Since the quantity ε on the right-hand side of equation B.16 is arbitrary, we may tighten the bound by choosing

$$\varepsilon = \frac{L_{3,M}}{M} \cdot \log\left[42 \cdot (L_{1,M}M)^{2rd}(L_{2,M}M)^{2K_M(s+1)+2}\right],$$

which yields the desired upper bound:

$$\begin{aligned}\mathbb{E}[T_{2,M}] &\leqslant \frac{L_{3,M}}{M}\Big[1 + \log(42) + 2rd\log(L_{1,M}) \\ &\qquad + 2(K_M(s+1)+1)\cdot\log(L_{2,M}) + 2(K_M(s+1)+1+rd)\cdot\log(M)\Big] \\ &\leqslant \frac{L_{3,M}(20K_M s + 5rd)\log(M)}{M} \qquad \text{(B.17)}\end{aligned}$$

when $M \geqslant \max(42 \cdot L_{1,M}^{2rd}, L_{2,M})$.

Step 4: Bounding $\mathbb{E}[T_{3,M}]$ via sub-Gaussian property. As $R_{g_\star}[X]_i \leqslant B_\phi < B_M$ and $R_{\widehat{g}_M}[X]_i \leqslant B_\phi < B_M$ a.s. We assume that the noise η is sub-Gaussian and that R_g is bounded for any $g \in \mathcal{G}_r^\beta$. Thus, using Lemma 2 in Kohler & Mehnert (2011) with Y_M and B_M given above, one can obtain that with

$$\begin{aligned}\Big|\mathbb{E}[T_{3,M}]\Big| &\leqslant \frac{1}{N}\sum_{i=1}^{N}\Big|\mathbb{E}\big[|Y_{i,M} - R_{g_\star}[X_i]|^2\big] - \mathbb{E}\big[|Y_i - R_{g_\star}[X_i]|^2\big]\Big| \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\Big|\mathbb{E}\big[|Y_{i,M} - R_{\widehat{g}_M}[X_i]|^2\big] - \mathbb{E}\big[|Y_i - R_{\widehat{g}_M}[X_i]|^2\big]\Big| \\ &\leqslant c_2\frac{\log(M)}{M}, \qquad \text{(B.18)}\end{aligned}$$

for some constant $c_2 > 0$ independent of M and N.

Step 5: Deriving the upper optimal rate. We now combine the bounds from equation B.10, equation B.16 and equation B.18, which control the terms $\mathbb{E}[T_{1,M}]$, $\mathbb{E}[T_{2,M}]$ and $\mathbb{E}[T_{3,M}]$, respectively, to obtain an upper bound on the total error in equation B.5:

$$\begin{aligned}&\mathbb{E}\Big[\|\widehat{g}_M - g_\star\|_{L_\rho^2}^2\Big] \\ &\leqslant c_{\mathcal{H}}^{-1}\left(\frac{2L^2\bar{a}^{2\beta}}{(s!)^2K_M^{2\beta}} + \frac{L_{3,M}(20K_M s + 5rd)\log(M)}{M} + c_2\frac{\log(M)}{M}\right) \qquad \text{(B.19)} \\ &\leqslant c_{\mathcal{H}}^{-1}\left(\frac{2L^2\bar{a}^{2\beta}}{(s!)^2K_M^{2\beta}} + \frac{L_{3,M}20K_M s(1+5L_{3,M})\log(M)}{M} + c_2\frac{\log(M)}{M}\right)\end{aligned}$$

using the assumption of the theorem $rd \leqslant \left(\frac{M}{\log M}\right)^{\frac{1}{2\beta+1}}$ and setting the value of K_M as

$$K_M = \left\lfloor \frac{1}{20L_{3,M}s}\left(\frac{M}{\log M}\right)^{\frac{1}{2\beta+1}}\right\rfloor. \qquad \text{(B.20)}$$

A relatively straightforward choice of K_M balances the terms in equation B.19 and leads to a desired upper bound. We note that a careful choice of K_M may affect the constants and the power of $\log(M)$ in the upper bound. Putting equation B.20 back into equation B.19 and noticing $c_{\mathcal{H}}^{-1} \leqslant N$, we get

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_M - g_\star\|_{L_\rho^2}^2\right] &\leqslant c_{\mathcal{H}}^{-1}\left[\frac{2L^2(20L_{3,M}s\bar{a})^{2\beta}}{(s!)^2}\left(\frac{\log M}{M}\right)^{\frac{2\beta}{2\beta+1}} + (1+5L_{3,M})\left(\frac{\log M}{M}\right)^{\frac{2\beta}{2\beta+1}} \right. \\ &\qquad\qquad \left. + \frac{c_2\log(M)}{M}\right] \\ &\leqslant N\left[\frac{2L^2(20sL_{3,M}\bar{a})^{2\beta}}{(s!)^2} + (1+5L_{3,M}) + c_2\right]\left(\frac{\log M}{M}\right)^{\frac{2\beta}{2\beta+1}} \qquad \text{(B.21)}\end{aligned}$$

when $M \geqslant \max(42 \cdot L_{1,M}^{2rd}, L_{2,M})$ with $L_{1,M} := 12r\bar{a}B_\phi \cdot 80B_M$, $L_{2,M} := 24eB_\phi \cdot 80B_M$. Recalling that $L_{3,M} = 24 \cdot 214 \cdot B_M^4 = 24 \cdot 214 \cdot c_1 \cdot \log(M)$, we get from equation B.21 that

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{g}_M - g_\star\|_{L^2_\rho}^2\right] &\leqslant N\frac{2L^2(2024 \cdot 214 \cdot c_1 \cdot s\bar{a})^{2\beta}}{(s!)^2} \cdot \frac{[\log(M)]^{\frac{2\beta}{2\beta+1}+8\beta}}{M^{\frac{2\beta}{2\beta+1}}} \\
&\quad + N[1 + 5 \cdot 24 \cdot 214 + c_2] \cdot \frac{(\log M)^{\frac{2\beta}{2\beta+1}+4}}{M^{\frac{2\beta}{2\beta+1}}} \\
&\leqslant N\left[C_1^\beta \frac{L^2(s\bar{a})^{2\beta}}{(s!)^2} + C_2\right] \cdot \frac{[\log(M)]^{\frac{2\beta}{2\beta+1}+4\max(2\beta,1)}}{M^{\frac{2\beta}{2\beta+1}}}
\end{aligned}$$

for some positive constant C_1, C_2. We complete the proof of Theorem 3.1 with $C_{N,L,\bar{a},\beta,s} = N[C_1^\beta \frac{L^2(s\bar{a})^{2\beta}}{(s!)^2} + C_2]$. □

Finally, to highlight the tradeoff between the parametric and the non-parametric part of the error, we present the following corollary. This corollary is directly derived from equation B.19 and equation B.20 not using the assumption $rd \leqslant \left(\frac{M}{\log M}\right)^{\frac{1}{2\beta+1}}$.

Corollary B.2 *Consider the estimator $\widehat{g}_M$ defined in equation 3.1 computed on data M i.i.d. observation satisfying Assumptions 2.1 and* (B1)*. Then, for $\widehat{g}_M$ defined in equation 3.1 it holds that*

$$\mathbb{E}\left[\|\widehat{g}_M - g_\star\|_{L^2_\rho}^2\right] \leqslant N\left[C_1^\beta \frac{L^2(s\bar{a})^{2\beta}}{(s!)^2} + C_2\right] \cdot \frac{[\log(M)]^{\frac{2\beta}{2\beta+1}+4\max(2\beta,1)}}{M^{\frac{2\beta}{2\beta+1}}} + C_3 rd \cdot \frac{(\log M)^2}{M}$$

where C_1, C_2 and C_3 are positive constants (maybe take different values than in the Theorem 3.1).

B.2 Auxiliary lemmas for the upper bound

Recall that the covering number $\mathcal{N}(\varepsilon, \mathcal{G}, d)$ is defined as the cardinality of the smallest ε-cover of $\mathcal{G}$ with respect to the metric d. When d is the Euclidean metric, we omit it from the notation and simply write $\mathcal{N}(\varepsilon, \mathcal{G})$. It is also common to take d to be an L^p-norm, either with respect to a probability measure ρ or its empirical counterpart ρ_M. In these cases, we write

$$\mathcal{N}_p(\varepsilon, \mathcal{G}, \rho) := \mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_{L^p_\rho}), \quad \mathcal{N}_p(\varepsilon, \mathcal{G}, \rho_M) := \mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_{L^p_{\rho_M}}).$$

We next derive an upper bound for $\mathcal{N}_1(\varepsilon, \mathcal{G}^s_{r,K_M}, \rho_M)$, i.e., $p = 1$, by covering the matrix component and the functional component separately. Our argument combines the covering number estimates for matrices from Vershynin (2018) with the results of Györfi et al. (2006) for function classes.

Lemma B.3 *Let $\mathcal{G}^s_{r,K_M}$ be defined in Definition 2.5. Assume that the sampled data $\{X_i^m\}_{i,m=1}^{N,M}$ are distributed according to Assumption 2.1. Then we have*

$$\mathcal{N}_1(\varepsilon, \mathcal{G}^s_{r,K_M}, \rho_M) \leqslant 3 \cdot \left(\frac{12r\bar{a}B_\phi}{\varepsilon}\right)^{2rd} \left(\frac{24eB_\phi}{\varepsilon}\right)^{2K_M(s+1)+2}. \tag{B.22}$$

Proof. Recall the matrix class defined in Definition 2.3:

$$\mathcal{A}_d(r, \bar{a}) := \left\{A \in \mathbb{R}^{d\times d} : \mathrm{rank}(A) \leqslant r,\ \|A\|_{\mathrm{op}} \leqslant \bar{a}\right\}.$$

Write $A = QK^\top$ via the truncated SVD, where $Q = U_r\Sigma_r^{1/2} \in \mathbb{R}^{d\times r}$ and $K = V_r\Sigma_r^{1/2} \in \mathbb{R}^{d\times r}$, with singular values belonging to $[0, \bar{a}]$, and $U_r, V_r^\top$ are semi unitary matrices of size $d \times r$, and Σ_r is a diagonal matrix of size $r \times r$. Then

$$\|Q\|_F^2 = \mathrm{Tr}(\Sigma_r) \leqslant r\bar{a}, \qquad \|K\|_F^2 \leqslant r\bar{a}.$$

Indeed, let $\delta > 0$ the δ-covering of the matrix class $\mathcal{Q}_{rd}(\sqrt{r\bar{a}}) := \{Q \in \mathbb{R}^{d\times r} : \|Q\|_F \leqslant \sqrt{r\bar{a}}\}$ is equivalent to the δ-covering of $B_{rd}(\sqrt{r\bar{a}})$, a centered ball with radius $\sqrt{r\bar{a}}$ in $\mathbb{R}^{rd}$ and (Vershynin, 2018, Corollary 4.2.11) implies that

$$n = \mathcal{N}(\epsilon, \mathcal{Q}_{rd}(\sqrt{r\bar{a}}), \|\cdot\|_F) = \mathcal{N}(\epsilon, B_{rd}(\sqrt{r\bar{a}})) \leqslant \left(\frac{3\sqrt{r\bar{a}}}{\epsilon}\right)^{rd}. \tag{B.23}$$

Notice that for $A_1 = Q_1 K_1^\top$ and $A_2 = Q_2 K_2^\top$ with $\{Q_i \in \mathcal{Q}_{rd}(\sqrt{r\bar{a}}), K_i \in \mathcal{Q}_{rd}(\sqrt{r\bar{a}})\}_{i=1}^2$ such that $\|Q_1 - Q_2\|_F \leqslant \delta/(2\sqrt{\bar{a}})$, $\|K_1 - K_2\|_F \leqslant \delta/(2\sqrt{\bar{a}})$, we have

$$\|A_1 - A_2\|_{\text{op}} = \|Q_1 K_1^\top - Q_2 K_2^\top\|_{\text{op}} \leqslant \|Q_1\|_{\text{op}} \|K_1 - K_2\|_F + \|Q_1 - Q_2\|_F \|K_2\|_{\text{op}} \leqslant \delta .$$

Moreover, by Assumption 2.1 that X_i, X_j lie within the unit unit ball and the assumption that $\phi \in \Phi^s_{K_M}$, a degree-s piecewise-polynomial approximation with K_M intervals, we get:

$$\left|\phi(\langle x, y\rangle_{A_1}) - \phi(\langle x, y\rangle_{A_2})\right| \leqslant B_\phi \|A_1 - A_2\|_{\text{op}} \leqslant B_\phi \delta$$

since $\|x\|, \|y\| \leqslant 1$. This proves that if A_1, A_2 are within δ in operator norm, the corresponding functions differ by at most $B_\phi \delta$. Thus,

$$\mathcal{N}_1(2B_\phi\, \delta, \mathcal{G}^s_{r,K_M}, \rho_M) \leqslant \sum\nolimits_{i,j\neq i}^n \mathcal{N}_1(B_\phi\, \delta, \{\phi(\langle x, y\rangle_{Q_i K_j^\top}) : \phi \in \Phi^s_{K_M}\}, \rho_M) . \tag{B.24}$$

On the other hand, (Györfi et al., 2006, Theorem. 9.4–9.5) shows the following bound

$$\mathcal{N}_1(B_\phi\, \delta, \Phi^s_{K_M}, \rho_M) \leqslant 3 \left(\frac{6e(B_\phi + 1)}{B_\phi\, \delta}\right)^{2K_M(s+1)+2} \leqslant 3 \left(\frac{12e}{\delta}\right)^{2K_M(s+1)+2}$$

for the empirical measure ρ_M in Definition 2.1 with $\{X^m \in \mathcal{C}_d\}_{m=1}^M$. Putting it back to equation B.24, we obtain that

$$\mathcal{N}_1(2B_\phi\, \delta, \mathcal{G}^s_{r,K_M}, \rho_M) \leqslant 3 \cdot \left(\frac{6r\bar{a}}{\delta}\right)^{2rd} \left(\frac{12e}{\delta}\right)^{2K_M(s+1)+2} . \tag{B.25}$$

Now re-parameterize by $\varepsilon = 2B_\phi \delta$, i.e. $\delta = \varepsilon/2B_\phi$, and absorb constants in equation B.25. This gives our desired estimate in equation B.22. □

Remark B.4 *As a by-product, we show that a $\frac{\delta}{2\sqrt{r\bar{a}}}$-cover for Q and K induces a δ-cover for $\mathcal{A}_d(r, \bar{a})$ in operator norm. Taking all pairs $Q_i K_j^\top$ and substituting $\epsilon = \frac{\delta}{2\sqrt{r\bar{a}}}$ in equation B.23 give that*

$$\mathcal{N}(\delta, \mathcal{A}(r, \bar{a}), \|\cdot\|_{\text{op}}) \leqslant \left(\frac{6r\bar{a}}{\delta}\right)^{2rd} .$$

C Appendix: lower bound proofs

Lemma C.1 (Continuous density of the bilinear form.) *Let $(X, Y) \in \mathbb{R}^{2d}$ have a joint density $p \in L^1(D)$ with $D \subset \mathbb{R}^{2d}$ being a bounded open set. Let $A \in \mathbb{R}^{d\times d}$ have rank $r \geqslant 1$, and define $U = X^\top A Y$. Then:*

(i) (Existence) *For every $r \geqslant 1$, the law of U is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$ with a density denoted by p_U.*

(ii) (Continuity) *If $r \geqslant 2$, then $p_U \in C(\mathbb{R})$.*

Note that $r \geqslant 2$ is *sharp* for p_U to be continuous: for $r = 1$, continuity at 0 may not hold: if X, Y are independent standard Gaussian in $\mathbb{R}$ and $A = I$, then $U = XY$ has density $p_U(u) = \frac{1}{\pi} K_0(|u|)$, where $K_0(x) \sim -\log x$ as $x \downarrow 0$, so p_U is singular at 0.

Proof. The proof consists of three steps: reduction to a canonical quadratic form on $\mathbb{R}^{2r}$, existence of the density, and continuity.

Step 1. Reduction to the canonical quadratic form on $\mathbb{R}^{2r}$. Let $A = W\Sigma V^\top$ be a singular value decomposition with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, where $\sigma_i > 0$ and $W, V \in \mathbb{R}^{d\times d}$ are orthonormal. Set $\alpha := W^\top X$, $\beta := V^\top Y$. Orthonormal changes preserve absolute continuity, so (α, β) has a joint density $\tilde{p}(\alpha, \beta) = p(W\alpha, V\beta)$ with support $\tilde{D} = \{(\alpha, \beta) = (W^{-1}x, V^{-1}y) : (x, y) \in D\}$, which is a bounded subset in $\mathbb{R}^{2d}$. Split $\alpha = (\alpha^{(r)}, \alpha^\perp)$ and $\beta = (\beta^{(r)}, \beta^\perp)$, where the superscript (r) denotes the first r coordinates. Then

$$U = X^\top A Y = \sum_{i=1}^r \sigma_i\, \alpha_i^{(r)}\, \beta_i^{(r)}.$$

Integrating out $(\alpha^{\perp}, \beta^{\perp})$ yields a marginal density $q \in L^1(\mathbb{R}^{2r})$ for $Z := (\alpha^{(r)}, \beta^{(r)})$ and it has a bounded support, which we denote by Ω. Thus it suffices to work in $\mathbb{R}^{2r}$ with

$$\Phi(\alpha, \beta) := \sum_{i=1}^{r} \sigma_i \alpha_i \beta_i, \qquad U = \Phi(Z), \qquad Z \sim q \in L^1(\Omega).$$

Step 2. Existence by the coarea formula. For any bounded measurable test function $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\varphi(U)] \;=\; \int_{\Omega} \varphi(\Phi(z))\, q(z)\, dz. \tag{C.1}$$

Note that Φ has gradient $\nabla\Phi(\alpha, \beta) = (\sigma_1\beta_1, \ldots, \sigma_r\beta_r, \sigma_1\alpha_1, \ldots, \sigma_r\alpha_r)$, which is Lipschitz continuous on Ω. Applying the Coarea formula (i.e., for any $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ Lipschitz and $g \in L^1_{\text{loc}}(\mathbb{R}^n)$, $\int_{\mathbb{R}^n} g(z)\, |\nabla f(z)|\, dz \;=\; \int_{\mathbb{R}} \left(\int_{f^{-1}(u)} g(z)\, d\mathcal{H}^{n-1}(z) \right) du$, where $\mathcal{H}^{n-1}(z)$ denotes the Hausdorff measure, see, e.g., Evans (2018)) to $f = \Phi$ with $g(z) = q(z)\, \varphi(\Phi(z))/|\nabla\Phi(z)|$ gives

$$\int_{\Omega} \varphi(\Phi(z))\, q(z)\, dz = \int_{\mathbb{R}} \left(\int_{\Phi^{-1}(u)} \frac{q(z)}{|\nabla\Phi(z)|}\, d\mathcal{H}^{2r-1}(z) \right) \varphi(u)\, du.$$

Hence, $p_U(u) = \int_{\Phi^{-1}(u)} \frac{q(z)}{|\nabla\Phi(z)|}\, d\mathcal{H}^{2r-1}(z)$ for $u \neq 0$.

Note that under the change of variables $z = \sqrt{|u|}\, w$, the Hausdorff surface measure scales by $|u|^{(2r-1)/2}$ and $|\nabla\Phi|$ by $|u|^{1/2}$. Then, for $u \neq 0$, the above equation can be written as

$$\int_{\Omega} \varphi(\Phi(z))\, q(z)\, dz = \int_{\mathbb{R}} |u|^{r-1} \left(\int_{\Phi(w)=\text{sign}(u)} \frac{q(\sqrt{|u|}\, w)}{|\nabla\Phi(w)|}\, d\mathcal{H}^{2r-1}(w) \right) \varphi(u)\, du. \tag{C.2}$$

Comparing equation C.1 and equation C.2, the push-forward measure is absolutely continuous with density

$$p_U(u) = |u|^{r-1} \int_{\Phi(w)=\text{sign}(u)} \frac{q(\sqrt{|u|}\, w)}{|\nabla\Phi(w)|}\, d\mathcal{H}^{2r-1}(w) \tag{C.3}$$

for all $u \neq 0$. This proves (i) for all $r \geqslant 1$.

Step 3. Continuity. Let $\xi_U(t) = \mathbb{E}[e^{itU}] = \int_{\mathbb{R}^{2r}} e^{it\Phi(z)}\, q(z)\, dz$ be the characteristic function. The phase $t\Phi(z)$ is a non-degenerate quadratic form with constant Hessian $H = t \begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix}$ (of full rank $2r$). By the standard stationary phase bound for quadratic phases (see, e.g., (Sogge, 2017, Theorem 1.1.4))

$$|\xi_U(t)| \;\leqslant\; C\,(1 + |t|)^{-r},$$

with C depending on q (e.g. if $q \in C_c^{\infty}$, then C depends on a finite number of derivatives; and it extends to general $q \in L^1$ since C_c^{∞} is dense in L^1). Hence, if $r \geqslant 2$, then $\xi_U \in L^1(\mathbb{R})$ and Fourier inversion yields a bounded continuous density $p_U(u) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itu} \xi_U(t)\, dt$. □

Next, we provide the proof of Lemma 4.1.

Proof of Lemma 4.1. Consider

$$U_{ij} = X_i^{\top} A_{\star} X_j, \qquad V_{ij} = X_i^{\top} \hat{A} X_j,$$

so that $\hat{g}(X_i, X_j) = \hat{\phi}(V_{ij})$ and $g^{\star}(X_i, X_j) = \phi^{\star}(U_{ij})$. Recall that $p_{U_{ij}}$ is the density of U_{ij} and $p_U = \frac{1}{N(N-1)} \sum_{i,j:i \neq j} p_{U_{ij}}$. Also, recall that the following functions are defined in equation 4.2:

$$\widehat{\psi}_{ij}(u) := \mathbb{E}\big[\widehat{\phi}(V_{ij}) \big| U_{ij} = u\big], \quad \hat{\psi}(u) := \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \frac{p_{U_{ij}(u)}}{N(N-1) p_U(u)} \hat{\psi}_{ij(u)}.$$

Since $\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \frac{p_{U_{ij}}(u)}{N(N-1)p_U(u)} = 1$, we have, by applying Jensen's inequality,

$$\begin{aligned} \big|\hat{\psi}(u) - \phi_{\star}(u)\big|^2 &= \Big|\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \frac{p_{U_{ij}}(u)}{N(N-1) p_U(u)} \widehat{\psi}_{ij}(u) - \phi_{\star}(u)\Big|^2 \\ &\leqslant \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \frac{p_{U_{ij}}(u)}{N(N-1) p_U(u)} \big|\widehat{\psi}_{ij}(u) - \phi_{\star}(u)\big|^2. \end{aligned} \tag{C.4}$$

Also, by applying Jensen's inequality to the conditional expectation, we have

$$\mathbb{E}\big[\big|\widehat{\phi}(V_{ij}) - \phi_\star(U_{ij})\big|^2\big] = \mathbb{E}[\mathbb{E}\big[\big|\widehat{\phi}(V_{ij}) - \phi_\star(U_{ij})\big|^2 | U_{ij}\big]]$$
$$\geqslant \mathbb{E}\big[\big|\mathbb{E}\big[\widehat{\phi}(V_{ij}) - \phi_\star(U_{ij}) | U_{ij}\big]\big|^2\big] = \int_{-\bar{a}}^{\bar{a}} \big|\widehat{\psi}_{ij}(u) - \phi_\star(u)\big|^2 p_{U_{ij}}(u) du.$$

Averaging over the pairs as in equation C.4, we have

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \mathbb{E}\big[\big|\widehat{\phi}(V_{ij}) - \phi_\star(U_{ij})\big|^2\big]$$
$$\geqslant \int_{-\bar{a}}^{\bar{a}} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \big|\widehat{\psi}_{ij}(u) - \phi_\star(u)\big|^2 \frac{p_{U_{ij}}(u)}{p_U(u)} p_U(u) du$$
$$\geqslant \int_{-\bar{a}}^{\bar{a}} \big|\hat{\psi}(u) - \phi_\star(u)\big|^2 p_U(u)\, du,$$

which is the desired inequality. □

Proof of Lemma 4.2 We construct $\bar{K} = \lceil c_{0,N} M^{\frac{1}{2\beta+1}} \rceil$ disjoint equidistance intervals.

$$\{\Delta_\ell = (r_\ell - h_M, r_\ell + h_M)\}_{\ell=1}^{\bar{K}}, \quad \text{with } h_M = \frac{L_0}{8 n_0 \bar{K}}, \tag{C.5}$$

where $\{r_\ell\}$, n_0 and L_0 are specific values that will be determined below. We will define the intervals by separating into two cases: one where the density of p_U is bounded below by $\underline{a}_0 > 0$ and one where it is not.

If $p_U(u) \geqslant \underline{a}_0 > 0$, we can simply use the uniform partition of supp(p_U) to obtain the desired $\{\Delta_\ell\}$. That is, we set $n_0 = 1$, $L_0 = 4$, and $r_\ell = -\bar{a} + (2\ell - 1)h_M$. If p_U is not bounded away from zero, we shall build the partition based on its continuity. Since p_U is continuous on $[-\bar{a}, \bar{a}]$, the constant $a_0 = \sup_{x\in[-\bar{a},\bar{a}]} p_U(x)$ exists, now consider $\underline{a}_0 < a_0 \wedge 1$. We can construct the $\bar{K}$ intervals described in equation C.5 which satisfy the following $\bigcup_\ell \Delta_\ell \subset A_0 := \{u \in [-\bar{a}, \bar{a}] : p_U(u) > \underline{a}_0\}$.

Let $L_0 := \frac{1-2\underline{a}_0}{a_0 - \underline{a}_0}$. Since for all $u \in A_0$, $p_U(u) \leqslant a_0$ and for all $u \in A_0^c$, $p_U(u) \leqslant \underline{a}_0$, together with the fact that $1 = \int_{A_0} p_U(u)du + \int_{A_0^c} p_U(u)du$, we get:

$$1 \leqslant a_0 \text{Leb}(A_0) + \underline{a}_0(2\bar{a} - \text{Leb}(A_0)) \Rightarrow L_0 \leqslant \text{Leb}(A_0) \leqslant 2\bar{a}. \tag{C.6}$$

Also, note that the set A_0 is open by continuity of p_U. Thus, there exist disjoint intervals (a_j, b_j) such that $A_0 = \bigcup_{j=1}^{\infty}(a_j, b_j)$. Without loss of generality, we assume that these intervals are descendingly ordered according to their length $b_j - a_j$. Let

$$n_0 = \min\{n : \sum_{j=1}^{n}(b_j - a_j) > \frac{L_0}{2}\}. \tag{C.7}$$

One can see that $n_0 > 1$. Now, we construct the first n_1 disjoint intervals $\{\Delta_\ell = (r_\ell - h_M, r_\ell + h_M)\}_{\ell=1}^{n_1} \subset (a_1, b_1)$ such that $r_\ell = a_1 + \ell h_M$ and $n_1 = \lfloor \frac{b_1 - a_1}{2h_M} \rfloor$. If $n_1 \geqslant \bar{K}$, we stop. Otherwise, we construct additional disjoint intervals $\{\Delta_\ell = (r_\ell - h_M, r_\ell + h_M)\}_{\ell=n_1+1}^{n_1+n_2} \subset (a_2, b_2)$ similarly, and continue to (a_j, b_j) until obtaining $\bar{K}$ intervals $\{\Delta_\ell\}$.

To show that we will at least obtain $\bar{K}$ such intervals, we show that $K_\star \geqslant \bar{K}$, where $K_\star$ is the total number of intervals $\{\Delta_\ell\}_{\ell=1}^{K_\star}$. Since the Lebesgue measure of $(a_j, b_j) \backslash \bigcup_{\ell=1}^{K_\star} \Delta_\ell$ is less than $2h_M$ for each j, the Lebesgue measure of the uncovered parts $\bigcup_{j=1}^{n_0}(a_j, b_j) \backslash \big(\bigcup_{\ell=1}^{K_\star} \Delta_\ell\big)$ is at most $2n_0 h_M$.

Thus, by equation C.7 the intervals $\{\Delta_\ell\}_{\ell=1}^{K_\star}$ must have a total length no less than $\frac{L_0}{2} - 2n_0 h_M$. And since each of them is in length of $2h_M$ the total number must satisfy:

$$K_\star \geqslant (\frac{L_0}{2} - 2n_0 h_M)/(2h_M)$$

and plugging in the definition of h_M from equation C.5 we get:

$$K_\star \geqslant 2\bar{K}n_0 - n_0 \geqslant \bar{K}.$$

Now we construct hypothesis functions satisfying Conditions (D1)–(D3). We first define $2^{\bar{K}}$ functions, from which we will select a subset of 2s-separated hypothesis functions,

$$\phi_\omega(u) = \sum_{\ell=1}^{\bar{K}} \omega_\ell \psi_{\ell,M}(u), \quad \omega = (\omega_1, \cdots, \omega_{\bar{K}}) \in \{0,1\}^{\bar{K}},$$

where the basis functions are

$$\psi_{\ell,M}(u) := L h_M^\beta \psi\Big(\tfrac{u-r_\ell}{h_M}\Big), \quad u \in [-\bar{a}, \bar{a}] \tag{C.8}$$

with $\psi(u) = e^{-\frac{1}{1-(2u)^2}}\mathbf{1}_{|u|\leqslant 1/2}$. Note that the support of $\psi_{\ell,M}(u)$ is Δ_ℓ, and $\int_{\Delta_\ell} |\psi_{\ell,M}(u)|^2 du = L^2 h_M^{2\beta+1}\|\psi\|_2^2$. By definition, these hypothesis functions satisfy Condition (D1), i.e., they are Holder continuous and

$$\|\psi_{\ell,M}\|_\infty \leqslant L h_M^\beta \leqslant L\|p_U\|_\infty^{-\beta} \leqslant L(2\bar{a})^{-\beta} \leqslant B_\phi$$

since $h_M = \frac{L_0}{8n_0\bar{K}} < L_0 \leqslant \frac{1}{a_0}$ with $a_0 = \|p_U\|_\infty$ and $\|p_U\|_\infty \leqslant \frac{1}{2\bar{a}}$.

Then, denoting $\phi_k(x) = \phi_{\omega^{(k)}}(x)$, we proceed to verify Conditions (D2)–(D3). Next, we select a subset of $2s_{N,M}$-separated functions $\{\phi_{k,M} := \phi_{\omega^{(k)}}\}_{k=1}^K$ satisfying Condition (D2), i.e., $\|\phi_{\omega^{(k)}} - \phi_{\omega^{(k')}}\|_{L^2_{p_U}} \geqslant 2s_{N,M}$ for any $k \neq k' \in \{1,\ldots,K\}$. Here $s_{N,M} = C_1 c_{0,N}^{-\beta} M^{-\frac{\beta}{2\beta+1}}$ with C_1 being a positive constant to be determined below. Since $\Delta_\ell = \text{supp}(\psi_{\ell,M}) \subseteq \Delta_\ell$ are disjoint, we have

$$\begin{aligned}\|\phi_\omega - \phi_{\omega'}\|_{L^2_{p_U}} &= \left(\int_{\mathbb{R}} \left|\sum_{\ell=1}^{\bar{K}} (\omega_\ell - \omega'_\ell)\psi_{\ell,M}(u)\right|^2 p_U(u)du\right)^{\frac{1}{2}} \\ &= \left(\sum_{\ell=1}^{\bar{K}} (\omega_\ell - \omega'_\ell)^2 \int_{\Delta_\ell} |\psi_{\ell,M}(u)|^2 p_U(u)du\right)^{\frac{1}{2}}.\end{aligned}$$

Since $p_U(u) \geqslant \underline{a}_0$ over each Δ_ℓ, we have

$$\int_{\Delta_\ell} |\psi_{\ell,M}(u)|^2 p_U(u)du \geqslant \underline{a}_0 \int_{\Delta_\ell} |\psi_{\ell,M}(u)|^2 du = \underline{a}_0 L^2 h_M^{2\beta+1}\|\psi\|_2^2.$$

Applying the Varshamov-Gilbert bound (Tsybakov, 2008, Lemma 2.9), one can obtain a subset $\{\omega^{(k)}\}_{k=1}^K$ with $K \geqslant 2^{\bar{K}/8}$ such that $\sum_\ell^{\bar{K}} (\omega_\ell^{(k)} - \omega_\ell^{(k')})^2 \geqslant \frac{\bar{K}}{8}$ for any $k \neq k' \in \{1,\ldots,K\}$. Thus,

$$\begin{aligned}\|\phi_\omega - \phi_{\omega'}\|_{L^2_{p_U}} &\geqslant \sqrt{\underline{a}_0} L\|\psi\|_2 \sqrt{\frac{\bar{K}}{8}} \left(\frac{L_0}{8n_0\bar{K}}\right)^{\beta+1/2} \\ &= \sqrt{\underline{a}_0} L\|\psi\|_2 \frac{\bar{K}^{1/2}}{2\sqrt{2}} \left(\frac{L_0}{8n_0}\right)^{\beta+1/2} \bar{K}^{-(\beta+1/2)} \\ &= \left(\frac{\sqrt{\underline{a}_0} L\|\psi\|_2}{2\sqrt{2}} \left(\frac{L_0}{8n_0}\right)^{\beta+1/2}\right) \bar{K}^{-\beta} = s_{N,M}\end{aligned}$$

where $s_{N,M} = C_1 c_{0,N}^{-\beta} M^{-\frac{\beta}{2\beta+1}}$ with

$$C_1 := \frac{\sqrt{\underline{a}_0} L\|\psi\|_2}{4\sqrt{2}} \left(\frac{L_0}{8n_0}\right)^{\beta+1/2}.$$

To verify condition (D3) for each fixed dataset $\{X^m\}_{m=1}^M$, we first compute the Kullback-Leibler (KL) divergence. Define $u_{ij}^m := (X_i^m)^\top A X_j^m$. Then for each m,

$$R_\phi[X^m]_i = \frac{1}{N-1}\sum_{j\neq i} \phi(u_{ij}^m).$$

Under the hypothesis $\phi_{k,M}$, the density of the outputs $\{Y^m\}_{m=1}^M$ is

$$p_k(y^1,\ldots,y^M) = \prod_{m=1}^M p_\eta\Big(y^m - R_{\phi_{k,M}}[X^m]\Big),$$

where $y^m \in \mathbb{R}^d$ represents the observed output Y^m. By definition of KL divergence and the i.i.d. noise assumption,

$$D_{\mathrm{KL}}\big(\bar{\mathbb{P}}_k, \bar{\mathbb{P}}_0\big) = \int\cdots\int \log \prod_{m=1}^M \frac{p_\eta(y^m)}{p_\eta\Big(y^m + R_{\phi_{k,M}}[X^m]\Big)} \prod_{m=1}^M p_\eta(y^m)\mathrm{d}y^m.$$

This simplifies to

$$D_{\mathrm{KL}}\big(\bar{\mathbb{P}}_k, \bar{\mathbb{P}}_0\big) = \sum_{m=1}^M \int_{\mathbb{R}^d} \log\Big[\tfrac{p_\eta(y^m)}{p_\eta\big(y^m + R_{\phi_{k,M}}[X^m]\big)}\Big] p_\eta(y^m)\mathrm{d}y^m.$$

Finally, by the noise smoothness assumption 2.2, for each m,

$$\int p_\eta(y)\log\Big[\tfrac{p_\eta(y)}{p_\eta(y+v)}\Big]\mathrm{d}y \leqslant c_\eta\|v\|^2,$$

where $v = R_{\phi_{k,M}}[X^m]$. Summing over $m = 1,\ldots,M$ yields

$$D_{\mathrm{KL}}\big(\mathbb{P}_k, \mathbb{P}_0\big) \leqslant c_\eta \sum_{m=1}^M \big\|R_{\phi_{k,M}}[X^m]\big\|^2, \tag{C.9}$$

Employing Jensen's inequality, we have

$$\|R_{\phi_{k,M}}[X^m]\|^2 = \sum_{i=1}^N\Big(\tfrac{1}{N-1}\sum_{j\neq i}\phi_{k,M}(u_{ij}^m)\Big)^2 \leqslant \sum_{i=1}^N \tfrac{1}{N-1}\sum_{j\neq i}\big|\phi_{k,M}(u_{ij}^m)\big|^2 = \tfrac{1}{N-1}\sum_{i=1}^N\sum_{j\neq i}\big|\phi_{k,M}(u_{ij}^m)\big|^2. \tag{C.10}$$

Recalling that $\phi_{k,M}(u_{ij}^m) = \sum_{\ell=1}^{\bar{K}} \omega_\ell^{(k)}\psi_{\ell,M}(u_{ij}^m)$, where $\mathrm{supp}(\psi_{\ell,M}) \subseteq \Delta_\ell$ are disjoint and $|\psi_{\ell,M}(u_{ij}^m)| = Lh_M^\beta\psi\left(\frac{u_{ij}^m - r_\ell}{h_M}\right) \leqslant Lh_M^\beta\|\psi\|_\infty \mathbf{1}_{\{u_{ij}^m\in\Delta_\ell\}}$, we have

$$|\phi_{k,M}(u_{ij}^m)|^2 = \sum_{\ell=1}^{\bar{K}}\omega_\ell^{(k)}\big|\psi_{\ell,M}(u_{ij}^m)\big|^2 \leqslant L^2 h_M^{2\beta}\|\psi\|_\infty^2 \sum_{\ell=1}^{\bar{K}}\mathbf{1}_{\{u_{ij}^m\in\Delta_\ell\}}, \tag{C.11}$$

where we have used the fact that $0 \leqslant \omega_\ell^{(k)} \leqslant 1$. By plugging in both equation C.10 and equation C.11 into equation C.9, we obtain

$$\begin{aligned}
D_{\mathrm{KL}}(\bar{\mathbb{P}}_k, \bar{\mathbb{P}}_0) &\leqslant \frac{c_\eta}{N-1}\sum_{m=1}^M\sum_{i=1}^N\sum_{j\neq i}\left(L^2 h_M^{2\beta}\|\psi\|_\infty^2 \sum_{\ell=1}^{\bar{K}}\mathbf{1}_{\{u_{ij}^m\in\Delta_\ell\}}\right)\\
&\leqslant \frac{c_\eta L^2\|\psi\|_\infty^2 h_M^{2\beta}}{N-1}\sum_{i,j,m}\left(\sum_{\ell=1}^{\bar{K}}\mathbf{1}_{\{u_{ij}^m\in\Delta_\ell\}}\right).
\end{aligned}$$

Since the intervals $\{\Delta_\ell\}$ are disjoint, the inner sum is at most 1. The total sum over i, j, m is therefore bounded by N^2M, which gives:

$$D_{\mathrm{KL}}(\bar{\mathbb{P}}_k, \bar{\mathbb{P}}_0) \leqslant c_\eta L^2\|\psi\|_\infty^2 NMh_M^{2\beta}.$$

Hence, by assigning $h_M = \frac{L_0}{8n_0\bar{K}}$ from equation C.5 and $\bar{K} = \lceil c_{0,N}M^{\frac{1}{2\beta+1}}\rceil$, we obtain

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^K D_{\mathrm{KL}}(\bar{\mathbb{P}}_k, \bar{\mathbb{P}}_0) &\leqslant \left(c_\eta L^2\|\psi\|_\infty^2\left(\frac{L_0}{8n_0}\right)^{2\beta}\right)N\left(\frac{\bar{K}}{c_{0,N}}\right)^{2\beta+1}\bar{K}^{-2\beta}\\
&= \left(\frac{c_\eta L^2\|\psi\|_\infty^2 N}{c_{0,N}^{2\beta+1}}\left(\frac{L_0}{8n_0}\right)^{2\beta}\right)\bar{K} \leqslant \alpha\log K
\end{aligned}$$

with $\alpha = \left(\frac{c_\eta L^2 \|\psi\|_\infty^2 N}{c_{0,N}^{2\beta+1}} \left(\frac{L_0}{8n_0}\right)^{2\beta} \right) \frac{8}{\log 2}$ since $K \geqslant 2^{\bar{K}/8}$. Thus, for condition (D3) to hold, i.e., $\alpha < 1/8$, we need

$$c_{0,N}^{2\beta+1} \geqslant 64 c_\eta L^2 \|\psi\|_\infty^2 N \left(\frac{L_0}{8n_0}\right)^{2\beta}.$$

Following $c_{0,N} = C_0 N^{\frac{1}{2\beta+1}}$, it suffices to set C_0 to be

$$C_0 := (32 c_\eta L^2 \|\psi\|_\infty^2 \left(\frac{L_0}{8n_0}\right)^{2\beta})^{\frac{1}{2\beta+1}}.$$

□

To prove the lower bound minimax rate, we will use the following lower bound for hypothesis test error, see e.g., Proposition 2.3 Tsybakov (2008) or Lemma 4.3 in Wang et al. (2025).

Lemma C.2 (Lower bound for hypothesis test error) *Let $\Theta = \{\theta_k\}_{k=0}^K$ with $K \geqslant 2$ be a set of $2s$-separated hypotheses, i.e., $d(\theta_k, \theta_{k'}) \geqslant 2s > 0$ for all $0 \leqslant k < k' \leqslant K$, for a given metric d on Θ. Denote $\mathbb{P}_k = \mathbb{P}_{\theta_k}$ and suppose they satisfy $\mathbb{P}_k \ll \mathbb{P}_0$ for each $k \geqslant 1$ and*

$$\frac{1}{K+1} \sum_{k=1}^K D_{\mathrm{KL}}(\mathbb{P}_k, \mathbb{P}_0) \leqslant \alpha \log(K), \quad \text{with } 0 < \alpha < 1/8. \tag{C.12}$$

Then, the average probability of the hypothesis testing error has a lower bound:

$$\inf_{k_{\text{test}}} \frac{1}{K+1} \sum_{k=0}^K \mathbb{P}_k\big(k_{\text{test}} \neq k\big) \geqslant \frac{\log(K+1) - \log(2)}{\log(K)} - \alpha, \tag{C.13}$$

where $\inf_{k_{\text{test}}}$ denotes the infimum over all tests.

Proof of Theorem 4.3 We aim to apply Tsybakov's method to simplify probability bounds by considering a finite set of hypothesis functions. Reducing the supremum over $\mathcal{C}^\beta(L, \bar{a})$ to the finite set of hypothesis functions, and applying the Markov inequality, we obtain

$$\begin{aligned}
&\sup_{\substack{\phi_\star \in \mathcal{C}^\beta(L,\bar{a}) \\ \|\phi_\star\|_\infty \leqslant B_\phi}} \mathbb{E}_{\phi_\star}\left[\|\widehat{\phi}_M - \phi_\star\|_{L^2_{p_U}}^2\right] \\
&\geqslant \max_{\phi_{k,M} \in \{\phi_{0,M} \dots \phi_{K,M}\}} \mathbb{E}_{\phi_{k,M}}\left[\|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}}^2\right] \\
&\geqslant \max_{\phi_{k,M} \in \{\phi_{0,M} \dots \phi_{K,M}\}} s_{N,M}^2 \mathbb{P}_{\phi_{k,M}}\left[\|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}} > s_{N,M}\right] \\
&\geqslant s_{N,M}^2 \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}_{X^1,\dots,X^M}\left[\mathbb{P}_{\phi_{k,M}}\left(\|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}} > s_{N,M} \Big| X^1, \dots, X^M\right)\right],
\end{aligned} \tag{C.14}$$

where the last inequality follows since the maximal value over the functions is no less than the average and since $\mathbb{P}(A) = \mathbb{E}[1_A] = \mathbb{E}_Z[\mathbb{E}[1_A|Z]] = \mathbb{E}[\mathbb{P}(A|Z)]$.

Next, we transform to bounds in the average probability of testing error of the $2s_{N,M}$-separated hypothesis functions. Define k_{test} as the minimum distance test:

$$k_{\text{test}} = \arg\min_{k=0,\dots,K} \|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}}.$$

Since $\phi_{k_{\text{test}},M}$ is the closest one, we have that $\|\widehat{\phi}_M - \phi_{k_{\text{test}},M}\|_{L^2_{p_U}} \leqslant \|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}}$ for all $k \neq k_{\text{test}}$. Using the fact that the function $\phi_{k,M}$ are built as $2s_{N,M}$ separated functions and using the triangle inequality we have:

$$2s_{N,M} \leqslant \|\phi_{k,M} - \phi_{k_{\text{test}},M}\|_{L^2_{p_U}} \leqslant \|\widehat{\phi}_M - \phi_{k_{\text{test}},M}\|_{L^2_{p_U}} + \|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}} \leqslant 2\|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}},$$

so $s_{N,M} \leqslant \|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}}$ for all $k \neq k_{test}$. Hence,

$$\mathbb{P}_{\phi_{k,M}}\left(\|\widehat{\phi}_M - \phi_{k,M}\|_{L^2_{p_U}} \geqslant s_{N,M} \Big| X^1, \cdots, X^M\right) \geqslant \mathbb{P}\left(k_{\text{test}} \neq k \big| X^1, \cdots, X^M\right). \tag{C.15}$$

Consequently,

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{P}_{\phi_{k,M}}\left(\|\widehat{\phi}_M-\phi_{k,M}\|_{L^2_{p_U}}\geqslant s_{N,M}\mid X^1,\cdots,X^M\right)$$
$$\geqslant\inf_{k_{\text{test}}}\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{P}_{\phi_{k,M}}\left(k_{\text{test}}\neq k\mid X^1,\cdots,X^M\right)=\inf_{k_{\text{test}}}\frac{1}{K+1}\sum_{k=0}^{K}\bar{\mathbb{P}}_k\left(k_{\text{test}}\neq k\right)\tag{C.16}$$

where $\bar{\mathbb{P}}_k(\cdot)=\mathbb{P}_{\phi_{k,M}}(\cdot\mid X^1,\ldots,X^M)$.

The Kullback divergence estimate in equation (D3) from Lemma 4.2 holds with $0<\alpha<1/8$, and by Lemma C.2 and the fact that $K=2^{\lceil c_{0,N}M^{\frac{1}{2\beta+1}}\rceil}$ in equation 4.3 increases exponentially in M, we have:

$$\inf_{k_{\text{test}}}\frac{1}{K+1}\sum_{k=0}^{K}\bar{\mathbb{P}}_k\left(k_{\text{test}}\neq k\right)\geqslant\frac{\log(K+1)-\log(2)}{\log(K)}-\alpha\geqslant\frac{1}{2}\tag{C.17}$$

if M is large. Note that the above lower bound of $\inf_{k_{\text{test}}}\frac{1}{K+1}\sum_{k=0}^{K}\bar{\mathbb{P}}_k\left(k_{\text{test}}\neq k\right)$ is independent of the dataset $\{X^m\}_{m=1}^M$. Using equation C.17 ,equation C.16 and equation C.14, we obtain with $c_0=\frac{1}{2}[C_1C_0^{-\beta}]^2$,

$$\sup_{\phi_\star\in\mathcal{C}^s(L,\bar{a})}\mathbb{E}_{\phi_\star}\left[\|\widehat{\phi}_M-\phi_\star\|^2_{L^2_{p_U}}\right]\geqslant\frac{s^2_{N,M}}{2}=c_0N^{-\frac{2\beta}{2\beta+1}}M^{-\frac{2\beta}{2\beta+1}}\tag{C.18}$$

for any estimator. Hence, the lower bound equation 4.3 holds. □

Proof of Theorem 4.4 First, we reduce the supremum over all $A_\star$ to a single one. Let $A^1\in\mathcal{A}_d(r,\bar{a})$ with $\operatorname{rank}(A^1)\geqslant 2$. Since

$$\mathcal{G}_{A^1}:=\left\{g_{\phi,A^1}(x,y)=\phi(x^\top A^1y)\ :\phi\in\mathcal{C}^\beta(L,\bar{a}),\ \|\phi\|_\infty\leqslant B_\phi\right\}\subseteq\mathcal{G}_r^\beta(L,B_\phi,\bar{a}),$$

we have for any $\widehat{g}$,

$$\sup_{g_\star\in\mathcal{G}_r^\beta(L,B_\phi,\bar{a})}\mathbb{E}\|\widehat{g}-g_\star\|^2_{L^2_\rho}\geqslant\sup_{g_\star\in\mathcal{G}_{A^1}}\mathbb{E}\|\widehat{g}-g_\star\|^2_{L^2_\rho}.\tag{C.19}$$

Thus, to prove equation 4.5, it suffices to prove it with $g_\star\in\mathcal{G}_{A^1}$.

Let U^1 be the random variable defined in (4.1) with $A_\star=A^1$. Then, Lemma 4.1 implies that

$$\|\widehat{g}-g_\star\|^2_{L^2_\rho}\geqslant\|\hat{\psi}-\phi_\star\|^2_{L^2_{p_{U^1}}}$$

for any $\widehat{g}(x,y):=\widehat{\phi}(x^\top\widehat{A}y)$ with $\widehat{\phi}\in L^2_{p_{U^1}}$ and $\widehat{A}\in\mathcal{A}_d(r,\bar{a})$ and any $g_\star\in\mathcal{G}_{A^1}$. Here, $\hat{\psi}$, defined in equation 4.2, varies according to $\widehat{g}$ since both $A_\star$ and the distribution of X are fixed. Taking first the expectation over $\widehat{g}$, then taking the supremum over $g_\star\in\mathcal{G}_{A^1}$ followed by the infimum over $\widehat{A}$ and $\widehat{\phi}$, we obtain

$$\inf_{\substack{\widehat{A}\in\mathcal{A}_d(r,\bar{a})\\ \widehat{\phi}\in L^2_{p_{U^1}}}}\sup_{g_\star\in\mathcal{G}_{A^1}}\mathbb{E}\|\widehat{g}-g_\star\|^2_{L^2_\rho}\geqslant\inf_{\hat{\psi}\in L^2_{p_{U^1}}}\sup_{\substack{\phi_\star\in\mathcal{C}^\beta(L,\bar{a})\\ \|\phi_\star\|_\infty\leqslant B_\phi}}\mathbb{E}\|\hat{\psi}-\phi_\star\|^2_{L^2_{p_{U^1}}}.\tag{C.20}$$

Meanwhile, Theorem 4.3 gives a lower bound

$$\liminf_{M\to\infty}\inf_{\hat{\psi}\in L^2_{p_{U^1}}}\sup_{\substack{\phi_\star\in\mathcal{C}^\beta(L,\bar{a})\\ \|\phi_\star\|_\infty\leqslant B_\phi}}M^{\frac{2\beta}{2\beta+1}}\mathbb{E}\left\|\hat{\psi}-\phi_\star\right\|^2_{L^2_{p_{U^1}}}\geqslant c_0N^{-\frac{2\beta}{2\beta+1}}\tag{C.21}$$

with $c_0>0$.

Combining (C.19)–(C.21), we then obtain:

$$\liminf_{M\to\infty}\inf_{\widehat{g}}\sup_{g_\star\in\mathcal{G}_r^\beta(L,B_\phi,\bar{a})}M^{\frac{2\beta}{2\beta+1}}\mathbb{E}\left[\|\widehat{g}-g_\star\|^2_{L^2_\rho}\right]$$
$$\geqslant\liminf_{M\to\infty}\inf_{\widehat{g}}\sup_{g_\star\in\mathcal{G}_{A^1}}M^{\frac{2\beta}{2\beta+1}}\mathbb{E}\left[\|\widehat{g}-g_\star\|^2_{L^2_\rho}\right]$$
$$\geqslant\liminf_{M\to\infty}\inf_{\hat{\psi}\in L^2_{p_{U^1}}}\sup_{\substack{\phi_\star\in\mathcal{C}^\beta(L,\bar{a})\\ \|\phi_\star\|_\infty\leqslant B_\phi}}M^{\frac{2\beta}{2\beta+1}}\mathbb{E}\left\|\hat{\psi}-\phi_\star\right\|^2_{L^2_{p_{U^1}}}\geqslant c_0N^{-\frac{2\beta}{2\beta+1}},$$

which gives the desired result in equation 4.5. □

D NUMERICAL SIMULATIONS CONFIGURATION

This section provides a detailed description of the simulations presented in Section 5.

D.1 DATA GENERATION

For each sample size M, we run a Monte Carlo simulation over different seeds as follows. We draw token arrays $X^{(m)} = (X_1^{(m)}, \ldots, X_N^{(m)}) \in \mathcal{C}_d^N$ i.i.d. with $X_i^{(m)} \sim \mathrm{Unif}[0,1]^d/\sqrt{d}$ sampled i.i.d and construct the $(X_i^{(m)})^\top A_\star X_j^{(m)}$ terms, evaluate the interaction via $\phi_\star$ (the sampling method of $\phi_\star$ and $A_\star$ is detailed below, and aggregate and add i.i.d. noise $\eta_i^{(m)} \sim \mathcal{N}(0, \sigma^2)$ as described in equation 2.1 to generate $Y_i^{(m)}$.

For each simulation, we sample the ground truth interaction $g_\star(x,y) = \phi_\star(x^\top A_\star y)$ by drawing random $\phi_\star$ and choosing $A_\star$. We represent $\phi_\star$ as a B-spline of degree $P_\star$ defined on an open uniform knots with K basis functions on $[-1, 1]$.

$$\phi_\star(u) = \sum_{k=1}^{K_\star} \theta_\star^k B_k(u).$$

For each seed, we draw $\theta_\star \sim \mathcal{N}(0, I_{K_\star})$ and then normalize it for $\|\theta_\star\| = \sqrt{K_\star}$.

D.2 ESTIMATOR

If $A_\star$ was known, the estimator can be computed by setting $\hat{A} = A_\star$ and setting $\hat{\phi}(u) = \sum_{k=1}^{K} \hat{\theta}_k B_k(u)$ with degree P_{est} and $\hat{\theta}$ chosen according to the ridge regression formula:

$$\hat{\theta} = \left(U^\top U + \lambda_\theta I\right)^{-1} U^\top y, \tag{D.1}$$

where $U = \left(U_{(m,i),k}\right) \in \mathbb{R}^{MN \times K}$ with

$$U_{(m,i),k} := \frac{1}{N-1} \sum_{j \neq i} B_k\left((X_i^{(m)})^\top A X_j^{(m)}\right) \tag{D.2}$$

and $y = \left(Y_i^{(m)}\right) \in \mathbb{R}^{MN \times 1}$.

However, since $A_\star$ is unknown, the joint estimation of (A, ϕ) is non-convex due to the composition $\phi(x^\top A y)$. To mitigate local minima, we use a hot start and an alternating scheme. We perform the hot start by setting $A^{(0)} = A_\star + \Delta_A$ with Δ_A being a perturbation specified in Table 1 and setting the initial $\theta^{(0)}$ as the matching ridge solution D.1. In the PyTorch implementation, the scheme includes a description of the function $\hat{\phi}$ as a neural network. This is because representing it as B-splines directly would require differentiating through the B-spline basis, which is cumbersome for automatic differentiation. To address this, we introduce a neural-network surrogate Φ_{net} that approximates the spline and can be used as a differentiable link in the A-step.

Alternating Optimization for $\hat{\phi}, \hat{A}$

1. **Hot start:** set $A^{(0)} = A_\star + \Delta_A$ and compute $\theta^{(0)} = \left(U^\top U + \lambda_\theta I\right)^{-1} U^\top y$ with U computed according to $A^{(0)}$ in equation D.2
2. **For** $t = 1, \ldots, T$**:**
   (a) *Approximate the current spline using a multilayer perceptron (MLP).* Fit an MLP $\Phi_{\text{net}}^{(t-1)}$ on a grid $\{u_\ell\}$ to minimize $\sum_\ell |\Phi_{\text{net}}^{(t-1)}(u_\ell) - \tilde{\phi}^{(t-1)}(u_\ell)|^2$

(b) *A-step through optimization.* Update A by minimizing the empirical loss with $\hat{\phi} = \Phi_{\text{net}}^{(t-1)}$ held fixed using the Adam optimizer

$$\min_A \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} \Big(\sum_{j \neq i} \hat{g}_{A, \Phi_{\text{net}}^{(t-1)}}(X_i^{(m)}, X_j^{(m)}) - Y_i^{(m)} \Big)^2 + \frac{\lambda_A}{2} \|A\|_F^2. \quad \text{(D.3)}$$

(c) *θ-step through closed form.* With A fixed at $A^{(t)}$, compute $\theta^{(t)}$ by ridge regression: stack $y \in \mathbb{R}^{MN}$ from $Y_i^{(m)}$, and build $U \in \mathbb{R}^{MN \times K}$ with

$$U_{(m,i),k} = \frac{1}{N-1} \sum_{j \neq i} B_k\big((X_i^{(m)})^\top A^{(t)} X_j^{(m)}\big)$$

and compute

$$\theta^{(t)} = \big(U^\top U + \lambda_\theta I\big)^{-1} U^\top y.$$

Choice of K_{est} and λ_θ. We set the number of spline coefficients by the bias variance trade-off for a β-Hölder smoothness as done in equation B.20

$$K_{\text{est}} = \text{round}\Big(K_{\text{scale}} (M/\log M)^{1/(2\beta+1)}\Big)$$

where K_{scale} is a chosen constant. For the ridge regularization constant λ_θ we follow the standard scaling for least squares models with MN responses and K coefficients, the variance of $\hat{\theta}$ should scale like $K/(MN)$, so we take

$$\lambda_\theta = \lambda_{\text{scale}} \frac{K_{\text{est}}}{M(N-1)},$$

D.3 Error estimate

We measure accuracy via the estimator test MSE, sampling never seen inputs $X^{(m)} \sim \text{Unif}[0,1]^d/\sqrt{d}$ and evaluating:

$$\text{MSE}_g = \frac{1}{N_{\text{test}}} \sum_{m=1}^{N_{\text{test}}} \frac{1}{N(N-1)} \sum_{i=1}^{N} \Big| \sum_{j \neq i} \hat{g}_{\hat{A}, \hat{\phi}}(X_i^{(m)}, X_j^{(m)}) - g_\star(X_i^{(m)}, X_j^{(m)}) \Big|^2.$$

D.4 Simulation parameters

The following table details the parameters used for the simulations described in Section 5.

Table 1: Chosen parameters for the simulation

Parameter	**Value**
Seeds	300
$A_\star$	Diagonal matrix with i.i.d. entries $A_{11} = 1,\ \forall i > 1\ A_{ii} \sim \mathrm{Unif}[-1, 1]$
Sample sizes M	[20000, 27355, 37416, 51177, 70000]
N	3
Gaussian noise std σ_η	0.07 (Gaussian)
Estimator degree	$P_{\text{est}} = P_\star$
$K_\star$	16
K_{scale}	[(a) and (b) for $P_\star = 3$]: 16 [(b) for $P_\star = 8$]: 30
Basis size K_{est}	[(a) and (b) for $P_\star = 3$]: $\{73, 78, 82, 87, 92\}$ (matching the M grid) [(b) for $P_\star = 8$]: $\{50, 51, 52, 53, 54\}$ (matching the M grid)
λ_A	10^{-5}
λ_{scale}	2
λ_θ	[(a) and (b) for $P_\star = 3$] $10^{-3} \times \{6.85, 5.30, 4.12, 3.19, 2.46\}$ [(b) for $P_\star = 8$]$\{2.50, 1.86, 1.39, 1.04, 0.77\}$ (matching the M grid)
Δ_A	Entry wise Gaussian noise with an std of $5/d \times 10^{-7}$
T	4
A-step optimizer	Adam, lr $= 10^{-8}$, 20 epochs.
$\Phi_{\text{net}}^{(t)}$ architecture	1-hidden layer of width 32 (GELU activation)
$\Phi_{\text{net}}^{(t)}$ optimization	1000 epochs (Adam) lr = 0.01
Test set	2000 samples