Novelty Assessment Report

Paper: Dimension-Free Minimax Rates for Learning Pairwise Interactions in Attention-Style Models
PDF URL: https://openreview.net/pdf?id=7Gfheg6seM
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

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.

Core Task Landscape

This paper addresses: **Learning Pairwise Interactions in Attention-Style Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Foundations and Statistical Properties**
- **Architectural Innovations and Mechanisms**
- **Application Domains**

Complete Taxonomy Tree

- Learning Pairwise Interactions in Attention-Style Models Survey Taxonomy
- Theoretical Foundations and Statistical Properties
  - Convergence and Generalization Theory ★ (2 papers)
  - [0] Dimension-Free Minimax Rates for Learning Pairwise Interactions in Attention-Style Models (Anon et al., 2026) View paper
  - [27] A Theoretical Study of (Hyper) Self-Attention through the Lens of Interactions: Representation, Training, Generalization (Qu, 2025) View paper
  - Representational Capacity and Equivalence (2 papers)
  - [6] On the Relationship between Self-Attention and Convolutional Layers (Jean-Baptiste Cordonnier, 2022) View paper
  - [9] The Origin of Self-Attention: Pairwise Affinity Matrices in Feature Selection and the Emergence of Self-Attention (Roffo, 2025) View paper
- Architectural Innovations and Mechanisms
  - Position Encoding and Structural Representations (2 papers)
  - [5] Unifying topological structure and self-attention mechanism for node classification in directed networks (Yue Peng, 2025) View paper
  - [48] PaTH Attention: Position Encoding via Accumulating Householder Transformations (yang songlin, 2025) View paper
  - General Attention Mechanisms and Variants (4 papers)
  - [14] Exploring self-attention for image recognition (Hengshuang Zhao, 2020) View paper
  - [21] Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions (Yichao Liu, 2021) View paper
  - [32] Sa-net: Shuffle attention for deep convolutional neural networks (Qing Long Zhang, 2021) View paper
  - [38] An attention mechanism module with spatial perception and channel information interaction (Yifan Wang, 2024) View paper
  - Graph-Based Attention and Relational Networks (3 papers)
  - [15] Hypergraph convolution and hypergraph attention (Song Bai, 2021) View paper
  - [20] Representing long-range context for graph neural networks with global attention (Wu, 2021) View paper
  - [31] Graph Structure Inference with BAM: Neural Dependency Processing via Bilinear Attention (Philipp Froehlich, 2024) View paper
  - Efficiency and Scalability Enhancements (2 papers)
  - [40] Longer Attention Span: Increasing Transformer Context Length With Sparse Graph Processing Techniques (Nathaniel Tomczak, 2025) View paper
  - [50] IAFormer: Interaction-Aware Transformer network for collider data analysis (Esmail, 2025) View paper
- Application Domains
  - Computer Vision Tasks
  - Fine-Grained Recognition and Pairwise Comparison (6 papers)
    - [1] Learning attentive pairwise interaction for fine-grained classification (Peiqin Zhuang, 2020) View paper
    - [2] Learning Pairwise Interaction for Generalizable DeepFake Detection (Ying Xu, 2023) View paper
    - [16] PAST: Pairwise attention swin transformer for offline signature verification: Y.-J. Xiong et al. (YJ Xiong, 2025) View paper
    - [17] PAST: Pairwise attention swin transformer for offline signature verification (Yu-Jie Xiong, 2025) View paper

    - [22] Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits (Raksit Raksasat, 2023) View paper
    - [47] Fine grained image recognition algorithm based on cnn-transformer and paired interaction (Yu Wang, 2024) View paper
  - General Image Recognition and Retrieval (4 papers)
    - [29] A Transformer Architecture based mutual attention for Image Anomaly Detection (Mengting Zhang, 2023) View paper
    - [33] All the attention you need: Global-local, spatial-channel attention for image retrieval (Song, 2022) View paper
    - [34] Pairwise dependency-based robust ensemble pruning for facial expression recognition (Xing Chen, 2024) View paper
    - [45] EG-VAN: A Global and Local Attention based Dual-Branch Ensemble Network with Advanced color balancing for Multi-Class Skin Cancer Recognition (Adnan Saeed, 2025) View paper
  - Object Detection and Interaction Recognition (3 papers)
    - [8] Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer (Frederic Z. Zhang, 2022) View paper
    - [35] Relation-mining self-attention network for skeleton-based human action recognition (Kumie Gedamu, 2023) View paper
    - [49] Relation-aware global attention for person re-identification (Zhizheng Zhang, 2020) View paper
  - Biomedical and Molecular Interaction Prediction (8 papers)
  - [3] Predicting drug-drug interaction with graph mutual interaction attention mechanism. (Xiao-ying Yan, 2024) View paper
  - [4] Predicting drug-target affinity by discovering pairwise interactions using cross attention network (S. Falahati, 2025) View paper
  - [7] AKA-SafeMed: A safe medication recommendation based on attention mechanism and knowledge augmentation (Xiaomei Yu, 2024) View paper
  - [13] Inferring genotype-phenotype maps using attention models (Rijal, 2025) View paper
  - [23] ReduMixDTI: Prediction of Drug-Target Interaction with Feature Redundancy Reduction and Interpretable Attention Mechanism (Mingqing Liu, 2024) View paper
  - [24] A prediction method of interaction based on Bilinear Attention Networks for designing polyphenol-protein complexes delivery systems (Zhipeng Wang, 2024) View paper
  - [30] A dual graph neural network for drugâ□□drug interactions prediction based on molecular structure and interactions (Mei Ma, 2023) View paper
  - [41] PiLSL: pairwise interaction learning-based graph neural network for synthetic lethality prediction in human cancers (Xin Liu, 2022) View paper
  - Multimodal and Cross-Domain Applications (4 papers)
  - [25] Murel: Multimodal relational reasoning for visual question answering (Cadene, 2019) View paper
  - [26] Exploring Pairwise Relationships Adaptively From Linguistic Context in Image Captioning (Zongjian Zhang, 2021) View paper
  - [37] Exploring global diverse attention via pairwise temporal relation for video summarization (Li Ping, 2021) View paper
  - [46] Exploring region relationships implicitly: Image captioning with visual relationship attention (Zongjian Zhang, 2021) View paper
  - Natural Language Processing and Structured Text (3 papers)
  - [39] Modeling dynamic pairwise attention for crime classification over legal articles (Pengfei Wang, 2018) View paper
  - [42] A Multi-channel Character Relationship Classification Model Based on Attention Mechanism (Yuhao Zhao, 2022) View paper
  - [44] Deep Biaffine Attention for Neural Dependency Parsing (Timothy Dozat, 2021) View paper
  - Temporal and Sequential Data Modeling (2 papers)
  - [19] Learning spatialâ□□temporal pairwise and high-order relationships for short-term passenger flow prediction in urban rail transit (Jinxin Wu, 2024) View paper
  - [43] Attention based vehicle trajectory prediction (Kaouther Messaoud, 2020) View paper
  - Recommendation and Personalization Systems (2 papers)
  - [18] Learn from others and be yourself in federated human activity recognition via attention-based pairwise collaborations (Can Bu, 2024) View paper
  - [28] Deep pairwise learning for user preferences via dual graph attention model in location-based social networks (Weihua Gong, 2023) View paper
  - Structured Data and Tensor Modeling (3 papers)
  - [10] Attention-mechanism-based neural latent-factorization-of-tensors model (Xiuqin Xu, 2025) View paper
  - [11] Not all features deserve attention: Graph-guided dependency learning for tabular data generation with language models (Zheyu Zhang, 2025) View paper
  - [12] Regularized Pairwise Relationship based Analytics for Structured Data (Zhaojing Luo, 2023) View paper
  - Neuroscience and Biological Systems (1 papers)
  - [36] Brain-state mediated modulation of inter-laminar dependencies in visual cortex (Anirban Das, 2024) View paper

Narrative

Core task: learning pairwise interactions in attention-style models. The field organizes around three main branches. Theoretical Foundations and Statistical Properties investigates convergence guarantees, generalization bounds, and the mathematical underpinnings of attention mechanisms that capture pairwise dependencies, as exemplified by Dimension-Free Minimax Pairwise[0] and Hyper Self-Attention Theory[27]. Architectural Innovations and Mechanisms explores novel designs for encoding interactions—ranging from bilinear attention modules like BAM Bilinear Attention[31] and Shuffle Attention[32] to unary-pairwise decompositions such as Unary-Pairwise Transformer[8] and specialized structures like Topological Self-Attention Networks[5]. Application Domains demonstrates how these interaction-learning techniques solve real-world problems in drug discovery (Drug Interaction Mutual Attention[3], SafeMed Attention Knowledge[7]), computer vision (Attentive Pairwise Fine-Grained[1], Pairwise DeepFake Detection[2]), and multimodal reasoning (Murel Multimodal Reasoning[25], Pairwise Linguistic Image Captioning[26]).

A particularly active line of work focuses on balancing expressiveness with computational efficiency: some methods adopt explicit pairwise representations to capture fine-grained relationships (Relation-Mining Self-Attention[35], IAFormer Interaction-Aware[50]), while others pursue factorized or low-rank approximations to scale gracefully (Attention Latent Factorization[10], Global-Local Spatial-Channel Attention[33]). Dimension-Free Minimax Pairwise[0] sits squarely within the convergence and generalization theory cluster, providing statistical guarantees that complement the empirical designs prevalent in architectural branches. Compared to nearby theoretical studies like Hyper Self-Attention Theory[27], which examines higher-order dependencies, Dimension-Free Minimax Pairwise[0] emphasizes dimension-independent bounds for classical pairwise attention, offering a rigorous foundation for understanding when and why these models generalize effectively across diverse interaction patterns.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. A Theoretical Study of (Hyper) Self-Attention through the Lens of Interactions: Representation, Training, Generalization

Authors: Qu, Guannan | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Self-attention has emerged as a core component of modern neural architectures, yet its theoretical underpinnings remain elusive. In this paper, we study self-attention through the lens of interacting entities, ranging from agents in multi-agent reinforcement learning to alleles in genetic sequences, and show that a single layer linear self-attention can efficiently represent, learn, and generalize functions capturing pairwise interactions, including out-of-distribution scenarios. Our analysis re...

Relationship Analysis

Both papers belong to the Convergence and Generalization Theory category, analyzing theoretical properties of learning pairwise interactions in attention-style models. They share overlapping focus on establishing convergence guarantees and generalization bounds for attention mechanisms that capture token-token interactions. However, the original paper establishes dimension-free minimax rates (M^(-2β/(2β+1))) for learning interaction functions in an interacting particle system framework with general nonlinear activations, while the candidate paper focuses on gradient flow convergence and length generalization for linear self-attention under a mutual interaction perspective, providing conditions for zero training error and out-of-distribution generalization rather than minimax optimality rates.

Contributions Analysis

Overall novelty summary. The paper establishes minimax convergence rates for learning pairwise interactions in single-layer attention models, proving a rate of M^(-2β/(2β+1)) that depends only on activation smoothness β and is independent of token count, ambient dimension, or weight matrix rank. Within the taxonomy, it resides in the Convergence and Generalization Theory leaf under Theoretical Foundations and Statistical Properties, alongside one sibling paper (Hyper Self-Attention Theory). This leaf represents a sparse research direction with only two papers, indicating that rigorous statistical analysis of attention mechanisms remains relatively underexplored compared to the broader field's emphasis on architectural innovations and applications.

The taxonomy reveals a field heavily weighted toward Architectural Innovations (fifteen papers across four sub-branches) and Application Domains (thirty-one papers across eight sub-branches), while Theoretical Foundations contains only four papers total. The sibling paper examines higher-order dependencies in hyper self-attention, whereas this work focuses on classical pairwise interactions with dimension-free guarantees. Neighboring branches like Representational Capacity and Equivalence (two papers) investigate what functions attention can express, but do not provide convergence rate analysis. The scope notes clarify that this leaf excludes empirical studies and application-specific models, concentrating purely on sample complexity and generalization bounds.

Among twenty-three candidates examined, the dimension-free minimax rate contribution shows no clear refutation across four candidates reviewed. However, the connection to interacting particle systems appears less novel, with three refutable candidates among ten examined, and the inverse problem well-posedness claim encounters two refutable candidates among nine reviewed. The limited search scope—top-K semantic matches plus citation expansion—means these statistics reflect a targeted sample rather than exhaustive coverage. The core statistical contribution appears more distinctive than the auxiliary theoretical connections, which have more substantial prior work in related mathematical frameworks.

Based on the limited literature search, the paper's primary novelty lies in establishing dimension-independent convergence guarantees for attention-style pairwise learning, addressing a gap in a sparsely populated theoretical branch. The auxiliary contributions on particle systems and inverse problems show greater overlap with existing work among the candidates examined. The analysis covers top-twenty-three semantic matches and does not claim exhaustive field coverage, particularly for adjacent mathematical literatures in statistical learning theory or dynamical systems.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Dimension-free minimax convergence rate for attention-style models

Description: The authors establish that the optimal convergence rate for learning pairwise interactions in attention-style models is M^{-2β/(2β+1)}, where this rate depends solely on the activation function's smoothness parameter β and is independent of embedding dimension, number of tokens, or weight matrix rank, demonstrating freedom from the curse of dimensionality.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning theory for inferring interaction kernels in second-order interacting agent systems

URL: View paper

Brief Assessment

Interaction Kernels Second-Order[70] focuses on second-order interacting particle systems with different mathematical structures (heterogeneous, multivariable models), not attention-style models with embedding matrices and activation functions as studied in the original paper.

---

2. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories

URL: View paper

Brief Assessment

Stochastic Interaction Kernels[73] focuses on learning interaction kernels in stochastic particle systems from trajectory data, not on attention mechanisms or transformer architectures. The paper addresses a fundamentally different problem domain (interacting particle systems) with different mathematical structures and does not challenge the novelty of dimension-free rates for attention-style models.

---

3. Minimax prediction in tree Ising models

URL: View paper

Brief Assessment

Minimax Tree Ising[72] focuses on parameter estimation in tree-structured Ising models for accurate marginal computation, not on learning pairwise interactions in attention-style neural network models with nonlinear activations.

---

4. High-dimensional adaptive minimax sparse estimation with interactions

URL: View paper

Brief Assessment

Adaptive Minimax Sparse[71] focuses on high-dimensional sparse linear regression with interaction effects under heredity conditions, not on attention-style models or token-based architectures. The dimension-free property in [71] relates to sparsity constraints in regression, not to embedding dimensions in attention mechanisms.

Contribution 2: Connection between transformers and interacting particle systems

Description: The authors formulate attention mechanisms as interacting particle systems where tokens are viewed as particles, enabling theoretical analysis of the inverse problem of recovering interaction functions from aggregated observations. This framework extends beyond standard independent, isotropic token distribution assumptions to handle dependent and anisotropic data.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. IAFormer: Interaction-Aware Transformer network for collider data analysis

URL: View paper

Brief Assessment

IAFormer Interaction-Aware[50] focuses on collider physics applications using transformer networks with pairwise particle interactions for jet tagging tasks, not on theoretical analysis of attention mechanisms as interacting particle systems or inverse problems for recovering interaction functions from aggregated observations.

---

2. Geometric Dynamics of Signal Propagation Predict Trainability of Transformers

URL: View paper

Brief Assessment

Geometric Dynamics Trainability[58] focuses on analyzing signal propagation dynamics and trainability conditions in transformers through a particle system lens, rather than addressing the inverse problem of learning pairwise interactions from aggregated observations that is central to the original paper's contribution.

---

3. Jet tagging with more-interaction particle transformer

URL: View paper

Brief Assessment

More-Interaction Particle Transformer[52] focuses on jet tagging in particle physics using attention mechanisms for particle interactions, not on the theoretical framework of transformers as interacting particle systems for inferring nonlinear attention interactions or statistical learning theory.

---

4. Attention to the strengths of physical interactions: Transformer and graph-based event classification for particle physics experiments

URL: View paper

Brief Assessment

Physical Interactions Transformer[53] applies transformers to particle physics event classification, not to the theoretical framework of viewing attention mechanisms as interacting particle systems for statistical inference. The candidate focuses on incorporating physics-motivated features into transformer architectures for experimental data analysis, while the original establishes a mathematical connection between transformers and IPS models to derive minimax convergence rates for learning pairwise interactions.

---

5. The Mean-Field Dynamics of Transformers

URL: View paper

Prior Art Analysis

Mean-Field Dynamics Transformers[57] demonstrates that the connection between transformers and interacting particle systems was established prior to the original paper's submission. The candidate explicitly develops a mathematical framework interpreting transformer attention as an interacting particle system and studies its continuum limits, directly addressing the same conceptual framework. Both papers view tokens as particles and formulate attention mechanisms through this lens, though they pursue different analytical goals (the candidate focuses on clustering dynamics while the original focuses on learning pairwise interactions).

Evidence

Evidence 1 - **Rationale**: Both papers explicitly establish the same conceptual connection between transformers and interacting particle systems, with tokens viewed as particles. The candidate's abstract directly states this framework was developed in their work, predating the original paper's claim to novelty. - **Original**: 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. - **Candidate**: we develop a mathematical framework that interprets transformer attention as an interacting particle system and studies its continuum (mean-field) limits.

Evidence 2 - **Rationale**: Both papers use the IPS framework to analyze transformer behavior beyond simple assumptions. The candidate demonstrates a fully developed mathematical framework connecting transformers to established IPS theory, showing this connection was already established in prior work. - **Original**: 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. - **Candidate**: by idealizing attention on the sphere, we connect transformer dynamics to wasserstein gradient flows, synchronization models (kuramoto), and mean-shift clustering.

Evidence 3 - **Rationale**: The original paper claims to establish this connection as a novel contribution, but the candidate paper already developed this mathematical framework, demonstrating that the connection between transformers and IPS was established in prior work. - **Original**: 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. - **Candidate**: we develop a mathematical framework that interprets transformer attention as an interacting particle system and studies its continuum (mean-field) limits.

---

6. Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View

URL: View paper

Prior Art Analysis

Multi-Particle Dynamic Transformers[54] demonstrates that the connection between transformers and interacting particle systems was established prior to the original paper. The candidate paper explicitly formulates the transformer architecture as a multi-particle dynamic system where tokens are viewed as particles, providing both theoretical analysis and a mathematical framework for understanding attention mechanisms through this lens. This directly challenges the novelty claim that the original authors were the first to establish this connection.

Evidence

Evidence 1 - **Rationale**: Both papers use the identical conceptual framework of viewing tokens as particles. The candidate paper explicitly describes this particle interpretation and uses it to analyze transformer dynamics, demonstrating prior establishment of this connection. - **Original**: tokens are viewed as "particles," and the self-attention aggregates pairwise interactions between them. - **Candidate**: the feature (a.k.a, embedding) of words in a sequence can be considered as the initial positions of a collection of particles,

and the latent representations abstracted in stacked transformer layers can be viewed as the location of particles moving in a high-dimensional space at different time point...

Evidence 2 - **Rationale**: The candidate paper explicitly claims to be the first to show the relationship between transformers and multi-particle dynamic systems, providing a framework for understanding transformer processing through particle dynamics, which directly refutes the original paper's novelty claim. - **Original**: 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. - **Candidate**: in this paper, we provide a novel perspective towards understanding the architecture. in particular, we are the first to show that the transformer architecture is inherently related to multi-particle dynamic system (mpds) in physics. mpds is a well-established research field which aims at modeling how...

---

7. A unified perspective on the dynamics of deep transformers

URL: View paper

Brief Assessment

Unified Dynamics Transformers[56] focuses on modeling transformers as interacting particle systems to analyze dynamics through layers, but does not address the inverse problem of recovering interaction functions from aggregated observations, which is the core novelty claim of the original paper.

---

8. Clustering in Causal Attention Masking

URL: View paper

Brief Assessment

Clustering Causal Attention[55] focuses on causal attention masking in transformers as interacting particle systems, specifically analyzing clustering dynamics. The original paper addresses the inverse problem of recovering interaction functions from aggregated observations with dependent/anisotropic data, which is a different theoretical contribution than analyzing clustering behavior under causal masking.

---

9. Multi-Particle Dynamical Systems Modeling Transformers

URL: View paper

Brief Assessment

Multi-Particle Dynamical Systems[59] focuses on modeling transformers as continuous dynamical systems to study convergence and clustering behavior, not on the inverse problem of recovering interaction functions from aggregated observations that the original paper addresses.

---

10. The emergence of clusters in self-attention dynamics

URL: View paper

Prior Art Analysis

Clusters Self-Attention Dynamics[51] demonstrates that the connection between transformers and interacting particle systems was established prior to the original paper. Both papers explicitly formulate attention mechanisms as interacting particle systems where tokens are viewed as particles. The candidate paper states 'viewing transformers as interacting particle systems' and models the dynamics as 'an interacting particle system' with the same mathematical framework. The candidate also addresses the inverse problem of understanding interaction functions, though from a different analytical perspective (clustering dynamics rather than statistical inference). This prior work refutes the novelty claim of being the first to establish this connection.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly establish the same conceptual connection between transformers and interacting particle systems, with tokens viewed as particles. The candidate paper was published earlier and demonstrates this framework. - **Original**: 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. - **Candidate**: viewing transformers as interacting particle systems, we describe the geometry of learned representations when the weights are not time-dependent. we show that particles, representing tokens, tend to cluster toward particular limiting objects as time tends to infinity.

Evidence 2 - **Rationale**: Both papers use identical terminology and conceptual framework: tokens as particles in an interacting particle system. The candidate paper provides the mathematical formulation of this connection. - **Original**: 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. - **Candidate**: we view tokens as particles, and the transformer dynamics as an interacting particle system of the form 9xiptq= nÿ j= 1 pijptqvx jptq,t pr 0, '8q

Evidence 3 - **Rationale**: Both papers analyze transformer dynamics through the IPS lens to understand token interactions beyond standard assumptions, though they focus on different aspects (the candidate on clustering dynamics, the original on statistical inference). - **Original**: 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. - **Candidate**: the transformer dynamics considered in(1) does not contain a layer normalization mechanism typically encountered in practice [vsp`17]. in absence of such a device, tokens may diverge to infinity as in theorem2.1. in fact, the norm of the tokensxiptq typically diverges exponentially toward '8 for anyd

---

Contribution 3: Well-posedness of the inverse problem under coercivity condition

Description: The authors prove that the inverse problem of inferring interaction functions is well-posed under a coercivity condition, which they establish holds for a broad class of input distributions satisfying exchangeability and continuity assumptions, addressing the fundamental challenge of nonlocal dependency in attention mechanisms.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Uniqueness for the SchrÃ¶dinger equation with an inverse square potential and application to controllability and inverse problems

URL: View paper

Brief Assessment

Inverse Square Potential[69] addresses uniqueness for inverse source problems in Schrödinger equations with singular potentials, not the well-posedness of inverse problems for learning interaction functions in attention mechanisms under coercivity conditions.

---

2. On the Identifiability of Nonlocal Interaction Kernels in First-Order Systems of Interacting Particles on Riemannian Manifolds

URL: View paper

Prior Art Analysis

Identifiability Nonlocal Kernels[66] demonstrates that well-posedness of inverse problems under coercivity conditions for interaction functions was established prior to the original paper's work. The candidate paper proves well-posedness by establishing strict positivity of a related integral operator and demonstrates that the coercivity condition holds for general function classes, specifically addressing nonlocal dependency in interacting particle systems. Both papers tackle the fundamental challenge of nonlocal dependency, but the candidate's work on Riemannian manifolds predates the original paper's attention mechanism context.

Evidence

Evidence 1 - **Rationale**: Both papers prove well-posedness of inverse problems through coercivity conditions. The candidate establishes this for interaction kernels in particle systems, while the original applies it to attention mechanisms. - **Original**: 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. - **Candidate**: our methodology involves casting the learning problem as a linear statistical inverse problem using an operator theoretical framework. we prove the well-posedness of the inverse problem by establishing the strict positivity of a related integral operator

Evidence 2 - **Rationale**: Both papers address the fundamental challenge of nonlocal dependency in inverse problems for interaction functions, though in different application contexts. - **Original**: nonlocal dependency. the nonlocal dependence presents a challenge in estimating the interaction function. the forward operator rgrxsdepends 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 h... - **Candidate**: in this paper, we approach the identifiability of ϕ in (1) from a linear statistical inverse learning problem perspective [frt21]. that is, the observational data comes in the form of positions of the n particles on the manifold and the (possible noisy) first derivatives of these positions with resp...

---

3. Identifiability of interaction kernels in mean-field equations of interacting particles

URL: View paper

Brief Assessment

Identifiability Interaction Kernels[65] studies mean-field equations with infinite particles, while the original paper addresses attention mechanisms with finite tokens. The coercivity conditions and problem settings differ fundamentally between these two contexts.

---

4. Uniqueness principle for fractional (non)-coercive anisotropic polyharmonic operators and applications to inverse problems

URL: View paper

Brief Assessment

Fractional Coercive Polyharmonic[68] addresses inverse problems for poly-fractional operators in PDEs (recovering potentials, source functions, coefficients), not the well-posedness of inverse problems for learning interaction functions in attention-style models under coercivity conditions for token distributions.

---

5. Optimal minimax rate of learning interaction kernels

URL: View paper

Prior Art Analysis

Minimax Interaction Kernels[60] demonstrates that well-posedness of inverse problems under coercivity conditions for interaction functions was established prior to the original paper. The candidate paper proves that the inverse problem in the large sample limit is well-posed under a coercivity condition for a broad class of exchangeable distributions, addressing the fundamental challenge of nonlocal dependency. Both papers establish coercivity conditions for well-posedness of inverse problems involving interaction functions, though applied to different model formulations (radial kernels vs. attention-style bilinear forms).

Evidence

Evidence 1 - **Rationale**: Both papers establish the same minimax rate under coercivity conditions. The candidate paper's establishment of this rate under coercivity conditions demonstrates prior work on well-posedness under such conditions. - **Original**: we prove that the rate of m´ 2β 2β `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. - **Candidate**: this study establishes that the rate of m´ 2β 2β` 1 is the optimal minimax convergence rate under a coercivity condition that ensures the well-posedness of the inverse problem in the large sample limit.

---

6. The -coercivity approach for mixed problems

URL: View paper

Brief Assessment

Coercivity Mixed Problems[64] addresses well-posedness of mixed problems using T-coercivity for saddle-point formulations in PDEs, not inverse problems for inferring interaction functions in attention mechanisms. The technical domains and problem structures are fundamentally different.

---

7. Coercivity-based analysis and its application to an inverse source problem for a subdiffusion equation with time-dependent principal parts

URL: View paper

Brief Assessment

Coercivity Inverse Subdiffusion[63] addresses inverse source problems for subdiffusion equations with time-dependent principal parts, not the inference of interaction functions in attention-style models or interacting particle systems.

---

8. Minimax rate for learning kernels in operators

URL: View paper

Brief Assessment

Minimax Kernels Operators[62] addresses learning kernels in operators from function-valued data, focusing on deconvolution inverse problems with compact normal operators. The original paper studies attention mechanisms with token interactions, where the inverse problem involves estimating interaction functions from aggregated observations. These are fundamentally different problem settings with distinct mathematical structures.

---

9. Interacting Particle Systems on Networks: joint inference of the network and the interaction kernel

URL: View paper

Brief Assessment

Interacting Particle Networks[61] addresses joint inference of networks and interaction kernels in particle systems on networks, focusing on graph weight matrices and interaction functions. The ORIGINAL paper studies attention mechanisms in transformers with a different mathematical framework (pairwise interactions through weight matrices and nonlinear activations), not network inference problems.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Dimension-Free Minimax Rates for Learning Pairwise Interactions in Attention-Style Models View paper
- [1] Learning attentive pairwise interaction for fine-grained classification View paper
- [2] Learning Pairwise Interaction for Generalizable DeepFake Detection View paper
- [3] Predicting drug-drug interaction with graph mutual interaction attention mechanism. View paper
- [4] Predicting drug-target affinity by discovering pairwise interactions using cross attention network View paper
- [5] Unifying topological structure and self-attention mechanism for node classification in directed networks View paper
- [6] On the Relationship between Self-Attention and Convolutional Layers View paper
- [7] AKA-SafeMed: A safe medication recommendation based on attention mechanism and knowledge augmentation View paper
- [8] Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer View paper
- [9] The Origin of Self-Attention: Pairwise Affinity Matrices in Feature Selection and the Emergence of Self-Attention View paper
- [10] Attention-mechanism-based neural latent-factorization-of-tensors model View paper
- [11] Not all features deserve attention: Graph-guided dependency learning for tabular data generation with language models View paper
- [12] Regularized Pairwise Relationship based Analytics for Structured Data View paper
- [13] Inferring genotype-phenotype maps using attention models View paper
- [14] Exploring self-attention for image recognition View paper
- [15] Hypergraph convolution and hypergraph attention View paper
- [16] PAST: Pairwise attention swin transformer for offline signature verification: Y.-J. Xiong et al. View paper
- [17] PAST: Pairwise attention swin transformer for offline signature verification View paper
- [18] Learn from others and be yourself in federated human activity recognition via attention-based pairwise collaborations View paper
- [19] Learning spatialâ□temporal pairwise and high-order relationships for short-term passenger flow prediction in urban rail transit View paper
- [20] Representing long-range context for graph neural networks with global attention View paper
- [21] Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions View paper
- [22] Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits View paper
- [23] ReduMixDTI: Prediction of Drug-Target Interaction with Feature Redundancy Reduction and Interpretable Attention Mechanism View paper
- [24] A prediction method of interaction based on Bilinear Attention Networks for designing polyphenol-protein complexes delivery systems View paper
- [25] Murel: Multimodal relational reasoning for visual question answering View paper
- [26] Exploring Pairwise Relationships Adaptively From Linguistic Context in Image Captioning View paper
- [27] A Theoretical Study of (Hyper) Self-Attention through the Lens of Interactions: Representation, Training, Generalization View paper
- [28] Deep pairwise learning for user preferences via dual graph attention model in location-based social networks View paper
- [29] A Transformer Architecture based mutual attention for Image Anomaly Detection View paper
- [30] A dual graph neural network for drugâ□drug interactions prediction based on molecular structure and interactions View paper
- [31] Graph Structure Inference with BAM: Neural Dependency Processing via Bilinear Attention View paper
- [32] Sa-net: Shuffle attention for deep convolutional neural networks View paper
- [33] All the attention you need: Global-local, spatial-channel attention for image retrieval View paper
- [34] Pairwise dependency-based robust ensemble pruning for facial expression recognition View paper
- [35] Relation-mining self-attention network for skeleton-based human action recognition View paper
- [36] Brain-state mediated modulation of inter-laminar dependencies in visual cortex View paper
- [37] Exploring global diverse attention via pairwise temporal relation for video summarization View paper
- [38] An attention mechanism module with spatial perception and channel information interaction View paper
- [39] Modeling dynamic pairwise attention for crime classification over legal articles View paper
- [40] Longer Attention Span: Increasing Transformer Context Length With Sparse Graph Processing Techniques View paper
- [41] PiLSL: pairwise interaction learning-based graph neural network for synthetic lethality prediction in human cancers View paper
- [42] A Multi-channel Character Relationship Classification Model Based on Attention Mechanism View paper
- [43] Attention based vehicle trajectory prediction View paper
- [44] Deep Biaffine Attention for Neural Dependency Parsing View paper
- [45] EG-VAN: A Global and Local Attention based Dual-Branch Ensemble Network with Advanced color balancing for Multi-Class Skin Cancer Recognition View paper
- [46] Exploring region relationships implicitly: Image captioning with visual relationship attention View paper
- [47] Fine grained image recognition algorithm based on cnn-transformer and paired interaction View paper
- [48] PaTH Attention: Position Encoding via Accumulating Householder Transformations View paper
- [49] Relation-aware global attention for person re-identification View paper
- [50] IAFormer: Interaction-Aware Transformer network for collider data analysis View paper
- [51] The emergence of clusters in self-attention dynamics View paper
- [52] Jet tagging with more-interaction particle transformer View paper
- [53] Attention to the strengths of physical interactions: Transformer and graph-based event classification for particle physics experiments View paper
- [54] Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View View paper
- [55] Clustering in Causal Attention Masking View paper
- [56] A unified perspective on the dynamics of deep transformers View paper
- [57] The Mean-Field Dynamics of Transformers View paper
- [58] Geometric Dynamics of Signal Propagation Predict Trainability of Transformers View paper
- [59] Multi-Particle Dynamical Systems Modeling Transformers View paper
- [60] Optimal minimax rate of learning interaction kernels View paper

- [61] Interacting Particle Systems on Networks: joint inference of the network and the interaction kernel View paper
- [62] Minimax rate for learning kernels in operators View paper
- [63] Coercivity-based analysis and its application to an inverse source problem for a subdiffusion equation with time-dependent principal parts View paper
- [64] The -coercivity approach for mixed problems View paper
- [65] Identifiability of interaction kernels in mean-field equations of interacting particles View paper
- [66] On the Identifiability of Nonlocal Interaction Kernels in First-Order Systems of Interacting Particles on Riemannian Manifolds View paper
- [67] Learning collective behaviors from observation View paper
- [68] Uniqueness principle for fractional (non)-coercive anisotropic polyharmonic operators and applications to inverse problems View paper
- [69] Uniqueness for the SchrÃ¶dinger equation with an inverse square potential and application to controllability and inverse problems View paper
- [70] Learning theory for inferring interaction kernels in second-order interacting agent systems View paper
- [71] High-dimensional adaptive minimax sparse estimation with interactions View paper
- [72] Minimax prediction in tree Ising models View paper
- [73] Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories View paper