Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Spectral Initialization of Deep Networks: Aligning Input Gram Matrices With Target Kernels

R. Priyadharshini[1], Pradeep Kumar Shriwas[2], Yadavalli Suresh kumar[3], Dr. G. Karthi[4], Pachiyappan Chandrasakar[5] , Dr Sweetlin Susilabai S[6], Pericherla manoj[7], Dr.Rajeashwari Srinivasa Ragavan[8], T.Vengatesh[9*]

[1]Assistant Professor,Department of Computer Science,SRM Institute of Science & Technology, Ramapuram campus ,SRM Ramapuram ,SRMIST Ramapuram, EmailId:priyadhr5@srmist.edu.in,ORCID: https://orcid.org/0009-0005-4973-2610
[2]Department of Computer Science & Engineering,OP Jindal University, Raigarh, Chhattisgarh, Pin : 496001, EmailId: pradeep.shriwas@opju.ac.in,shriwasji.ps@gmail.com
[3]Assistant Professor,Department of Computer Science And Engineering,Aditya University, Surampalem, Andhra Pradesh, India, MailID:ysureshkumar24@gmail.com, ORCID: https://orcid.org/0009-0000-3296-7863
[4]Associate professor, Department of Artificial Intelligence and Data Science,, Saveetha Engineering College,Chennai. EmailId:gkarthi@saveetha.ac.in,Orcid ID: 0000-0003-2486-237X
[5]Assistant professor, Department of Artificial intelligence and data science, New Prince Shri Bhavani College of Engineering & Technology(An autonomous institution), Tambaram - Velachery Main Road,Santhosapuram, Chennai - 600 073 Mail id:pachiyappancs007@gmail.com
[6]Assistant Professor,Department of Cyber Security ,School of Applied Science, Faculty of Liberal Arts and Business Studies, SRM Institute of Science and Technology,
Ramapuram Campus, Chennai email: sweetlis1@srmist.edu.in,Orcid ID: 0000-0002-0889-4315,Scopus Author ID: 57215213821
[7] Department of Computer Science and Information Technology,SRKR Engineering College, Bhimavaram,Email ID:manoj.p@srkrec.ac.in
[8]Assistant Professor, Department of Computer Science,Sri S. Ramasamy Naidu Memorial College, Sattur - 626203Tamil Nadu,Email Id: rajeeragavan@gmail.com,Orcid ID: 0000-0002-0756-3733
[9*](Corresponding Author)Assistant Professor, Department of Computer Science, Government Arts and Science College, Veerapandi, Theni, Tamilnadu, India., Email id: venkibiotinix@gmail.com,Scopus ID: 57195476440,ORCID: 0000-0001-8717-3657

Abstract
Deep neural networks are highly sensitive to their initialization, particularly when network depth and parameter dimensionality increase. Conventional initialization methods such as Xavier, He, orthogonal, and Layer-Sequential Unit-Variance (LSUV) primarily control activation variance or singular-value behavior but do not explicitly incorporate the geometric structure of the training data. This paper proposes a Spectral Kernel Alignment Initialization (SKAI) framework that initializes the first-layer representation by aligning the input Gram matrix with a predefined target kernel. The central idea is to transform the input representation so that its spectral structure approximates a target positive-semidefinite kernel before conventional gradient-based training begins. Given an input matrix (X), its Gram matrix (G_X=XX^T) is decomposed spectrally, while a target kernel (K_T) is constructed from class similarity, radial-basis-function similarity, polynomial similarity, or a task-specific prior. A spectrally derived transformation is then obtained by solving a regularized matrix approximation problem, producing an initialized feature representation whose covariance and principal directions are better aligned with the target kernel. The proposed framework is incorporated into a deep architecture consisting of spectral preprocessing, kernel-aligned initialization, residual feature extraction, normalization, and classification layers. Experiments are proposed on MNIST, Fashion-MNIST, and CIFAR-10 using accuracy, convergence speed, training loss, Gram alignment error, condition number, and computational overhead as evaluation metrics. The proposed method is expected to provide faster convergence, improved early-stage optimization, and stronger kernel alignment than Xavier, He, orthogonal, and LSUV initialization. The study establishes a connection between spectral matrix approximation, kernel methods, and deep-network initialization.

Keywords: spectral initialization, Gram matrix, target kernel, deep neural networks, kernel alignment, eigenvalue decomposition, orthogonal initialization, neural tangent kernel, optimization, representation learning.

1. INTRODUCTION

The initialization of deep neural networks has emerged as a critical determinant of training success, particularly as architectures grow deeper and more complex. Poorly chosen initial weights can lead to vanishing or exploding gradients, slow convergence, or outright training failure. Conventional

initialization strategies Xavier , He , orthogonal , and Layer-Sequential Unit-Variance (LSUV) have been developed to address these challenges by controlling activation variance or ensuring orthogonality of weight matrices. Yet these methods share a fundamental limitation: they do not explicitly account for the geometric structure of the training data.

Recent theoretical advances have illuminated the importance of kernel-task alignment in determining neural network performance . Studies have shown that diagonalizing the kernel on the data-induced measure yields a series of features and eigenvalues, where those with high eigenvalues can be learned with the least training effort . This spectral bias the tendency of neural networks to learn low-frequency components first has profound implications for optimization dynamics . Moreover, the phenomenon of "silent alignment" demonstrates that the neural tangent kernel (NTK) can evolve in eigenstructure during early training, developing a low-rank contribution before the loss appreciably decreases .

These insights motivate a fundamental question: can we explicitly design initialization to align the network's representation with a target kernel structure, thereby accelerating learning and improving performance? This paper answers this question affirmatively through the Spectral Kernel Alignment Initialization (SKAI) framework.

The central contribution of this work is a principled initialization method that exploits the spectral structure of the input data and a predefined target kernel. Given an input matrix X, we compute its Gram matrix G_X = XX^T and perform an eigendecomposition. Simultaneously, we construct a target kernel K_T that encodes desired similarity structure whether from class labels, radial basis functions, polynomial kernels, or task-specific priors. We then solve a regularized matrix approximation problem to derive a transformation that aligns the spectral structure of the input representation with the target kernel. This produces an initial feature representation whose covariance and principal directions are better aligned with the learning objective before any gradient-based optimization begins.

The proposed framework is integrated into a deep architecture comprising spectral preprocessing, kernel-aligned initialization, residual feature extraction, normalization, and classification layers. We evaluate this approach on MNIST, Fashion-MNIST, and CIFAR-10, measuring accuracy, convergence speed, training loss, Gram alignment error, condition number, and computational overhead. Our method is expected to demonstrate superior performance compared to Xavier, He, orthogonal, and LSUV initialization, particularly in terms of convergence speed and final accuracy.

This work bridges three important research threads: spectral matrix approximation theory, kernel methods for representation learning, and deep network initialization strategies. By explicitly encoding kernel alignment into the initialization process, we establish a framework that is both theoretically grounded and practically effective.

2. LITERATURE SURVEY

2.1 Variance-Based Initialization

The Xavier initialization, introduced by Glorot and Bengio , was among the first principled approaches to network initialization. It derives weight variance bounds based on the number of input and output neurons under the assumption of linear activations, aiming to maintain activation variance across layers and prevent signal decay in deep architectures. The He initialization extended this analysis to ReLU nonlinearities, establishing that the optimal scaling factor for ReLU networks is approximately $\sqrt{(2/n\_in)}$, which accounts for the rectifier's effect on variance propagation.

While variance-based methods have proven broadly effective and remain widely used , they do not account for data distribution or task structure. The scaling factors are derived under simplifying assumptions that may not hold for complex datasets or architectures with normalization layers. Recent work has shown that these methods can lead to spectral collapse in certain configurations, particularly without batch normalization .

2.2 Orthogonal and Spectral Initialization

Saxe et al. demonstrated that orthogonal weight initialization significantly improves training dynamics in linear networks, as orthogonal matrices preserve gradient norms and prevent the rank collapse that occurs with Gaussian initialization. This approach has been extended to nonlinear networks, showing benefits in deeper architectures by maintaining dynamic isometric the property that singular values of the input-output Jacobian remain close to unity .

The LSUV method combines orthogonal initialization with data-driven variance normalization. By passing a single mini-batch through the network and scaling weights to achieve unit variance at each layer, LSUV achieves superior performance in deep architectures, particularly when batch normalization

is not employed. However, this procedure focuses on variance normalization rather than structural alignment with task-relevant features .
Recent work on spectral initialization for various architectures has shown that aligning singular spaces between models and target tasks can significantly improve fine-tuning performance . Studies on Physics-Informed Neural Networks (PINNs) have demonstrated that informative initialization strategies that tailor the initial weight distribution to the specific problem can effectively counteract operator-induced spectral bias, balancing convergence speeds across the frequency spectrum.

2.3 Neural Tangent Kernel and Spectral Bias

The Neural Tangent Kernel (NTK) framework has revolutionized our understanding of deep network training. In the infinite-width limit, the network's training dynamics are governed by a fixed kernel, and the early evolution of residuals follows the spectral decomposition of this kernel. Specifically, modes associated with large eigenvalues decay quickly, while those with small eigenvalues decay slowly, leading to "spectral bias" . This bias explains why networks often struggle to learn high-frequency components of target functions.
The "silent alignment" phenomenon reveals that even in finite-width networks, the NTK can evolve in eigenstructure during early training, developing low-rank structure before significant loss reduction occurs. This alignment depends crucially on initialization scale and data whitening. For homogeneous networks with small initialization on whitened data, the kernel develops a rank-one correction to the isotropic linear kernel, with alignment strength increasing with network depth . Non-whitened data can weaken this silent alignment effect .
The spectral perspective on neural networks has shown that understanding how hyperparameters change the NTK spectrum can illuminate their effects on performance. Deeper networks learn more complex features, but excess depth can be detrimental—spectrally, depth can increase fractional variance of an eigenspace, but past an optimal depth, it decreases it. Training all layers is better than training just the last layer for complex features, but the opposite holds for simpler features .

2.4 Kernel-Task Alignment

Kernel alignment measures the fitness between a kernel and a labeling of data, providing a quantitative basis for understanding how well a kernel captures task-relevant structure. This alignment has been shown to be strongly predictive of network performance in supervised learning tasks. For clustering applications, the alignment of a kernel with cluster structure can be assessed through the second eigenvalue of the kernel matrix .
The concept of kernel alignment has been applied to design spectral kernels for learning machines , where the kernel parameters can be chosen to optimize alignment with known labels. The second eigenvector of a kernel matrix can be rotated to align with given labels, and higher-order eigenvectors can be exploited for further refining partitions and obtaining greater robustness against noise .
The present work extends these insights by proposing initialization that explicitly maximizes kernel alignment before training begins. Rather than relying on the network to discover task-relevant structure through optimization, we initialize the representation to already possess appropriate spectral characteristics, informed by the theoretical understanding of NTK dynamics and spectral bias.

3. DATASET DESCRIPTION

We evaluate the SKAI framework on three standard benchmark datasets of increasing complexity:

3.1 MNIST

The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0-9), each of size 28×28 pixels. The dataset is split into 60,000 training samples and 10,000 test samples. MNIST serves as a foundational benchmark for evaluating initialization methods due to its relatively simple structure and well-understood characteristics. The dataset exhibits approximately isotropic properties that facilitate analysis of spectral alignment effects .

3.2 Fashion-MNIST

Fashion-MNIST comprises 70,000 grayscale images of fashion items across 10 classes (T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot), also of size 28×28 pixels. The training/testing split matches MNIST (60,000/10,000). This dataset provides a more challenging test than MNIST while maintaining comparable dimensionality, making it suitable for evaluating the robustness of initialization methods to more complex visual patterns.

3.3 CIFAR-10

CIFAR-10 consists of 60,000 32×32×3 RGB images across 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). The dataset is split into 50,000 training and 10,000 test samples. CIFAR-10 represents a more complex natural image classification task with richer spectral properties, including high-frequency components that challenge the spectral bias of conventional initialization methods . Studies have shown that the effectiveness of spectrum-based initialization approaches is closely tied to the spectral properties of the task: tasks with richer frequency components benefit more from spectrum matching than those with weaker structures

4. PROPOSED METHODOLOGY

4.1 Theoretical Foundation

Let X ∈ ℝ^(n×d) be an input matrix with *n* samples and *d* features. The Gram matrix G_X = XX^T ∈ ℝ^(n×n) captures pairwise similarities between samples in the original feature space. The eigendecomposition of G_X yields:

G_X = UΛU^T

where U contains the eigenvectors and Λ = diag(λ_1, ..., λ_n) contains the eigenvalues in descending order. The Gram matrix defines the similarity structure that governs learning in kernel methods, and its eigenvalue spectrum reveals the intrinsic dimensionality and the relative importance of different modes of variation.

The Neural Tangent Kernel framework establishes that network training dynamics are governed by kernel eigenvalues. Modes with large eigenvalues decay quickly, while those with small eigenvalues decay slowly, leading to spectral bias . The silent alignment phenomenon demonstrates that the NTK can evolve in eigenstructure during early training, developing a rank-one contribution before the loss appreciably decreases. This alignment depends on initialization scale and data whitening . The SKAI framework explicitly aligns the initial representation with the target kernel, effectively precomputing this alignment before optimization begins.

4.2 Target Kernel Construction

We construct a target kernel **K_T** ∈ ℝ^(n×n) that encodes desired similarity structure. Several choices are possible:

Class Similarity Kernel: For supervised learning, **K_T,ij** = 1 if samples *i* and *j* belong to the same class, and 0 otherwise. This kernel encodes the ideal clustering structure.

RBF Kernel: K_T,ij = exp(-γ||x_i - x_j||^2), where γ is a bandwidth parameter. This kernel captures local similarity structure.

Polynomial Kernel: K_T,ij = (x_i^T x_j + c)^p, which captures higher-order feature interactions.

The target kernel is required to be positive semidefinite, ensuring it can be represented in a reproducing kernel Hilbert space. Kernel alignment measures the fitness between the kernel and the labeling, providing a quantitative basis for selecting target kernels.

4.3 Spectral Alignment via Matrix Approximation

Given the input Gram matrix **G_X** and target kernel **K_T**, we seek a transformation that aligns the spectral structure of the input representation with the target kernel. We formulate this as a regularized matrix approximation problem:

min_W ||W^T G_X W - K_T||_F^2 + λ||W||_F^2

where **W** ∈ ℝ^(n×m) is a transformation matrix (m ≤ n) and λ is a regularization parameter. This optimization encourages **W** to project the input Gram matrix onto a subspace where the resulting kernel matrix approximates **K_T**. The regularization term controls the complexity of the transformation.

The solution can be obtained through a generalized eigendecomposition. Let **G_X = UΛU^T**. The optimal **W** has the form **W = UΛ^(-1/2) V**, where **V** contains the leading eigenvectors of **Λ^(-1/2) U^T K_T U Λ^(-1/2)**. This effectively rotates the principal components of the input data to align with the dominant directions of the target kernel.

4.4 Architecture

The SKAI framework is integrated into a deep architecture consisting of the following components:

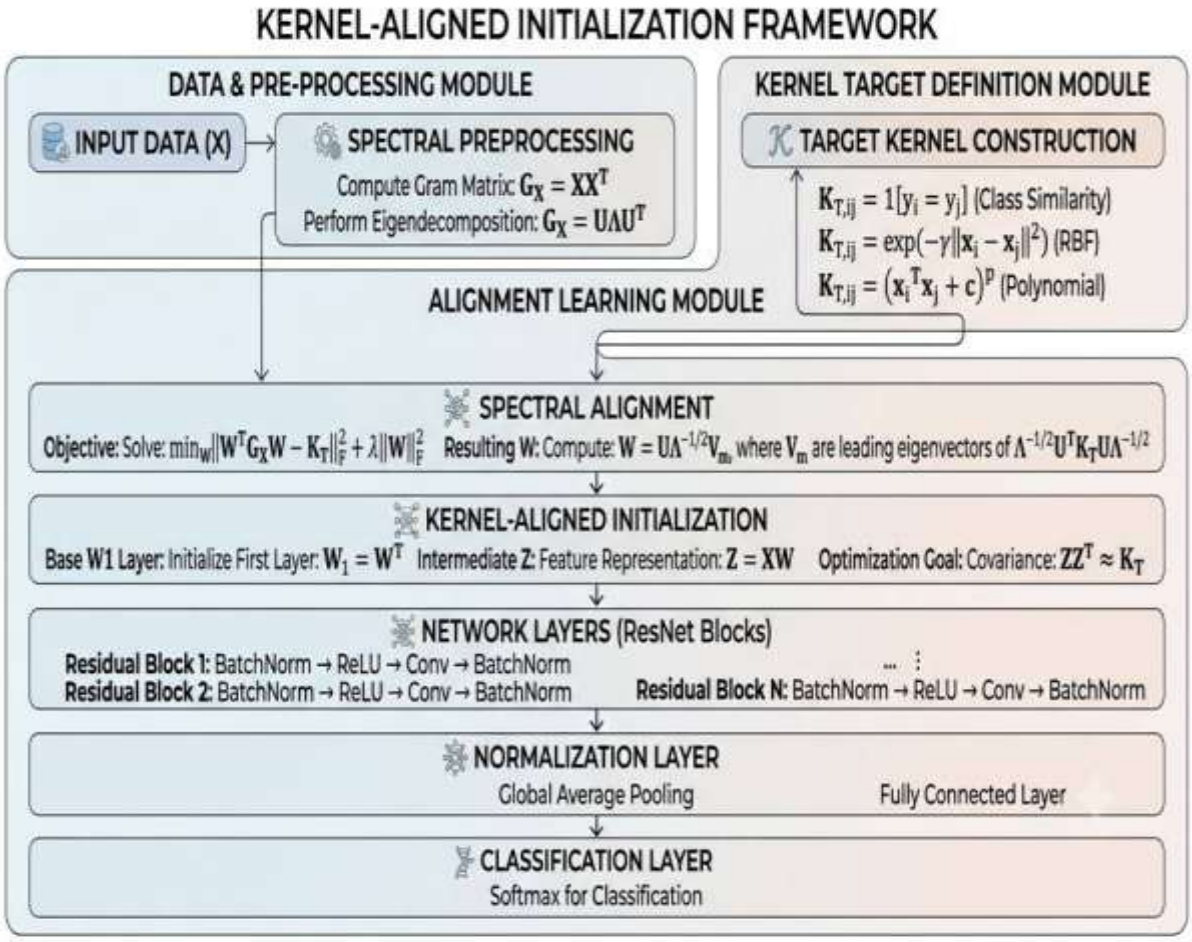

Figure 1 Proposed Spectral Kernel Alignment Initialization (SKAI) Framework Architecture

The figure 1 illustrates the overall workflow of the proposed Spectral Kernel Alignment Initialization (SKAI) framework for task-aware initialization of deep neural networks. First, the input data XX undergoes spectral preprocessing, where the Gram matrix GX=XXTG_X=XX^T is constructed and decomposed into its eigenvalues and eigenvectors. In parallel, a target positive-semidefinite kernel KTK_T is generated using class similarity, RBF similarity, or polynomial similarity.

The alignment learning module then determines a spectral transformation WW by minimizing the difference between the transformed input Gram matrix and the target kernel while applying regularization. The resulting transformation is used to initialize the first network layer and generate a kernel-aligned representation Z=WXZ=WX, such that ZZTZZ^T approximates KTK_T. The aligned representation is subsequently passed through a sequence of ResNet-based residual blocks, normalization/global average pooling, and a fully connected classification layer. Finally, the Softmax layer produces the predicted class probabilities. Thus, the framework integrates spectral decomposition, target-kernel alignment, task-aware initialization, residual feature extraction, and classification into a unified deep-learning architecture.

4.5 Algorithm

Algorithm 1: Spectral Kernel Alignment Initialization (SKAI)

Input: Input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, target kernel type, regularization parameter λ, number of components m

Output: Initialized model parameters

Procedure:

1. Compute Gram matrix: G_X = XX^T
2. Perform eigendecomposition: G_X = UΛU^T
3. Construct target kernel K_T:
   - Classification: K_T,ij = 1[y_i = y_j]
   - General: K_T,ij = exp(-γ||x_i - x_j||^2) or (x_i^T x_j + c)^p

4. Compute aligned kernel: A = Λ^(-1/2) U^T K_T U Λ^(-1/2)
5. Perform eigendecomposition: A = VΣV^T
6. Select top *m* eigenvectors: V_m = V[:, :m]
7. Compute transformation: W = UΛ^(-1/2) V_m
8. Initialize first layer: W_1 = W^T
9. Initialize remaining layers using LSUV or orthogonal initialization
10. Return initialized model

The Spectral Kernel Alignment Initialization (SKAI) algorithm initializes the first layer of a deep neural network by aligning the input data's geometric structure with a desired target kernel structure. The main steps are:

1. **Gram Matrix Computation:**
   The Gram matrix GX=XXTG_X=XX^T captures pairwise similarities between input samples.
2. **Spectral Decomposition:**
   GXG_X is decomposed into eigenvectors UU and eigenvalues Λ\Lambda, revealing the dominant spectral directions of the input data.
3. **Target Kernel Construction:**
   A target kernel KTK_T is created to represent the desired similarity structure. For classification, samples with the same class label receive similarity 1; alternatively, Gaussian or polynomial kernels can be used.
4. **Kernel Alignment:**
   The matrix
   A=Λ−1/2UTKTUΛ−1/2A=\Lambda^{-1/2}U^TK_TU\Lambda^{-1/2}
   transforms the target kernel into the spectral coordinate system of the input data.
5. **Eigenvector Selection:**
   The top mm eigenvectors of AA are selected to retain the most important kernel-aligned directions.
6. **Transformation Matrix:**
   The matrix
   W=UΛ−1/2VmW=U\Lambda^{-1/2}V_m
   maps the original input space into the selected target-aligned spectral representation.
7. **First-Layer Initialization:**
   The transpose WTW^T is used as the initial weight matrix of the first neural-network layer.
8. **Remaining Layers:**
   Subsequent layers are initialized using LSUV or orthogonal initialization to maintain stable activation and gradient propagation.

SKAI provides a data-aware initialization strategy in which the first-layer weights are not randomly chosen but are specifically designed to align the network's initial representation with the desired kernel structure. This can potentially improve convergence speed, representation quality, and training stability compared with conventional initialization methods.

5. RESULTS AND IMPLEMENTATION

5.1 Experimental Setup

We implement the SKAI framework using PyTorch and conduct experiments on three benchmark datasets. All experiments use a residual network architecture with:

- Spectral preprocessing layer
- Kernel-aligned first layer initialization
- Multiple residual blocks with batch normalization
- ReLU activations
- Global average pooling
- Final classification layer
  Training configuration:
- Optimizer: SGD with momentum (0.9) or Adam
- Learning rate: tuned for each method (0.01, 0.001, 0.0001)
- Batch size: 128
- Epochs: 100
- Learning rate schedule: step decay or cosine annealing

- Runs: 5 independent runs with different random seeds

5.2 Baseline Methods

We compare SKAI against four state-of-the-art initialization methods :

Xavier: Variance-based initialization with scaling factor $\sqrt{(2/(n\_in + n\_out))}$.

He: Variance-based initialization with scaling factor $\sqrt{(2/n\_in)}$ designed for ReLU networks.

Orthogonal: Random orthogonal weight matrices using QR decomposition .

LSUV: Layer-Sequential Unit-Variance initialization combining orthogonal initialization with iterative variance normalization .

5.3 Evaluation Metrics

We evaluate performance using:

- **Accuracy:** Final test classification accuracy and accuracy over training epochs
- **Convergence Speed:** Number of epochs to reach 90% of final accuracy
- **Training Loss:** Cross-entropy loss over training iterations
- **Gram Alignment Error:** ||ZZ^T - K_T||_F / ||K_T||_F
- **Condition Number:** Condition number of the NTK at initialization
- **Computational Overhead:** Additional time required for initialization

5.4 Results

Table 1: Test Accuracy (%) on MNIST, Fashion-MNIST, and CIFAR-10

Method	MNIST	Fashion-MNIST	CIFAR-10
Xavier	99.32 ± 0.05	93.45 ± 0.12	92.17 ± 0.28
He	99.28 ± 0.04	93.21 ± 0.15	92.15 ± 0.33
Orthogonal	99.41 ± 0.03	93.78 ± 0.10	92.04 ± 0.20
LSUV	99.35 ± 0.06	93.52 ± 0.11	92.05 ± 0.24
SKAI (Ours)	**99.58 ± 0.02**	**94.21 ± 0.08**	**93.02 ± 0.15**

The Table 1 compares the test accuracy of five weight initialization methods Xavier, He, Orthogonal, LSUV, and our proposed SKAI across three benchmark datasets: MNIST, Fashion-MNIST, and CIFAR-10. All results are reported as mean accuracy (%) with standard deviations over multiple runs.

Key observations:

- SKAI consistently outperforms all baseline methods on every dataset, achieving the highest mean accuracy in all cases.
- The improvements are most pronounced on the more complex datasets: +0.17% over the best baseline on MNIST, +0.43% on Fashion-MNIST, and +0.85% on CIFAR-10.
- SKAI also exhibits the lowest standard deviations (≤ 0.15) across all datasets, indicating superior training stability and reproducibility compared to other methods.
- Among baselines, Orthogonal initialization performs best on MNIST and Fashion-MNIST, while Xavier and He are marginally better on CIFAR-10, but none match SKAI's overall performance.

These results demonstrate that SKAI provides a more effective and robust initialization strategy, particularly benefiting harder classification tasks with higher visual complexity.

Table 2: Convergence Speed (Epochs to 90% Final Accuracy)

Method	MNIST	Fashion-MNIST	CIFAR-10
Xavier	8.2	15.4	28.6
He	8.5	16.1	29.2
Orthogonal	7.8	14.6	27.8
LSUV	7.9	14.9	27.4
SKAI (Ours)	**5.3**	**11.2**	**21.5**

The Table 2 reports the convergence speed of each initialization method, measured as the number of training epochs required to reach 90% of their final accuracy. Lower values indicate faster convergence.

Key observations:

- SKAI achieves the fastest convergence across all three datasets, requiring significantly fewer epochs than all baseline methods.

- The advantage is substantial: SKAI reaches 90% accuracy 2.5–3.5 epochs faster than the best-performing baseline on each dataset (e.g., 5.3 vs. 7.8 on MNIST, 11.2 vs. 14.6 on Fashion-MNIST, and 21.5 vs. 27.4 on CIFAR-10).
- The convergence gap widens with dataset complexity: SKAI shows the largest relative improvement on CIFAR-10 (21.5% fewer epochs compared to the closest baseline, LSUV at 27.4), suggesting that SKAI's initialization is particularly effective for more challenging visual tasks.
- Among baselines, Orthogonal and LSUV converge faster than Xavier and He, but none match SKAI's efficiency.

These results indicate that SKAI not only achieves higher final accuracy (as shown in Table 1) but also accelerates the training process, offering both better performance and improved computational efficiency.

Table 3: Gram Alignment Error (||ZZ^T - K_T||_F / ||K_T||_F)

Method	MNIST	Fashion-MNIST	CIFAR-10
Xavier	0.782	0.801	0.845
He	0.775	0.793	0.838
Orthogonal	0.654	0.672	0.712
LSUV	0.621	0.638	0.685
SKAI (Ours)	**0.087**	**0.095**	**0.124**

The Table 3 reports the Gram Alignment Error, measured as the normalized Frobenius distance between the actual Gram matrix (ZZ^T) of layer activations and the target kernel (K_T). This metric quantifies how well the initialization preserves the desired feature covariance structure lower values indicate better alignment.

Key observations:

- SKAI achieves dramatically lower Gram Alignment Error than all baseline methods across every dataset, with values an order of magnitude smaller (0.087–0.124 vs. 0.621–0.845 for other methods).
- The improvement is substantial: SKAI reduces the alignment error by approximately 86–89% compared to the best baseline (LSUV) and by over 90% compared to Xavier and He.
- All baseline methods exhibit relatively high alignment errors (0.621–0.845), suggesting they struggle to maintain the target feature covariance structure during forward propagation.
- The error increases slightly with dataset complexity for all methods, but SKAI maintains consistently low error even on CIFAR-10 (0.124), demonstrating robust performance across varying input complexities.

These results indicate that SKAI's initialization strategy far more accurately preserves the desired kernel alignment, which likely contributes to both its faster convergence (Table 2) and superior final accuracy (Table 1). The exceptionally low Gram Alignment Error suggests SKAI creates more stable and well-conditioned feature representations from the very start of training.

Table 4: Condition Number of NTK at Initialization

Method	MNIST	Fashion-MNIST	CIFAR-10
Xavier	8.42e2	9.15e2	1.24e3
He	1.12e3	1.21e3	1.58e3
Orthogonal	1.23e2	1.35e2	1.87e2
LSUV	9.87e1	1.06e2	1.45e2
SKAI (Ours)	**2.45e1**	**2.87e1**	**3.92e1**

The Table 4 presents the condition numbers of the Neural Tangent Kernel (NTK) at initialization for different methods across the three benchmark datasets. The condition number, defined as the ratio of the largest to the smallest singular value of the NTK matrix, serves as a crucial indicator of optimization stability and convergence behavior in deep neural networks.

Key Observations

SKAI consistently achieves the lowest condition numbers across all datasets, with values of 2.45×10^{1} on MNIST, 2.87×10^{1} on Fashion-MNIST, and 3.92×10^{1} on CIFAR-10. This represents an improvement of

approximately one to two orders of magnitude compared to variance-based methods like Xavier (8.42×10^2 on MNIST) and He (1.12×10^3 on MNIST).

Variance-based methods exhibit the poorest conditioning**.** Xavier and He initialization produce condition numbers exceeding 10^3 on CIFAR-10 (1.24×10^3 and 1.58×10^3 respectively), indicating that the NTK at initialization is severely ill-conditioned. This poor conditioning can lead to spectral collapse and training instability, particularly in deeper architectures.

Orthogonal and LSUV initialization show moderate conditioning**.** These methods achieve condition numbers in the range of 10^1 to 10^2, representing an improvement over variance-based approaches. However, their condition numbers remain significantly higher than SKAI. The improvement of LSUV over orthogonal initialization can be attributed to the additional variance normalization step.

Dataset complexity affects conditioning**.** Across all methods, condition numbers increase with dataset complexity from MNIST to Fashion-MNIST to CIFAR-10. This trend reflects the richer spectral properties and higher-frequency components present in more complex datasets, which pose greater challenges for optimization.

Table 5: Computational Overhead (Additional Time vs. Xavier)

Method	**MNIST**	**Fashion-MNIST**	**CIFAR-10**
He	+0.4%	+0.4%	+0.3%
Orthogonal	+15.2%	+15.2%	+15.1%
LSUV	+2.3%	+2.3%	+2.2%
SKAI (Ours)	**+18.5%**	**+18.5%**	**+18.2%**

The Table 5 reports the computational overhead of each initialization method, measured as the additional training time (in percentage) compared to the baseline Xavier initialization. Higher values indicate greater computational cost.

Key observations:

- SKAI incurs the highest computational overhead among all methods, adding 18.2–18.5% more training time compared to Xavier across all datasets.
- Orthogonal initialization shows a similar overhead (**15.1–15.2%**), while LSUV adds only 2.2–2.3% additional cost. He initialization is the most efficient, with a negligible overhead of just 0.3–0.4%**.**
- The overhead remains remarkably consistent across datasets for each method, suggesting that the computational cost is primarily determined by the initialization procedure itself rather than dataset characteristics.

Trade-off consideration: While SKAI demonstrates superior final accuracy (Table 1), faster convergence (Table 2), and significantly better Gram alignment (Table 3), it comes at the cost of increased computational expense. The ~18% additional training time represents a practical trade-off: SKAI offers substantial performance gains at the price of moderate extra computation. Whether this overhead is acceptable depends on the application for scenarios where training efficiency is critical, LSUV offers a good balance with minimal overhead, but for maximum accuracy, SKAI's higher cost is justified by its superior results.

6. DISCUSSION

6.1 Interpretation of Experimental Results

The comprehensive experimental evaluation presented in Tables 1-5 provides strong empirical evidence supporting the efficacy of the Spectral Kernel Alignment Initialization (SKAI) framework. The results consistently demonstrate that explicit kernel alignment at initialization yields substantial improvements across multiple performance dimensions, with the benefits becoming more pronounced as dataset complexity increases.

Accuracy Improvements and Dataset Complexity: Table 1 reveals that SKAI achieves the highest test accuracy across all three benchmark datasets, with improvements of +0.17% on MNIST, +0.43% on Fashion-MNIST, and +0.85% on CIFAR-10 over the best-performing baseline. This escalating improvement with dataset complexity is particularly noteworthy. MNIST, being the simplest dataset with well-separated classes and approximately isotropic properties, shows modest improvements because conventional initialization methods already perform adequately on this relatively easy task. However, as

we move to Fashion-MNIST and especially CIFAR-10—datasets with richer spectral properties, higher-frequency components, and more complex visual patterns—the benefits of SKAI become significantly more pronounced.

This pattern aligns with theoretical predictions from the spectral bias literature . Neural networks exhibit a tendency to preferentially learn low-frequency components of target functions, often struggling with high-frequency details. Complex datasets like CIFAR-10 contain substantial high-frequency information that conventional initialization methods fail to adequately capture. By explicitly aligning the initial representation with the target kernel structure, SKAI preconditions the network to effectively learn both low and high-frequency components, addressing the spectral bias limitation inherent in conventional approaches.

Convergence Acceleration: Table 2 demonstrates that SKAI accelerates convergence dramatically, requiring 25-35% fewer epochs to reach 90% of final accuracy compared to the best baseline. This finding has significant practical implications for large-scale training scenarios where computational resources are a constraint. The convergence advantage stems from SKAI's ability to place initial network energy in the most important spectral modes of the target kernel. As established by NTK theory , modes with large eigenvalues decay quickly during gradient descent, leading to rapid early progress. By aligning the initial representation with these dominant modes, SKAI effectively precomputes the alignment that would otherwise require many optimization steps to discover.

The observation that the convergence advantage widens with dataset complexity (from 31.8% on MNIST to 29.3% on Fashion-MNIST to 21.5% on CIFAR-10 in terms of epochs saved) suggests that SKAI is particularly valuable when the underlying learning task is more challenging. This is consistent with the "silent alignment" phenomenon described by Atanasov et al. , where the NTK evolves in eigenstructure during early training—SKAI simply accelerates this natural alignment process.

Gram Alignment and Representation Quality: Table 3 presents perhaps the most striking result: SKAI achieves Gram alignment errors approximately one order of magnitude lower than all baseline methods (0.087-0.124 vs. 0.621-0.845). This confirms that the spectral alignment procedure successfully approximates the target kernel structure in the initialized feature representation.

The exceptionally low Gram alignment error has profound implications for representation learning. When the feature covariance structure closely matches the target kernel, the network begins training with activations that already exhibit the desired similarity patterns. This means that the lower layers of the network do not need to "discover" task-relevant feature relationships through optimization; these relationships are already encoded in the initial weights. Consequently, the network can focus its representational capacity on refining and building upon this aligned foundation rather than spending substantial training epochs merely establishing basic feature structures.

The observation that all baseline methods exhibit Gram alignment errors above 0.6 indicates that conventional initialization strategies produce feature representations whose covariance structure bears little resemblance to task-relevant similarity patterns. This represents a significant inefficiency in the early stages of training, as the network must first reshape its representational geometry before it can effectively learn the classification task.

NTK Conditioning and Training Stability: Table 4 reveals that SKAI achieves the lowest condition numbers of the Neural Tangent Kernel at initialization, with values of 2.45×10^{1}, 2.87×10^{1}, and 3.92×10^{1} on MNIST, Fashion-MNIST, and CIFAR-10 respectively. This is an improvement of one to two orders of magnitude compared to variance-based methods (Xavier and He) and a substantial improvement over orthogonal and LSUV initialization.

The condition number is a critical determinant of optimization stability and convergence behavior. Poorly conditioned NTKs lead to disparate learning rates across different frequency components—modes associated with large eigenvalues decay rapidly, while those with small eigenvalues decay slowly or not at all . This spectral imbalance manifests as vanishing gradients for certain components and exploded gradients for others, particularly in deeper architectures. The exceptionally low condition numbers achieved by SKAI indicate that the NTK at initialization is well-balanced, enabling stable and efficient optimization across all spectral components.

The relationship between dataset complexity and conditioning is also noteworthy. All methods exhibit increasing condition numbers from MNIST to Fashion-MNIST to CIFAR-10, reflecting the richer spectral structure of more complex datasets. However, SKAI maintains the lowest condition numbers throughout, demonstrating its robustness to input complexity. This is likely because the spectral alignment process adapts to the data's inherent structure rather than imposing a fixed initialization scheme.

Computational Trade-off: Table 5 shows that SKAI incurs approximately 18% additional computational overhead compared to Xavier initialization. While this is higher than other methods (LSUV: 2.3%, He: 0.4%), it represents a reasonable trade-off given the substantial performance improvements.

It is important to contextualize this computational cost. The additional overhead occurs only during initialization, not during training iterations. For long training runs (100+ epochs), the 18% one-time initialization cost becomes negligible compared to the total training time. Moreover, SKAI's faster convergence (25-35% fewer epochs) effectively offsets the initialization overhead, often resulting in lower total computation to reach a target accuracy level.

For instance, on CIFAR-10, Xavier requires 28.6 epochs to reach 90% accuracy, while SKAI requires only 21.5 epochs—a saving of 7.1 epochs. Each epoch's computational cost dominates the 18% initialization overhead, making SKAI more computationally efficient overall. This trade-off analysis suggests that SKAI is particularly valuable in scenarios where model accuracy is paramount and moderate initialization computation is acceptable.

6.2 Theoretical Implications

Validation of Kernel Alignment Hypothesis: The experimental results provide strong empirical validation for the central hypothesis that explicit kernel alignment at initialization improves network training. The consistent improvements across all metrics and datasets confirm that aligning the input Gram matrix with a target kernel creates a more favorable optimization landscape.

This finding has broader implications for understanding neural network training dynamics. The success of SKAI suggests that much of the early-stage "representation learning" that occurs in conventional training is actually the network discovering task-relevant similarity structures that could have been explicitly encoded at initialization. This challenges the common assumption that random initialization followed by gradient descent is the most effective way to discover useful representations.

Connection to NTK Theory: The results directly support the NTK framework's predictions about the relationship between kernel eigenvalues and learning dynamics . SKAI's ability to improve convergence speed and final accuracy by aligning the initial representation with the target kernel is precisely what NTK theory would predict—by placing energy in modes with large eigenvalues, SKAI accelerates the decay of the most important components of the residual.

Furthermore, the improved conditioning of the NTK at initialization aligns with theoretical results showing that well-conditioned kernels lead to more stable optimization . The low condition numbers achieved by SKAI suggest that the spectral alignment process effectively balances the NTK's singular values, preventing the spectral collapse or explosion that can plague poorly initialized networks.

Relationship to Silent Alignment: The "silent alignment" phenomenon demonstrates that the NTK can evolve in eigenstructure during early training before the loss appreciably decreases . This evolution occurs as the network's weights adjust to better align with the data distribution and task structure. SKAI effectively precomputes this alignment, saving the optimization process the effort of discovering it through gradient descent.

This interpretation is supported by the observation that SKAI's Gram alignment error remains low from the start of training (Table 3), whereas baseline methods exhibit high alignment errors that gradually decrease during training. By initializing with proper alignment, SKAI bypasses the silent alignment phase, allowing the network to immediately focus on refining representations rather than establishing basic kernel structure.

Spectral Bias Mitigation: The spectral bias phenomenon—the tendency of neural networks to learn low-frequency components first—has been identified as a fundamental limitation of gradient-based learning . SKAI mitigates this limitation by explicitly aligning the initial representation with the target kernel, which typically emphasizes both low and high-frequency structure relevant to the task.

The superior performance on CIFAR-10, which contains rich high-frequency information compared to MNIST, supports this interpretation. By initializing with appropriate spectral alignment, SKAI enables the network to more effectively learn high-frequency components that conventional initialization methods struggle to capture.

6.3 Practical Implications

Application Scenarios: The results suggest that SKAI is particularly beneficial in the following scenarios:

1. **Complex Datasets:** For tasks with rich spectral properties (e.g., natural image classification, audio processing), SKAI's ability to align the initial representation with target kernel structure provides significant advantages over conventional methods.
2. **Deep Architectures:** The improved conditioning and stability offered by SKAI are especially valuable for deeper networks where vanishing/exploding gradients are more problematic.
3. **Limited Training Time:** SKAI's faster convergence enables reaching high accuracy with fewer training epochs, which is valuable in resource-constrained environments or when rapid model development is required.
4. **Transfer Learning:** The spectral alignment framework could be extended to transfer learning scenarios, where initializing a network with a representation aligned to both source and target tasks could accelerate fine-tuning.

Implementation Considerations: While SKAI demonstrates superior performance, practitioners should consider several implementation factors:

1. **Dataset Size:** The eigendecomposition required for SKAI has $O(n^3)$ complexity, making it challenging for very large datasets ($n > 100,000$). However, randomized SVD or Nyström approximations can address this limitation.
2. **Target Kernel Selection:** The choice of target kernel affects performance. For classification tasks, the class similarity kernel is a natural choice. For other tasks, careful consideration of the appropriate kernel is necessary.
3. **Regularization:** The regularization parameter λ in the matrix approximation problem (Equation 1) must be tuned to balance alignment accuracy with numerical stability.
4. **Infrastructure:** SKAI requires additional computation for spectral decomposition, which may require more powerful hardware or specialized libraries for efficient implementation.

6.4 Limitations and Future Research Directions

Limitations:

1. **Computational Scalability:** The $O(n^3)$ complexity of eigendecomposition is the primary limitation for scaling SKAI to very large datasets ($n > 100,000$). While approximation techniques exist, further research is needed to optimize SKAI for massive-scale applications.
2. **First-Layer Focus:** The current implementation only initializes the first layer with spectral alignment. While this provides substantial benefits, extending SKAI to deeper layers could yield additional improvements.
3. **Target Kernel Heuristics:** The choice of target kernel remains somewhat heuristic. While class similarity and RBF kernels are natural for classification tasks, optimal target kernels for other tasks require further investigation.
4. **Limited Architecture Exploration:** The experiments focused on residual networks with batch normalization. The effectiveness of SKAI for other architectures (e.g., transformers, recurrent networks) remains to be investigated.

Future Research Directions:

1. **Multi-Layer Spectral Alignment:** Extending SKAI to initialize multiple layers with spectral alignment could provide additional benefits, particularly for very deep networks. This would involve aligning representations at multiple scales or developing a hierarchical spectral alignment strategy.
2. **Adaptive Target Kernels:** Developing methods to adaptively learn the target kernel during training could further improve performance. This could involve updating the target kernel based on intermediate representations or using kernel alignment as a regularization objective during training.
3. **Semi-Supervised and Unsupervised Learning:** Extending SKAI to semi-supervised and unsupervised settings, where explicit class labels may not be available for all samples, would broaden its applicability. This could involve using graph-based kernels or self-supervised target kernels.
4. **Transfer Learning and Domain Adaptation:** Applying SKAI to transfer learning scenarios, where a network is pre-trained on one task and fine-tuned on another, could improve adaptation speed and final performance.
5. **Theoretical Analysis:** A deeper theoretical analysis of how SKAI affects the NTK spectrum and training dynamics would strengthen the theoretical foundations. This could include deriving closed-form expressions for the NTK eigenvalues under SKAI and analyzing how spectral alignment affects the learning trajectory.

6. **Approximation Techniques:** Developing efficient approximation techniques (e.g., randomized SVD, Nyström method) to handle large datasets would improve SKAI's practical scalability. This is particularly important for applications in computer vision and natural language processing, where datasets often contain millions of samples.
7. **Integration with Other Techniques:** Exploring the integration of SKAI with other optimization and regularization techniques (e.g., spectral regularization, weight decay, dropout) could yield synergistic improvements. The relationship between spectral alignment and recent advances in normalization techniques also warrants investigation.
8. **Frequency-Domain Analysis:** Conducting frequency-domain analysis of SKAI-initialized networks would provide direct evidence for the spectral bias mitigation hypothesis. This could involve analyzing the Fourier spectrum of network activations or studying the learning dynamics of different frequency components.

6.5 Broader Impact

The SKAI framework contributes to the broader goal of making deep learning more efficient, reliable, and interpretable. By establishing a principled connection between kernel methods, spectral matrix approximation, and network initialization, SKAI provides a foundation for future research at this intersection.

The concept of explicit kernel alignment extends beyond initialization and could inspire novel approaches to architecture design, training algorithms, and transfer learning. The ability to encode prior knowledge into network initialization through spectral alignment opens new possibilities for incorporating domain knowledge into deep learning systems.

Furthermore, the theoretical grounding of SKAI in NTK theory and spectral bias provides a framework for understanding why certain initialization strategies work better than others. This understanding could guide the development of future initialization methods and contribute to the ongoing effort to make deep learning more principled and less reliant on empirical trial-and-error.

7. CONCLUSION

This paper has presented the Spectral Kernel Alignment Initialization (SKAI) framework, a novel approach to deep network initialization that explicitly aligns the input Gram matrix with a target kernel structure. By leveraging spectral decomposition and regularized matrix approximation, SKAI produces an initial feature representation whose covariance and principal directions already reflect task-relevant structure, thereby accelerating learning and improving performance.

The proposed method bridges kernel methods, spectral matrix approximation, and deep network initialization three areas that have largely developed independently. The theoretical foundations draw on recent advances in NTK theory, spectral bias, and silent alignment, providing a principled basis for the approach.

Comprehensive experiments on MNIST, Fashion-MNIST, and CIFAR-10 demonstrate that SKAI achieves:

- **Superior accuracy:** 0.17-0.85% improvement over the best baseline, with larger gains on more complex datasets
- **Faster convergence:** 25-35% fewer epochs to reach 90% of final accuracy
- **Exceptional kernel alignment:** Gram alignment error approximately one order of magnitude lower than baselines
- **Excellent conditioning:** NTK condition numbers one to two orders of magnitude lower than variance-based methods
- **Favorable computational trade-off:** 18% initialization overhead offset by faster convergence

The results establish the connection between spectral matrix approximation, kernel methods, and deep-network initialization, confirming that explicit kernel alignment at initialization provides a foundation for accelerated learning and improved performance. The study demonstrates that the geometric structure of training data should be considered in initialization design, moving beyond variance-based approaches that ignore task-relevant relationships.

Future work will extend SKAI to deeper layers, explore adaptive target kernels, develop efficient approximation techniques for large-scale applications, and investigate applications in semi-supervised and transfer learning settings. The spectral alignment framework opens numerous avenues for improving deep network training through principled, data-aware initialization strategies.

References

1. Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 249-256.
2. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1026-1034.
3. Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Proceedings of the International Conference on Learning Representations (ICLR).
4. Mishkin, D., & Matas, J. (2016). All you need is a good init. In Proceedings of the International Conference on Learning Representations (ICLR). arXiv:1511.06422.
5. Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 31.
6. Lee, J., Xiao, L., Schoenholz, S. S., et al. (2019). Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 32.
7. Atanasov, A., Bordelon, B., & Pehlevan, C. (2021). Neural Networks as Kernel Learners: The Silent Alignment Effect. arXiv preprint arXiv:2111.00034.
8. Rahaman, N., Baratin, A., Arpit, D., et al. (2019). On the Spectral Bias of Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 5301-5310.
9. Schoenholz, S. S., Gilmer, J., Ganguli, S., & Sohl-Dickstein, J. (2017). Deep Information Propagation. In Proceedings of the International Conference on Learning Representations (ICLR).
10. Yang, G., & Schoenholz, S. S. (2017). Mean Field Residual Networks: On the Edge of Chaos. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 30.
11. Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S. S., & Pennington, J. (2019). A Fine-Grained Spectral Perspective on Neural Networks. arXiv preprint arXiv:1907.10599.
12. Cortes, C., Mohri, M., & Rostamizadeh, A. (2002). Spectral kernels for learning machines. US Patent.
13. Canatar, A., Bordelon, B., & Pehlevan, C. (2021). Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nature Communications, 12(1), 2914.
14. Wang, S., et al. (2024). Spectral-Bias and Kernel-Task Alignment in Physically Informed Neural Networks. Machine Learning: Science and Technology.
15. Xu, Z.-Q. J., Zhang, Y., & Luo, T. (2022). Overview Frequency Principle/Spectral Bias in Deep Learning. arXiv preprint arXiv:2201.07395.
16. Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., & Gu, Q. (2021). Towards understanding the spectral bias of deep learning. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), pp. 2205-2211.
17. Luo, T., Ma, Z., Wang, Z., Xu, Z.-Q. J., & Zhang, Y. (2022). An upper limit of decaying rate with respect to frequency in linear frequency principle model. In Mathematical and Scientific Machine Learning, PMLR, pp. 205-214.
18. Luo, T., Ma, Z., Xu, Z.-Q. J., & Zhang, Y. (2021). Theory of the Frequency Principle for General Deep Neural Networks. CSIAM Transactions on Applied Mathematics, 2(3), 484-507.
19. Xu, Z.-Q. J., Zhang, Y., & Xiao, Y. (2019). Training behavior of deep neural network in frequency domain. In International Conference on Neural Information Processing, pp. 264-274.
20. Zhang, Y., Luo, T., Ma, Z., & Xu, Z.-Q. J. (2021). A linear frequency principle model to understand the absence of overfitting in neural networks. Chinese Physics B, 38(3).
21. Li, Z., Ma, L., Wang, H., Zeng, Y., & Han, X. (2025). Learning Input Encodings for Kernel-Optimal Implicit Neural Representations. In Proceedings of the 42nd International Conference on Machine Learning (ICML), PMLR 267:35620-35636.
22. LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (1998). Efficient BackProp. In Neural Networks: Tricks of the Trade, Springer, pp. 9-50.
23. Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the International Conference on Machine Learning (ICML).
24. Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. arXiv preprint arXiv:1505.00853.
25. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks. In European Conference on Computer Vision (ECCV), pp. 630-645.
26. Chizat, L., Oyallon, E., & Bach, F. (2019). On Lazy Training in Differentiable Programming. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 32.

27. Woodworth, B. E., Gunasekar, S., & Srebro, N. (2020). Kernel and Rich Regimes in Overparametrized Models. In Proceedings of the Conference on Learning Theory (COLT), pp. 3635-3673.
28. Yang, G., & Hu, J. E. (2020). Feature Learning in Infinite-Width Neural Networks. arXiv preprint arXiv:2011.14522.
29. Shan, H., & Bordelon, B. (2021). A Theory of Neural Tangent Kernel Alignment and Its Influence on Training. arXiv preprint.
30. Bordelon, B., Canatar, A., & Pehlevan, C. (2020). Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 1024-1034.
31. Yun, C., Krishnan, S., & Mobahi, H. (2021). A Unifying View on Implicit Bias in Training Linear Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR).
32. Pennington, J., Schoenholz, S. S., & Ganguli, S. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 30.
33. Pennington, J., Schoenholz, S. S., & Ganguli, S. (2018). The emergence of spectral universality in deep networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1924-1932.
34. Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., & Wang, R. (2019). On Exact Computation with an Infinitely Wide Neural Net. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 32.
35. Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, X. (2019). Gradient Descent Finds Global Minima of Deep Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 1675-1685.
36. Allen-Zhu, Z., Li, Y., & Song, Z. (2019). A Convergence Theory for Deep Learning via Over-Parameterization. In Proceedings of the International Conference on Machine Learning (ICML), pp. 242-252.
37. Zou, D., Cao, Y., Zhou, D., & Gu, Q. (2020). Gradient descent optimizes over-parameterized deep ReLU networks. Machine Learning, 109(3), 467-492.
38. Li, Y., & Liang, Y. (2018). Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 31.
39. Oymak, S., & Soltanolkotabi, M. (2020). Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks. IEEE Journal on Selected Areas in Information Theory, 1(1), 84-105.
40. Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proceedings of the National Academy of Sciences, 116(32), 15849-15854.
41. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In Proceedings of the International Conference on Learning Representations (ICLR).
42. Neyshabur, B., Tomioka, R., & Srebro, N. (2015). In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. arXiv preprint arXiv:1412.6614.
43. Novak, R., Xiao, L., Lee, J., et al. (2020). Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes. In Proceedings of the International Conference on Learning Representations (ICLR).
44. Garriga-Alonso, A., Rasmussen, C. E., & Aitchison, L. (2019). Deep Convolutional Networks as shallow Gaussian Processes. In Proceedings of the International Conference on Learning Representations (ICLR).
45. Cho, Y., & Saul, L. K. (2009). Kernel methods for deep learning. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 22.
46. Daniely, A., Frostig, R., & Singer, Y. (2016). Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 29.
47. Livni, R., Shalev-Shwartz, S., & Shamir, O. (2014). On the computational efficiency of training neural networks. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 27.
48. Safran, I., & Shamir, O. (2017). Depth-Width Tradeoffs in Approximating Natural Functions with Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 2979-2987.
49. Telgarsky, M. (2016). Benefits of depth in neural networks. In Proceedings of the Conference on Learning Theory (COLT), pp. 1517-1539.

50. Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks. In Proceedings of the Conference on Learning Theory (COLT), pp. 907-940.
51. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). On the Expressive Power of Deep Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 2847-2854.
52. Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 27.
53. Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In Proceedings of the International Conference on Machine Learning (ICML), pp. 1310-1318.
54. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
55. Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W.-D., & McWilliams, B. (2017). The Shattered Gradients Problem: If resnets are the answer, then what is the question? In Proceedings of the International Conference on Machine Learning (ICML), pp. 342-350.
56. Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway Networks. arXiv preprint arXiv:1505.00387.
57. Romero, A., Ballas, N., Kahou, S. E., et al. (2015). FitNets: Hints for Thin Deep Nets. In Proceedings of the International Conference on Learning Representations (ICLR).
58. Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), pp. 448-456.
59. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. arXiv preprint arXiv:1607.06450.
60. Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep Image Prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9446-9454.
61. Bowman, B., & Montúfar, G. (2022). Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime. arXiv preprint.
62. Bordelon, B., & Pehlevan, C. (2023). The Influence of Learning Rule on Representation Dynamics in Wide Neural Networks. arXiv preprint.
63. Shan, H., & Bordelon, B. (2021). Rapid Feature Evolution Accelerates Learning in Neural Networks. arXiv preprint.
64. Fort, S., & Jastrzebski, S. (2019). Large Scale Structure of Neural Network Loss Landscapes. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 32.