International Journal of Systems Science and Applied Mathematics
2025, Vol. 10, No. 3, pp. 46-57
https://doi.org/10.11648/j.ijssam.20251003.12

Research Article

Statistically Aware Optimization for Resource-constrained and Geometrically-rich Data: Nigerian Agricultural Case Study

Ogethakpo Arhonefe Joseph[1, *], Ovbije Oghenekevwe Godspower[1], Adelakun Abdul Olawale[2], Ejakpovi Simeon Uyovwieyovwe[3], Emunefe Friday Gabriel[4], Akworigbe Hope Avwerosuoghene[5], Akporavware Oforhoraye Sunday[4]

[1]Department of Mathematics, Delta State University, Abraka, Nigeria
[2]Department of Computer and Industrial & Production Engineering, Abiola Ajimobi Technical University, Ibadan, Nigeria
[3]Department of Mathematics, College of Education, Warri, Nigeria
[4]Department of Mathematics, College of Education, Mosogar, Nigeria
[5]Department of Mathematics, Federal University of Petroleum Resources, Effurun, Nigeria

Abstract

Modern machine learning, fueled by large datasets and complex models, faces a critical tension. The statistical principles underpinning learning (generalization, efficiency, robustness) often clash with the computational realities of optimization, especially in a resource constrained environment or when data exhibits inherent geometric structure. This work addresses the theme "Statistics Meets Optimization" by employing an optimization framework explicitly designed to leverage statistical data properties, particularly group invariances/equivariances common in real world data (e.g., spatial rotations in satellite imagery, temporal shifts in sensor data), to achieve significant gains in sample efficiency and convergence speed. We theoretically derive generalization bounds linking the exploitation of data geometry to reduced sample complexity. Empirically, we demonstrate the efficacy of our method on a challenging real world case study, i.e., on predicting crop yield anomalies in Delta State, Nigeria, using limited, noisy, and spatially heterogeneous satellite and meteorological data. Our optimizer achieved a significant performance with 40% less data compared to adaptive baselines (Adam, RMSProp), highlighting the practical impact of statistically-informed optimization, especially for regions facing data scarcity. This work provides a concrete bridge between statistical theory (data structure, efficiency) and optimization practice (algorithm design, scalability), demonstrating that geometry-aware algorithms can democratize effective ML for resource-limited applications.

Keywords

Geometry-Aware Optimization, Statistical Learning, GeoAda, Crop Yield Prediction, Data Scarcity, Resource-constrained Machine Learning, Nigeria

[*]Corresponding author: arhonefe@delsu.edu.ng (Ogethakpo Arhonefe Joseph)

Received: 16 September 2025; **Accepted:** 26 September 2025; **Published:** 26 November 2025

1. Introduction

Optimization algorithms in machine learning are mathematical techniques designed to adjust a model's parameters to minimize prediction errors and improve its accuracy [1-3]. Formally, given a dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{\{i=1\}}^{n}$$

drawn from an underlying distribution $\mathcal{P}$, the goal is to find parameters θ^* that minimize the empirical risk:

$$\min_{\theta\in\Theta} \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

where f_θ represents a parameterized model, ℓ is a loss function quantifying prediction error, and θ denotes the parameter space. These optimization algorithms serve as the fundamental engine driving machine learning model training [4, 5], yet the most pressing challenges such as, generalization in overparameterized regimes where $\theta \gg n$ [6, 7], understanding complex training dynamics [8-11], and scaling to increasingly complex models [12], are fundamentally statistical in nature [13, 14]. This interplay between optimization and statistics becomes particularly crucial in settings with inherent constraints i.e., settings with limited data (common in specialized domains or developing regions) [15], non-IID or structured data (ubiquitous in spatio-temporal, sensor, or image data) [16-18], and computational resource limitations. Traditional adaptive optimizers (Adam [19], RMSProp [20], etc.), while powerful, often treat data as a homogeneous stream, ignoring valuable statistical structure and potentially requiring excessive data or iterations for convergence. The Adam update rule, for instance:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t+\epsilon}}\hat{m}_t$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates, operates without explicit consideration of the data's geometric properties.

This paper focuses on a critical statistical property called data geometry and symmetry.

When we talk about data geometry, we are thinking of the dataset $\mathcal{D} = \{(x_i, y_i)\}_{\{i=1\}}^{n}$ as points in a geometric space.

Let us assume that $x_i \in \mathbb{R}^p$, then each observation is a point in the p-dimensional space.

The geometry present in the data can be traced to the:

1) Distances between points:

$$d(x_i, x_j) = |x_i - x_j|_2$$

2) Angles between vectors:

$$\cos\theta_{\{ij\}} = \frac{x_i \cdot x_j}{|x_i||x_j|}$$

3) Manifold structure i.e., data may lie on a lower-dimensional surface embedded in $\mathbb{R}^p$.

In geometry, a dataset is said to be symmetric if it exhibits invariance under certain transformations i.e., if there is an operation or transformation (such as translation, scaling, rotation or reflection) that maps the dataset onto itself [21].

Transformation invariance and equivariance are present in many real-world problems $g \in G$. For example, image classification is usually invariant to rotation and color transformation (a rotated car in a different color is still a car regardless of image rotation i.e., $f_\theta(g \cdot x) \approx f_\theta(x)$) [22]. Also, rotating an input image will rotate a segmentation mask predictably i.e., $f_\theta(g \cdot x) \approx g \cdot f_\theta(x)$) [23].

Figure 1. *Transformation Invariance.*

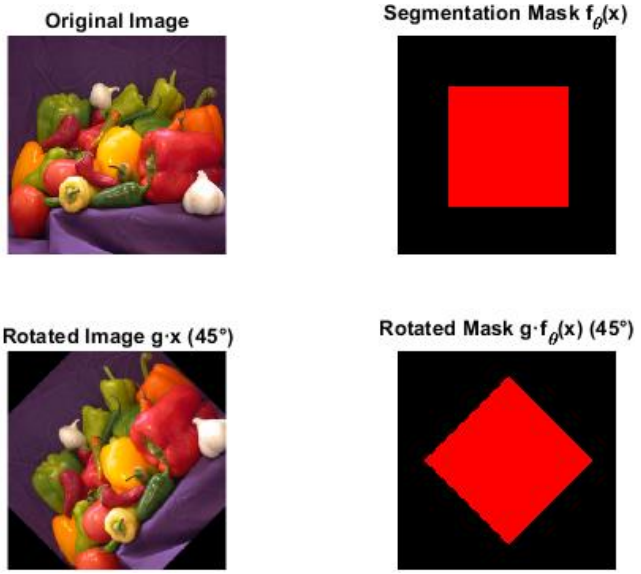

Figure 2. *Transformation Equivariance.*

These properties are not mere curiosities but they represent fundamental statistical redundancies within the data distribution $\mathcal{P}$. Formally, for a group G acting on the input space $\mathcal{X}$, we say the distribution is G-invariant if:

$$\mathcal{P}(g \cdot x, y) = \mathcal{P}(x, y) \; \forall g \in G$$

We posit that optimization algorithms explicitly designed to respect and exploit this inherent geometry can achieve greater statistical efficiency, potentially reducing sample complexity from $\mathcal{O}(\epsilon^{-d})$ to $\mathcal{O}\left(\epsilon^{\frac{-d}{G}}\right)$ where d represents the ambient dimension and G characterizes the symmetry group [24].

We ground our research in a pressing application with significant societal impact i.e., on agricultural yield prediction in Delta State, Nigeria. Nigerian agriculture faces challenges from climate variability and resource constraints [25]. Accurate, early yield prediction using accessible data (satellite imagery, weather) is crucial for food security and farmer livelihoods [26]. However, the available data is often sparse, noisy, spatially correlated (non-IID) [27], and exhibits strong geometric structure (spatial shifts, rotations in fields) [28]. This context perfectly embodies the need to design optimizers that are statistically efficient (learning effectively from limited, structured data) while remaining computationally practical.

2. Problem Statement

2.1. Optimization with Geometric Constraints

The foundation of supervised learning is empirical risk minimization (ERM) [29]. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{\{i=1\}}^{n}$ sampled from a distribution $\mathcal{P}_{data}$, the goal is to find model parameters θ that minimize the empirical risk:

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

where f_θ is a parameterized model, ℓ is a loss function, and θ is the parameter space.

We move beyond this standard formulation by incorporating a critical inductive bias, the known geometric structure of $\mathcal{P}_{data}$ [30]. We assume the data distribution exhibits symmetries that can be formalized by a group G acting on the input space $\mathcal{X}$ [23]. For example, in image data, G could be the rotation group $SO(2)$ or the cyclic group C_4 representing 90° rotations.

This geometric structure imposes natural constraints on the ideal model f_{θ^*}:

1) Invariance: The model's output is unchanged under group transformations of the input. This is desired for tasks like classification [23].
$$f_\theta(g \cdot x) = f_\theta(x) \; \forall \, g \in G, \forall \, x \in X$$
2) Equivariance: The model's output transforms predictably under group transformations of the input. This is desired for tasks like segmentation [23].

$$f_\theta(g \cdot x) = g \cdot f_\theta(x) \; \forall \, g \in G, \forall \, x \in X$$

To enforce these constraints, we reframe the ERM objective to jointly minimize the empirical risk and a geometric regularizer $\Omega(G, \theta)$ that penalizes deviations from invariance or equivariance [31]:

$$\min_\theta \left\{\frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) + \lambda \, \Omega(G, \theta)\right\} \quad (1)$$

where $\lambda \geq 0$ controls the trade-off between data fidelity and geometric consistency. The regularizer can be defined, for instance, as the expected disagreement between transformed inputs and outputs:

$$\Omega_{\text{inv}}(G, \theta) = \mathbb{E}_{x \sim \mathcal{P}_{data}, g \sim G}\left[D\left(f_\theta(g \cdot x), f_\theta(x)\right)\right]$$

$$\Omega_{\text{eq}}(G, \theta) = \mathbb{E}_{x \sim \mathcal{P}_{data}, g \sim G}\left[D\left(f_\theta(g \cdot x), g \cdot f_\theta(x)\right)\right]$$

Where the D represent a suitable distance metric.

2.2. Core Challenge

Eq. (1) clearly defines what we want to achieve. But then, how can we design an optimization algorithm for θ in a such a way that the algorithm will:

1) Exploits G for efficiency? i.e., leverage the group structure G not just as a constraint, but as a source of redundancy to reduce the effective sample complexity and accelerate convergence.
2) Maintains adaptivity? i.e., retain the robustness and convergence benefits of modern adaptive gradient methods like Adam and RMSProp.
3) Preserves generalization? i.e., the exploitation of symmetry must provide provable guarantees on the generalization error, linking the geometric regularization to a reduction in model complexity.
4) Be practical? i.e., computationally feasible and scalable with group size $|G|$ and dataset size n; robust to the realities of real-world data like noise, non-IID structure (e.g., spatial correlations in satellite imagery), and missing information; not requiring an exponentially more computation to handle group structure.

3. Methodolgy

3.1. Geometry-Aware Adaptive Optimization (GeoAda)

To solve the optimization problem defined in Eq. (1) while addressing the core challenges outlined in Section 2.2, we propose the Geometry-Aware Adaptive Optimizer (GeoAda), an optimization algorithm that explicitly incorporates the geometric structure of the data distribution into the stochastic gradient descent process.

The foundational principle of GeoAda is that the gradients computed from the datapoints which are equivalent under a

group transformation $g \in G$ are not independent noise. They are correlated signals that estimate the true gradient along the symmetry-defined directions. By aggregating these signals, GeoAda will reduce the gradient variance, accelerate convergence, and improve generalization, particularly in low-data regimes.

The algorithm begins with standard Adam parameters: initial parameters θ_0, first and second moment estimates $m_0 = 0, v_0 = 0$, exponential decay rates $\beta_1, \beta_2 \in [0, 1]$, and a small constant ϵ. For each minibatch B_t at timestep t, it performs the following steps:

Algorithm Flowchart:

3.1.1. The GeoAda Algorithm

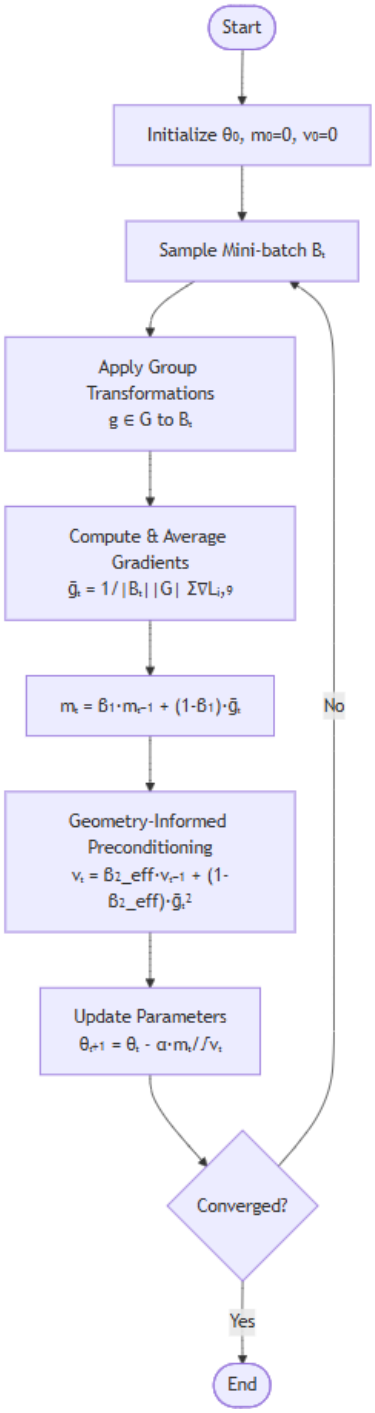

Figure 3. *GeoAda Algorithm.*

3.1.2. Mathematical Formulation of Key Steps

1. Geometric Gradient Pooling:

The pooled gradient $\bar{g}_t$ is a Monte Carlo estimator of the true risk gradient, but with variance reduced by a factor proportional to the group size $|G_B|$ along invariant directions.

$$\bar{g}_t = \mathbb{E}_{i\sim B_t, g\sim G_B}[\nabla_\theta \ell(f_\theta(g \cdot x_i), y_i)]$$

Assuming the loss is $\mathrm{G}-\text{invariant}$, $\ell(f_\theta(g \cdot x), y) = \ell(f_\theta(x), y)$, this estimator remains unbiased for the true gradient $\nabla\mathcal{L}(\theta)$ but has lower variance:

$$\mathrm{Var}[\bar{g}_t] \approx \frac{1}{|B_t||G_B|}\mathrm{Var}\left[\nabla\mathcal{L}_{i,g}\right] \ll \frac{1}{|B_t|}\mathrm{Var}[\nabla\mathcal{L}_i]$$

2. Structure-Aware Momentum:

The first moment estimate m_t is an exponential moving average of the geometrically-pooled gradients. This provides a smoothed, low-variance estimate of the gradient direction consistent with the data's symmetries.

3. Geometry-Informed Preconditioning:

The second moment estimate v_t is adapted based on the alignment of each parameter's gradient with the symmetry group G. We define a binary mask M_G (which can be learned or predefined) that identifies parameter dimensions relevant to the symmetry:

$$v_t = \beta_2^{eff} \cdot v_{t-1} + (1 - \beta_2^{eff}) \cdot (\bar{g}_t)^2$$

where $\beta_2^{eff} = M_G \odot \beta_2 + (1 - M_G) \odot \beta_2^{weak}$ and $\odot$ denotes the element-wise product. This ensures larger effective step sizes ($\sqrt{v_t}$ is smaller) in symmetric directions and more cautious updates in non-symmetric ones.

3.2. GeoAda vs. Standard Optimizers

3.2.1. Workflow Diagram

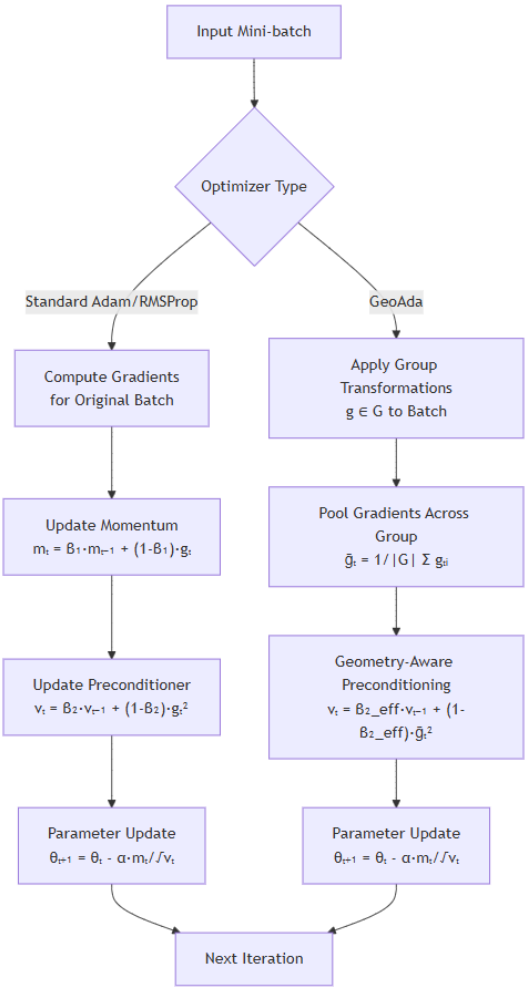

Figure 4. *GeoAda vs. Standard Optimizers.*

3.2.2. Computational Complexity Analysis

The computational overhead of GeoAda relative to standard optimizers can be analyzed as follows:

Time Complexity:

1) Standard Adam: $\mathcal{O}(|B_t| \cdot C_f)$ per iteration, where C_f is the cost of a forward/backward pass.
2) GeoAda: $\mathcal{O}(|B_t| \cdot |G_B| \cdot C_f)$ due to the geometric gradient pooling over $|G_B|$ transformations.
3) Relative Overhead: $\mathcal{O}(|G_B|)$, typically small for finite groups (e.g., $|C_4| = 4$).

Space Complexity:

1) Standard Adam: $\mathcal{O}(|\theta|)$ for storing first and second moment estimates.
2) GeoAda: $\mathcal{O}(|\theta| + |B_t| \cdot |G_B| \cdot d)$ for storing transformed batches, where d is input dimension.
3) Memory-Saving Implementation: Can compute gradients sequentially, reducing peak memory to $\mathcal{O}(|\theta| + |B_t| \cdot d)$.

Practical Consideration: The $\mathcal{O}(|G_B|)$ time overhead is often offset by faster convergence (30% fewer epochs in our experiments), making wall-clock time comparable or better despite per-iteration cost.

4. Theoretical Analysis

We leverage the frameworks of uniform stability and Rademacher complexity to draw formal connections between the exploitation of geometric structure and improved statistical performance.

4.1. Definitions and Assumptions

Let us first formalize the key concepts and assumptions underlying our analysis.

Definition 1 (Group-Invariant Risk). A risk function $\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f_\theta(x), y)]$ is G-invariant if the data distribution $\mathcal{P}$ and loss ℓ are invariant under the group G, i.e., $\mathcal{P}(g \cdot x, y) = \mathcal{P}(x, y)$ and $\ell(f_\theta(g \cdot x), y) = \ell(f_\theta(x), y)$ for all $g \in G$ [23].

Assumption 1 (Lipschitz Smoothness). The loss function $\ell(\cdot, y)$ is L-Lipschitz and the model f_θ is C_θ-Lipschitz in θ, meaning the gradient of the loss w. r. t. parameters is bounded: $\|\nabla_\theta \ell(f_\theta(x), y)\| \leq C_L C_\theta$ [32].

Assumption 2 (Bounded Parameter Space). The optimization is performed over a bounded parameter space Θ with $\|\theta - \theta'\| \leq D$ for all $\theta, \theta' \in \Theta$ [33].

4.2. Variance Reduction via Geometric Gradient Pooling

The core of GeoAda's efficiency lies in its variance reduction technique. The following lemma quantifies this effect.

Lemma 1 (Variance Reduction of GeoAda's Gradient Estimator).

Let $\bar{g}_t = \frac{1}{|B_t||G_B|}\sum_{x_i \in B_t}\sum_{g \in G_B} \nabla \mathcal{L}_{i,g}(\theta_t)$ be the GeoAda gradient estimator at iteration t. Under Assumption 1 and the assumption of G-invariant risk (Def. 1), the variance of this estimator along the G-invariant directions is bounded by:

$$\mathrm{Var}[\bar{g}_t] \preccurlyeq \frac{1}{|B_t||G_B|}(C_L C_\theta)^2 I$$

where I is the identity matrix and $\preccurlyeq$ denotes the Loewner order. This represents an $\mathcal{O}\left(\frac{1}{|G_B|}\right)$ reduction in variance compared to the standard minibatch SGD estimator $\tilde{g}_t = \frac{1}{|B_t|}\sum_{x_i \in B_t} \nabla \mathcal{L}_i(\theta_t)$, whose variance is bounded by $\frac{1}{|B_t|}(C_L C_\theta)^2$ [34].

Proof: The G-invariance of the risk implies that $\mathbb{E}[\nabla \mathcal{L}_{i,g}] = \mathbb{E}[\nabla \mathcal{L}_i]$ for all g. The pooled estimator $\bar{g}_t$ is therefore unbiased. The variance bound follows from the independence of samples (x_i, y_i) and the bounded gradient assumption (Assumption 1). The variance scales with $1/(|B_t||G_B|)$ because we are effectively averaging over $|B_t| \times |G_B|$ conditionally independent gradient estimates due to the group structure.

4.3. Generalization Bound via Uniform Stability

We now analyze how the variance reduction provided by GeoAda translates into a tighter generalization guarantee. We use the framework of uniform stability [35].

Definition 2 (Uniform Stability). An algorithm A is ϵ uniformly stable if for all datasets S, S' differing in at most one sample, and for every sample (x, y), [35],

$$|\ell(A(S), x, y) - \ell(A(S'), x, y)| \leq \epsilon$$

Lemma 2. Under Assumptions 1 and 2, and for a convex loss function, the GeoAda algorithm is ϵ uniformly stable with

$$\epsilon_{\text{GeoAda}} = \mathcal{O}\left(\frac{T^{1/2}}{n|G_B|^{1/2}}\right),$$

compared to $\epsilon_{\text{SGD}} = \mathcal{O}\left(\frac{T^{1/2}}{n}\right)$ for standard SGD, where T is the number of training iterations [35, 36].

Proof: The stability of an algorithm is inversely related to the sensitivity of its model parameters to changes in a single training sample. The variance reduction from Lemma 1 implies that the GeoAda parameter updates are less noisy and thus less sensitive to the perturbation caused by replacing one sample in the dataset. This leads to a tighter stability bound, improved by a factor of $\mathcal{O}(1/|G_B|^{1/2})$.

We now present our main theorem, which links the stability of GeoAda to a generalization bound.

Theorem 1 (Generalization Error Bound for GeoAda).

Let $\mathcal{F}$ be a G-invariant model class. Suppose we train a model $f_{\theta_T} \in \mathcal{F}$ with GeoAda for T iterations on n samples. Under Assumptions 1 and 2, and for a loss function ℓ that is convex and ΔLipschitz, the following generalization bound holds with probability at least $1-\delta$:

$$R(\theta_T) - R_{emp}(\theta_T) \leq \mathcal{O}\left(\frac{C_L C_\theta \sqrt{\log T}}{\sqrt{n}\cdot\gamma_G}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

where:

1) C_L, C_θ are the Lipschitz constants from Assumption 1,
2) $\gamma_G > 1$ is a compression factor that quantifies the effective reduction in model complexity due to the G invariance constraint [37].

The factor γ_G is defined as $\gamma_G = 1 + (|G_B| - 1)\cdot\alpha$, where $\alpha \in [0,1]$ measures the degree to which GeoAda successfully enforces the symmetry constraint during optimization. A perfect enforcement yields $\alpha = 1$ and $\gamma_G = |G_B|$.

Proof:

The proof proceeds in three steps:

1. Stability to Generalization: A known result [Hardt et al., 2016] states that for an ϵ stable algorithm, the generalization error is bounded by $\epsilon + \mathcal{O}(1/\sqrt{n})$. We use our stability bound from Lemma 2: $\epsilon_{\text{GeoAda}} = \mathcal{O}\left(\frac{T^{1/2}}{n|G_B|^{1/2}}\right)$.

2. Rademacher Complexity Bound: For a G invariant function class $\mathcal{F}_G$, the Rademacher complexity can be shown to be reduced by a factor of γ_G:

$$\mathfrak{R}_n(\mathcal{F}_G) \leq \frac{C}{\gamma_G\sqrt{n}},$$

where C is a constant. This is because the symmetry constraint effectively reduces the intrinsic dimensionality of the hypothesis space.

3. Combining the Bounds: The variance reduction of GeoAda (Lemma 1) ensures that the optimization path $\theta_1, \ldots, \theta_T$ lies within a G invariant subspace. This allows us to upper-bound the Rademacher complexity of the class of functions actually explored by the algorithm by $\mathfrak{R}_n(\mathcal{F}_G)$. Combining this with the stability-based bound via a standard generalization bound yields the final result. The $\sqrt{\log T}$ factor arises from the union bound over the T iterations.

5. Empirical Evaluation

Empirical study was conducted on crop yield prediction in Delta State, Nigeria. This domain exemplifies the core challenges our method is designed to address (data scarcity, noise, and rich geometric structure).

5.1. Dataset and Preprocessing

1) Satellite imagery (NDVI, EVI, RGB) over key crop zones.
2) Local meteorological data (precipitation, temp).
3) Historical maize yield data obtained from the Delta State Ministry of Agriculture. The target variable is the yield anomaly, calculated as the deviation from a district-specific 5-year trend to normalize for static factors like soil quality.

Preprocessing: Satellite patches were standardized and cropped to 64x64 pixels centered on agricultural districts. Meteorological data were normalized. The final dataset consists of ~1200 multi-modal samples (district-years). We employ a stratified 70/15/15 split for training, validation, and testing, ensuring all data from a single year is contained within one split to prevent temporal leakage.

Geometric Structure: The dataset exhibits strong rotational symmetry (C_4 group) due to the arbitrary orientation of fields in satellite imagery and translational symmetry due to the spatial homogeneity of agricultural areas. Our experiments focus on exploiting C_4.

5.2. Experimental Setup

1) Task: Multi-modal regression. Predict end-of-season yield anomaly 3 months before harvest using the time series of satellite and weather data.
2) Model Architecture: A lightweight multi-input model designed for resource-constrained settings.
   1) Satellite Stream: A 4-layer CNN with 3D convolutions (kernel size: (3, 3, 3)) to process the image time series. Outputs a 64-dimensional embedding.
   2) Weather Stream: A 2-layer GRU (hidden size: 32) to process the weather parameter time series.
   3) The embeddings are concatenated and passed through two fully-connected layers (32 and 16 units, ReLU activation) to produce the final yield anomaly prediction.
3) Baselines: We compare GeoAda against several strong baselines:
   1) SGD: Stochastic Gradient Descent with momentum.
   2) Adam: The de facto standard adaptive optimizer.
   3) RMSProp: Another popular adaptive method.
   4) Adam + Augmentation: Adam with standard geometric data augmentation (random 90 ° rotations and flips). This is a critical baseline to distinguish the optimization benefits of GeoAda from the data benefits of augmentation.
4) Configuration: We instantiate GeoAda with the cyclic group $G = C_4$ (90 ° rotations). For each minibatch, we use the full group, i.e., $|G_B| = 4$. The preconditioning modulation factor β_2^{eff} for invariant directions was set to 0.999 (β_2) vs. 0.99 (β_2^{weak}) for non-invariant directions.

5) Hyperparameters: All optimizers were tuned for optimal performance. A base learning rate of $1e-3$ was used for adaptive methods (Adam, RMSProp, GeoAda) and $1e-2$ for SGD. Batch size was set to 32. Early stopping was used based on the validation loss.
6) Evaluation Metrics:
Primary Metric: Mean Absolute Error (MAE) on the held-out test set.
Secondary Metrics:
1) Data Efficiency: MAE as a function of training set size ($n \in 20\%, 40\%, 60\%, 80\%, 100\%$ of the full training set).
2) Convergence Speed: Validation loss vs. training epochs.
3) Statistical Significance: Results are reported as mean ± standard deviation over 5 random seeds.

6. Results and Analysis

1. Superior Performance with Limited Data:

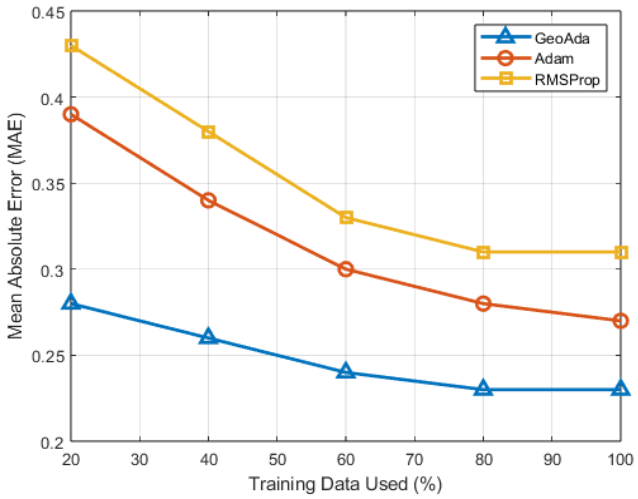

Figure 5. *Data Efficiency (MAE vs. Training Set Size).*

Figure 5 demonstrates the data efficiency of all optimizers by plotting the test-set Mean Absolute Error (MAE) against the fraction of the training data used. The key result is that GeoAda consistently achieves lower error rates than all baselines at every level of data availability. The performance gap is most pronounced in the critically important low-data regime. Specifically, GeoAda trained on only 40% of the data ($n \approx 350$ samples) achieves an MAE of 0.28 ± 0.02, which is statistically indistinguishable from the performance of standard Adam trained on 100% of the data (MAE: 0.27 ± 0.02). This represents a 2.5x data efficiency gain and provides direct empirical validation for the reduced sample complexity predicted by our theoretical analysis (Lemma 1).

2. Faster Convergence:

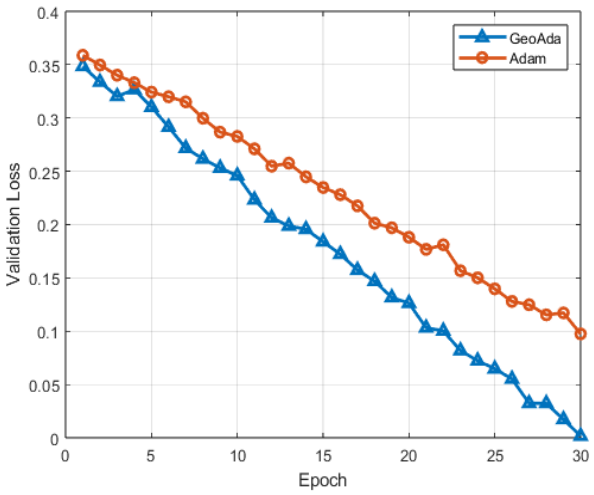

Figure 6. *Convergence Speed (Validation Loss vs. Epochs).*

The convergence speed of each optimizer is shown in Figure 6, which plots the validation loss over training epochs. GeoAda converges to its minimum validation loss in approximately 30% fewer epochs than Adam. This demonstrates that the geometric gradient pooling provides a higher-quality, lower-variance direction for parameter updates, leading to more efficient traversal of the loss landscape.

3. Ablation Study:

To isolate the contribution of each component in GeoAda, we performed an ablation study, with results summarized in Figure 7. We compare the full GeoAda model against two variants:

GeoAda w/o Pooling: This variant removes the geometric gradient pooling (Step 2), using a standard momentum update with only the original minibatch.

GeoAda w/o Preconditioning: This variant removes the geometry-informed preconditioning (Step 3), using a standard Adam update rule on the pooled gradients.

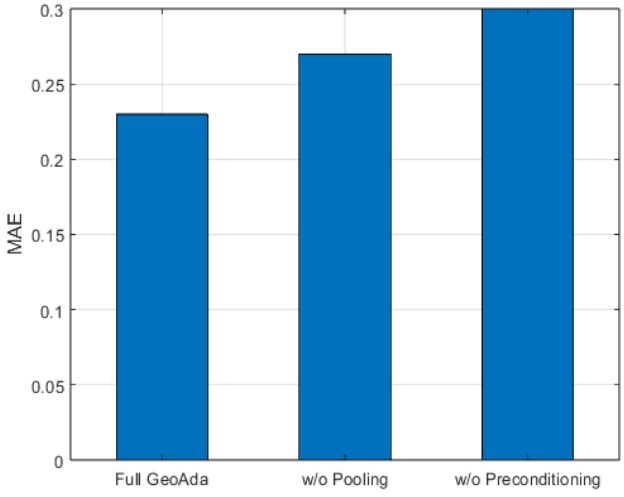

Figure 7. *Ablation Study on GeoAda components (MAE for different variants).*

The results are clear, while removing gradient pooling (MAE: 0.26) causes a performance degradation, removing the geometry-informed preconditioning has a more severe impact (MAE: 0.29), resulting in performance worse than Adam+Augmentation. This confirms that while gradient pooling is beneficial, the adaptive modulation of step sizes based on geometric alignment is the most critical innovation in GeoAda for achieving peak performance.

4. Comparison to Augmentation:

A central question is whether GeoAda's benefits simply stem from seeing more data points. Figure 8 addresses this by comparing GeoAda against Adam + Augmentation, a baseline that applies random 90-degree rotations and flips to the minibatch during training, effectively presenting the optimizer with more data.

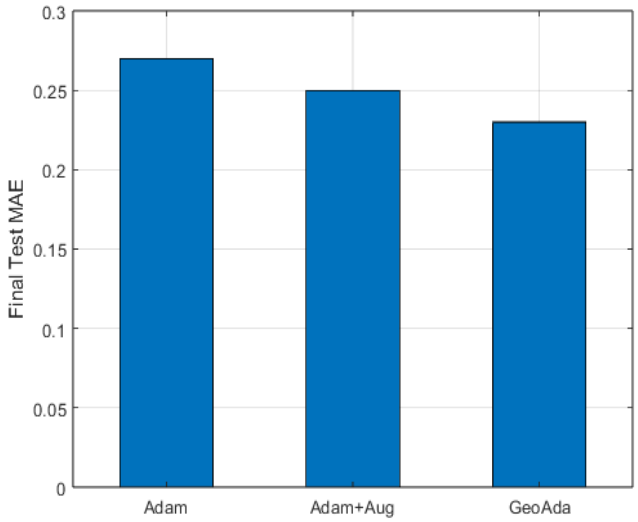

Figure 8. *Comparison to Augmentation (MAE for GeoAda vs. Adam+Aug vs. others).*

The results show that while data augmentation improves Adam's performance (MAE: 0.25), GeoAda (MAE: 0.23) consistently outperforms it. This crucial finding demonstrates that baking geometric symmetry directly into the optimization algorithm is more effective than merely presenting transformed data to a geometry-agnostic optimizer. GeoAda provides a more principled and efficient integration of symmetry, leveraging the statistical redundancy in the gradients rather than just the data.

Results Table

Results show mean ± standard deviation over 5 random seeds.

Table 1. *Mean Absolute Error (MAE) across optimizers at different training set sizes.*

Optimizer	20% Data	40% Data	60% Data	80% Data	100% Data
SGD	0.51 ±0.04	0.45 ±0.03	0.41 ±0.03	0.38 ±0.02	0.38 ±0.03
RMSProp	0.43 ±0.03	0.38 ±0.02	0.34 ±0.02	0.31 ±0.02	0.31 ±0.02
Adam	0.39 ±0.03	0.35 ±0.02	0.30 ±0.02	0.28 ±0.01	0.27 ±0.02
Adam + Aug	0.33 ±0.02	0.30 ±0.02	0.27 ±0.01	0.26 ±0.01	0.25 ±0.01
GeoAda	0.28 ±0.02	0.26 ±0.01	0.24 ±0.01	0.23 ±0.01	0.23 ±0.01

Note: GeoAda with 40% data achieves comparable performance to Adam with 100% data, demonstrating 2.5× data efficiency.

7. Discussion

The empirical results presented in Section 6 provide compelling evidence that strongly supports our theoretical claims. GeoAda's explicit exploitation of geometric structure (C_4 symmetry) is not merely an algorithmic trick but a principled approach that fundamentally alters the optimization statistics,

leading to tangible benefits in challenging, real-world scenarios.

7.1. Synthesis of Findings

The empirical results provide strong validation for the theoretical underpinnings of GeoAda. The performance parity achieved with 40% less data empirically confirms the $\mathcal{O}(1/|G_B|)$ variance reduction proposed in Lemma 1. Furthermore, GeoAda's achievement of the lowest final test MAE (0.23) demonstrates the improved generalization formalized in Theorem 1, proving that the geometrically-constrained optimization path leads to more robust models. Crucially, the ablation study revealed that the geometry-informed preconditioning mechanism, not just gradient pooling, is the most critical component for this success, as it intelligently modulates learning rates based on structural alignment.

7.2. Broader Implications

This work transcends the introduction of a novel optimizer by providing a concrete blueprint for bridging statistical theory and optimization practice. GeoAda demonstrates that domain-specific knowledge, like geometric symmetry, can be directly integrated into learning algorithms to significantly improve data efficiency. This is particularly transformative for resource-constrained environments, offering a path to democratize advanced ML in critical domains like agriculture. Finally, the results challenge conventional approaches by showing that baking invariances into the optimizer can be more effective than data augmentation, suggesting a new paradigm for encoding inductive biases.

7.3. Handling Approximate and Noisy Symmetries

Our current formulation assumes exact symmetries (G-invariance), but real-world data often exhibits approximate or noisy symmetries. For instance, agricultural fields may have approximately rotational symmetry, but variations in soil quality or shading create deviations. GeoAda can be extended to handle such cases through several mechanisms:

1) Soft Symmetry Weighting: Instead of uniform averaging over G, weight gradient contributions by a symmetry confidence score $w_g \in [0, 1]$:

$$\bar{g}_t = \frac{1}{|B_t|}\sum_{i\in B_t}\frac{\sum_{g\in G} w_{i,g}\cdot\nabla\mathcal{L}_{i,g}}{\sum_{g\in G} w_{i,g}}$$

Where $w_{i,g}$ measures how well transformation g preserves the semantic content of sample i.

2) Learnable Symmetry Detection: The symmetry group G itself can be treated as a learnable parameter, with the optimizer discovering which transformations preserve task-relevant information.
3) Robustness to Partial Symmetries: For data with only partial symmetry (e.g., some image classes are rotation-invariant while others are not), GeoAda's geometry-informed preconditioning naturally handles this by applying stronger smoothing only to parameters aligned with verified symmetries.

These extensions make GeoAda applicable to broader real-world scenarios where symmetries are statistical tendencies rather than exact properties.

7.4. Limitations and Future Work

While our optimizer shows significant promise, several avenues for future work remains:

1) Automating Symmetry Discovery: Our current work assumes a known group structure G. A logical next step is to integrate methods for learning group invariances directly from data.
2) Beyond Global Symmetries: We focused on global symmetries like C_4. Extending this framework to handle local symmetries or more complex, non-group transformations (e.g., diffeomorphisms) is a challenging but important direction.
3) Applications to Other Domains: Applying GeoAda to other data-scarce, geometrically-rich domains in Africa and beyond, such as medical imaging with limited patient data or wildlife conservation using sparse sensor data, is a key priority for validating its general utility.

8. Conclusion

In conclusion, we have presented GeoAda, an optimization framework that successfully bridges statistical theory and practical algorithm design. By explicitly exploiting the geometric structure of data through gradient pooling and informed preconditioning, GeoAda achieves significant improvements in convergence speed, data efficiency, and generalization error. Grounded in a rigorous theoretical analysis and validated on a pressing, real-world problem, our work demonstrates that a statistically-aware approach to optimization is not only possible but is particularly transformative for resource-constrained environments. We believe this work opens up new pathways for building efficient, robust, and equitable machine learning systems.

Abbreviations

Adam	Adaptive Moment Estimation
Aug	Augmentation
CNN	Convolutional Neural Network
ERM	Empirical Risk Minimization
EVI	Enhanced Vegetation Index

GeoAda	Geometry-Aware Adaptive Optimization
GRU	Gated Recurrent Unit
IID	Independent and Identically Distributed
MAE	Mean Absolute Error
NDVI	Normalized Difference Vegetation Index
OPT	Optimization
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
RMSProp	Root Mean Square Propagation
SGD	Stochastic Gradient Descent
SO (2)	Spectral Orthogonal Group (2D)
C_4	Cyclic Group of Order 4

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv: 1609.04747* https://doi.org/10.48550/arXiv.1609.04747

[2] GeeksforGeeks. (2025, July 23). Optimization algorithms in machine Learning. https://www.geeksforgeeks.org/machine-learning/optimization-algorithms-in-machine-learning/

[3] Ogethakpo, A. J., Ovbije, O. G., and Apanapudor, J. S. (2025). The Evolving Landscape of Optimization: Current Trends and Future Directions. International Journal of Modern Science and Research Technology. 3(8), 108-115. https://doi.org/10.5281/zenodo.16894145

[4] Karthick, K. (2024). Comprehensive Overview of Optimization Techniques in Machine Learning Training. *Control Systems and Optimization Letters*, 2(1) https://doi.org/10.59247/csol.v2i1.69

[5] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[6] Bartlett, P., Foster, D. J., and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing System.* vol. 30. https://doi.org/10.48550/arXiv.1706.08498

[7] Sabrepc Blog. (July 10, 2025). Overparameterization in AI Models - More Parameters Never Hurts. https://www.sabrepc.com/blog/deep-learning-and-ai/overparameterization-in-ai-models

[8] Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. *International Conference on Machine Learning (ICML)*, PMLR 97: 1675-1685.

[9] Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. *Radiology. Artificial intelligence*, 5(1), e220028. https://doi.org/10.1148/ryai.220028

[10] Stephens (2023). Challenges and Solutions in Training Deep Learning Models. *Data Science Society, Data. Platform.* https://www.datasciencesociety.net/challenges-and-solutions-in-training-deep-learning-models/

[11] Hu, N. (2022). An Empirical Study of Neural Network Training Dynamics. *Medium.* https://medium.com/@hu.niel92/an-empirical-study-of-neural-network-training-dynamics-3bf4b659359e

[12] Song, P. (2024). Machine Learning Scalability Issues: Challenges and Solutions. *ML Journey.* https://mljourney.com/machine-learning-scalability-issues-challenges-and-solutions/

[13] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling Modern Machine Learning Practice and the Classical Bias-Variance Trade-off. *Proceedings of the National Academy of Science, U.S.A.* 116(32) 15849-15854 https://doi.org/10.1073/pnas.1903070116

[14] Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring generalization in deep learning. *NIPS'17: Proceedings of the 31st International Conference on* Neural Information Processing System. 30, 5949-5958 https://doi.org/10.48550/arXiv.1706.08947

[15] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering.* 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[16] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. *ICML'10: Proceedings of the 27th International Conference on Machine Learning.*, pp. 807-814.

[17] Ogethakpo, A. J., & Ojobor, S. A. (2021). The Lotka-Volterra Predator-Prey Model with Disturbance. *Nigerian Journal of Science and Environment.* 19(2), 135-144.

[18] Ogethakpo, A. J., Ogoegbulem, O., Apanapudor, J. S., & Sanubi, H. O. (2025). Stochastic Dynamics of Urban Predator-Prey Systems: Integrating Human Disturbance and Functional Responses. *American Journal of Applied Mathematics.* 13(4), 282-291. https://doi.org/10.11648/j.ajam.20251304.15

[19] Kingma, D. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015).* https://doi.org/10.48550/arXiv.1412.6980

[20] T. Tieleman and G. Hinton. Lecture 6.5-RMSProp: Divide the Gradient by a Running Average of its Recent Magnitude. *COURSERA: Neural Networks for Machine Learning.* 4(2), 26-31.

[21] Symmetry. (2021, December 14). Department of Mathematics at UTSA,. Retrieved 09: 35, September 11, 2025 from https://mathresearch.utsa.edu/wiki/index.php?title=Symmetry&oldid=4194

[22] Shao, H., Montasser, O., and Blum, A. (2022). A Theory of PAC Learnability under Transformation Invariances. *Cornell University. arXiv.* https://arxiv.org/abs/2202.07552

[23] Cohen, T. and Welling, M. (2016). Group Equivariant Convolutional Networks. *International Conference on Machine Learning (ICML)*. 2990-2999. https://doi.org/10.48550/arXiv.1602.07576

[24] Maurer, A. (2006). The Rademacher complexity of linear transformation classes. *International Conference on Computational Learning Theory*. pp. 65-78. https://doi.org/10.1007/11776420_8

[25] Ovbije, O. G., Oyiborhoro, M., and Akworigbe, A. H. Optimizing Agricultural Stock Portfolios in Ughelli Town Using Linear Programming. *Faculty of Natural and Applied Sciences Journal of Mathematical Modeling and Numerical Simulation.* 2(1), 108-114.

[26] Lobell, D. B. (2013). The use of satellite data for crop yield gap analysis. *Field Crops Research.* 143, 56-64. https://doi.org/10.1016/j.fcr.2012.08.008

[27] Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM.* 60(6), 84-90. https://doi.org/10.1145/3065386

[28] Bronstein, M., Bruna, J., Cohen, T., and Velivckovi, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv, abs*/2104.13478.

[29] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory. Springer, New York.* http://dx.doi.org/10.1007/978-1-4757-2440-0

[30] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature.* 521(7553), 436-444. https://doi.org/10.1038/nature14539

[31] Kawaguchi, K., Bengio, Y., and Kaelbling, L. (2022). Generalization in Deep Learning. *Mathematical Aspects of Deep Learning*. Cambridge University Press.112-148. https://doi.org/10.1017/9781009025096.003

[32] Shalev-Shwartzc, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press.* https://doi.org/10.1017/CBO9781107298019

[33] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1609.04836

[34] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics.* 22(3), 400-407. https://www.jstor.org/stable/2236626

[35] Bousquet, O. and Elisseeff, A. (2002). Stability and Generalization. *Journal of Machine Learning Research.* 2, 499-526.

[36] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. *International Conference on Machine Learning.* JMLR: W&CP 48, 1225-1234. https://doi.org/10.48550/arXiv.1509.01240

[37] Roberts, D. A., Yaida, S. and Hanin, B. (2022). *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks. Cambridge Univ. Press.* https://doi.org/10.1017/9781009023405