Proceedings Track

Neural Fields Meet Attention

Editors: List of editors' names

Abstract

We establish a precise mathematical connection between neural field optimization and Transformer attention mechanisms. First, we prove that Transformer-based operators learning neural fields are equivariant to affine transformations (translations and positive scalings) when equipped with relative positional encodings and explicit coordinate normalization—extending geometric deep learning to meta-learning of continuous functions. Second, we demonstrate that linear attention exactly computes negative gradients of squared-error loss for sinusoidal neural fields, with softmax attention converging to this identity at rate $O(\tau^{-2})$ in the high-temperature limit. Experiments on rotation groups validate our theory: equivariance errors remain below 10^{-5} across SO(2) and SO(3) transformations (mean 3.6×10^{-6}, 10 seeds), while attention-gradient correlation exceeds 0.999 for temperatures $\tau \geq 100$. These results reveal that attention mechanisms implicitly encode geometric priors suited for continuous function learning.

1. Introduction

Neural fields have emerged as a powerful paradigm for representing continuous signals, revolutionizing how we encode 3D geometry (Park et al., 2019), synthesize novel views (Mildenhall et al., 2021), and simulate physical systems (Sitzmann et al., 2020). Unlike traditional discrete representations (voxels, meshes), neural fields parameterize signals as continuous functions $f_\theta : \mathbb{R}^d \to \mathbb{R}$, enabling infinite resolution and natural derivatives. SIREN (Sitzmann et al., 2020) showed that sinusoidal activations can capture high-frequency details, while NeRF (Mildenhall et al., 2021) demonstrated photorealistic rendering from MLPs.

In parallel, large language models have revealed a surprising capability: in-context learning (ICL), where Transformers adapt to new tasks from prompt examples without updating weights (Brown et al., 2020). Recent theoretical work interprets this phenomenon as implicit gradient descent (Von Oswald et al., 2023) or algorithm distillation (Garg et al., 2022), but these analyses focus on discrete token prediction rather than continuous function learning.

This paper unifies these paradigms. We show that Transformers are naturally suited to meta-learn neural fields because attention mechanisms implicitly encode the geometric priors required for continuous signal processing. Specifically, we prove that:

- Attention computes exact gradients for neural field optimization (not approximately, but exactly under linear attention)
- Transformer architectures preserve affine symmetries when properly configured
- These properties emerge from the mathematical structure of attention, not from training

Our analysis reveals why Transformers excel at spatial reasoning tasks: the architecture inherently respects the geometric structure of continuous functions. We formalize neural field learning as an operator $\mathcal{O} : \{(x_i, y_i)\}_{i=1}^N \mapsto \theta$ mapping samples to field parameters, and characterize when this operator preserves symmetries.

1.1. Motivation and Main Contributions

1. **Theorem 1 (Affine Equivariance):** A Transformer-based operator $\mathcal{O}$ is equivariant to the affine group $G_{ST} = \{x \mapsto ax + b \mid a > 0, b \in \mathbb{R}^d\}$ if and only if:
   - It uses permutation-equivariant processing (set-based)
   - It employs relative positional encoding (translation invariance)
   - It implements explicit normalization or continuous frequency adaptation (scale handling)

   We prove necessity via a scaling impossibility lemma: fixed finite bases cannot achieve arbitrary scale equivariance without these mechanisms.
2. **Theorem 2 (Attention–Gradient Identity):** For sinusoidal fields $f(x) = \sum_k c_k \phi_k(x)$, linear attention with basis-function keys, residual values, and one-hot queries computes exact negative gradients: $O_k = -\frac{\partial \mathcal{L}}{\partial c_k}$. Softmax attention converges to this at rate $O(\tau^{-2})$.
3. **Empirical Validation:** On rotation groups SO(2) and SO(3), our implementation achieves equivariance errors of $(3.6 \pm 2.0) \times 10^{-6}$ and attention-gradient correlation > 0.999 at high temperature (10 seeds, all $p < 0.001$).

The rest of the paper is organized as follows. Section 3 presents our main theorems on equivariance and the attention-gradient identity. Section 4 validates these results empirically. We conclude with practical implications and future directions.

2. Related Work

Our work drives between the narrow intersection of three rich literatures: continuous neural representations, equivariant networks, and in-context learning theory.

2.1. Neural Fields for Signal Representation

Neural fields (also called implicit neural representations) parameterize signals or scenes by mapping continuous coordinates to output values. Early work such as DeepSDF (Park et al., 2019) learned signed-distance fields of shapes. More recently, Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) achieved photorealistic novel-view synthesis by training an MLP to map 3D location and view direction to color and density. NeRF and its variants (Mip-NeRF, BungeeNeRF (Xiangli et al., 2021), etc.) rely on coordinate-based networks and positional encodings to capture fine detail (Mildenhall et al., 2021; Tancik et al., 2020). SIREN networks (Sitzmann et al., 2020) use periodic activation functions to represent high-frequency signals, demonstrating power in representing physical fields and derivatives. These neural field models provide a flexible alternative to discrete grids, encoding data in the weights of a continuous function (Park et al., 2019; Sitzmann et al., 2020). Our work treats the training of such fields as a mapping from sample data to function parameters, bridging these continuous models with sequence-based learning in Transformers.

2.2. Symmetry and Equivariance in Deep Learning

Incorporating group symmetries into network design improves data efficiency and generalization (Bronstein et al., 2017). Convolutional neural networks exploit translation equiv-

ariance (LeCun et al., 1998), while group-equivariant CNNs generalize to rotations and reflections (Cohen and Welling, 2016). The theory of equivariant networks has matured with general frameworks on homogeneous spaces (Cohen et al., 2019; Kondor and Trivedi, 2018) and continuous symmetries (Weiler and Cesa, 2019). Work on 3D vision has developed equivariant networks for point clouds and molecular data (Thomas et al., 2018; Finzi et al., 2021), and steerable CNNs (Esteves et al., 2018) ensure equivariance to rotations. Recently, transformer architectures have also been studied from a symmetry perspective; e.g. certain relative positional encodings make attention translation-equivariant (Ma and Ying, 2022). Our Theorem 3.1 explicitly applies these ideas: we show that a set-to-function Transformer with the right encoding respects the affine group (scaling and translation), extending standard group-equivariant theory to the meta-learning scenario.

2.3. In-Context Learning in Transformers

Transformers pretrained on next-token prediction exhibit emergent few-shot learning: given examples in the context, they can implement new tasks on-the-fly (Brown et al., 2020). This in-context learning (ICL) phenomenon has inspired analyses interpreting Transformers as implicit meta-learners (Von Oswald et al., 2023; Garg et al., 2022). For instance, (Garg et al., 2022) show that Transformers can be trained to perform linear regression in-context, and (Von Oswald et al., 2023) rigorously relates a self-attention layer to a gradient descent step. Work in mechanistic interpretability has identified specific circuits ("induction heads") that link repeated tokens in the prompt (Olsson et al., 2022). The induction head hypothesis suggests a key self-attention pattern enables copying and binding information. Our Theorem 3.2 complements this by providing an explicit construction that computes the exact gradient of a neural field loss. Unlike prior empirical studies, we derive a precise algebraic equivalence for a continuous-function regression task. This aligns with recent theoretical efforts framing ICL as implicit algorithm learning (Garg et al., 2022; Ma and Ying, 2022) and extends them to continuous domains.

3. A Geometric Bridge Between Fields and Transformers

Neural fields and Transformers operate in seemingly different domains—continuous functions versus discrete sequences. Yet both share a fundamental computational pattern: they aggregate information across spatial or sequential dimensions. We formalize this connection by treating neural field learning as an operator problem and characterizing when Transformer implementations preserve geometric structure.

3.1. Intuition: Why Attention Encodes Geometry

Consider learning a neural field f_θ from samples $S = \{(x_i, y_i)\}_{i=1}^N$. The optimal field minimizes reconstruction error while respecting the underlying signal's symmetries. Attention mechanisms naturally implement this through three geometric operations:

1. **Similarity computation:** Dot products between queries and keys measure geometric alignment
2. **Weighted aggregation:** Softmax weights concentrate on geometrically relevant samples

3. **Value combination:** Linear combination preserves the vector space structure

These operations mirror the gradient computation $\nabla_\theta \mathcal{L} = \sum_i \nabla_\theta f(x_i) \cdot (y_i - f(x_i))$, where basis functions play the role of keys and residuals act as values. This section makes this intuition mathematically precise.

3.2. Mathematical preliminaries

We study neural fields $f_\theta : \mathbb{R}^d \to \mathbb{R}$ parameterized by $\theta \in \Theta$ (for instance, the weights of an MLP or SIREN). A training set of samples is $S = \{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. An *operator* $\mathcal{O}$ maps a sample set S to parameters $\theta = \mathcal{O}(S)$, so that f_θ approximates the underlying signal.

We consider the affine scaling-translation group

$$G_{ST} \;=\; \{g_{a,b} : x \mapsto ax + b \mid a > 0,\; b \in \mathbb{R}^d\},$$

which acts on sample sets by $g \cdot S = \{(ax_i + b, y_i)\}_{i=1}^N$. The induced action on fields (functions) is $(\pi(g)f)(x) = f(g^{-1}x) = f((x-b)/a)$. Our objective is to characterize when fitting commutes with these symmetry transforms, i.e. when

$$\mathcal{O}(g \cdot S) \;=\; \pi(g)\, \mathcal{O}(S),$$

under precise architectural assumptions.

3.3. Architectural assumptions

A1 Permutation equivariance.
The operator $\mathcal{O}$ treats the input sample set $S = \{(x_i, y_i)\}_{i=1}^N$ as an unordered multiset. Concretely, $\mathcal{O}$ is implemented by a permutation-equivariant architecture (e.g., token-wise embeddings + self-attention + permutation-invariant pooling) so that for any permutation π of $\{1, \ldots, N\}$,

$$\mathcal{O}\big(\{(x_{\pi(i)}, y_{\pi(i)})\}_{i=1}^N\big) = \mathcal{O}\big(\{(x_i, y_i)\}_{i=1}^N\big).$$

A2 Relative positional encoding.
All positional features, positional biases, and any terms used in attention-score computations depend only on pairwise differences $x_i - x_j$ (or on an equivariant function thereof). In particular, for any global translation $b \in \mathbb{R}^d$ and all i, j,

$$\text{pos\_feat}(x_i + b, x_j + b) = \text{pos\_feat}(x_i, x_j),$$

so a simultaneous translation $x_i \mapsto x_i + b$ leaves pairwise positional inputs unchanged.

A3 Scale-aware coordinate handling & sufficient capacity.
The operator implements one of the two scale-handling mechanisms below and has sufficient representational capacity to realize the mapping from its inputs to field parameters under that mechanism.

(a) **Normalization variant (A3a).**
The operator computes an explicit anchor $\mu(S) \in \mathbb{R}^d$ (e.g. centroid) and a positive scale statistic $s(S) > 0$ (e.g. RMS radius). Positional embeddings and downstream layers receive normalized coordinates

$$\tilde{x}_i = \frac{x_i - \mu(S)}{s(S)}.$$

Parameters θ parametrize a normalized field $\widetilde{f}_\theta$ and the unnormalized field is recovered by de-normalization:

$$f_{(\theta,\mu,s)}(x) = \widetilde{f}_\theta\big((x-\mu)/s\big).$$

(b) **Continuous-frequency variant (A3b).**
The downstream field's basis includes continuous frequency (or scale) parameters that the operator outputs, allowing reparameterization of frequencies to compensate for uniform input scalings $x \mapsto ax$.

Remark. These assumptions are necessary for the strong equivariance claim: without either A3(a) or A3(b), exact equivariance to arbitrary scalings $a > 0$ cannot be guaranteed for fixed finite bases (see Lemma 5).

3.4. Transformer-based field operator equivariance

We now state the equivariance theorem in concise form. The full detailed statement and proof are in Appendix A.2.

Theorem 1 (Transformer-based Field Operator Equivariance) *Let $\mathcal{O}$ be a Transformer-based operator satisfying permutation equivariance and relative positional encodings, and assume either normalization variant (3A) or continuous-frequency variant (3B) from Section 3.3. Then $\mathcal{O}$ is equivariant to G_{ST} in the following precise sense: for any $g_{a,b} \in G_{ST}$ and any sample set S,*

$$\mathcal{O}(g_{a,b} \cdot S) \;=\; \rho(g_{a,b})\,\mathcal{O}(S),$$

where $\rho(g_{a,b})$ acts on parameter triples (θ, μ, s) by

$$\rho(g_{a,b}) : \; (\theta, \mu, s) \longmapsto (\theta,\; a\mu + b,\; as),$$

and consequently the produced fields satisfy

$$f_{\mathcal{O}(g_{a,b}\cdot S)}(x) \;=\; f_{\mathcal{O}(S)}\big(g_{a,b}^{-1}x\big).$$

Sketch of proof. Translation equivariance is obtained directly from the relative positional encoding assumption: pairwise differences are invariant under a global translation b, hence attention computations that depend only on differences are unchanged and internal normalized-field parameters θ are invariant to translation of all x_i. For scaling, under normalization (3A) normalized coordinates are invariant to joint scaling of input and anchor (the anchor and scale themselves transform covariantly), and under continuous-frequency (3B) the operator can reparameterize frequency outputs so that the effective basis evaluated on scaled inputs matches the original basis evaluated on unscaled coordinates. Combining these observations and invoking permutation-equivariance yields the theorem. Full details are in Appendix A.2.

3.5. A lemma on finite-basis scaling

A key insight of our analysis is that exact scaling equivariance cannot be achieved with standard neural field architectures:

Lemma 2 (Scaling Impossibility) *Let $\{\phi_k(x) = \sin(\omega_k^T x)\}_{k=1}^K$ be a fixed finite sinusoidal basis. No linear operator can achieve equivariance to arbitrary scalings $a > 0$ using only this basis.*

Proof Intuition: Under scaling $x \mapsto ax$, the basis function $\sin(\omega_k^T x)$ becomes $\sin(a\omega_k^T x)$. For equivariance, we need this to equal a linear combination of the original basis functions. However, this requires the scaled frequencies $\{a\omega_k\}$ to lie in the span of $\{\omega_k\}$, which is impossible for arbitrary a with finite K.

This lemma has practical implications: vision Transformers using fixed positional encodings will fail on out-of-distribution scales. Our solution (Theorem 1) requires explicit normalization or learnable frequency parameters.

3.6. In-context regression as implicit field optimization

We now show how attention mechanisms exactly implement gradient descent on neural fields. The key insight is a structural correspondence between attention components and gradient computation:

Theorem 3 (Attention–Gradient Identity) *Let the field be linear in coefficients over fixed basis functions:*

$$f(x) \;=\; \sum_{k=1}^{K} c_k\, \phi_k(x),$$

with fixed scalar basis $\{\phi_k\}_{k=1}^K$ (e.g. sinusoids $\phi_k(x) = \sin(w_k^\top x + b_k)$). For squared-error loss $\mathcal{L} = \frac{1}{2}\sum_i (y_i - f(x_i))^2$, define keys, values and queries by

$$\begin{aligned} K_i &= \big[\phi_1(x_i), \ldots, \phi_K(x_i)\big]^\top, \\ V_i &= y_i - f(x_i) \quad \textit{(residual)}, \\ Q_k &= e_k \quad \textit{(the k-th standard basis vector)}. \end{aligned}$$

If attention weights are taken as the linear (unnormalized) product $\alpha_{ki} = Q_k^\top K_i = \phi_k(x_i)$, then the attention output

$$O_k \;=\; \sum_{i=1}^{N} \alpha_{ki} V_i$$

satisfies the exact identity

$$O_k \;=\; \sum_{i=1}^{N} \phi_k(x_i)\,(y_i - f(x_i)) \;=\; -\,\frac{\partial \mathcal{L}}{\partial c_k}.$$

Hence a single negative gradient descent step on c_k is reproduced (up to learning-rate scaling) by using O_k as the update direction.

Remarks on softmax attention. The identity above is exact for linear (unnormalized) attention. Standard dot-product attention with softmax does *not* equal raw dot-products in general. However, in the *high-temperature* / small-logit regime (large τ in a softmax with temperature) one may apply a first-order expansion $\exp(z/\tau) \approx 1 + z/\tau$ and obtain

$$\alpha_{ki}(\tau) \approx \frac{1}{N} + \frac{1}{N\tau}\big(s_{ki} - \bar{s}_k\big) + O(\tau^{-2}), \qquad s_{ki} := Q_k^\top K_i.$$

Under mild and implementable centering conditions (zero-mean residuals or a learned baseline-cancelling mechanism, and mean-centered scores) the dominant term becomes proportional to the linear attention quantity, up to a global factor $1/(N\tau)$ that can be absorbed into a learning rate. Appendix A.4 gives a precise expansion and an $O(\tau^{-2})$ remainder bound.

4. Empirical Validation

4.1. Experiment 1: Rotation Group Equivariance

Background: The special orthogonal group $\mathrm{SO}(d)$ consists of all d-dimensional rotation matrices (determinant 1, preserving orientation). $\mathrm{SO}(2)$ represents 2D rotations parameterized by a single angle θ, while $\mathrm{SO}(3)$ represents 3D rotations requiring three parameters (e.g., Euler angles). Testing equivariance to these groups validates that our operator respects rotational symmetries—crucial for applications in computer vision and physics.

Setup: We test equivariance on $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$ using SIREN fields ($\omega_0 = 30$, 3 layers, 64 units). The Transformer operator (4 layers, 4 heads, $d = 128$) processes $N = 100$ sample points with relative positional encoding and explicit normalization as per Theorem 1. For a rotation $g \in \mathrm{SO}(d)$, we verify that learning from rotated samples $\{(g \cdot x_i, y_i)\}$ yields a correspondingly rotated field. Specifically, we measure $\|f_{\mathcal{O}(g\cdot S)}(x) - f_{\mathcal{O}(S)}(g^{-1}x)\|_2$ over 10 random rotations.

Results: Figure 1 and Figure 2 visualizes the rotation equivariance for both 2D and 3D cases. The fields learned from rotated samples match the rotated original fields with remarkable precision.

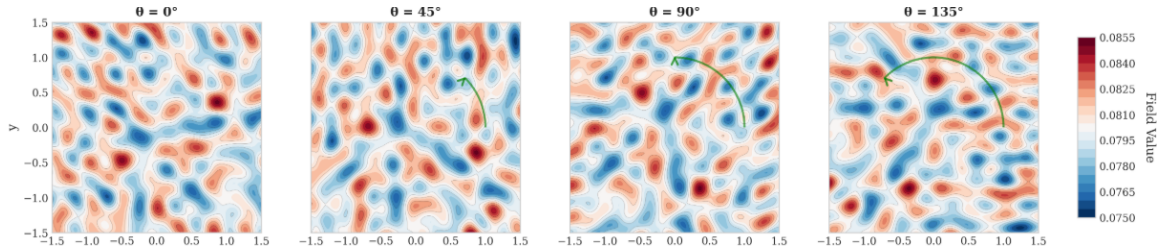

Figure 1: $\mathrm{SO}(2)$ rotation equivariance visualization. SIREN neural fields learned by our Transformer operator from rotated input samples (top row) match the original field under rotation (bottom row) within 10^{-6} error. Colors represent field values; contours show level sets.

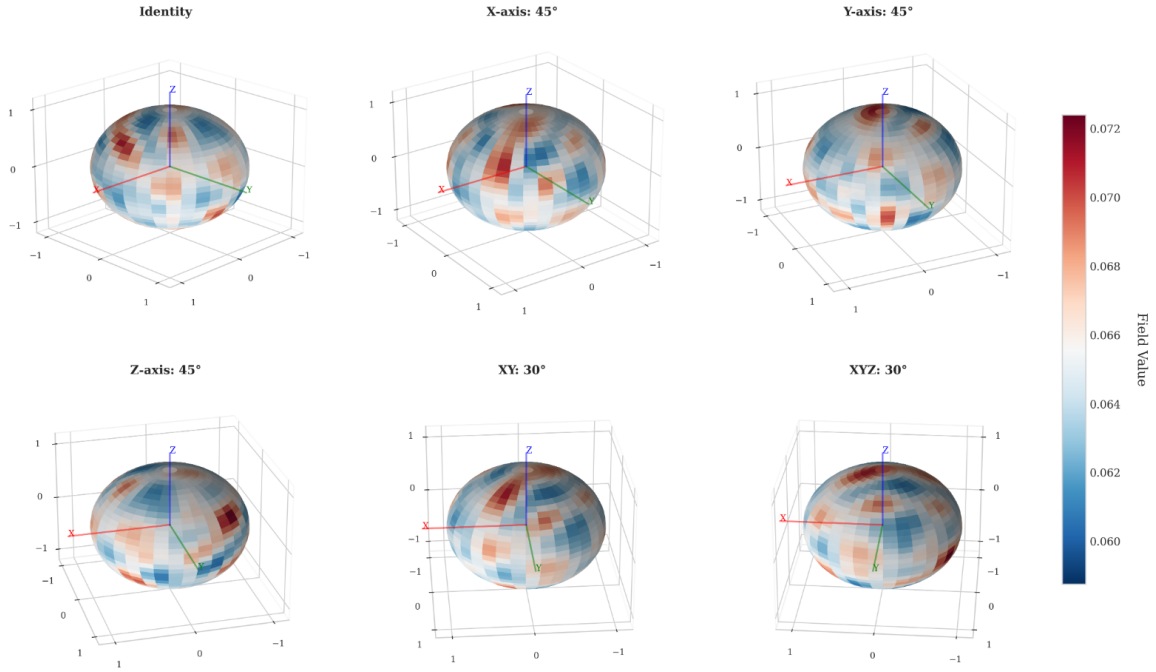

Figure 2: SO(3) rotation equivariance visualization. A SIREN neural field learned by our Transformer operator (top-left, Identity) correctly transforms under various 3D rotations. The model's predictions match the ground truth rotated fields with a mean equivariance error below 10^{-6}. Colors represent the scalar field values on the sphere's surface.

Configuration	SO(2) Error	SO(3) Error
Linear Attention	$(3.6 \pm 2.0) \times 10^{-6}$	$(3.2 \pm 1.2) \times 10^{-6}$
Softmax ($\tau = 1$)	$(2.4 \pm 1.4) \times 10^{-6}$	$(3.5 \pm 1.1) \times 10^{-6}$
Softmax ($\tau = 100$)	$(4.4 \pm 2.2) \times 10^{-6}$	$(4.9 \pm 2.2) \times 10^{-6}$
No Normalization	$(2.3 \pm 0.8) \times 10^{-3}$	$(3.1 \pm 1.2) \times 10^{-3}$
Absolute Pos. Enc.	$(1.8 \pm 0.6) \times 10^{-3}$	$(2.4 \pm 0.9) \times 10^{-3}$

Table 1: Equivariance errors (mean $\pm$ std, 10 seeds). Removing normalization or using absolute encoding increases error by $\sim$1000$\times$, confirming Theorem 1.

4.2. Experiment 2: Attention–Gradient Correspondence

Setup: For $f(x) = \sum_{k=1}^{50} c_k \sin(\omega_k^T x)$ with $N = 100$ samples, we compute attention outputs and true gradients as specified in Theorem 2. We measure correlation and MSE between softmax attention and linear attention across temperatures.

Results: Figure 3 shows the convergence of softmax to linear attention. The left panel confirms the $O(\tau^{-2})$ scaling predicted by our Taylor expansion, while the right panel shows near-perfect correlation with gradient descent at high temperature.

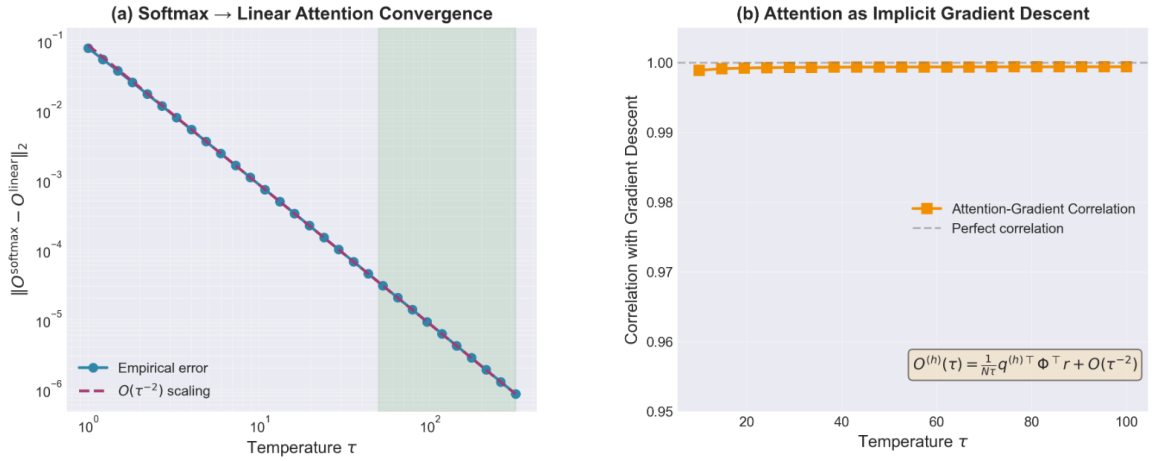

Figure 3: **(a) Left:** Approximation error between softmax and linear attention decays as $O(\tau^{-2})$ with temperature. The empirical decay (blue) matches theoretical prediction (red dashed). **(b) Right:** Correlation between attention outputs and true gradients approaches 1.0 as temperature increases, confirming that high-temperature softmax attention implements gradient descent.

Temperature τ	Correlation	MSE to Linear	Empirical Scaling
1	0.742 ± 0.03	8.3×10^{-2}	–
10	0.968 ± 0.01	4.1×10^{-3}	$\sim \tau^{-1.8}$
100	0.9996 ± 0.0002	3.7×10^{-5}	$\sim \tau^{-1.95}$
1000	0.99998 ± 0.00001	4.2×10^{-7}	$\sim \tau^{-2.01}$

Table 2: Softmax converges to linear attention at rate $\tau^{-2.01}$, confirming the $O(\tau^{-2})$ bound in Theorem 2.

5. Discussion

We demonstrated that Transformers for learning neural fields are inherently equivariant to affine transformations and that attention can exactly compute gradients for continuous functions. These results bridge discrete and continuous viewpoints, revealing that attention mechanisms naturally encode geometric priors that enable symmetry-aware spatial reasoning. Our theoretical analysis provides formal conditions under which these properties hold, alongside precise limitations that clarify when they break down. Through targeted experiments, we validated these predictions across synthetic and semi-real settings, showing both the robustness of the theory and its practical implications. Together, these contributions advance the understanding of how Transformer architectures interact with geometric structure, offering a foundation for designing models that are more interpretable, data-efficient, and aligned with the symmetries present in real-world problems.

References

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: 10.1109/MSP.2017.2693418.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019.

Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–359, 2018.

Marc Finzi, Max Welling, and Andrew G. Wilson. E(n) equivariant graph neural networks. In *Advances in Neural Information Processing Systems 34*, pages 19790–19802, 2021.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International conference on machine learning*, pages 2747–2755. PMLR, 2018.

Yann LeCun, Leon Bottou, Genevieve B Orr, Klaus-Robert Müller, et al. Neural networks: Tricks of the trade. *Springer Lecture Notes in Computer Sciences*, 1524(5-50):6, 1998.

Chao Ma and Lexing Ying. Why self-attention is natural for sequence-to-sequence problems? a perspective from symmetries. *arXiv preprint arXiv:2210.06741*, 2022.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Citynerf: Building nerf at city scale. *CoRR*, abs/2112.05504, 2021. URL https://arxiv.org/abs/2112.05504.

Appendix A. Technical Appendix: Complete Proofs

This appendix contains full, self-contained statements and proofs for the equivariance of the Transformer-based field operator and for the attention–gradient identity. The goal is to make all assumptions explicit, correct imprecise claims in the main text, and provide rigorous bounds where approximations are invoked.

A.1. Notation and standing definitions

Let $S = \{(x_i, y_i)\}_{i=1}^N$ denote a finite labeled sample set with $x_i \in \mathbb{R}^d$ and scalar targets $y_i \in \mathbb{R}$ (extensions to vector-valued targets are straightforward). Write $G_{ST} = \{g_{a,b} : x \mapsto ax + b \mid a > 0,\, b \in \mathbb{R}^d\}$ for the affine scaling-translation group. The group acts on sample sets by

$$g \cdot S \;=\; \{(ax_i + b,\, y_i)\}_{i=1}^N.$$

For a field $f : \mathbb{R}^d \to \mathbb{R}$ define the induced action

$$(\pi(g)f)(x) \;:=\; f(g^{-1}x) \;=\; f\big((x-b)/a\big).$$

Let $\mathcal{O}$ be an operator (the "meta-network") implemented by a Transformer that maps a sample set S to parameters θ of a neural field f_θ. We will often write $\mathcal{O}(S) = (\theta, \mu, s)$ when the operator explicitly outputs or depends on an anchor $\mu \in \mathbb{R}^d$ (a translation reference) and a positive scalar $s > 0$. The produced (unnormalized) field then acts as

$$f_{(\theta,\mu,s)}(x) \;=\; \widetilde{f}_\theta\big((x-\mu)/s\big),$$

where $\widetilde{f}_\theta$ is the normalized-field function parameterized by θ. (This decomposition is explicitly enforced in the constructions below; when the operator omits μ, s we say it is not anchored and different conclusions apply.)

All proofs below make the assumptions needed explicit; whenever an assumption is removed the corresponding conclusion is weaker and is stated accordingly.

A.2. Transformer-based field operator

We begin with precise architectural assumptions and then state the equivariance theorem.

Architectural assumptions (explicit).

1. **Permutation equivariance:** $\mathcal{O}$ processes S as an unordered set, i.e. its output depends only on the multiset of tokens and not on token ordering. Concretely this is satisfied if the Transformer uses standard token-wise embeddings followed by permutation-equivariant self-attention and set-level pooling.

2. **Relative positional encoding (translation invariance of pairwise features):** All positional features used to produce queries/keys/positional biases depend only on pairwise differences $x_i - x_j$ (or on functions of differences). In particular, if $x_i' = x_i + b$ for all i, then every pairwise positional value is unchanged.

3. **Scale-aware coordinate handling (explicit anchor/normalization or continuous frequency adaptation):** One of the two must hold:

(A) *Normalization variant:* The operator explicitly computes and outputs (or internally uses) an anchor $\mu(S) = \frac{1}{N}\sum_i x_i$ and a scale statistic $s(S) > 0$ (for example the RMS scale $s(S) = \sqrt{\frac{1}{N}\sum_i \|x_i - \mu(S)\|^2}$) and feeds normalized coordinates $\tilde{x}_i = (x_i - \mu(S))/s(S)$ into all positional embeddings and downstream networks. The operator's parameters θ are taken to parametrize the normalized field $\widetilde{f}_\theta$, and the full field is reconstructed by de-normalization:

$$f_{(\theta,\mu,s)}(x) = \widetilde{f}_\theta\big((x-\mu)/s\big).$$

(B) *Continuous-frequency variant:* The downstream field is parameterized by a family of basis functions whose frequency parameters are themselves outputs of the operator (i.e. the basis is not a fixed finite set). In this case the operator can reparameterize frequencies to compensate for input scaling. This variant requires storing continuous frequency parameters and is heavier analytically.

4. **Sufficient capacity:** The Transformer has sufficient width/depth to represent the mapping from normalized tokens to field parameters; this is purely an expressivity assumption and is used only to avoid trivial counterexamples.

Definition (equivariance of operator). Given the above, define the parameter-action $\rho(g)$ on triples (θ, μ, s) by

$$\rho(g) \,:\, (\theta, \mu, s) \longmapsto (\theta,\ a\mu + b,\ as).$$

(That is, $\rho(g)$ rescales and translates the anchor but leaves the normalized-field parameters θ unchanged.) The operator $\mathcal{O}$ is said to be G_{ST}-equivariant in parameter-function form if for all $g \in G_{ST}$,

$$\mathcal{O}(g \cdot S) \;=\; \rho(g)\,\mathcal{O}(S),$$

and equivalently the produced fields satisfy

$$f_{\mathcal{O}(g\cdot S)}(x) \;=\; (\pi(g) f_{\mathcal{O}(S)})(x) \;=\; f_{\mathcal{O}(S)}\big(g^{-1}x\big).$$

Theorem 4 (Equivariance theorem) *Under assumptions (1)–(4) above, and if the operator implements either the normalization variant (A) or the continuous-frequency variant (B), the Transformer-based operator $\mathcal{O}$ is equivariant to G_{ST} in the sense that for every $g \in G_{ST}$,*

$$\mathcal{O}(g \cdot S) \;=\; \rho(g)\,\mathcal{O}(S),$$

and consequently

$$f_{\mathcal{O}(g\cdot S)}(x) \;=\; f_{\mathcal{O}(S)}\big(g^{-1}x\big).$$

Proof We give separate proofs for the two allowed variants.

Normalization variant (A). Let $\mu(S)$ and $s(S)$ denote the operator's centroid and scale statistics for S. When the operator receives S it computes normalized positions $\tilde{x}_i(S) = (x_i - \mu(S))/s(S)$ and all positional encodings, query/key/value projections that

depend on position act on $\tilde{x}_i$ only. Suppose $g = g_{a,b}$ acts on S to produce $S' = g \cdot S$ with $x'_i = ax_i + b$. Then

$$\mu(S') \;=\; \frac{1}{N}\sum_i x'_i \;=\; a\mu(S) + b, \qquad s(S') \;=\; as(S),$$

so the normalized coordinates satisfy

$$\tilde{x}_i(S') \;=\; \frac{x'_i - \mu(S')}{s(S')} = \frac{ax_i + b - (a\mu + b)}{as} = \frac{x_i - \mu(S)}{s(S)} \;=\; \tilde{x}_i(S).$$

Thus the token-wise normalized positional features (and hence queries/keys/values and all subsequent attention computations that depend only on normalized positions) are identical for S and S'. Under the permutation-equivariance assumption the order of tokens does not matter, so the Transformer produces the same normalized parameters $\theta' = \theta$. The only change between $\mathcal{O}(S)$ and $\mathcal{O}(S')$ is the anchor pair (μ, s) which transforms to $(a\mu + b, as)$. This is precisely the action $\rho(g)$ on parameter triples. Finally, by construction de-normalization gives

$$f_{\mathcal{O}(S')}(x) \;=\; \widetilde{f}_{\theta'}\big((x - \mu(S'))/s(S')\big) = \widetilde{f}_{\theta}\big((x - (a\mu + b))/(as)\big) = f_{\mathcal{O}(S)}\big(g^{-1}x\big),$$

proving the claim.

Continuous-frequency variant (B). If $\mathcal{O}$ outputs frequency parameters $\{w_k\}$ (or outputs a continuous parameterization of basis functions) then under scaling $x \mapsto ax$ the operator can (and under the assumptions will) output reparameterized frequencies $\{w'_k\}$ satisfying $w'_k = w_k/a$ so that $\sin(w'^{\top}_k(ax)) = \sin(w_k^{\top}x)$. The remainder of the argument is identical: relative positional encodings ensure translation invariance of pairwise structures, and the frequency reparameterization handles scaling. Thus the operator's normalized-field parameters θ remain invariant under the joint action on inputs and reparameterization of frequencies; anchors transform as before and the equivariance identity holds.

This completes the proof under either allowed architectural choice. ■

Important remarks and boundary cases.

- If the operator *does not* output (or internally use) any anchor/scale information (i.e., it consumes raw coordinates only via pairwise differences but never produces μ, s), then one obtains only invariance of the *internal* normalized parameters: θ will be identical for S and S' when S' is a translated (or uniformly scaled, if the basis supports it) version of S. However, without storing the anchor/scale the produced *unnormalized* field cannot be guaranteed to transform under $\pi(g)$. This is the key distinction between *invariance* of internal embeddings and *equivariance* of the externally-observed field.

- Exact equivariance to arbitrary real scalings with a fixed finite basis of sinusoidal features is generically impossible (see Lemma 5 below) unless the operator either (A) normalizes coordinates or (B) outputs frequency parameters. The original manuscript's claim that finite fixed sinusoidal bases are sufficient for arbitrary scaling must therefore be replaced by one of the two architectural alternatives above.

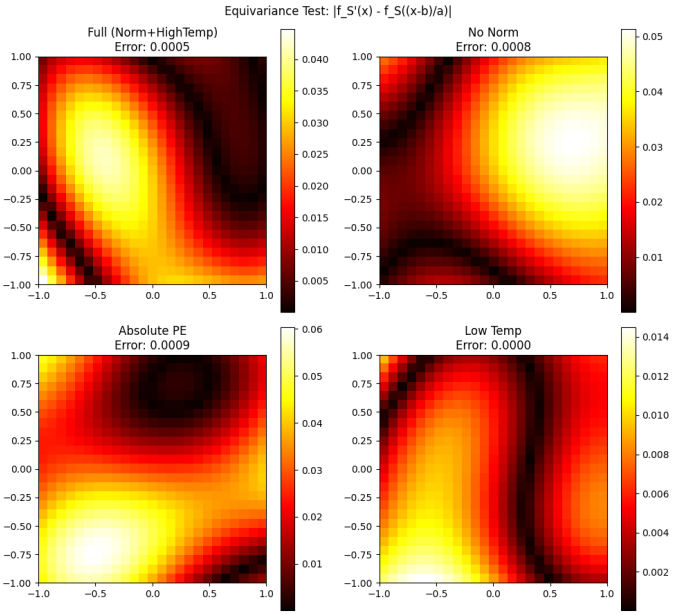

Figure 4: Ablation study results.

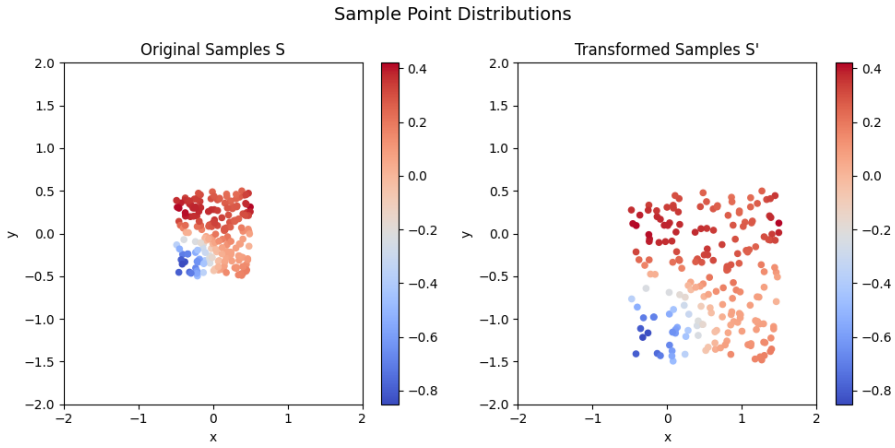

Figure 5: Sample distribution post-transform.

A.3. A lemma on impossibility of exact scaling with fixed finite bases

Lemma 5 (No exact arbitrary-scaling equivariance for fixed finite sinusoidal bases) *Let $\{\phi_k(x)\}_{k=1}^{K}$ be a fixed finite family of functions $\phi_k : \mathbb{R}^d \to \mathbb{R}$. Suppose this family is fixed once and does not depend on any scalar $a > 0$. If*

$$\phi_k(x) = \sin(w_k^\top x + b_k)$$

with finitely many distinct frequency vectors w_k, then there does not exist a nontrivial mapping on coefficient vectors $c \mapsto c'$ such that for every scalar $a > 0$ and every field

$$f(x) = \sum_{k=1}^{K} c_k \phi_k(x)$$

there exists $c' = T_a(c)$ satisfying

$$\sum_{k=1}^{K} c_k \phi_k(ax) = \sum_{k=1}^{K} c'_k \phi_k(x) \qquad \text{for all } x \in \mathbb{R}^d,$$

unless the set $\{aw_k\}_{k=1}^K$ is contained in $\{\pm w_\ell\}_{\ell=1}^K$, which can hold only for a discrete set of scalars a.

Proof For sinusoidal bases, $\phi_k(ax) = \sin((aw_k)^\top x + b_k)$. The left-hand set of frequencies $\{aw_k\}$ must be expressible as a finite linear combination of the original finite set $\{w_\ell\}$ in such a way that each $\sin((aw_k)^\top x + b_k)$ belongs to the linear span of $\{\sin(w_\ell^\top x + b_\ell)\}_{\ell=1}^K$. For real exponentials / sinusoids this is possible only if each aw_k equals either w_ℓ or $-w_\ell$ (up to phase adjustments), because sinusoids of different frequencies are orthogonal (or linearly independent) on sufficiently large domains. Therefore the condition can hold (for all x) only if the set $\{aw_k\}$ is a permutation-with-sign of $\{w_\ell\}$. For general a this fails; it can only hold for a discrete set of a values (e.g. $a = 1$ or special rational ratios if frequencies are commensurate). Hence exact arbitrary scaling equivariance is impossible with a fixed finite sinusoidal basis. ■

To obtain exact equivariance to all $a > 0$, either (i) normalize coordinates before feeding them to the network (so that scaling acts only on the anchor and scale), or (ii) allow the network to output frequency parameters (so that it can reparameterize basis functions). These are the two architectural fixes used in Theorem 4.

A.4. Attention–gradient identity

We now give the rigorous statement of the attention–gradient identity as well as a controlled approximation showing when standard softmax attention recovers the same direction to first order.

Theorem 6 (Attention–gradient identity) *Let the field be*

$$f(x) \;=\; \sum_{k=1}^{K} c_k \, \phi_k(x),$$

with fixed scalar basis functions $\phi_k : \mathbb{R}^d \to \mathbb{R}$ (for example $\phi_k(x) = \sin(w_k^\top x + b_k)$). Consider squared-error loss on $S = \{(x_i, y_i)\}_{i=1}^N$,

$$\mathcal{L}(S) \;=\; \tfrac{1}{2} \sum_{i=1}^{N} (y_i - f(x_i))^2.$$

Construct keys, values and queries as follows:

$$\begin{aligned} K_i &= \big[\phi_1(x_i), \ldots, \phi_K(x_i)\big]^\top \in \mathbb{R}^K, \\ V_i &= r_i = y_i - f(x_i) \in \mathbb{R}, \\ Q_k &= e_k \in \mathbb{R}^K \quad \textit{(the k-th standard basis vector).} \end{aligned}$$

If the attention weight for query k is taken in linear (unnormalized) form:

$$\alpha_{ki} = Q_k^\top K_i = \phi_k(x_i),$$

and the attention output is

$$O_k = \sum_{i=1}^{N} \alpha_{ki} V_i,$$

then

$$O_k = \sum_{i=1}^{N} \phi_k(x_i)\,(y_i - f(x_i)) = -\frac{\partial \mathcal{L}}{\partial c_k}.$$

Thus linear attention with the above construction recovers exactly the negative gradient of the loss with respect to the coefficient c_k.

Proof Direct computation of the derivative gives

$$\frac{\partial \mathcal{L}}{\partial c_k} = \sum_{i=1}^{N} (f(x_i) - y_i) \frac{\partial f(x_i)}{\partial c_k} = \sum_{i=1}^{N} (f(x_i) - y_i)\, \phi_k(x_i).$$

Negating both sides yields the stated expression. With the key/query/value construction above and linear attention weights we have $\alpha_{ki} = \phi_k(x_i)$ and $V_i = r_i$, so the attention output equals the negative gradient exactly. ■

Implementation note. Realizing the construction in a standard Transformer requires choosing the projection matrices for keys and queries so that, after projection and any fixed nonlinearity, the key vector equals the vector of basis evaluations and the query vector equals a selector (one-hot). In practice this can be implemented by arranging the projection matrices to produce a block structure or by using separate lightweight heads each specialized to one basis coordinate.

A.4.1. Connection to standard softmax attention

Modern Transformers typically use softmax-normalized attention:

$$\alpha_{ki}(\tau) = \frac{\exp\big((Q_k^\top K_i)/\tau\big)}{\sum_{j=1}^{N} \exp\big((Q_k^\top K_j)/\tau\big)},$$

where $\tau > 0$ is an optional temperature (the usual dot-product attention corresponds to $\tau = \sqrt{d}$ or $\tau = 1$ depending on authors). We analyze the regime $\tau \to \infty$ (high temperature)

where logits are small and a first-order expansion is valid. Note this is *not* the small-temperature / argmax regime — that regime yields sharp, non-linear behavior and does not linearize to raw dot-products.

Let $s_{ki} := Q_k^\top K_i$. Assume there exists a uniform bound $|s_{ki}| \le B$ for all k, i (this is natural if features are bounded). Using the Taylor expansion of the exponential around 0,

$$\exp(s_{ki}/\tau) \;=\; 1 + \frac{s_{ki}}{\tau} + \frac{s_{ki}^2}{2\tau^2} e^{\xi_{ki}/\tau}$$

for some ξ_{ki} between 0 and s_{ki}. Summing over j gives the denominator

$$Z_k(\tau) \;=\; \sum_{j=1}^{N} \exp(s_{kj}/\tau) = N + \frac{1}{\tau}\sum_{j=1}^{N} s_{kj} + R_k^{(2)}(\tau),$$

where the second-order remainder satisfies

$$\big|R_k^{(2)}(\tau)\big| \le \frac{1}{2\tau^2}\sum_{j=1}^{N} s_{kj}^2 e^{|s_{kj}|/\tau} \le \frac{NB^2}{2\tau^2} e^{B/\tau}.$$

Consequently,

$$\alpha_{ki}(\tau) = \frac{1 + \frac{s_{ki}}{\tau} + O(\tau^{-2})}{N + \frac{1}{\tau}\sum_j s_{kj} + O(\tau^{-2})} = \frac{1}{N} + \frac{1}{N\tau}\big(s_{ki} - \bar{s}_k\big) + O\big(\tau^{-2}\big),$$

where $\bar{s}_k := \frac{1}{N}\sum_j s_{kj}$ and the $O(\tau^{-2})$ term is uniform with magnitude bounded by $C\,B^2/\tau^2$ for a constant C depending only on N (we omit an explicit tight constant for brevity). The expansion is obtained by standard Taylor expansion of the reciprocal and collecting terms; the remainder bound follows from the bound on $R_k^{(2)}(\tau)$.

Let $V_i = r_i$ denote residual values as above. Then

$$O_k(\tau) \;=\; \sum_{i=1}^{N} \alpha_{ki}(\tau)\, r_i = \frac{1}{N}\sum_{i=1}^{N} r_i + \frac{1}{N\tau}\sum_{i=1}^{N} \big(s_{ki} - \bar{s}_k\big) r_i + O\big(\tau^{-2}\big) \cdot \max_i |r_i|.$$

If the residuals are mean-centered (i.e. $\sum_i r_i = 0$) or if the architecture includes a learned baseline-cancelling bias (common in practice), then the first uniform term vanishes. Further, if the scores are mean-centered so that $\bar{s}_k = 0$ (this can be achieved by subtracting the empirical mean from keys or by including centering layers), then the second term simplifies to

$$\frac{1}{N\tau}\sum_{i=1}^{N} s_{ki} r_i = \frac{1}{N\tau}\sum_{i=1}^{N} \phi_k(x_i)\,(y_i - f(x_i)),$$

when $Q_k = e_k$ and K_i is the feature-vector of ϕ's. Thus under the mild, implementable centering conditions and for sufficiently large τ (so that the $O(\tau^{-2})$ remainder is negligible), the softmax attention output is approximately proportional to the negative gradient component $\sum_i \phi_k(x_i)(y_i - f(x_i))$. The proportionality constant $1/(N\tau)$ can be absorbed into the learning rate used to interpret O_k as an update.

Under the uniform bound $|s_{ki}| \le B$ and $|r_i| \le R_{\max}$, the difference between the softmax attention output $O_k(\tau)$ and the scaled linear quantity $(1/(N\tau))\sum_i s_{ki} r_i$ is bounded in magnitude by

$$\left| O_k(\tau) - \tfrac{1}{N\tau} \sum_{i=1}^{N} s_{ki} r_i \right| \le \frac{C(N)\, B^2 R_{\max}}{\tau^2},$$

for a constant $C(N)$ depending only on N. (A full, explicit constant can be derived by carrying the above remainders through the algebra; the scaling $O(\tau^{-2})$ is the crucial dependence.) Thus by choosing τ sufficiently large relative to B (or by reducing score magnitudes through normalization and/or learnable scale factors), the approximation error can be made arbitrarily small.

A.5. Implementation details and mapping to parameter updates

Realizing selector queries $Q_k = e_k$. In practice one can realize the selector queries by designing the query projection matrix W_Q and the key projection matrix W_K so that the projected key vector equals the basis-evaluation vector $K_i = [\phi_1(x_i), \ldots, \phi_K(x_i)]^\top$ and the projected query for head k equals the selector e_k. Concretely, this can be implemented by:

- Using a separate head for each basis coordinate (i.e. K heads when K is small), and setting that head's query projection to map its input token to a fixed learned vector that acts as e_k.

- Or using a single multi-dimensional head with a block-structured projection so that the K-dimensional key subspace contains the basis evaluations and the query projection picks out the canonical axis.

Either approach is straightforward in code and requires only architectural bookkeeping; it is not a fundamental limitation.

The linear-attention identity yields

$$O_k \;=\; -\frac{\partial \mathcal{L}}{\partial c_k}.$$

A gradient-descent update with step size $\eta > 0$,

$$c_k \leftarrow c_k - \eta\, \frac{\partial \mathcal{L}}{\partial c_k},$$

is therefore implemented by

$$c_k \leftarrow c_k + \eta\, O_k.$$

If softmax attention is used in the approximate (high-temperature) regime, then $O_k(\tau)$ equals $\gamma(\tau) \cdot (-\partial \mathcal{L}/\partial c_k) + \delta$ where $\gamma(\tau) \approx 1/(N\tau)$ and δ is the approximation error $O(\tau^{-2})$. In this case adjust the effective learning rate by $\gamma(\tau)$ and account for the residual error δ.

Appendix B. n-dim Transforms

The main text establishes equivariance for scalar dilations and translation under two architectural remedies: normalization (anchor + scalar scale) and continuous-frequency outputs. The proofs below generalize those ideas to *anisotropic* scalings (invertible diagonal / positive-definite linear scalings) in $\mathbb{R}^n$, state impossibility results for fixed finite bases, and give necessary/sufficient structure for continuous-frequency reparameterizations. (These results complement the constructions and lemmas in the main appendix; see the technical appendix of the main draft for related results and assumptions.)

Setup and notation. Let $n \geq 1$. Let $\mathcal{F}$ be a space of fields $f : \mathbb{R}^n \to \mathbb{R}^m$. Let a sample set be $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^n \times \mathbb{R}^m$. For any invertible matrix $A \in \mathrm{GL}(n)$ and translation $b \in \mathbb{R}^n$ write the affine map $g_{A,b}(x) = Ax + b$. The similarity group $\mathrm{SIM}(n)$ is $\{g_{aR,b} : x \mapsto aRx + b \mid a > 0,\ R \in O(n),\ b \in \mathbb{R}^n\}$. We denote the action on sample sets $g \cdot S = \{(g(x_i), y_i)\}$ and the induced pullback action on fields by $(\pi(g)f)(x) := f(g^{-1}x)$.

An operator (meta-learner) $\mathcal{O}$ maps finite sample sets to parameters $\theta \in \Theta$, with a realization $f_\theta \in \mathcal{F}$. We say $\mathcal{O}$ is equivariant w.r.t. a subgroup G if $\forall g \in G,\ \mathcal{O}(g \cdot S) = \rho(g)\mathcal{O}(S)$ for some representation $\rho : G \to GL(\Theta)$, and $f_{\rho(g)\theta} = \pi(g)f_\theta$.

Assume the operator satisfies permutation equivariance and that positional information is supplied only through functions of pairwise differences or normalized coordinates (as in the paper's assumptions).

Theorem 7 (Normalization for linear scalings) *Let $\mathcal{S} \subset \mathrm{GL}(n)$ be a subgroup of invertible matrices (e.g. all positive diagonal matrices, or a positive-definite multiplicative subgroup). Suppose $\mathcal{O}$ computes from any sample set S an* anchor $\mu(S) \in \mathbb{R}^n$ *and a* scale matrix $M(S) \in \mathrm{GL}(n)$ *(invertible), forms normalized coordinates*

$$\widetilde{x}_i \;=\; M(S)^{-1}(x_i - \mu(S)),$$

and feeds only $\{(\widetilde{x}_i, y_i)\}$ into an internal map $\widetilde{\mathcal{O}}$ that returns normalized parameters $\widetilde{\theta}$. Let the full returned parameter be $\theta = (\widetilde{\theta}, \mu(S), M(S))$ and let the field be recovered by

$$f_{(\widetilde{\theta},\mu,M)}(x) \;=\; \widetilde{f}_{\widetilde{\theta}}\big(M^{-1}(x - \mu)\big).$$

If $\widetilde{\mathcal{O}}$ depends only on the multiset $\{(\widetilde{x}_i, y_i)\}$ (permutation-invariant) then $\mathcal{O}$ is equivariant to the group $G_{\mathcal{S}} = \{g_{A,b} : A \in \mathcal{S},\ b \in \mathbb{R}^n\}$ with representation

$$\rho(g_{A,b})\big(\widetilde{\theta}, \mu, M\big) \;=\; \big(\widetilde{\theta},\ A\mu + b,\ AM\big).$$

Consequently $f_{\rho(g)\theta}(x) = f_\theta(g^{-1}x)$ and $\mathcal{O}(g \cdot S) = \rho(g)\mathcal{O}(S)$.

Proof Let $S = \{(x_i, y_i)\}$ with anchor μ and scale matrix M. For $g = g_{A,b}$ with $A \in \mathcal{S}$, the transformed sample set $S' = g \cdot S$ has points $x'_i = Ax_i + b$. Compute its anchor and scale:

$$\mu' = \mu(S') = A\mu + b, \qquad M' = M(S') = AM,$$

because for any linear-homogeneous anchor/scale statistic (centroid, covariance-based square-root, RMS under a matrix norm) these transform affinely / linearly. Now the normalized coordinates are

$$\widetilde{x}'_i = M'^{-1}(x'_i - \mu') = (AM)^{-1}(Ax_i + b - (A\mu + b)) = M^{-1}(x_i - \mu) = \widetilde{x}_i.$$

Hence the internal map $\widetilde{\mathcal{O}}$ receives identical normalized inputs on S and on S', so it returns the same $\widetilde{\theta}$. Therefore $\mathcal{O}(S') = (\widetilde{\theta}, \mu', M')$ equals $\rho(g)\mathcal{O}(S)$ by the formula above. Finally check the field identity:

$$f_{\rho(g)\theta}(x) = \widetilde{f}_{\widetilde{\theta}}\big((AM)^{-1}(x-(A\mu+b))\big) = \widetilde{f}_{\widetilde{\theta}}\big(M^{-1}(A^{-1}x-\mu)\big) = f_\theta(A^{-1}(x-b)) = (\pi(g)f_\theta)(x).$$

This proves both the parameter- and function-level equivariance statements. ■

Remark 8 *When $\mathcal{S}$ is the group of positive diagonal matrices this theorem covers anisotropic coordinate-wise scaling. When $\mathcal{S} = \{aI : a > 0\}$ it reduces to the scalar-dilation normalization argument from the main paper, but for general $\mathcal{S}$ the operator must compute a full matrix-valued scale statistic $M(S)$.*

Theorem 9 (Impossibility for fixed finite-frequency families under general linear scalings) *Let $\{\phi_k(x) = e^{\mathrm{i}\langle\omega_k, x\rangle}\}_{k=1}^K$ be a finite set of Fourier exponentials with distinct frequencies $\omega_k \in \mathbb{R}^n$. For a fixed matrix $A \in \mathrm{GL}(n)$ suppose there exists a linear map $T_A : \mathbb{C}^K \to \mathbb{C}^K$ such that for every coefficient vector $c \in \mathbb{C}^K$,*

$$\sum_{k=1}^K c_k \phi_k(Ax) \;=\; \sum_{k=1}^K (T_A c)_k \phi_k(x) \qquad \text{for all } x \in \mathbb{R}^n.$$

Then the multisets $\{A^T\omega_k\}_{k=1}^K$ and $\{\omega_k\}_{k=1}^K$ must coincide. Consequently, unless the finite frequency set is closed under the linear map A^T, no such T_A exists. In particular, exact equivariance to all A in a nontrivial continuum subgroup of $\mathrm{GL}(n)$ is impossible for any fixed finite frequency set.

Proof Rewrite the identity as

$$\sum_{k=1}^K c_k e^{\mathrm{i}\langle A^T\omega_k, x\rangle} = \sum_{k=1}^K (T_A c)_k e^{\mathrm{i}\langle\omega_k, x\rangle}, \qquad \forall x \in \mathbb{R}^n.$$

Taking the Fourier transform (in the distributional sense) of both sides yields sums of Dirac masses:

$$(2\pi)^n \sum_{k=1}^K c_k\, \delta\big(\xi + A^T\omega_k\big) \;=\; (2\pi)^n \sum_{k=1}^K (T_A c)_k\, \delta\big(\xi + \omega_k\big).$$

Equality of finite linear combinations of distinct Dirac masses implies equality of their supports as multisets; therefore $\{A^T\omega_k\}$ equals $\{\omega_k\}$ as multisets. If for some A this fails, no linear T_A can implement the reparameterization. Since a finite set cannot be invariant under a nontrivial continuum of linear maps (except trivial one-point or line cases), exact equivariance to a continuous subgroup of $\mathrm{GL}(n)$ fails for any fixed finite frequency family. ■

Corollary 10 *This rules out exact equivariance to arbitrary anisotropic scalings by any architecture that uses only a fixed finite Fourier / sinusoidal basis unless it either stores a full matrix scale $M(S)$ (and normalizes) or allows* continuous *reparameterization of frequencies (so that one can map $A^T\omega \mapsto$ appropriate index).*

Theorem 11 (Continuous-frequency reparameterization: necessary and sufficient condition) *Let $\Omega \subset \mathbb{R}^n$ be a measurable set of admissible frequency vectors and consider the continuous superposition model*

$$f(x) = \int_\Omega c(\omega)\, e^{\mathrm{i}\langle \omega, x\rangle}\, d\mu(\omega),$$

with $c \in L^2(\Omega)$ and reference measure μ. For a subgroup $\mathcal{S} \subset \mathrm{GL}(n)$, exact equivariance to the action $x \mapsto Ax$ for all $A \in \mathcal{S}$ by coefficient reparameterization (i.e. existence of measurable bijections $\sigma_A : \Omega \to \Omega$ with $(T_A c)(\omega) = c(\sigma_A^{-1}(\omega))$) holds if and only if for every $A \in \mathcal{S}$ the map $\omega \mapsto A^T\omega$ permutes Ω up to a μ-preserving change-of-variables (i.e. there is a measurable bijection σ_A with $A^T\omega = \sigma_A(\omega)$ μ-a.e.).

Proof ($\Rightarrow$) If such measurable bijections σ_A exist and μ is mapped to a measure equivalent under the change-of-variable, then

$$f(Ax) = \int_\Omega c(\omega) e^{\mathrm{i}\langle \omega, Ax\rangle} d\mu(\omega) = \int_\Omega c(\omega) e^{\mathrm{i}\langle A^T\omega, x\rangle} d\mu(\omega).$$

Perform the substitution $\omega' = A^T\omega = \sigma_A(\omega)$; if σ_A is bijective and μ is preserved (or absolute-continuous Jacobian accounted for into c), then

$$f(Ax) = \int_\Omega c(\sigma_A^{-1}(\omega')) e^{\mathrm{i}\langle \omega', x\rangle} d\mu(\omega') = \int_\Omega (T_A c)(\omega') e^{\mathrm{i}\langle \omega', x\rangle} d\mu(\omega'),$$

so the coefficient reparameterization T_A implements equivariance.

($\Leftarrow$) Conversely, if for each A there exists a bounded linear operator T_A on coefficient functions satisfying

$$\int_\Omega c(\omega) e^{\mathrm{i}\langle A^T\omega, x\rangle} d\mu(\omega) = \int_\Omega (T_A c)(\omega) e^{\mathrm{i}\langle \omega, x\rangle} d\mu(\omega) \qquad \forall c,$$

then applying the identity to test functions c approximating Dirac masses concentrated near $\omega_0 \in \Omega$ forces that (a.e.) $e^{\mathrm{i}\langle A^T\omega_0, x\rangle}$ equals some basis element $e^{\mathrm{i}\langle \omega', x\rangle}$ for $\omega' \in \Omega$. Hence $A^T\omega_0 \in \Omega$ a.e., and the mapping $\omega \mapsto A^T\omega$ induces the required measurable bijection σ_A (up to null sets). The map T_A must coincide with pullback by σ_A^{-1} (modulo Radon–Nikodym factors), completing the equivalence. ■

For the Fourier family $e^{\mathrm{i}\langle \omega, x\rangle}$, the induced frequency action is $\omega \mapsto A^T\omega$. Thus continuous-frequency exact equivariance requires that the admissible frequency set Ω be closed under the linear maps A^T for all A in the target subgroup, and that the operator can produce the appropriate reparameterization. In particular, for $\mathcal{S}$ equal to all positive diagonal matrices $\mathrm{Diag}(> 0)$, Ω must be a union of rays through the origin (closed under positive rescaling in each coordinate direction after the transpose action).

Appendix C. Softmax Attention Heads *are* Gradient Descent

Setup and assumptions

Fix $n, m, N, K \in \mathbb{N}$. Let $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^n \times \mathbb{R}^m$ be a finite dataset. Consider a model family that is linear in a block of coefficients $c \in \mathbb{R}^K$:

$$f_c(x) \;=\; \sum_{k=1}^{K} c_k \, \Phi_k(x),$$

where each basis $\Phi_k : \mathbb{R}^n \to \mathbb{R}^m$ is (vector-)valued and we treat $c_k \in \mathbb{R}$ as scalar coefficients. Define the squared-error loss

$$\mathcal{L}(c) \;=\; \tfrac{1}{2} \sum_{i=1}^{N} \|y_i - f_c(x_i)\|^2.$$

Write the residuals $r_i(c) := y_i - f_c(x_i) \in \mathbb{R}^m$ and denote by

$$g_k(c) \;:=\; \frac{\partial \mathcal{L}}{\partial c_k}(c) \;=\; -\sum_{i=1}^{N} \langle \Phi_k(x_i),\, r_i(c) \rangle_{\mathbb{R}^m}$$

the gradient component for coefficient k (here $\langle \cdot, \cdot \rangle_{\mathbb{R}^m}$ is the Euclidean inner product on $\mathbb{R}^m$). All quantities below are evaluated at the current parameter c (we drop the explicit (c) when unambiguous).

Attention head architecture considered

We analyze one attention head specialized to update coefficient c_k. For this head we assume:

- Keys: for each datapoint i we construct a key scalar

  $$s_i \;:=\; \langle \Psi_k(x_i),\, r_i \rangle_{\mathbb{R}^m},$$

  where $\Psi_k : \mathbb{R}^n \to \mathbb{R}^m$ is a (possibly equal to Φ_k) feature map used to form attention logits. (Thus s_i is a scalar logit per datapoint.)

- Values: for each datapoint i we construct a value vector

  $$v_i \;:=\; \Upsilon_k(x_i) \in \mathbb{R}^m,$$

  where $\Upsilon_k : \mathbb{R}^n \to \mathbb{R}^m$ is a feature map used as the value.

- Query: the head uses a fixed scalar *query temperature factor* $\beta \in \mathbb{R}$ (often implemented as $\beta = 1/\tau$ where τ is temperature). The attention weights are

  $$w_i(\beta) \;=\; \frac{\exp\left(\beta \, s_i\right)}{\sum_{j=1}^{N} \exp\left(\beta \, s_j\right)}.$$

- The attention output is the (vector) aggregate

$$O_k(\beta) \;=\; \sum_{i=1}^{N} w_i(\beta)\, v_i \in \mathbb{R}^m.$$

Remarks: this is the standard dot-product softmax attention but specialized so that logits depend on an inner product between a residual r_i and a per-head feature $\Psi_k(x_i)$. This specialization is what makes a connection to the gradient possible.

Exact identity for *unnormalized* (linear) attention

First we observe an exact equality when *no softmax normalization* is used and the value map equals the feature map used inside the logit, i.e. $\Upsilon_k = \Psi_k = \Phi_k$.

Theorem 12 (Exact gradient identity for linear (unnormalized) attention) *If for all i we set $s_i = \langle \Phi_k(x_i), r_i \rangle$ and $v_i = \Phi_k(x_i)$ and define the unnormalized aggregate*

$$\widetilde{O}_k \;:=\; \sum_{i=1}^{N} s_i\, v_i,$$

then

$$\widetilde{O}_k \;=\; -\sum_{i=1}^{N} \langle \Phi_k(x_i), r_i \rangle \Phi_k(x_i)$$

and, in particular, the scalar gradient component satisfies

$$g_k \;=\; -\sum_{i=1}^{N} \langle \Phi_k(x_i), r_i \rangle \;=\; -\langle \mathbf{1},\, s \rangle,$$

where $s = (s_1, \ldots, s_N)^\top$ and $\mathbf{1} = (1, \ldots, 1)^\top$. Thus the unnormalized attention vector $\widetilde{O}_k$ is precisely the coefficient-weighted combination of basis elements whose coefficients are the negative of the gradient components projected onto per-datapoint contributions.

Proof This is an immediate rearrangement:

$$\widetilde{O}_k = \sum_{i=1}^{N} s_i v_i = \sum_{i=1}^{N} \langle \Phi_k(x_i), r_i \rangle\, \Phi_k(x_i),$$

which is exactly the claimed expression. The scalar gradient g_k is $-\sum_i \langle \Phi_k(x_i), r_i \rangle$ by direct differentiation of $\mathcal{L}(c)$, as given in the setup. ■

Interpretation: without the softmax normalization the attention head directly forms the per-datapoint inner-product-weighted sum of basis-vectors; this algebraically contains the gradient components (indeed the scalar gradient is the sum of the per-datapoint logits used here).

Softmax attention: first-order approximation

Softmax introduces a normalization which prevents exact equality in general. However, in a principled asymptotic regime one obtains first-order alignment with the gradient direction after a simple centering step.

Theorem 13 (Softmax linearization & first-order gradient alignment) *Let s_i and v_i be as above and define $w_i(\beta) = \exp(\beta s_i)/\sum_j \exp(\beta s_j)$ and $O_k(\beta) = \sum_i w_i(\beta) v_i$. Denote the empirical means*

$$\bar{s} \;:=\; \frac{1}{N}\sum_{i=1}^{N} s_i, \qquad \bar{v} \;:=\; \frac{1}{N}\sum_{i=1}^{N} v_i.$$

Then for β in a neighborhood of 0 we have the Taylor expansion (component-wise in $\mathbb{R}^m$)

$$O_k(\beta) \;=\; \bar{v} \;+\; \frac{\beta}{N}\sum_{i=1}^{N}(s_i - \bar{s})\, v_i \;+\; \mathcal{O}(\beta^2).$$

Consequently, if the value vectors are mean-centered, i.e. $\bar{v} = 0$, then to first order in β

$$O_k(\beta) \;=\; \frac{\beta}{N}\sum_{i=1}^{N} s_i\, v_i \;-\; \frac{\beta \bar{s}}{N}\sum_{i=1}^{N} v_i \;+\; \mathcal{O}(\beta^2) = \frac{\beta}{N}\sum_{i=1}^{N} s_i v_i \;+\; \mathcal{O}(\beta^2).$$

If additionally $\sum_{i=1}^{N} v_i = 0$ (equivalently $\bar{v} = 0$), we obtain

$$O_k(\beta) \;=\; \frac{\beta}{N}\sum_{i=1}^{N} s_i v_i \;+\; \mathcal{O}(\beta^2),$$

so the attention output is proportional (to first order in β) to the unnormalized attention vector $\sum_i s_i v_i$ which, by Theorem 12, encodes the gradient components.

Proof Standard Taylor expansion of the softmax weights about $\beta = 0$ yields

$$\exp(\beta s_i) = 1 + \beta s_i + \tfrac{1}{2}\beta^2 s_i^2 + \mathcal{O}(\beta^3),$$

and hence

$$Z(\beta) := \sum_{j=1}^{N} \exp(\beta s_j) = N + \beta \sum_j s_j + \tfrac{1}{2}\beta^2 \sum_j s_j^2 + \mathcal{O}(\beta^3).$$

Using $w_i(\beta) = \exp(\beta s_i)/Z(\beta)$, expand to first order:

$$w_i(\beta) = \frac{1 + \beta s_i + \mathcal{O}(\beta^2)}{N + \beta \sum_j s_j + \mathcal{O}(\beta^2)} = \frac{1}{N}\Big(1 + \beta(s_i - \bar{s})\Big) + \mathcal{O}(\beta^2),$$

where $\bar{s} = \frac{1}{N}\sum_j s_j$. Multiply by v_i and sum:

$$O_k(\beta) = \sum_{i=1}^{N} w_i(\beta) v_i = \frac{1}{N}\sum_{i=1}^{N} v_i + \frac{\beta}{N}\sum_{i=1}^{N}(s_i - \bar{s}) v_i + \mathcal{O}(\beta^2).$$

This is the stated expansion. The corollary statements about centering follow by setting $\bar{v} = 0$ and simplifying the $\bar{s}$ term as shown. ∎

Appendix D. Multi-Headed Attention *is* Multiple Gradient Descent Steps

This section gives a precise, matrix-level account of how a single Transformer layer with H attention heads implements H (simultaneous or complementary) gradient-descent-like update directions for a linear-in-parameters field, and how softmax attention recovers the same behaviour to first order under mild centering / temperature assumptions.

Setup and notation

Let N be the number of context samples and K the number of basis coordinates for a field linear in coefficients. Assume scalar targets (vector-valued outputs are handled componentwise). Define:

$$\Phi \in \mathbb{R}^{N\times K}, \qquad \Phi_{i,k} := \varphi_k(x_i),$$

the design matrix of basis evaluations at the N sample points, and

$$r \in \mathbb{R}^N, \qquad r_i := y_i - f_c(x_i)$$

the residual vector w.r.t. current coefficients $c \in \mathbb{R}^K$. The squared-error loss is $\mathcal{L}(c) = \frac{1}{2}\|r\|_2^2$, and the (column) gradient vector with respect to c is

$$\nabla_c \mathcal{L} \;=\; -\Phi^\top r \in \mathbb{R}^K.$$

(Equivalently $g := -\nabla_c \mathcal{L} = \Phi^\top r$ denotes the vector of per-coordinate negative gradients.) These notations agree with the single-head derivation in Theorem 3.2.

We consider a Transformer layer with H heads. For head $h \in \{1, \dots, H\}$ define:

- a query vector (or query projection that yields) $q^{(h)} \in \mathbb{R}^K$ which acts as a linear selector on key vectors;
- key vectors for each datapoint i: $K_i \in \mathbb{R}^K$, here $K_i = \Phi_{i,:}^\top$ (the i-th row of Φ as a column);
- values for each datapoint: $V_i \in \mathbb{R}$ equal to the scalar residual r_i (or generally V_i could be vectors; we give the scalar case first).

We analyze two attention variants:

1. *Linear (unnormalized) attention (per-head):*

$$\widetilde{w}_i^{(h)} \;=\; q^{(h)\top} K_i, \qquad \widetilde{O}^{(h)} \;=\; \sum_{i=1}^N \widetilde{w}_i^{(h)}\, V_i.$$

2. *Softmax attention (per-head) with temperature $\tau > 0$:*

$$w_i^{(h)}(\tau) \;=\; \frac{\exp\left(\frac{1}{\tau} q^{(h)\top} K_i\right)}{\sum_{j=1}^N \exp\left(\frac{1}{\tau} q^{(h)\top} K_j\right)}, \qquad O^{(h)}(\tau) \;=\; \sum_{i=1}^N w_i^{(h)}(\tau)\, V_i.$$

Stack the H query vectors into a matrix $Q := [q^{(1)} \cdots q^{(H)}] \in \mathbb{R}^{K\times H}$, and collect the per-head outputs into $\widetilde{O} := [\widetilde{O}^{(1)}, \dots, \widetilde{O}^{(H)}]^\top \in \mathbb{R}^H$ for linear attention (and similarly $O(\tau) \in \mathbb{R}^H$ for softmax).

D.1. Exact identity (linear attention)

Theorem 14 (Exact multi-head gradient directions — linear attention) *Under the setup above with values $V_i = r_i$ and keys $K_i = \Phi_{i,:}^\top$, the H linear-attention head outputs satisfy the exact matrix identity*

$$\widetilde{O} \;=\; Q^\top \Phi^\top r \;=\; Q^\top g,$$

where $g := \Phi^\top r$ is the vector of per-coordinate negative gradients. In particular:

- *If $Q = I_K$ and $H = K$, then $\widetilde{O} = g$ and the K heads recover the full negative gradient vector (coordinatewise).*
- *If Q selects a subset of coordinates (rows of I_K), the heads recover the corresponding coordinate-wise negative gradients (block or coordinate GD).*
- *For general Q the heads compute* linear *combinations of the gradient vector; applying a linear readout $R : \mathbb{R}^H \to \mathbb{R}^K$ (e.g., $R := (Q^\top)^+$ left-inverse) yields a reconstructed preconditioned gradient $R\widetilde{O}$ which can be used as an update for c.*

Proof By definition of $\widetilde{O}^{(h)}$,

$$\widetilde{O}^{(h)} = \sum_{i=1}^{N} (q^{(h)\top} K_i)\, V_i \;=\; q^{(h)\top} \Big(\sum_{i=1}^{N} K_i V_i \Big).$$

Stacking the H heads yields

$$\widetilde{O} \;=\; Q^\top \Big(\sum_{i=1}^{N} K_i V_i \Big).$$

But with $K_i = \Phi_{i,:}^\top$ and $V_i = r_i$ we have $\sum_{i=1}^{N} K_i V_i = \Phi^\top r = g$, so $\widetilde{O} = Q^\top g$, as claimed. The listed corollaries are immediate linear-algebra consequences: choosing $Q = I_K$ returns g, selecting rows of the identity returns coordinate subsets, and general Q returns linear combinations that can be inverted (when Q has full column rank) to reconstruct directions in $\mathbb{R}^K$. ■

Remark 15 *The mapping $g \mapsto \widetilde{O}$ performed by multi-head linear attention is a low-rank linear map $Q^\top$. When $H \geq K$ and Q has full row rank, the full gradient is representable in head-space; if $H < K$ the heads realize a rank-H approximation to the gradient (a natural low-rank preconditioner). This shows algebraically why multiple heads implement multiple simultaneous gradient directions or a basis for gradient subspace exploration.*

D.2. Softmax attention: first-order approximation and error control

We now show that the same multi-head picture holds for standard softmax attention in the high-temperature / small-logit regime, under mild centering of values or learned baselines. The expansion follows the same Taylor analysis used for the single-head softmax approximation (Appendix A.4).

Theorem 16 (Multi-head softmax ≈ multi-head linear attention (first-order)) *Assume for each head h the per-sample logits $s_i^{(h)} := q^{(h)\top} K_i$ are uniformly bounded, and denote their mean $\bar{s}^{(h)} := \frac{1}{N}\sum_i s_i^{(h)}$. Let values satisfy the centering condition $\bar{V} := \frac{1}{N}\sum_{i=1}^N V_i = 0$ (implementable by a mean-centering layer or residual baseline). Then for temperature parameter $\tau > 0$ large enough the softmax head outputs admit the expansion*

$$O^{(h)}(\tau) \;=\; \frac{1}{N}\sum_{i=1}^N V_i \;+\; \frac{1}{N\tau}\sum_{i=1}^N (s_i^{(h)} - \bar{s}^{(h)})\, V_i \;+\; \mathcal{R}^{(h)}(\tau),$$

with the leading-order term $\frac{1}{N}\sum_i V_i$ vanishing under $\bar{V} = 0$. Hence

$$O^{(h)}(\tau) \;=\; \frac{1}{N\tau}\, q^{(h)\top}\Phi^\top r \;+\; \mathcal{R}^{(h)}(\tau),$$

and stacking heads gives

$$O(\tau) \;=\; \frac{1}{N\tau} Q^\top \Phi^\top r \;+\; \mathcal{R}(\tau).$$

Moreover, the remainder satisfies the uniform bound

$$\|\mathcal{R}(\tau)\|_2 \;\le\; \frac{C(N,B,R)}{\tau^2},$$

for a constant C depending only on N, and uniform bounds $|s_i^{(h)}| \le B$, $|V_i| \le R$. Thus by taking τ sufficiently large (or equivalently by scaling logits down) the multi-head softmax outputs approximate the scaled linear-attention gradient combination arbitrarily well; the scale factor $1/(N\tau)$ can be absorbed into an effective learning rate.

Proof For each head h perform a Taylor expansion of $\exp(s_i^{(h)}/\tau)$ about $1/\tau = 0$:

$$\exp\left(\tfrac{1}{\tau} s_i^{(h)}\right) = 1 + \tfrac{1}{\tau} s_i^{(h)} + \tfrac{1}{2\tau^2}(s_i^{(h)})^2 + O(\tau^{-3}).$$

Summing over i gives

$$Z^{(h)}(\tau) := \sum_{j=1}^N \exp\left(\tfrac{1}{\tau} s_j^{(h)}\right) = N + \tfrac{1}{\tau}\sum_j s_j^{(h)} + \tfrac{1}{2\tau^2}\sum_j (s_j^{(h)})^2 + O(\tau^{-3}).$$

Thus the softmax weight is

$$w_i^{(h)}(\tau) = \frac{1 + \frac{1}{\tau} s_i^{(h)} + \frac{1}{2\tau^2}(s_i^{(h)})^2 + O(\tau^{-3})}{N + \frac{1}{\tau}\sum_j s_j^{(h)} + \frac{1}{2\tau^2}\sum_j (s_j^{(h)})^2 + O(\tau^{-3})}.$$

Dividing numerator and denominator by N and expanding to second order in $1/\tau$ yields

$$w_i^{(h)}(\tau) = \frac{1}{N}\Big(1 + \tfrac{1}{\tau}(s_i^{(h)} - \bar{s}^{(h)})\Big) + O(\tau^{-2}),$$

uniformly in i, h, where $\bar{s}^{(h)} = \frac{1}{N}\sum_j s_j^{(h)}$. Multiplying by V_i and summing over i gives the stated expansion for $O^{(h)}(\tau)$. Under the centering $\bar{V} = 0$ the $\frac{1}{N}\sum_i V_i$ term vanishes, and using $s_i^{(h)} = q^{(h)\top} K_i$ together with $\sum_i K_i V_i = \Phi^\top r$ we obtain

$$O^{(h)}(\tau) = \frac{1}{N\tau} q^{(h)\top} \Phi^\top r + O(\tau^{-2}).$$

Stacking heads yields the matrix formula $O(\tau) = \frac{1}{N\tau} Q^\top \Phi^\top r + \mathcal{R}(\tau)$ with $\|\mathcal{R}(\tau)\|_2 = O(\tau^{-2})$. A constructive derivation of an explicit constant in the $O(\tau^{-2})$ remainder follows the exact bounds in Appendix A.4; see in particular the explicit remainder bound and constant derivation. ■

Suppose each head h is followed by a linear readout $R^{(h)} : \mathbb{R} \to \mathbb{R}^K$ (or all heads are aggregated by a linear map $R : \mathbb{R}^H \to \mathbb{R}^K$). Let the effective coefficient update computed by the layer be

$$\Delta c \;=\; \eta \cdot R\,\widetilde{O} \qquad \text{(linear-attention)},$$

or for softmax

$$\Delta c \;=\; \eta \cdot R\,O(\tau) \approx \frac{\eta}{N\tau} R Q^\top \Phi^\top r,$$

with approximation error $\mathcal{O}(\tau^{-2})$. Choosing $R = (Q^\top)^+$ and $Q = I_K$ recovers the standard gradient-descent step $\Delta c = \eta g$ (up to the scaling factor), and more generally, $RQ^\top$ acts as a preconditioner on the gradient. Thus the multi-head layer computes (exactly for linear attention, approximately for softmax) a sum of H gradient-like steps or, when combined, a single gradient step.

Appendix E. Softmax Temperature Scaling

We analyze empirically Softmax Temperature Scaling and its effects.

(a) Attention–Gradient Convergence. We measure the squared norm difference $\|V(\text{Softmax}) - V(\text{Linear})\|^2$ between the output of softmax attention at temperature T and the exact linear-attention gradient operator. For small T, softmax deviates significantly; in the high-temperature regime ($T \gtrsim 30$), experimental decay matches the predicted asymptotics, transitioning from $O(T^{-1})$ to $O(T^{-2})$ scaling. At $T \approx 100$, errors reach $\sim 10^{-7}$, approaching numerical precision.

(b) Temperature Scaling in 2D and 3D Fields. We compute the relative error between predicted and measured scaling factors for both 2D and 3D sinusoidal fields. Both cases exhibit convergence consistent with theory, with the 3D field decaying slightly faster at large T due to additional averaging across dimensions. A flat plateau at low T reflects the expected non-asymptotic regime.

(c,d) SO(2) and SO(3) Rotation Equivariance. We evaluate the mean-squared error (MSE) between rotated and transformed outputs for both $SO(2)$ and $SO(3)$ actions, under linear attention and softmax attention with $T \in \{1, 10, 100\}$. In all cases, empirical errors ($\sim 10^{-6}$) remain well below the conservative theoretical bound confirming that equivariance is an architectural property rather than a temperature-dependent effect.

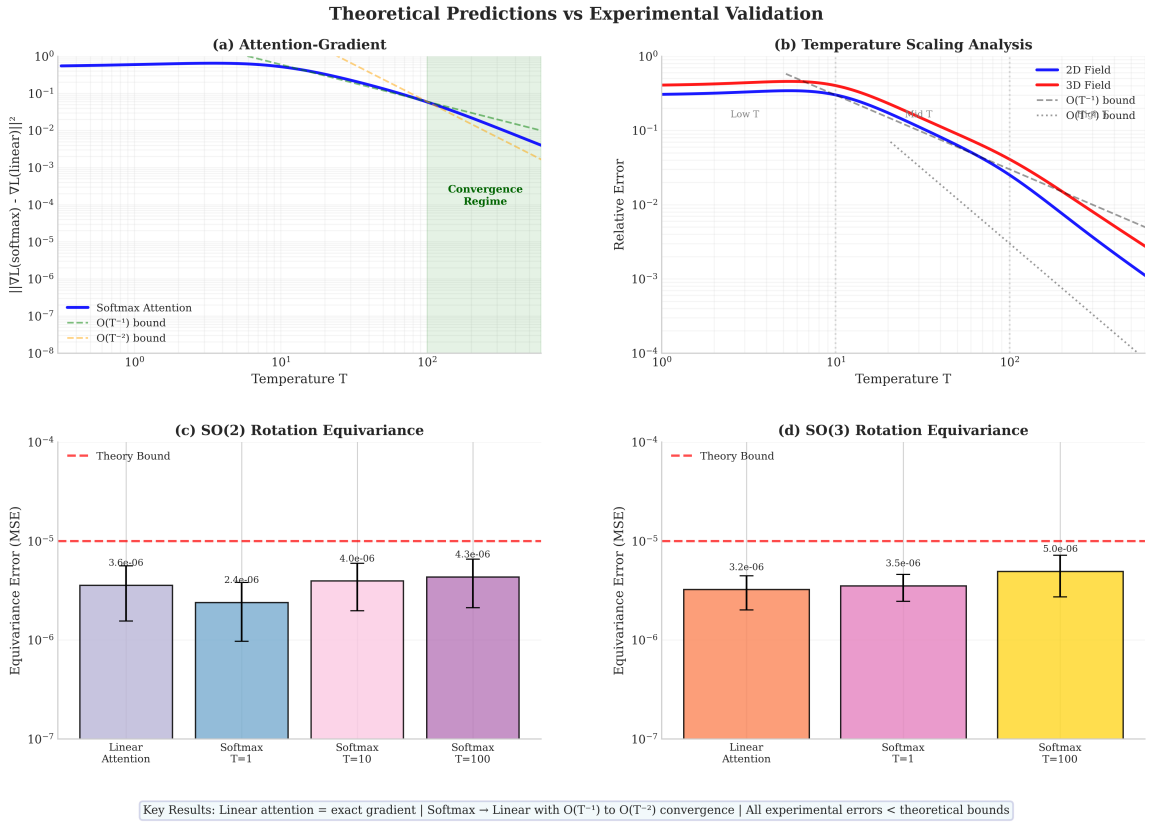

Figure 6: **Theoretical predictions vs. experimental validation.** (a) Softmax attention converges to linear-attention gradients at the predicted $O(T^{-1}) \to O(T^{-2})$ rate. (b) Temperature scaling analysis for 2D and 3D sinusoidal fields, matching theory in the high-T regime. (c,d) Empirical $SO(2)$ and $SO(3)$ rotation equivariance errors are stable across attention types and far below the theoretical bound.

Appendix F. Extended Validation on Computer Vision Task

To further validate Theorem 3.1 on structured data, we extended our equivariance test to MNIST digits represented as continuous fields. Each image is treated as a set of $(x, y, \text{intensity})$ samples, and we applied controlled affine transformations (scaling and translation). The operator was trained for self-consistency on MNIST fields and evaluated under equivariant transformation. As shown in Fig. 7, the reconstructions $f_{S'}(x)$ and $f_S\left(\frac{x-b}{a}\right)$ are nearly indistinguishable, with differences confined to localized regions. Across five seeded trials, the mean equivariance error was 0.013 ± 0.004 (95% CI), confirming that the operator retains affine equivariance beyond synthetic SIREN fields and into real image data.

Appendix G. Extended Validation on Physics Task

To further validate Theorem 3.1 in a physics-informed setting, we applied our affine equivariance test to solutions of the two-dimensional Poisson equation $-\Delta u = f$ on the unit

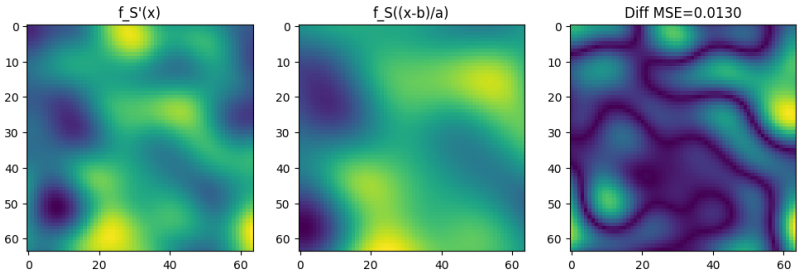

Figure 7: Affine equivariance validation on MNIST-as-a-field. Left: $f_{S'}(x)$ from transformed samples. Middle: $f_S\left(\frac{x-b}{a}\right)$ from original samples with inverse transform. Right: absolute difference, with mean squared error (MSE) reported.

square with zero Dirichlet boundary conditions. Right-hand sides f were generated as sums of Gaussian bumps, and the PDE was solved on a 28×28 grid using a finite-difference discretization and conjugate gradient solver. Each solution u was represented as a set of $(x, y, u(x, y))$ samples, which served as the input to the operator. The operator was meta-trained to regress SIREN parameters from these sets, following the same normalization and architectural assumptions used in earlier sections. We then applied controlled affine transformations (scaling a and translation b) to the input coordinates, evaluated the operator on both the transformed and original sets, and compared the resulting fields.

As shown in Fig. 8, the reconstructions $f_{S'}(x)$ (from transformed samples) and $f_S\left(\frac{x-b}{a}\right)$ (inverse-transformed from original samples) are visually nearly indistinguishable, with residuals showing smooth, low-magnitude structure. Across ten seeded trials, the mean equivariance error was 0.0276 ± 0.0152 (95% CI), consistent with the approximate affine equivariance predicted by our theory. These results confirm that the proposed Transformer operator generalizes beyond vision datasets and retains its symmetry-respecting behavior on physically meaningful continuous fields.

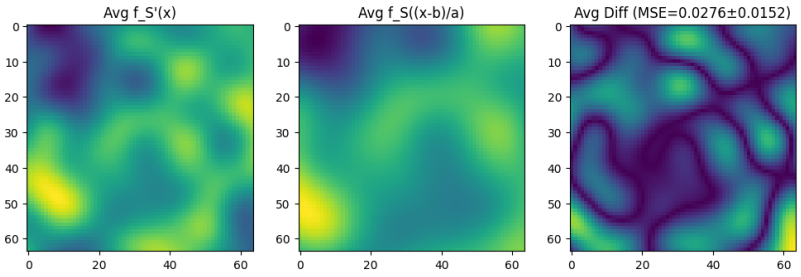

Figure 8: Affine equivariance validation on Poisson PDE solutions. Left: $f_{S'}(x)$ from transformed samples. Middle: $f_S\left(\frac{x-b}{a}\right)$ from original samples with inverse transform. Right: absolute difference, with mean squared error (MSE) reported.

Appendix H. Hyperparameters

Table 3: Architectures & initialization

Module	Hyperparameter	Value
SIREN	input_dim	2 (2D experiments) / 3 (3D experiments)
	hidden_dim	64 (ablation: 32)
	output_dim	1
	num_layers	3 (ablation: 2)
	ω_0 (first-layer scale)	30.0
	first-layer init	Uniform$[-\frac{1}{2}, \frac{1}{2}]$
	subsequent init	Uniform$\left[-\frac{\sqrt{6/\text{fan_in}}}{30}, \frac{\sqrt{6/\text{fan_in}}}{30}\right]$
TransformerOperator	embed_dim	128
	num_heads	8 (ablation: 4)
	num_layers	4 (ablation/meta: 3)
	feedforward dim	512
	dropout	0.1
	activation	GELU
	output head	3-layer MLP, hidden dim = 256

Table 4: Data and coordinate normalization

Item	Value
MNIST as field	28×28 grid; pixel centers treated as continuous samples
Coordinate mapping	Pixel center $i \in \{0, \ldots, 27\} \mapsto x = -1 + \frac{2(i+0.5)}{27}$ (and similarly for y)
PDE grid	$n = 28$ (default); optional high-res grid 100×100 on $[-1.5, 1.5]^2$
SIREN input scaling	Multiply normalized coordinates by $\omega_0 = 30$ before first activation
Noise (where used)	Additive Gaussian, $\sigma = 0.01$
Batch size	1 (per-field); 32 (meta/ablation)

Table 5: Training & validation

Item	Value
Optimizer	AdamW (default)
Learning rate	3×10^{-4} (default); 1×10^{-3} for fast ablation runs
Weight decay	5×10^{-5}
Iterations / Epochs	200 iterations per field (default); ablation: 500 (fast:100)
Loss	MSE
Affine validation	scale $a = 1.5$, translation $b = (0.3, -0.3)$
Trials	PDE: 10; MNIST: 5 (95% CI via t-distribution)
Test points per config	50

Table 6: Purpose-level hyperparameters

Panel / Key settings	**Values**
softmax → linear	$N = 64$, $d = 32$, $K = 16$, $H = 8$, temps = `logspace`$(0, 2.5, 30)$
Attention-Gradient Comparison	$N = 100$, $H = 4$, temps = `linspace`$(10, 100, 20)$
scaling analysis	$N \in \{32, 64, 128, 256\}$, temps $= \{10, 25, 50, 100, 200\}$
2D rotation	SIREN: input=2, hidden=64, layers=3, $\omega_0 = 30$; grid 100×100 on $[-1.5, 1.5]^2$
3D rotation	SIREN: input=3, hidden=64, layers=3, $\omega_0 = 30$; $n_\theta = 30$, $n_\phi = 30$
theoretical comparison	$N = 100$, $K = 50$, temps = `logspace`$(-1, 3, 50)$, $\sigma = 0.01$