Towards Isotropic Deep Learning
A New Default Inductive Bias

George Bird
Department of Computer Science
University of Manchester
Manchester, UK
george.bird@postgrad.manchester.ac.uk

Abstract

This position paper explores the often overlooked geometric implications of current functional forms used in deep learning and proposes a new paradigm to be termed *Isotropic Deep Learning*. It has been demonstrated that the functional forms of current deep learning influence the activation distributions. Through training, broken symmetries in functional forms can induce broken symmetries in embedded representations. Thus producing a geometric artefact in representations which is not task-necessitated, and solely due to human-imposed choices of functional forms. There appears to be no strong a priori justification for why such a representation or functional form is desirable, while this paper proposes several detrimental effects of the current formulation. As a result, a modified framework for functional forms will be explored with the goal of unconstraining representations by elevating a rotational symmetry-inductive bias throughout the network. This framework is encouraged to be adopted as a new default. This direction is proposed to improve network performance. Preliminary functions are proposed, including activation functions. Since this overhauls almost all functional forms characterising modern deep learning, it is suggested that this shift may constitute a new branch of deep learning.

1 Introduction

Current deep learning models typically employ an elementwise functional form. This is particularly evident in activation functions, sometimes called ridged activation functions. These functions are often displayed univariately as shown in *Eqn.* 1, with σ being a specific activation function implemented, e.g. ReLU, Tanh, etc.

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto f\left(x\right) = \sigma\left(x\right) \tag{1}$$

However, this display choice obfuscates a crucial (standard) basis dependence. This is explicitly displayed in, what should be considered a more implementation-correct multivariate form, of *Eqn.* 2. This reveals the functional form's usually hidden $\hat{e}_i$ basis dependence. The multivariate form is depicted for an n neuron layer, with activation vector $\vec{x} \in \mathbb{R}^n$. This standard basis dependence is arbitrary and appears as a historical precedent, rather than appropriate inductive bias, discussed further in *App* A.

$$\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n, \quad \vec{x} \mapsto \mathbf{f}\left(\vec{x}\right) = \sum_{i=1}^{n} \sigma\left(\vec{x} \cdot \hat{e}_i\right) \hat{e}_i \tag{2}$$

Due to this basis dependence, non-linear transformations differ angularly in effect. Therefore, this will be termed an *anisotropic function*, indicating this rotational asymmetry. Due to the pervasive use of these functional forms, current deep learning as a paradigm may consequently be termed *anisotropic deep learning*. Despite its implications, this *choice* of basis-dependent anisotropy appears

Preprint. Under review.

unappreciated and incidental in the development of most contemporary models. Anisotropic forms are almost treated as axiomatic to deep learning. Hence, re-evaluating and systematically reformulating this foundational aspect of modern deep learning is felt to constitute a new branch: *Isotropic Deep Learning*.

This asymmetry in the non-linear transform is particularly about the standard (Kronecker) basis vectors and their negative $\{+\hat{e}_i, -\hat{e}_i\}_{\forall i}$, due to the elementwise nature. Therefore, it can be said to distinguish the standard basis - a *distinguished basis*[1]. For example, the standard basis is visible in *Fig.* 1 showing the mapping of elementwise-$\tanh$ on a variety of test shapes. Most generally, this *choice* of functional forms can be considered to break *continuous* rotational symmetry, and reduce it to a *discrete* rotational symmetry — a permutation symmetry of the standard basis. In effect, if the function is treated in its multivariate form, it is equivariant to a permutation of the components of its vector decomposed in the standard basis. For an element of the permutation group $\mathbf{P} \in \mathcal{S}_n$, the following equivariance relation holds $f\left(\mathbf{P}\vec{x}\right) = \mathbf{P}f\left(\vec{x}\right)$.

Figure 1: [Insert caption here]

Non-linearities are usually pivotal to the network's ability to achieve a desired computation, as seen through the universal approximation theorem's explicit dependence on the activation function . The non-linearities produce differing local transformations, such as stretching, compressing, and generally reshaping a manifold. Consequently, the network may be expected to adapt by moving representations to geometries about these distinguished directions, to use specific local transforms to achieve the desired computation. Hence, an anisotropy about a distinguished basis is induced into the activation distribution.

This has been empirically demonstrated: training results in the broken symmetry of the functional forms, inducing a broken symmetry in the activations. Since these non-linear zones are centred around the distinguished bases, the embedded representations are expected to move to beneficial angular arrangements about the arbitrarily imposed privileged basis's geometry. For example, they appear to move towards the non-linearities' extremums, aligned, anti-aligned or other geometries, through training. This may correspond to a local, dense or sparse coding and superposition, respectively.

Therefore, the network has adapted its representations through training for various reasons due to these functional form choices. This is the causal hypothesis proposed. This hypothesis aids in explaining the observed tendency of privileged-basis alignment. This is a human-caused representational collapse onto the privileged basis, and it is proposed thaty this is frequently not a task-necessitated collapse. There appears to be little justification as to why this is universally desirable. Without a priori justification, this inductive bias may be detrimental to computation, so unconstraining the activation appears preferable. Several key negative implications are discussed in *Sec.* 2.

Overall, historic and frequently overlooked functional forms for modern deep learning directly influence the models' activations and therefore behaviour. There exists a functional form dependence

[1] This is generalised from a *'privileged basis'*. The change to *'distinguished basis'* reflects that the basis may be more-or-less aligned to the representation; whereas *'privileged basis'* will suggest a basis more aligned to the embedded activations. The term 'basis' will be retained even though the set of *distinguished vectors* may also be under-/over complete for spanning the whole space. There may be multiple distinguished bases, such as aligned and anti-aligned with the standard basis for the model.

on a basis which appears entirely arbitrary and overlooked. A causal link between this arbitrary basis and activations has been empirically demonstrated, and hence, a resultant effect on the final performance of the model is hypothesised. These are often underappreciated choices which have consequences and should be well-justified and studied. This is the position of the authors.

Throughout the rest of this position paper, **it is argued that a departure from this anisotropic functional form paradigm towards the isotropic paradigm is preferable**, unless otherwise justified. It encourages the reader to be conscious of these choices when designing a model, as well as the usual architectural tool kit. Particularly, isotropic choices, equivalent to basis independence, may be thought to unconstrain the representations into more optimal arrangements for a task. The tenets of this paradigm are suggested for all architectures on general tasks.

2 Hypothesised Problems of Anisotropy

This section argues that current functional forms impose unintended anisotropic performance detriments, indicating that *Isotropic Deep Learning*, once substantially developed, is proposed as the default inductive bias unless an alternative is task-necessitated.

This section lays out a non-exhaustive set of arguments discussing some implications that anisotropic functional forms may cause. These mainly centre on the role of the activation functions, since this is the area the author has primarily explored in their PhD thus far. To the author's knowledge, some of these failure modes are newly characterised phenomena, such as the so-called '*neural refractive problem*'.

2.1 The Neural Refractive Problem

The '*neural refractive problem*' describes how linear trajectories of activations may converge or diverge from their initial consistent path after an activation function is applied. This is analogous to a light ray refracting through an optically varying medium or boundary.

It appears to be a phenomenon in all anisotropic activation functions to date. The 'refraction' typically occurs more significantly at larger magnitudes — potentially a failure mode under network extrapolation. Mathematically, this has several representations, a magnitude-varying 'dynamic refraction' shown in *Eqn.* 3 or differentially in *Eqn.* 4. Also defined is a 'static refraction' definition shown in *Eqn.* 5. These are described for a multivariate activation function $\mathbf{f}$ and vector $\vec{x} = \alpha\hat{x}$ where $\hat{x}$ is a unit-vector. This relation may be satisfied for a single direction, a subset of the space or all directions $\hat{x} \in \mathcal{X} \subseteq S^n$. The relations generally show how the activation function alters the direction of its input vector in a anisotropic mannor.

$$\exists\hat{x} \in \mathcal{S}^{n-1}, \exists\alpha_1 \neq \alpha_2 > 0 : \frac{\mathbf{f}\left(\alpha_1\hat{x}\right)}{\left\|\mathbf{f}\left(\alpha_1\hat{x}\right)\right\|} \neq \frac{\mathbf{f}\left(\alpha_2\hat{x}\right)}{\left\|\mathbf{f}\left(\alpha_2\hat{x}\right)\right\|} \tag{3}$$

$$\exists\hat{x} \in \mathcal{S}^{n-1}, \exists\alpha_0 : \frac{\partial}{\partial\alpha}\left.\frac{\mathbf{f}\left(\alpha_1\hat{x}\right)}{\left\|\mathbf{f}\left(\alpha_1\hat{x}\right)\right\|}\right|_{\alpha_0} \neq \vec{0} \tag{4}$$

$$\exists\hat{x} \in \mathcal{S}^{n-1}, \exists\alpha : \frac{\mathbf{f}\left(\alpha\hat{x}\right)}{\left\|\mathbf{f}\left(\alpha\hat{x}\right)\right\|} \neq \hat{x} \tag{5}$$

It can be seen that along a straight-line trajectory in direction $\hat{x}$, the result of the activation function is a curved line if dynamically refracted. Therefore, if the linear feature hypothesis is followed, *every* linear feature, in these directions, becomes curved following the activation function. The network may exploit some of this curvature to construct new linear features in the subsequent layers; however, there may be many instances where this curvature is detrimental to an established semantic. The network may loose semantic separability, produce magnitude-based semantic inconsistency or compensatory maladaptions in later layers, due to this refraction. This may hinder network performance and is resolved by isotropic choices.

Particularly detrimental, in both refraction cases, can be the loss of semantic separability. If two distinct trajectories, representing different semantics, are transformed into curves which intersect or converge, then the separability of these concepts is lost. For example, suppose one direction is a linear feature for the presence of a dog in an image, whilst the other is for a horse. In that case, if these activations are of a magnitude where the activation function causes convergence, the identity of the activation's meaning can be misrepresented.

This may be particularly consequential for functions such as Sigmoid and Tanh, since large magnitude inputs end up at particular limit points (discussed as trivial representational alignments in Bird [2025]). For example, Tanh produces the limit points shown in *Eqn.* 6 when $\hat{x} \cdot \hat{e}_i \neq 0$ for all i. If there exists an i, s.t. $\hat{x} \cdot \hat{e}_i = 0$, then the corresponding index also has a 0 as a limit point. A fully-connected layer can only effectively separate two such converging directions at a time, which are then further curved by a subsequent activation function.

$$\lim_{\alpha \to \infty} \mathbf{f}\left(\alpha \hat{x}\right) = \sum_{i=1}^{N} \tanh\left(\alpha \hat{x} \cdot \hat{e}_i\right) \hat{e}_i \approx \sum_{i=1}^{N} \pm \hat{e}_i = \left(\pm 1, \cdots, \pm 1\right)^T \tag{6}$$

Consequently, semantic separability is lost except for 3^n discrete limit points for Tanh and Sigmoid. Therefore, embedded activations may be expected to align with these limit points, an empirically observed tendency Bird [2025]. Similarly, ReLU has one distinct limit point, $\vec{0}$, but otherwise an orthant unaffected by neural refraction. The authors speculate whether this is an additional reason for the success of ReLU, due to only a subset of directions experiencing the neural refraction phenomena. Furthermore, this would suggest an advantage of Leaky-ReLU: despite featuring static-refraction, directions do not become overlapped, so semantic separability is retained. Otherwise, the network may expend training time on producing robuster semantic separability, a needless compensatory adaption, which may lower representational capacity or extend training as a result.

More generally, dynamic deflection of trajectories may cause semantic ambiguity for the network, where only samples interpolable from training samples are reliably semantically identifiable. Particularly, *the larger the deflection, the greater the semantic ambiguity expected.* More considerable deflections typically occur at larger magnitudes in many current functions. Therefore, a magnitude-dependent semantic inconsistency may arise due to such deflections. A deflection function can be a trivial diagnostic measure defined by *Eqn.* 7 for a particular activation function.

$$\theta\left(\alpha; \hat{x}, \mathbf{f}\right) = \arccos\left(\frac{\mathbf{f}\left(\alpha \hat{x}\right) \cdot \hat{x}}{\left\|\mathbf{f}\left(\alpha \hat{x}\right)\right\|}\right) \tag{7}$$

This may explain why the network may perform excessively poorly on out-of-training-distribution samples. For example, suppose a linear feature roughly represents the quantity of cows in a field. In that case, the network may fail to extrapolate its function when an anomalous amount of cows are present, as this would be a very large magnitude of the linear feature and therefore the deflection is unprecedented and uninterpretable. The activation function would result in a loss of semantic consistency. Consequently, a network seeking to preserve linear features may constrain activation magnitudes, through training, to regions where the non-linear response is approximately predictable and stable to avoid the damaging consequences of neural refractions.

Angular anisotropies fundamentally cause the refraction phenomenon. If compression and rarefaction of certain angular regions occur, linear features will be deflected in various ways. A fix for this is introducing isotropy — the initial motivation for developing the paradigm. This does not prevent compression and rarefaction of activation distributions in general, as a bias can be added to reintroduce these useful phenomena predictably. It is argued that these are only an issue when they affect linear, not affine, features in a potentially unpredictable and thus semantically uninterpretable way.

The phenomenon is eliminated from networks by rearranging *Eqn.* 5 shown in *Eqn.* 8, then applying the simplification $\left\|\mathbf{f}\left(\alpha \hat{x}\right)\right\| = \sigma\left(\alpha\right)$ in *Eqn.* 9.

$$\mathbf{f}\left(\alpha \hat{x}\right) = \left\|\mathbf{f}\left(\alpha \hat{x}\right)\right\| \hat{x}' \tag{8}$$

$$\mathbf{f}\left(\alpha \hat{x}\right) = \sigma\left(\alpha\right) \hat{x}' \tag{9}$$

Finally choosing $\hat{x}' = \mathbf{R}\hat{x}$ for isotropy and $\mathbf{R}\hat{x} = \mathrm{I}_n \hat{x} = \hat{x}$ for simplicity, shown in *Eqn.* 10.

$$\mathbf{f}\left(\alpha \hat{x}\right) = \sigma\left(\alpha\right) \hat{x} \tag{10}$$

In standard notation, *Eqn.* 10 can be rewritten into the *final functional form for isotropic activation functions* shown in *Eqn.* 11.

$$\mathbf{f}\left(\vec{x}\right) = \sigma\left(\left\|\vec{x}\right\|\right) \hat{x} \tag{11}$$

This can be generalised as a result from rotational equivariance of the function. This can be expressed as a condition in *Eqn.* 12, which uses a commutator bracket for convenience, with $\forall \mathbf{R} \in \mathrm{SO}\left(n\right)$.

This bracket can be used to similarly define the current permutation-anistropic paradigm, by using the transform $\forall \mathbf{P} \in \mathcal{S}_n$ instead of the rotation. This may be recognised as superficially similar to equivariant neural networks, due to an analogous equivariance relation; however, the differences in both implementation and motivations are discussed in *App* **??**.

$$[\mathbf{R}, \mathbf{f}] = (\mathbf{Rf} - \mathbf{Rf}) = \vec{0} \tag{12}$$

The relation may be more familiar as $\mathbf{f}\left(\mathbf{R}\vec{x}\right) = \mathbf{R}\mathbf{f}\left(\vec{x}\right)$. This relation only applies to single-argument functions and requires generalising to more circumstances. A preliminary condition may be $\mathbf{f}\left(\mathbf{R}\vec{a}_1, \cdots, \mathbf{R}\vec{a}_N\right) = \mathbf{R}\mathbf{f}\left(\vec{a}_1, \cdots, \vec{a}_N\right)$ for $\mathbf{f}: \bigotimes_N \mathbb{R}^n \to \mathbb{R}^n$.

This introduces the general isotropic functional form for activation functions given in *Eqn* 13. This should be a piecewise function, defined using the identity at $\vec{x} = \vec{0}$, but this is suppressed for simplicity. This functional form is $\mathcal{O}\left(n\right)$ time for $\mathbb{R}^n$. Future work is establishing a universal approximation theorem for this functional form, as this is ongoing research for the author's PhD.

$$\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n, \quad \vec{x} \mapsto \mathbf{f}\left(\vec{x}\right) = \sigma\left(\|\vec{x}\|\right)\hat{x} \tag{13}$$

This is not to be confused with the radial-basis functional form displayed in *Eqn.* 14.

$$\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n, \quad \vec{x} \mapsto \mathbf{f}\left(\vec{x}\right) = \sum_{i=1}^{N} \sigma\left(\|\vec{x} - \vec{c}_i\|\right)\hat{e}_i \tag{14}$$

Overall, the 'neural refractive problem' outlines how semantic meanings may become intertwined or ambiguous due to current functional forms consistently skewing linear features in undesirable ways. A hypothesis is developed that this may be especially detrimental for out-of-distribution activations, which are likely to be most deflected and hence most semantically corrupted. Thus, the network's generalisation then fails excessively. Finally, it may be expected that the network produces compensatory adaptions for the phenomena. Since neural refraction is an non-linear and anisotropic phenomena, it cannot be inverted by a single subsequent layer, potentially wasting training time on corrections to the activation distributions due to unintended refraction. These refraction corrections may be limited in scope and produce maladaptive behaviour outside of the original training data.

2.2 Weight Locking, Optimisation Barriers and Disconnected Basins

'*Weight locking'* is a term to describe how particulary the weight parameter[2] may suffer from being stuck in local-minima found further into loss valleys encountered after a sufficient amount of training. This arises only due to the anisotropic functional form's *discrete* permutation symmetry.

Qualitatively, this is because the semantically meaningful linear features tend to become aligned with geometric positions about the distinguished bases. Any small perturbation to a parameter may misalign activations to the network's existing 'understood' semantics. In effect, further small perturbations to the parameters may move activations from a semantically-aligned to a semantically-dislocated state, thus the activation's meaning becomes ambiguous, forfeiting performance and resulting in a 'false' local-minima created only by the *discrete* nature of the permutation symmetry. It may also suggest that the optimisation barrier may be some function of the angular separation of the semantic directions. Consequently, creating a plethora of architectural local minima in the space. Only sufficiently large perturbations to a parameter may move activations between two differing semantically aligned directions.

This semantic dislocation is an emergent consequence of breaking the continuous symmetric forms, as without continuous rotational symmetry, it results in the dual phenomena of connectivity of many minima basins being lost. In effect, enforcing the isotropy constraints results in sets of continuously connected local minima which can be smoothly transformed into one another, by corresponding parameter rotations shown in *Eqn.* 15, a consequence of *Eqn.* 11. If this is downgraded to discrete rotational symmetry (i.e. permutation symmetry), then artificial optimisation barriers may reemerge in these basins. In this case, only a sufficiently large perturbation to the parameters may dislodge the network into a more optimal minima, while more minor perturbations are insufficient. Effectively, the discrete permutation symmetry may result in a discretised lattice solution for the parameters, much like how it breaks the symmetry of activations through training too Bird [2025].

$$\forall \mathbf{R} \in \mathrm{SO}\left(n\right): \underbrace{\mathbf{W}^l \mathbf{R}^\top}_{\mathbf{W}'^l} \mathbf{f}\Big(\underbrace{\mathbf{R}\mathbf{W}^{l-1}}_{\mathbf{W}'^{l-1}}\vec{x} + \underbrace{\mathbf{R}\vec{b}}_{\vec{b}'}\Big) = \mathbf{W}^l \mathbf{f}\left(\mathbf{W}^{l-1}\vec{x} + \vec{b}\right) \tag{15}$$

[2] Though similarly applies to a 'locking' of the bias to $\vec{0}$.

This exists as a qualitative intuition since until robust methods to determine semantically meaningful directions are produced, this hypothesis remains difficult to verify. Nevertheless, steps can be taken immediately to counteract the problem, and this is to introduce isotropy to connect these minima.

2.3 Emergence of Linear Features and Semantic Interpolatability

As previously mentioned, symmetry-broken functional forms induce symmetry-broken representations. Thus, *approximately* discrete embedding directions are tended towards. It may be conjectured that semantically meaningful linear directions may also be encouraged to discretise, aligning with these anisotropic embeddings. This generally appears to be the case, with notable counterexamples. Moreover, these counterexamples are for networks which *do not feature anisotropic functional forms*.

However, many real-life semantics are continuums: colours, positions of objects, broad morphology, even within a single species. Representational collapse onto a single discrete semantic may lose this important nuance. It appears a poor inductive bias to have functional forms encourage discretised representations. Isotropic functions do not prevent discrete semantics, but they don't encourage them either, — enable continuous ones since they generally unconstrain the representations. Therefore, moving towards isotropy is hypothesised to encourage embeddings to be more smoothly distributed, taking on intermediate values between typically discrete linear features and substantially enlarging the expressivity and representation capacity of networks — only limited by concept interferences.

In this case, the discrete concept of representation capacity may become irrelevant; each layer may express different continuous arrangements, where differing concepts are angularly suppressed and expressed in analogy to the linear features hypothesis. Instead, the '*magnitude-direction hypothesis*' is proposed as a continuous extension, magnitudes indicating the amount of stimulus present, direction indicating the particular concept. Activations then populate this more continuous manifold.

This may also produce a better organised semantic map at each layer of the network, since intermediate representations may now relate otherwise discrete features and therefore this connectivity can bring them into proximity continuously (which 'weight locking' may typically prevent). This may additionally aid researchers in comparing representational alignment between models and biology[3].

Therefore, in general settings, the inductive bias of isotropy appears more appropriate default unless justification for anisotropy is present. It is an inductive bias that meaningful semantics are often continuous and interpolatable, whilst still retaining discrete semantics when task neccessitated as opposed to human imposition. Hence, it generalises the discrete linear features paradigm into a more continuous setting. It suggests that a continuous rotational symmetry allows a continuous embedding, since direction-based symmetry breaking is not induced by functional forms. It is hoped this will enable networks to acquire a more optimal and natural representation embedding for the given task.

3 Alternative View: Embedding Folding

4 Isotropic Implementations

This position paper argues for the implementation of isotropic functional forms into neural networks as a default inductive bias. Near-term adoption may be rate-limited due to the development of suitable functions, particularly since *anisotropic deep learning* has a substantial head start and analogues to existing functions are not so trivial to produce. Despite this, several preliminary implementations are outlined in this section as a starting point; however, these functions are far from optimal and substantial research and development is required to bring isotropic deep learning into practicality.

Nevertheless, below shares a non-exhaustive list of activation functions and some consequences of them. It is hoped that a directed search for new functions may occur.

[3] Though there is no guarantee that *all* basins are connected, so therefore would not necessarily be alignable through continuous rotation transforms. Furthermore, the success of ensemble methods with diverse constituent models would suggest that diverse disconnected basins remain, complicating representational alignment even under isotropy conditions, though isotropy may help somewhat.

4.1 Activation Functions

As stated, an isotropic functional form for activation functions is given in *Eqn.* 13 and compared with current functional forms in *Tab.* 4.1 — where its basis-independence and relative simplicity is clear. Further criteria, in addition to isotropy, are also a performance-neccessity, but will be outlined in future work.

Beginning from this functional form, four familiar analogoues to elementwise functions are developed: Isotropic-Tanh, Isotropic-ReLU, Isotropic-Leaky-ReLU and Isotropic-Exponential-ReLU. However, it is hoped that development of the paradigm will produce further activation functions which are not just mere analogues of existing activation functions, but exploit the novel properties of isotropy for optimal performance.

Radial Basis Form	**Elementwise Form**	**Isotropic Form**
$\mathbf{f}(\vec{x}) = \sum_{i=1}^{N} \sigma\left(\|\vec{x} - \vec{c}_i\|\right) \hat{e}_i$	$\mathbf{f}(\vec{x}) = \sum_{i=1}^{n} \sigma\left(\vec{x} \cdot \hat{e}_i\right) \hat{e}_i$	$\mathbf{f}(\vec{x}) = \sigma\left(\|\vec{x}\|\right) \hat{x}$

Isotropic-Tanh is described in *Eqn.* 16. In basis directions, $\hat{e}_i$, it is equal in function to standard elementwise-$\tanh$, as indicated by its name. It is bounded, up to a norm of one, but does not angularly saturate like standard $\tanh$, allowing activation to continue semantically shifting. It is approximately, linear about the origin.

$$\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n, \quad \vec{x} \mapsto \mathbf{f}\left(\vec{x}\right) = \tanh\left(\|\vec{x}\|\right) \hat{x} \tag{16}$$

This function is reasonably cheap, computation of $r = \|\vec{x}\|$, $\tanh\left(r\right)$ and $\operatorname{sech}^2\left(r\right)$, need only be computed once (including for backward-pass) rather than per-component. The vector-norms are naturally constrained to $[0, 1)$ acting as an implicit normaliser. Around the origin, the transform is approximately the identity: $\lim_{r \to 0} \mathbf{J}\left(r\hat{x}\right) = \mathrm{I}_n$, justifying $\mathbf{f}\left(\vec{0}\right) = \vec{0}$, to preserve a smooth gradient. It is also globally 1-lipschitz.

Isotropic-ReLU is shown in *Eqn.* 17, an analogue to its traditional implementation.

$$\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n, \quad \vec{x} \mapsto \mathbf{f}\left(\vec{x}\right) = \max\left(\|\vec{x}\| - R, 0\right) \hat{x} \tag{17}$$

Isotropic-Leaky-ReLU

Isotropic-Soft-ReLU Soft-ReLU can be derived from integrating a $\tanh$-like radial function, to $\ln\cosh x$ which may be approximated to $\ln\left(1 + \exp x\right)$, in anaology to Softplus.

Isotropic-Sinusoids , is a possibility which does not have an anologue within the current anisotropic paradigm — demonstrating a broad array of possibilities. With the aforementioned isotropic-ReLU-like functions, the networks may normalise magnitudes such that the distributions 'escape' the non-linear regions of the function, due to utilising the non-linearity constructivly can initially be an unpredictable learning hurdle. Therefore, a function which introduces a non-linearity throughout the space may be desirable. Proposed is 'isotropic-Sinusoids', which allows for distributions to be compressed, rarefied and folded (for $|\lambda_m| > 1$) in a predictable mannor. It is hoped the network can utilise this for effective computation. This activation function is demonstrated in *Eqn.* 18. It includes a monotonicity-violating parameter $\lambda_m \in \mathbb{R}$, which may be useful.

$$\mathbf{f}\left(\vec{x}\right) = \vec{x} + \lambda_m \sin\left(\|\vec{x}\|\right) \hat{x} \tag{18}$$

If monotonicity-violating functions are useful for 'folding' a distribution, then $\alpha x + \sin\left(x\right)$ may be desirable.

4.1.1 Quasi-Isotropic Activation Functions

It is felt that a small symmetry breaking may be useful for semantically meaningful directions to develop from. This is achieved, by introducing many small purturbations to the isotropy.

One method is the quasi-quantize method, shown in *Eqn.* 19, where $[\cdot]$ indicates rounding and $\Phi\left(\vec{x}\right) \neq \vec{x}$. In fact, the anisotropic purturbation may be implemented as simply as: $\Phi\left(\vec{x}\right) = \beta\vec{x}$ for

$\beta \neq 1$. The overall angular term is approximately unit-normed, but can be trivially modified to be exactly norm-1.

$$\Phi\left(\hat{x};\alpha\right) = \frac{[\alpha\hat{x}]}{\alpha} + \phi\left(\hat{x} - \frac{[\alpha\hat{x}]}{\alpha}\right) \tag{19}$$

This produces a quasi-isotropic functional form shown in *Eqn.* 20, with an isotropy-breaking parameter α. Slight anisotropic refraction is added, independent of magnitude, such that it is predictable and thus extrapolatable to the network. Due to the angular rarefaction and compression by the proposed non-linearity, this may cause representation over and underdensitities, where semanticity may begin to be assigned. However, for $\alpha \to \infty$ isotropy is reintroduced continuously and could perhaps be an optimisable parameter.

$$\mathbf{f}\left(\vec{x}\right) = \sigma\left(\|\vec{x}\|\right)\Phi\left(\hat{x}\right) \tag{20}$$

4.2 Dynamical Networks

4.3 Normalisers, Regularisers and Optimisers

Do not wish to discourage anisotropic distributions if they are beneficial, afterall, the goal is to unconstrain the network and regularisers should reflect this. The scale factor α controls how many anisotropic-centres there are

Can imagine a dense thick layer

5 Applications

6 Conclusion

References

George Bird. The spotlight resonance method: Resolving the alignment of embedded activations. In *Second Workshop on Representational Alignment at ICLR 2025*, 2025. URL `https://openreview.net/forum?id=alxPpqVRzX`.

A Historical Precedent

B Distinction from Equivariant Networks of Geometric Deep Learning

Both isotropic deep learning and equivariant networks share a similar equivariant relation, which may make them appear superficially similar. However, they differ substantially in how this relation is implemented, motivated and consequences. This section will briefly outline those differences.

C Taxonomy of Functional Forms

D Stochastic Isotropy — Producing Immediate Anisotropic Analogs