Preliminary Steps Towards Isotropic Deep Learning

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 functional forms and proposes a new paradigm to be termed *Isotropic Deep Learning*. It has been shown that 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 neccessitated, and solely due to human imposed choices of functional forms. There appears no strong a-priori justification why such a representation or functional form is desirable, whilst in this paper several detrimental effects of the current formulation are proposed. As a result, a modified framework for functional forms will be be explored with the goal of unconstraining representations by elevating a rotational symmetry inducive bias throughout the network. It is encouraging the adoption of this framework as a new default. This direction is proposed to improve performance of networks. Preliminary activation functions, regularisers, normalisation and optimsers are proposed. Since this is an overhaul to almost all functional forms characterising modern deep learning, it is suggested that the magnitude of this shift may constitute a new branch of the deep learning.

1 Introduction

Current deep learning models typically employ an elementwise functional form. This is particularly evident in activation functions, sometimes referred specifically to as ridged activation functions, with their form often displayed univariatly as shown in *Eqn.* 1. σ is specific activation function implemented, say ReLU, Tanh, etc.

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

However, this display choice obsfuscates a crucial (standard) basis dependence, which is better displayed in their more correct multivariate form in *Eqn.* 2, revealing the $\hat{e}_i$ basis dependence of the functional form. 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 precedence, discussed further in *App* **??**.

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

Due to this dependence, non-linear transformation differ angularly in effect. Therefore, this will be termed an *anisotropic function*, indicating this rotational asymmetry. Current deep learning as a paradigm may consequently be termed *anisotropic deep learning* due to the pervasive use of these functional forms. This *choice* appears unappreciated and anisotropic form incidental in the development of the majority of current models — almost treated as axiomatic to deep learning. Hence

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

questioning and overhauling this core underpinning of modern deep learning is argued to constitute a new branch: *Isotropic Deep Learning*.

This assymetric uneveness in the non-linear transform is particularly about the standard (Kronecker) basis vectors and its 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 the model. For example, the standard basis is clearly 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, particularly permutation symmetry about the standard basis. In effect, if the function is treated in its multivariate form, then it is equivariant to a permutation to the components of its vector decomposed in the standard basis. Mathematicall this is: for an element of the permutation group $\mathbf{P} \in \mathcal{S}_n$ the following equivariance relation holds $f(\mathbf{P}\vec{x}) = \mathbf{P}f(\vec{x})$.

Figure 1: [Insert caption here]

Non-linearities are usually pivotal to the networks ability to achieve a desired computation, such as a universal approximator. The non-linearities produce differing local transformations such as stretching, compressing and generally reshaping a manifold to achieve the desired computation and it is these localised transforms which are anisotropic about a particular basis. Consequently, the network may be expected to adapt its computation, by moving representations, about these distinguished directions.

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, it is expected that the embedded representations then move to useful angular arangements about the arbitrarily-imposed privilidged basis' geometry. For example, they appear to move towards the non-linearities's extremums, aligned, anti-aligned or other aligned geometries, through training. This may correspond to a local, dense or sparse coding respectively, where the latter includes the superposition phenomena.

Therefore the network has adapted its representations through training due to these functional form choices. This is the causal hypothesis proposed. This tendency of neuron alignment is observed in several studies — with this hypothesis aiding in explaining the general case of privilidged-basis alignment. This is a human caused representational collapse onto the privileged basis, usually not a task necessitated collapse. There appears little justification as to why this is universally desirable. Without a-priori justification, this inductive bias may be detrimental to computation. Several key negative implications are discussed in *Sec.* 2.

Overall, historic and frequently overlooked functional forms for modern deep learning are having a direct influence on the models activations and therefore behaviour. To stress, there exists a functional form depenence 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

[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, *'privlidged basis'* will be used to 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 to the standard-basis

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 functional form paradigm towards the isotropic paradigm is preferable, unless otherwise justified. It is to encourage the reader to be conscious of these choices when designing a model, as well as the usual architectural tool kit. Particularly, isotropic choices, equivilant to basis indepent to any arbitrary basis, may be thought to unconstrain the representations into more optimal arangments for a task. This tenants of this paradigm are suggested for all architectures on general tasks.

2 Problems of Anisotropy

It is argued that current functional forms impose unintended anisotropic biases due to privilidged bases. This section argues how these may be detrimental to performance in general; therefore, arguing that *Isotropic Deep Learning*, once substantially developed, should be the default inductive bias chosen unless an alternative is task neccessitated. Adoption of the paradigm may be limited in the near-term due to development of suitable functions, particularly since *anisotropic deep learning* has a substantial head start and analogs to existing functions are not so easily produced.

In this section several arguments are layed out which demonstrate the detrimental implications anisotropic functional forms may cause. To the best of the author's knowledge some of these failure modes are newly characterised phenomena, such as the so-called '*neural refractive problem*' — initially the motivating realisation for this paradigm shift.

This list is far from exhaustive and particularly focuses around the consequences of activation functions, since this is the area primarily explored by the author thus far.

2.1 The Neural Refractive Problem

The '*neural refractive problem*' describes how linear trajectories of activations may converge or diverge from their initial trajectories after an activation function is applied. This is in analogy to a light-ray refracting through a optically-varying media or boundary. This appears to be a detrimental side-effect of all anisotropic activation functions to date and typically occurs by a greater amount at larger magnitudes in many current activation functions. Mathematically, this is represented through *Eqn.* 3, for a multivariate activation function f and vector $\vec{x} = \alpha\hat{x}$ where $\hat{x}$ is a unit-vector. This relation may be satisified for a single direction, subset of the space or all directions $\hat{x} \in \mathcal{X} \subseteq S^n$.

$$\exists\hat{x} : \mathbf{f}\left(\alpha\hat{x}\right) \neq \sigma\left(\alpha\right)\hat{x} \tag{3}$$

It can be seen that along a straight-line trajectory in direction $\hat{x}$, the result of the activation function may be a curved line. Therefore, if the linear feature hypothesis is followed, linear features become curved. This may be utilised by the network to construct new linear features in the following layers; however, current functional forms also introduce several undesirable nuances which are not featured in isotropic functional forms.

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

For functions such as Sigmoid and Tanh, this may be particularly consequential since large magnitude inputs end up at particular limit points (discussed as trivial representational alignments in Bird [2025]). For example, Tanh produces the stable limit points shown in *Eqn.* 4 when $\hat{x} = \sum_i a_i\hat{e}_i$ for $a_i \neq 0$ for all i. Where a component is 0 then the corresponding limit point has a 0 in the matching component to, these could deemed 'unstable' limit points.

$$\lim_{\alpha\to\infty}\mathbf{f}\left(\alpha\hat{x}\right) = \left(\pm1, \cdots, \pm1\right)^T \tag{4}$$

Consequently, semantic separability is entirely lost except for 3^n discrete limit points for Tanh and Sigmoid. Therefore, embedded activations may be expected to align with these limit points, which is an empirically observed tendencyBird [2025]. Similarly, ReLU has one distinct limit point, $\vec{0}$, but otherwise a 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.

More generally, deflection of trajectories may cause semantic ambiguity for the network, where only samples interpolatable from training samples are robustly semantically identifiable. Particularly, *larger the deflection; we may expect greater semantic ambiguity*. Larger deflections typically occur at larger magnitudes in many current functions. Deflection functions can be a trivial diagnostic measure defined by *Eqn.* 5 for a particular activation function.

$$\theta\left(\alpha;\hat{x},\mathbf{f}\right)=\arccos\left(\frac{\mathbf{f}\left(\alpha\hat{x}\right)\cdot\hat{x}}{\left\|\mathbf{f}\left(\alpha\hat{x}\right)\right\|}\right) \tag{5}$$

This may explain why the network may perform excessivly poorly on out-of-training-distribution samples. For example, if a linear feature roughly represents the quantity of cows in a field, then the network may fail to extrapolate its function when an anomalous amount of cows are present, as this would be a large magnitude of the linear feature and therefore is detrimentally deflected by the activation function's non-linearity and lose its semantic meaning. Consequently, it may be expected that through training, if a network intends to preserve a linear feature, then it may reduce magnitudes to within a zone of predictable non-linear behaviour to avoid the damaging consequences of neural refractions.

This phenomena is fundamentally caused by angular anisotropies. If compression and rarefaction of certain angular regions occurs, then linear features will be consistently deflected in various ways. The unique fix for this is introducing isotropy, which is initial motivation for development of 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 in a predictable way. It is argued that these are only an issue when it affects linear, not affine, features in a potentially unpredictable and thus semantically uninterpretable way.

The phenomena is eliminated from networks by changing *Eqn.* 3 to an equality. This then becomes the standard relation for rotational equivariance of the function. This can be expressed as a condition in *Eqn.* 6, which uses an commutator bracket for convenience and $\forall\mathbf{R}\in\mathrm{SO}\left(n\right)$.

$$\left[\mathbf{R},\mathbf{f}\right]=\left(\mathbf{R}\mathbf{f}-\mathbf{R}\mathbf{f}\right)=\vec{0} \tag{6}$$

It 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 for single argument functions and requires generalising for more circumstances. This may be recognised as superficially similar to equivariant neural networks, due to an analogous equivariance relation; however, the differences are discussed in *App* **??**.

This introduces the general isotropic functional form for activation functions given in *Eqn* 7. This is $\mathcal{O}\left(n\right)$ for $\mathbb{R}^n$. Future work is establishing a universal approximation theorem for this functional form, as this is on-going research for the author's PhD.

$$\mathbf{f}:\mathbb{R}^n\rightarrow\mathbb{R}^n,\quad\vec{x}\mapsto\mathbf{f}\left(\vec{x}\right)=\sigma\left(\left\|\vec{x}\right\|\right)\hat{x} \tag{7}$$

Overall, the 'neural refractive problem' outlines how semantic meanings may become intertwined or ambigious 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 and thus the network's generalisation then fails excessivly.

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 to geometric positions about the distinguished bases, then any small purturbation to a parameter may misalign activations to an existing 'understood' semantic of the network. In effect, further small

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

purturbations to the parameters may move activations from a semantically-aligned to a semantically-dislocated state, thus the activation's meaning becomes ambigious, forfeighting 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 about the space. Only sufficiently large purturbations 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-minimas which can be smoothly transformed into one-another, by corresponding parameter rotations shown in *Eqn.* 8, a consequence of *Eqn.* 6. If this is downgraded to discrete rotational symmetry (i.e. permutation symmetry), then artificial optimisation barriers may reemerge in these basins. In which case, only a sufficiently large purturbation to the parameters may dislodge the network into a more optimal minima, whilst smaller purturbations are insufficient. Effectively, the discrete permutation symmetry may result in a discretised lattice solutions for the parameters, much like how it breaks the symmetry of activations through training too Bird [2025].

$$\forall \mathbf{R} \in \mathrm{SO}\left(n\right) : \underbrace{\mathbf{W}^l \mathbf{R}^\top}_{\mathbf{W}'^l} \mathbf{f}\left(\underbrace{\mathbf{R}\mathbf{W}^{l-1}}_{\mathbf{W}'^{l-1}} \vec{x} + \underbrace{\mathbf{R}\vec{b}}_{\vec{b}'}\right) = \mathbf{W}^l \mathbf{f}\left(\mathbf{W}^{l-1}\vec{x} + \vec{b}\right) \tag{8}$$

This exists as a qualatative intuition, since until robust methods to determine semanticically meaningful directions are produced, this remains a difficult-to-verify hypothesis. Nevertheless, steps can be taken immediatly 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 may be tended towards. Therefore 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. It appears a poor inductive bias to have functional forms encourage discrete representations. Isotropic functions do not prevent discrete semantics, but they don't encourage them either — enable continuous ones since they are generally unconstraining the representations. Therefore, moving towards isotropy would encourage embeddings to be more smoothly distributed, being able to take on intermediate values between typically discrete linear features and hence substantially enlarge the expressivity and representation capacity of networks — only limited by concept interferences.

In this case, the discrete concept of representation capacity may become irrellevant; each layer may express different continuous arangements, 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 indicates the particular concept.

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

Therefore, in generality the inductive bias of isotropy appears more appropriate as a default, unless justification for anisotropy is present.

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

conjecture that semantics are also discretised into linear features. Few samples populate the intermediate

A continuous symmetry may also enable continuously interpolatable semantics. Representations may then occupy intermediate unaligned states which are semantically interpolatable, rather than discrete semantic directions suggested by the linear features hypothesis. This may advantage the model by allowing

As mentioned, there is little justification that privilidged-basis alignment is desirable to begin with. Allowing continous semantic interpolation may allow the representations to become unaligned from any privilidged basis and produce a more effective distribution - one which suits the data or the task, rather than a human imposition.

2.4 Non-Linear Corrections

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`.