The Affine Divergence
Aligning Activation Updates Beyond Normalisation

George Bird — University of Manchester
george.bird@postgrad.manchester.ac.uk

Replacing Parameter Steepest Descent with Activation Steepest Descent → New Normalisations (PatchNorm) and Affine Layers

Overview

Parameter steepest descent in affine (and convolutional) layers produces a mismatch in activation steepest descent. Correcting this yields two solutions: one is the classical (parameterless) L_2-norm, the other a replacement for the affine map. Both perform strongly across different comparable MLP networks in an ablation test, with the affine-like solution often outperforming other maps, e.g. Batch/Layer/RMS-Norm. It is especially interesting given that the affine-like map does not exhibit scale invariance, so it offers some counter evidence for this explanation.

$$\mathbf{f}\left(\vec{x};\mathbf{W},\vec{b}\right)=\mathbf{W}\frac{\vec{x}}{\|\vec{x}\|}+\vec{b}=\mathbf{W}\hat{x}+\vec{b}$$

Solution 1: **Norm-Like** (Parameterless L_2-Normalisation)

$$\mathbf{f}\left(\vec{x};\mathbf{W},\vec{b}\right)=\frac{\mathbf{W}\vec{x}+\vec{b}}{\sqrt{\|\vec{x}\|^2+1}}$$

Solution 2: **Affine Replacement** (In-built normalisation)

$$y_{i,j,d}=\frac{\sum_{u,v}\sum_{c}W_{u,v,c,d}\,x_{i+u,\,j+v,\,c}+b_d}{\sqrt{\sum_{u,v}\sum_{c}x^2_{i+u,\,j+v,\,c}+1}}$$

Convolution Solution: **PatchNorm**

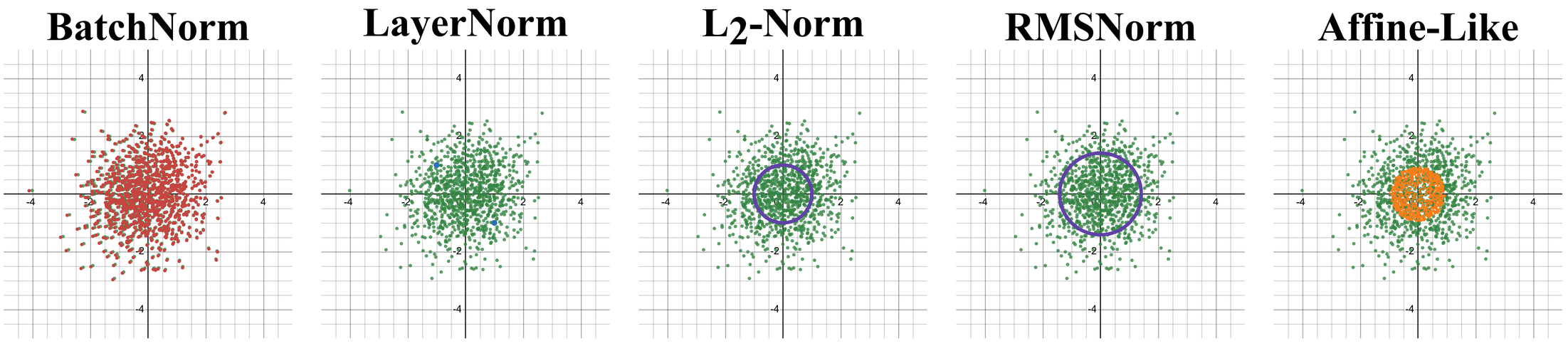
Highlights

1. Unexpected first-principles derivation of the L_2-normaliser from the premise of steepest-descent activation updates.
2. Replacement for the affine map, predicted from theory. Outperforms alternatives, doesn't feature scale invariance, enabling a mechanistic ablation.
3. Predicts, then verifies, a surprising negative correlation between performance & batchsize for these new normalisers — acts as secondary, independent confirmation of this mechanistic hypothesis.
4. New convolutional normalisers "**PatchNorm**", derived from the same approach but applied to convolutional layers. Preserves locality similar to BatchNorm, acting per feature patch.
5. Can be further generalised across many scenarios, enabling a new avenue of exploration.
6. Speculative explanation for the prenormalisation step in query-key attention (by applying theory to queries and keys independently).
7. Considers the geometric over statistical interpretation of normalisers, arguing for category unification with activation functions.

Discovery

1. During backpropagation, various gradient quantities are determined with respect to the loss.
2. It is typical to modify the parameters by subtracting their scaled gradient: the steepest descent step (in parameter space).
3. But, in turn, this modifies activations by a small correction.
4. This small correction is shown to *not* be the activation's steepest descent, due to a quadratic-scaling term.
5. This quadratic term produces a geometric distortion, suggesting it may be pathological.
6. Removing it requires changes to the forward map, producing two distinct solutions: L_2-Norm and a replacement for the affine map.
7. Both solutions produce the steepest descent in **both** the parameter and activation spaces.
8. This generalises to convolution through a new functional form "**PatchNorm**".

Overall: **Propagate parameter updates into activation corrections, then compare to analytical gradient — modify until they equal.**



Shows the effect of various maps on a standard multivariate normal point cloud (green), for various normalisation maps $\mathbf{f}:\mathbb{R}^2\to\mathbb{R}^2$ as different coloured point clouds overlayed.

Takeaways

1. It appears worthwhile to explore novel forward maps that exhibit steepest-descent activation.
2. Provides some counter evidence for scale-invariance in normalisation, through the affine-like solution.
3. New forms of affine layers and convolution are derived from first principles in this approach.