Milestone 007 — Derive before training

Tyler J.H.G.

Abstract

What can we prove about normalization before spending compute?

Research period: September 2–3, 2026 and later formal hardening
Historical anchors: #196, #199, #200, #252

The unit constraint left several possible mechanisms open. It could remove an unhelpful magnitude channel, change optimization, suppress a shortcut, preserve some directions while contracting others, or alter recurrent dynamics. Training a new model for each possibility would be an expensive way to discover which premises were already mathematically false.

I started deriving the mechanisms first. If an explanation requires normalization to behave a particular way, we can establish that behavior before asking what the trained model does with it. The remaining experiment then has a more precise job.

> **Formal checkpoint:** for a nonzero vector `x`, normalization is `N(x) = x / ||x||`. If `u = x / ||x||`, its derivative is
>
> `DN_x = (1 / ||x||)(I - uuᵀ)`
>
> The checked version is available in the public Theorem Library normalization proof.

= Visual intuition

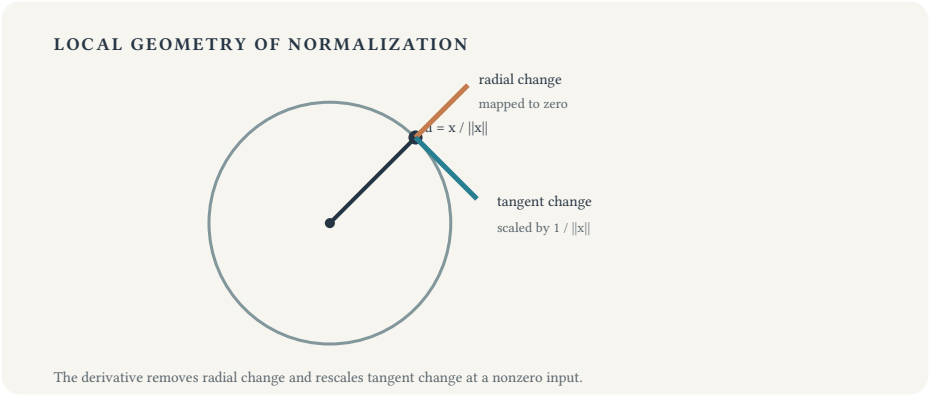


Figure 1: At a nonzero input, the derivative of normalization removes radial change and rescales tangent change.

Published Sep 29, 2026

1. Radial and tangent directions

The matrix `I - uuᵀ` projects away the component parallel to `u`. That gives us two local cases.

Change only the magnitude of `x`, and the normalized direction stays fixed to first order: the radial derivative is zero. Change `x` perpendicular to its radius, and the perturbation survives with scale `1 / ||x||`. At unit radius, its tangent component is preserved to first order.

> **Sticky note — tangent direction:** at a point on a sphere, a tangent direction is perpendicular to the radius. It describes local movement along the surface.

This establishes exactly how normalization acts on small perturbations. We still have to identify what those perturbations carry in the model: a geometric direction acquires a task-specific meaning through the representation and data.

2. Try a small example

A two-dimensional example separates radial and tangent perturbations. Try changing the point and perturbation vectors.

```
import numpy as np
x = np.array([3., 4.]); r=np.linalg.norm(x); u=x/r
DN = (np.eye(2)-np.outer(u,u))/r
radial = u
tangent = np.array([-u[1], u[0]])
print('radial derivative:', DN @ radial)
print('tangent gain:', np.linalg.norm(DN @ tangent))
```

Saved output:

```
radial derivative: [-8.39328607e-18 -1.45217172e-17]
tangent gain: 0.2
```

3. What this changed

Suppose our explanation is that magnitude carries nuisance variation while direction carries useful information. The derivative establishes the local removal of radial variation. To complete the explanation, we also need to measure whether radial variation is nuisance and whether tangent variation carries the relevant task distinctions.

Writing the argument this way separates its premises. Some can be derived, some need an experiment, and some may turn out to be inconsistent with the construction. A proof or counterexample can then remove a branch before we build a model for it.

The workflow became:

```
hypothesis → derivation / theorem boundary → surviving empirical
question → experiment
```

Each step tells us what the next one still needs to establish.

4. Research context

Normalized Transformer states and hyperspherical optimization already appear in work such as **nGPT**. That literature provided constructions to compare against and helped locate the mathematical questions in this project.

My immediate question concerned the mechanism behind the measurements I already had: which effects follow from normalization itself, and which depend on this trained representation? Existing work supplies part of the first answer; the formal calculation makes the local constraint explicit, leaving the task-dependent premises for measurement.

5. Interpretation

The derivative establishes a local, first-order effect: radial perturbations vanish, and tangent perturbations scale by inverse radius. Whether either direction carries useful task information depends on the trained model.

Recurrence adds another question. After several nonlinear updates, what survives from a perturbation introduced at the beginning? That required finite-horizon diagnostics on the frozen model.

6. Public formal sources

- Normalization
- NormalizationSpectrum

These pinned sources state the assumptions and checked mathematical conclusions.

7. Sources and chronology

- Milestone notebook 007 — preserved chronological source
- Research chronology
- Publication dispositions and Atlas-to-article map

The matching Atlas theme is **Normalization derivative**. The earlier Visual Intuition Atlas notebook preserves the source visualization.

The next chronological article is **Milestone 008 — What does the sphere actually do?**.