Notes on the paper “A Geometric Modeling of Occam's Razor in Deep Learning”

By Paul Thomson, June 7, 2025

Relevant questions regarding our understanding of the Deep Learning models

Why are we taught by statisticians to prioritize simpler models yet we still use 1-billion-parameter neural nets to diagnose MRIs and a 1-trillion-parameter ChatGPT?

Why do all methods seem to disobey Occam’s razor - the principle that simpler models tend to generalize best?

Answer: Spend a couple hours with this insane paper that just dropped yesterday [1], I will try to explain it a bit as it's very cool.

Bonuses: if you stick around to the end, you will see how optimization trajectories in deep nets often follow "light-like paths" in general relativity. OK here’s the explainer, starting with a high-level tricky explanation then trying to make it a bit simpler:

The authors treat the parameter space of deep neural nets as a singular semi‑Riemannian manifold (will explain this in a minute!), but the Fisher information matrix (FIM) is highly degenerate and does not have as many meaningful or helpful parameters as it seems

The number of significant (non-zero) eigenvalues of the FIM reflects how many directions in parameter space meaningfully affect predictions, and it's not very many, it's way less than the dimension of the neural net

Using minimum description length theory, they derive an MDL-style complexity term where singular (low-eigenvalue) dimensions contribute a “negative complexity.” This explains why massively over-parameterized networks can still generalize well.

Let’s unpack this slowly, as the last part on "negative complexity" is not very intuitive (at least to me):

The Fisher information matrix (FIM) tells you how sensitive the likelihood of your data (strictly, the log-likelihood under your model) is to changes in your model parameters that you are optimizing. If you are fitting a generative model to a batch of training data like brain MRIs or photos of faces or images of hand-written digits, the model might fit better if you tweak 100 of the neural net’s parameters, but not all 1,000,000. The other parameters might be redundant or “degenerate”.

If you have a model with parameters θ, and you observe data x, the FIM measures the curvature of the log-likelihood function:

This is a positive semi-definite matrix that captures *how much each parameter influences the predictions, and how parameters interact* (via the off-diagonal terms).

High curvature (large eigenvalues) means small changes in that direction of parameter space make a big change to the likelihood hence the parameter is "well-identified".

Low curvature (small or 0 eigenvalues) means the model is "flat" in that direction so that parameter doesn’t affect the output much.

So people have been studying the FIM spectrum (eigenvalue set) for a while.

Generalization is still possible in large models, as the "effective" number of parameters (rank of FIM) may be small. In neural networks, the FIM can be very high-dimensional (e.g. millions × millions), is often low-rank: many parameters don’t significantly affect the loss (especially in overparameterized models).

At first I did not see why they say in the paper that singular (low-eigenvalue) dimensions contribute a “negative complexity.” Surely they shouldn’t contribute anything at all, i.e. the complexity is not less.

The paper claims that having extra, ineffective parameters can actually reduce the “cost” of encoding the model, because they collapse the posterior volume without adding information (negative complexity). This is why massively overparameterized deep nets can still generalize.

The complexity of a model includes a term for the best-fit likelihood, and a complexity penalty for the number of “effective parameters”.

So if you train an autoencoder on MNIST digits, VAEs with massive decoders and encoders generalize just fine, because the effective model capacity (as measured by FIM spectrum) is low.

Most encoder parameters don’t matter: In a VAE with a 10D latent space and a large encoder, only a small number of directions in parameter space significantly affect the output; the rest are nearly "flat."

The Fisher Information Matrix (FIM) has many tiny eigenvalues in these flat directions, which means the model is insensitive to changes along them.

In the MDL formula, small eigenvalues contribute large negative log terms — so these unused directions lower the total model complexity, which is called “negative complexity.”

Bonus: if you like Riemannian geometry, the paper also draws an analogy to light-like geodesics from general relativity to describe the geometry of deep neural networks and how their parameters relate to the log-likelihood of the data (model fit).

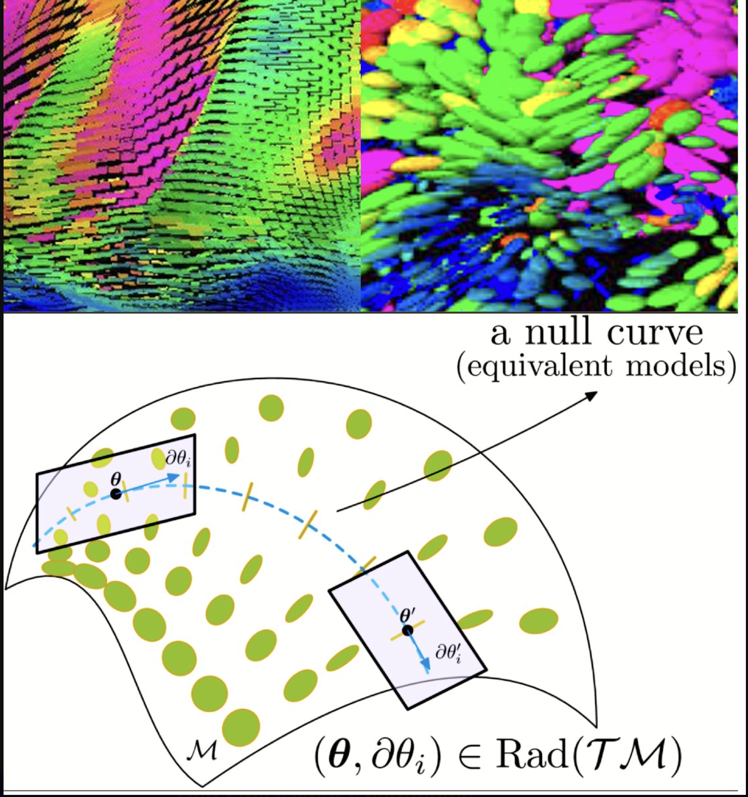
In general relativity, a light-like or null geodesic is a path that light takes through space-time. These paths have zero length in the spacetime metric (i.e., they’re on the edge between time-like and space-like).

The paper applies this mathematical idea to deep networks: they treat the parameter space of a neural network as a semi-Riemannian manifold, where the Fisher Information Matrix (FIM) plays the role of a local metric.

Singular directions in the FIM (those with near-zero curvature) behave like light-like directions: you can move along them in parameter space without changing the model’s predictions (the log-likelihood remains constant).

These are the flat directions of the loss landscape, they contribute zero “distance” (or change in information) even though the parameters move.

​​So optimization trajectories in deep nets often follow these "light-like paths", proceeding through large regions of parameter space where nothing changes, and this helps explain generalization when there's a lot of parameter redundancy.



References

[1] [A Geometric Modeling of Occam's Razor in Deep Learning, Ke Sun, Frank Nielsen, 2025](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/geometric_deep_learning/A_Geometric_Modeling_of_Occams_Razor_in_Deep_Learning_Sun_2025.pdf)

[2] [Note on X by Paul Thompson on the paper, Jun 7th, 2025](https://x.com/PTenigma/status/1931431292383756664)