The Goldilocks Zone and Geometric Features of High-Dimensional Parameter Spaces

Jeffrey Chang, Department of Physics, Stanford University
jeffjar@stanford.edu

Summary

How can we think about the high-dimensional parameter spaces of neural networks?

One hypothesis[1] is that the good solutions lie within a hyper-annulus ("Goldilocks Zone").

This project

- verifies the existence of the Goldilocks Zone when fitting to the CIFAR-10 dataset.
- gives simple geometric arguments that explain some of the observed behaviors

Background

In modern neural networks, the number of parameters $D \sim 10^{5+}$ is extremely large.

The parameters live in a D-dimensional vector space. When a network is trained, it searches for solutions by tracing out a trajectory $\vec{r}(t)$ in parameter-space. A good solution has low loss $J(\vec{r})$ and high generalization accuracy.

Some regions of parameter-space may contain more good solutions than others.

Related Work

Neural networks are heavily overparameterized.

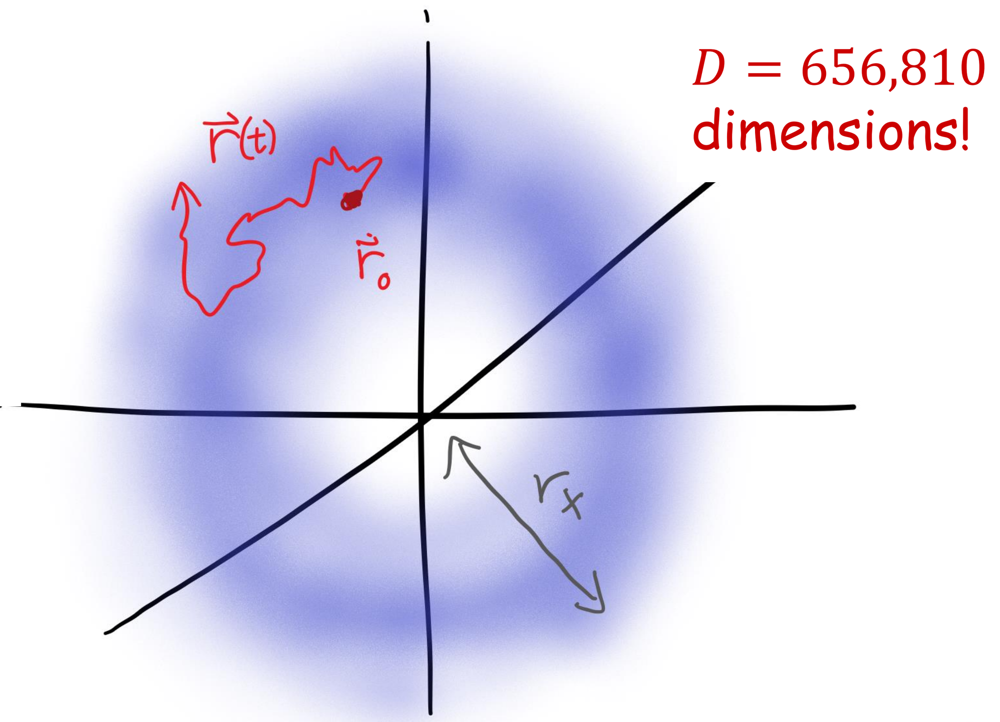
It is possible[2] to find good solutions when the training trajectory is restricted to random d-dimensional hyperplanes—*even if* $d \ll D$.

Further work[1] has demonstrated that the accuracy achieved on a hyperplane *depends on the radial distance* $r \equiv \|\vec{r}\|_2$.

For fully connected networks trained on MNIST, good solutions appear to be common in the hyper-annulus $\frac{1}{10} r_X \lesssim r \lesssim 10\, r_X$, where r_X is the typical radial distance (i.e., L2 norm of weights) of successful initialization schemes (e.g. Xavier, He).

Does the same result hold for CIFAR-10?

Parameter Space



$$r_i = W_{ab}^{[\ell]}, \quad |\vec{r}|^2 = \sum_{\ell,a,b} \left(W_{ab}^{[\ell]}\right)^2$$

where $\boldsymbol{W}^{[\ell]}$ are the weight matrices.

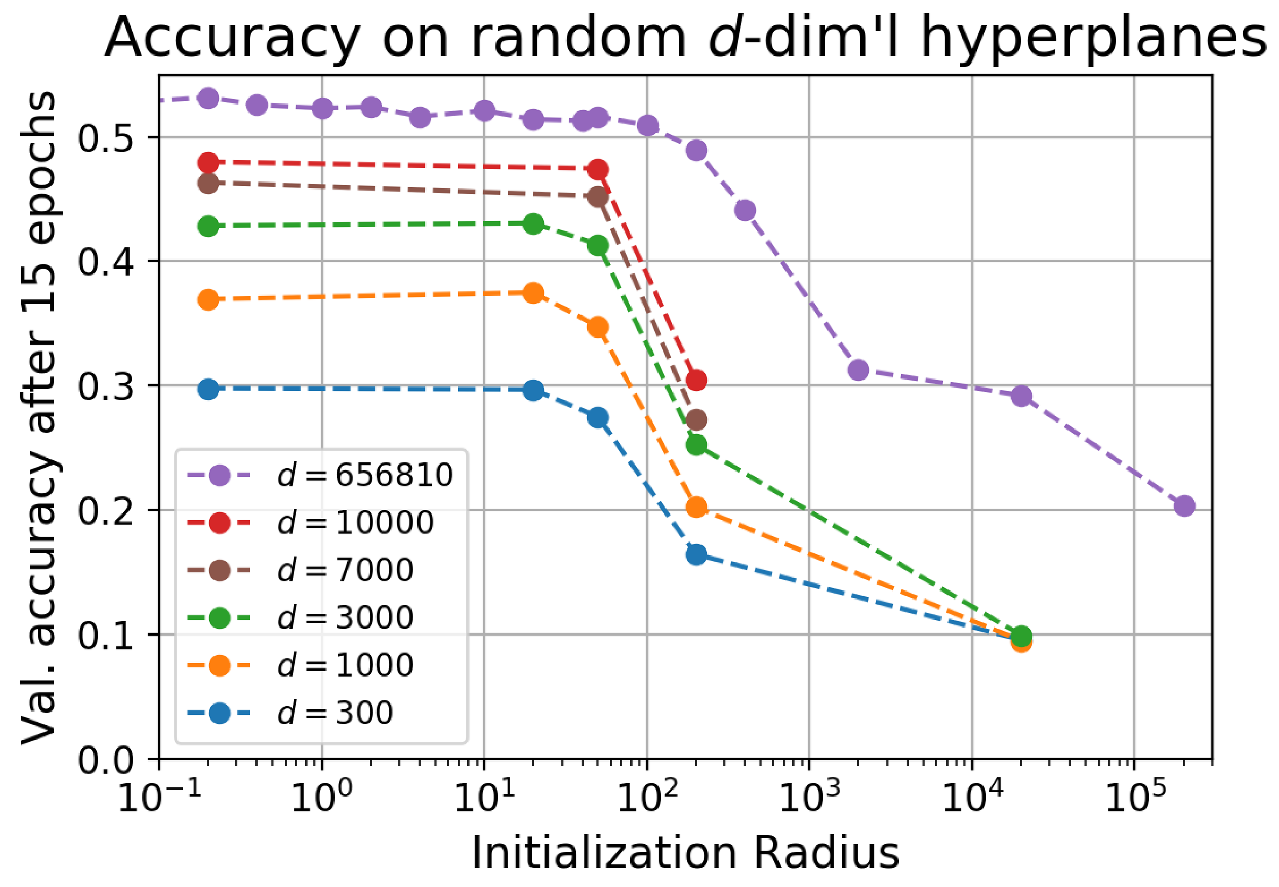
Random d-dimensional hyperplane

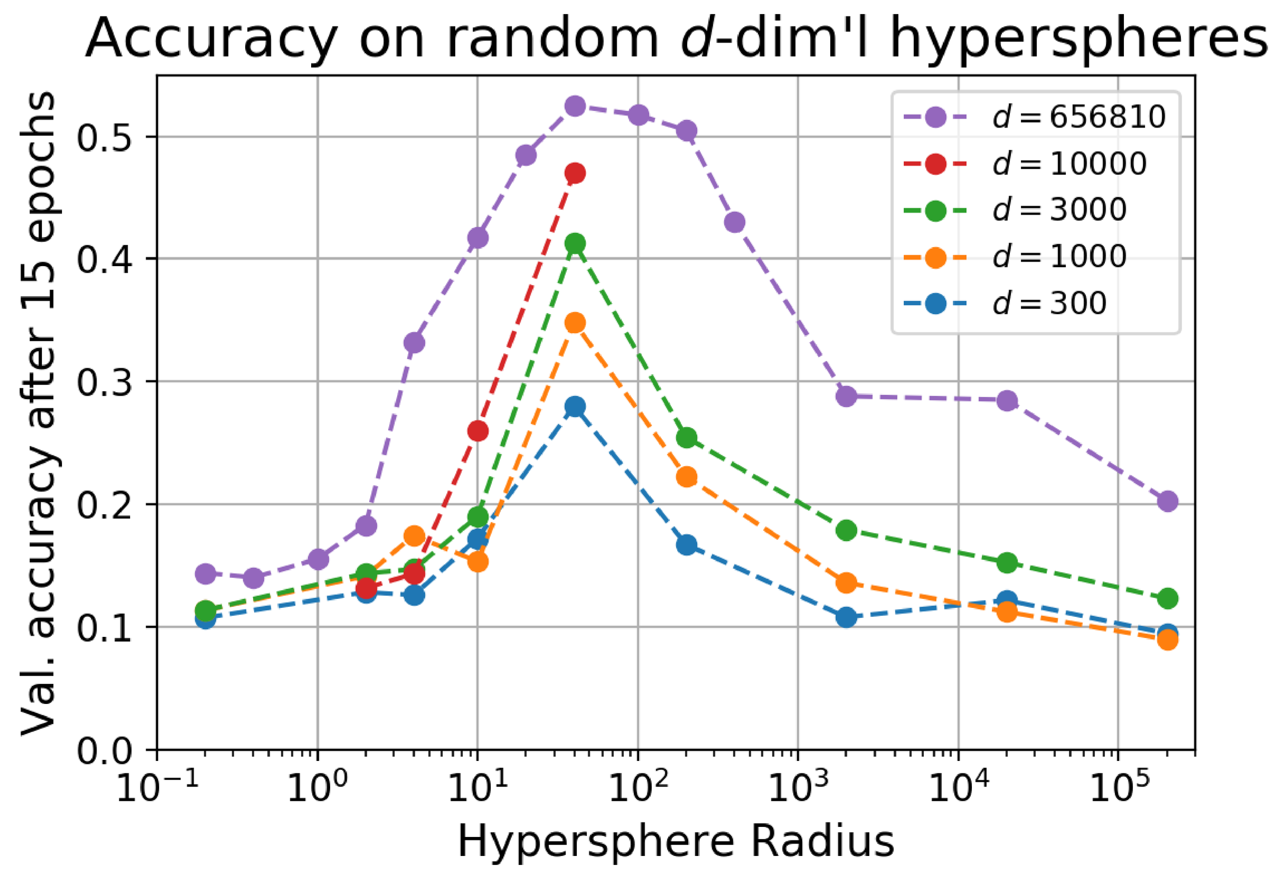
$$\vec{r} = \vec{a} + \boldsymbol{P}\vec{\theta}, \qquad \theta \in \mathbb{R}^d$$

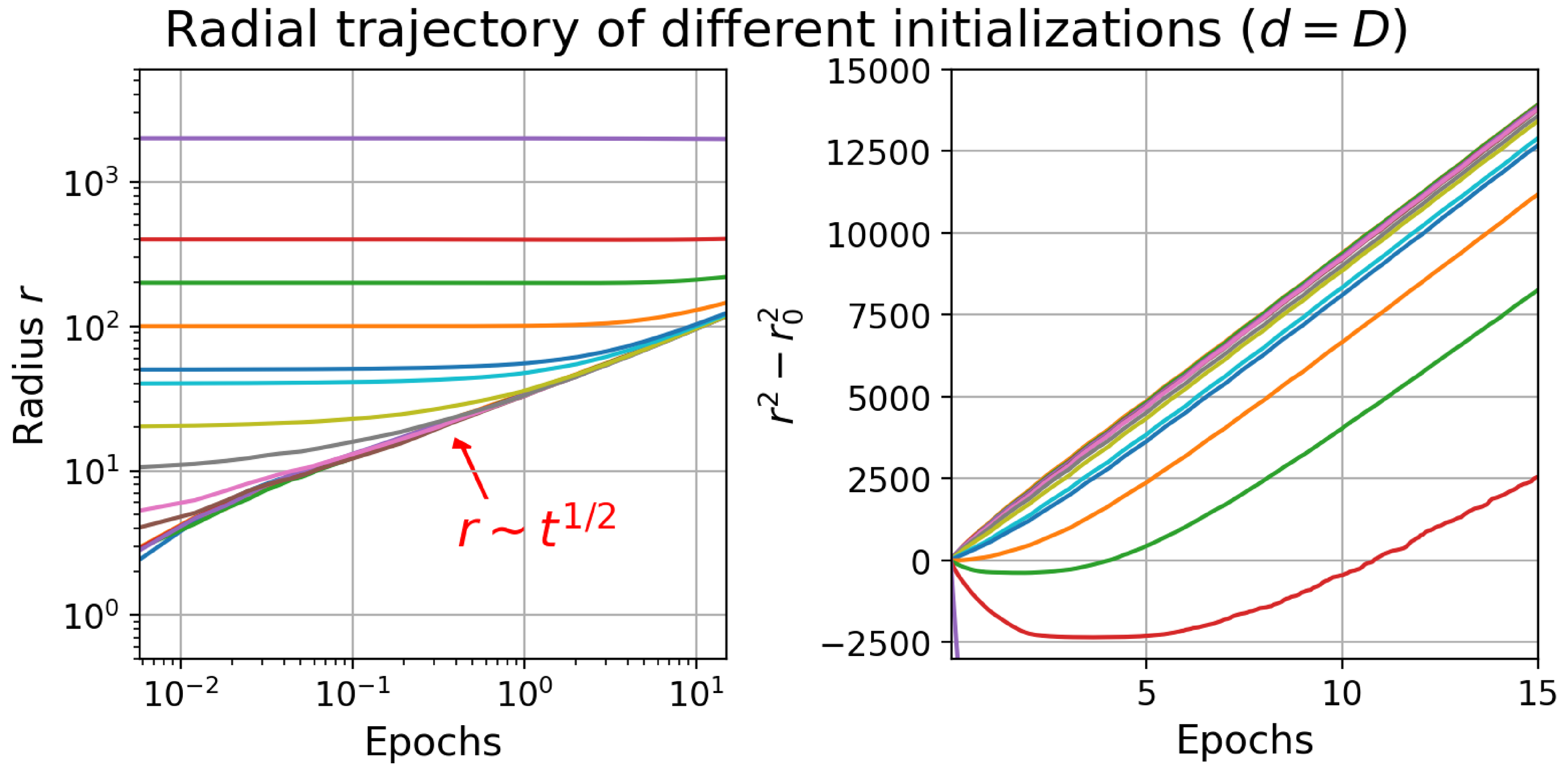
Random d-dimensional hypersphere

$$\vec{r} = \boldsymbol{P}\vec{\theta},\ \|\vec{r}\|_2 = r_0, \qquad \theta \in \mathbb{R}^d$$

Fix a random $\vec{a} \in \mathbb{R}^D$ and $\boldsymbol{P} \in \mathbb{R}^{D\times d}$ with orthonormal columns, and only train the d parameters θ_i.







Details

A fully connected neural network (3702 → 200 → 200 → 10) was trained using the Adam optimizer on the CIFAR-10 dataset (60,000 RGB images of resolution 32x32 in 10 classes). The validation accuracy was measured after 15 epochs.

The projection matrix P was implemented as a sparse matrix with $O(\sqrt{D}d)$ nonzero entries. Each entry was chosen to be ± 1 with probability $1/\sqrt{D}$. The columns were subsequently normalized.

The radius was constrained by subtracting off the radial component $\hat{r}\cdot\nabla_\theta \ell$ from the gradient, and then re-normalizing the radius $\vec{r} \coloneqq \vec{r}\, r_0/|\vec{r}|$ after each time step.

Works Cited

[1] Fort, Stanislav, and Adam Scherlis. "The Goldilocks zone: Towards better understanding of neural network loss landscapes." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019.

[2] Li, Chunyuan, et al. "Measuring the intrinsic dimension of objective landscapes." *arXiv preprint arXiv:1804.08838* (2018).

Findings

Fully-connected neural networks trained on CIFAR-10 exhibit a Goldilocks Zone.

- When initialized at $r_0 < r_X$, the radius grows as $r \propto \sqrt{t}$, and a good solution is found near $r \simeq r_X$ after 15 epochs of training.
- When initialized at $r_0 > r_X$, the radius does not change appreciably over training, and no good solution is found.
- If the radius is constrained to a fixed r_0, a good solution can only be found if $r \simeq r_X$.

Explanations

Some of these observations can be explained by the peculiar properties of high-dimensional spaces.

(1) There is much more volume where r is large.

$$\int \ldots\, d^D\vec{r} = \int \ldots\, r^{D-1} dr$$

So all else being equal, regions of large r are more likely to contain solutions.

(2) The radius increases along the vast majority of directions in high-dimensional space.

- *Consider a random walk $\vec{r}(t) = \vec{r}_0 + \vec{s}$, where $s_i \sim \mathcal{N}(0, \sigma^2 t)$.*
- *Then $\langle |\vec{r}(t)|^2 \rangle = r_0^2 + D\sigma^2 t$.*

In this light, it is unsurprising that the radius of the training trajectory grows as $r \propto \sqrt{t}$ when $r_0 < r_X$.

Further Questions

- Why are solutions hard to find when $r_0 \gg r_X$?
- How similar is a training trajectory to a random walk? What about for $d \ll D$?

Acknowledgements

The author would like to thank Stanislav Fort for inspiring this project and for providing stimulating discussions.

Stanford University