Measuring the Intrinsic Dimension of Objective Landscapes

Chunyuan Li*, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski

UBER AI Labs · Duke University

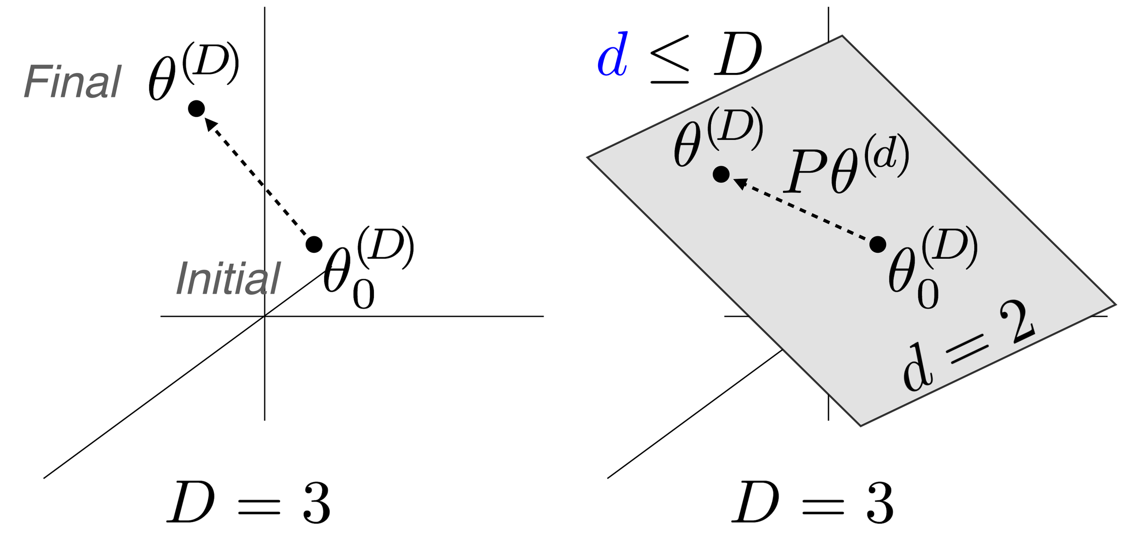
Blog:

Motivation

- Find the minimum number of trainable parameters for a specific task
- A quantitative metric to compare task difficulty across different domains

Method

- **Direct Training**: Optimization in the naive parameter space
- **Subspace Training**: Optimization in a random subspace of lower dim.



$$\theta^{(D)} = \theta_0^{(D)} + P\theta^{(d)}$$

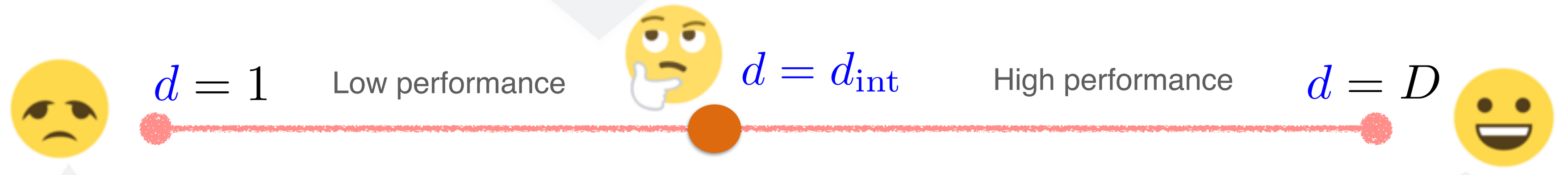
Projection matrix

Subspace parameters

Dimension of solution set

$\longrightarrow S \geq D - d_{\text{int}}$

As we increase d, we generally observe a transition of network performance, at which we define **Intrinsic Dimension !**



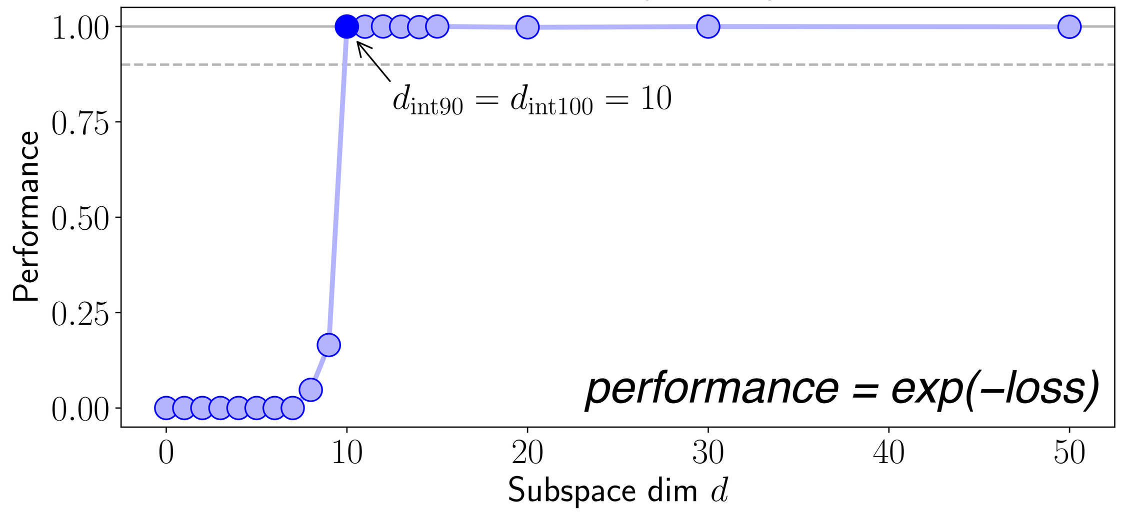
1d random line search; Hard to find a good solution

The entire space is spanned; Any available solutions can be discovered

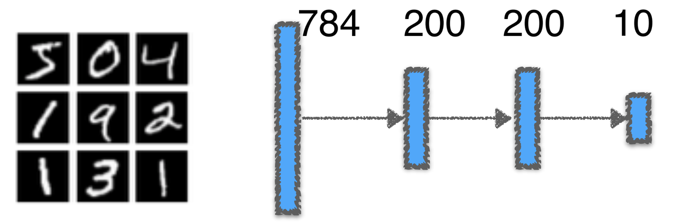
Toy Problem

- $D = 1000$
- $d_{\text{int}} = 10$
- $s = 990$

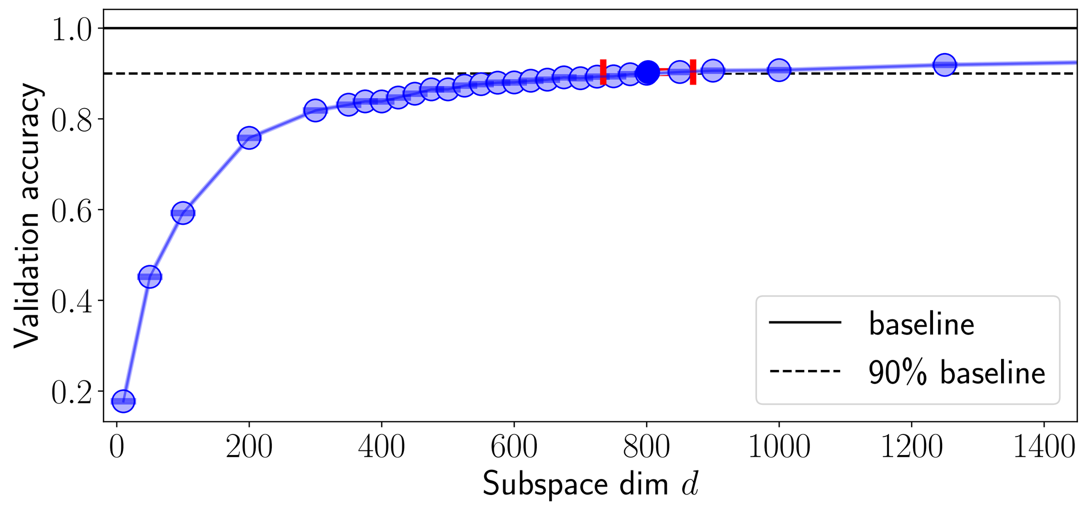
Constraints: the sum of every 100 parameters is 1



FC on MNIST



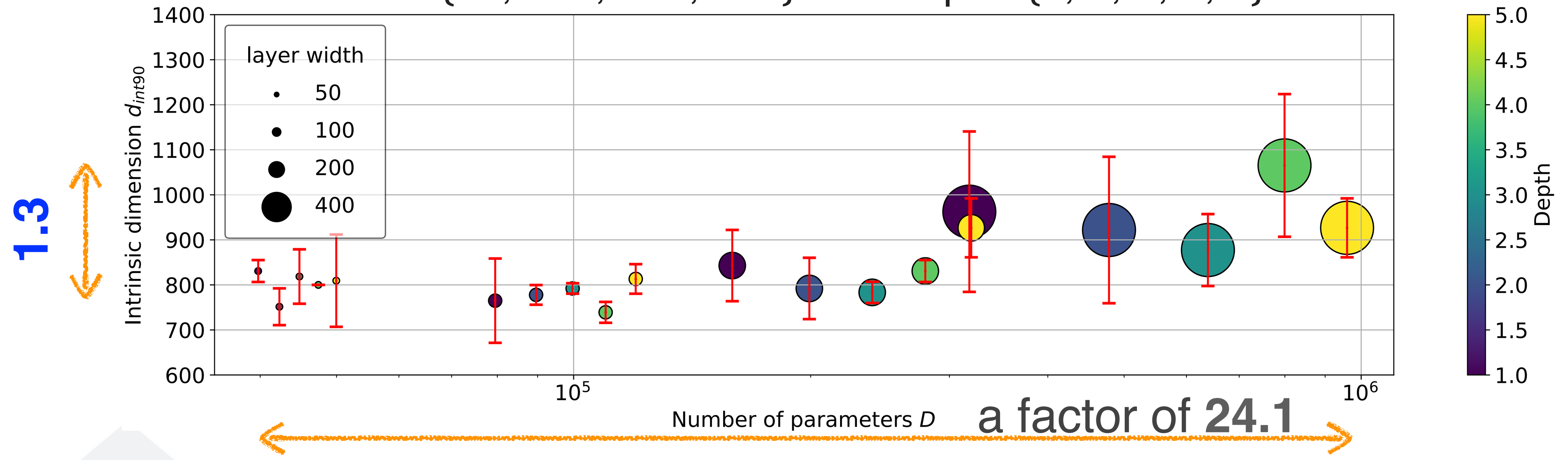
$D = 199,210$ $\quad d_{\text{int}} = 750$



- **750 is smaller** than the input pixel 784; A lot of image pixels are always black
- **Highly redundant solution set**: S > 199,210 - 750 = 198,460
- **High compression rate**: 0.4%; Storage only requires 750 parameters + 1 seed

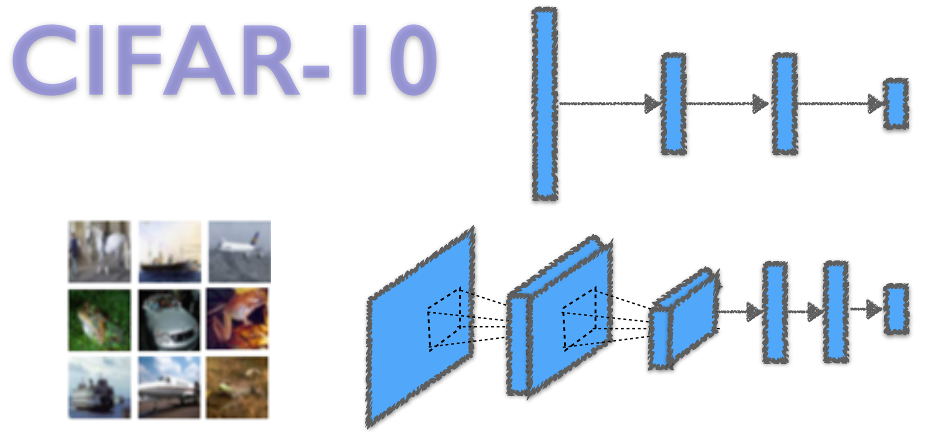
Wider or Taller FC on MNIST

Width = {50, 100, 200, 400} and Depth={1, 2, 3, 4, 5}



- A stable metric across a family of models
- Every extra parameter added to the native space just goes directly toward increasing the redundancy of the solution set

CIFAR-10



$d_{\text{int}} = 9,000$

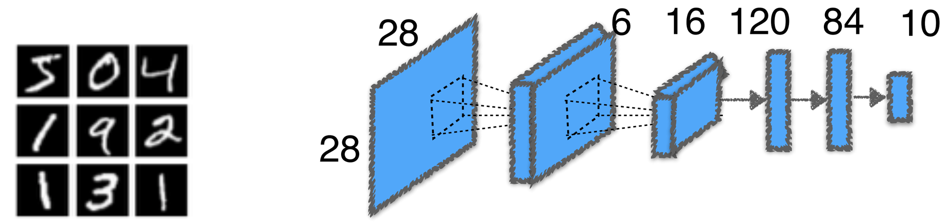
$d_{\text{int}} = 2,900$

10 times harder than MNIST, for both FC and CNN

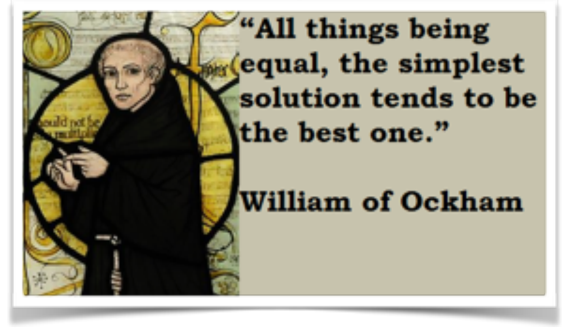
ImageNet $d_{\text{int}} = 800,000$

Fastfood transform for efficient projection

Are CNNs always better than FC?



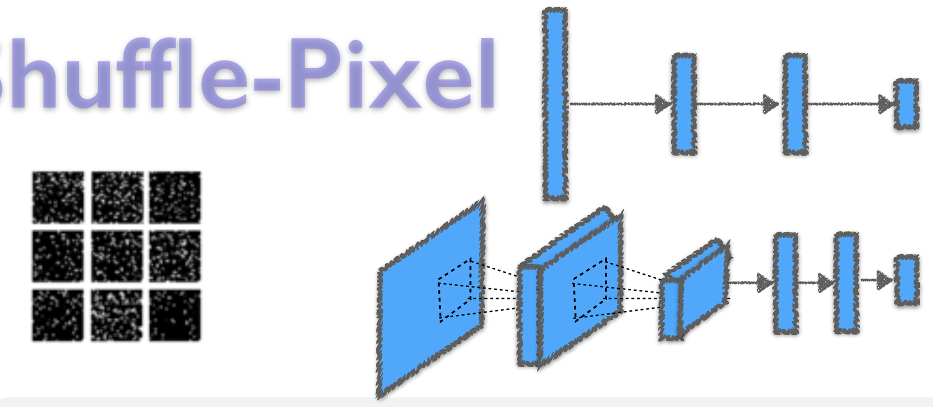
$D = 44,426$ $\quad d_{\text{int}} = 290 < 750$



Occam's Razor

Intrinsic Dimension $\geq$ **Minimum Description Length (MDL)**

Shuffle-Pixel

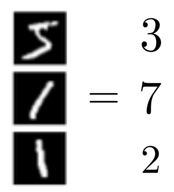


$d_{\text{int}} = 750 = 750$

$d_{\text{int}} = 1,400 > 290$

CNNs are better until the assumption of local structure is broken, after which they're measurably worse

Shuffle-Label



Dataset size	>> 750	#Para./label
5K	$d_{\text{int}} = 90,000$	18
50K	$d_{\text{int}} = 190,000$	3.8

Training on random labels forces the network to set up a base infrastructure to make further memorization more efficient

Reinforcement learning

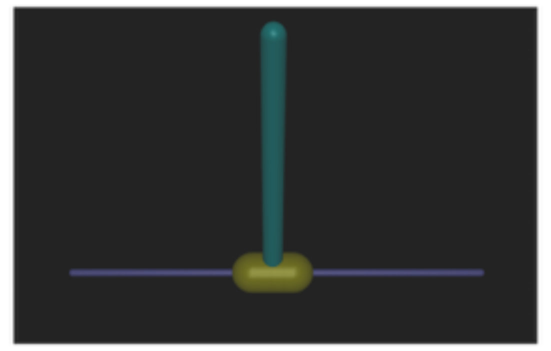
$d_{\text{int}} = 6,000$ ~ CIFAR

Atari Pong

$d_{\text{int}} = 700$ ~ MNIST

Humanoid

$d_{\text{int}} = 4$



Inverted Pendulum

The low d_{int} suggests why random search and gradient-free methods work

* Work performed as an intern at Uber AI Labs