Deep Learning: Optimization and Generalization

1

Yuan YAO

HKUST

Some Theories are limited but help:

- *Approximation Theory and Harmonic Analysis* : **What functions are represented well by deep neural networks, without suffering the curse of dimensionality and better than shallow networks?**
  - Sparse (local), hierarchical (multiscale), compositional functions avoid the curse dimensionality
  - Group (translation, rotational, scaling, deformation) invariances achieved as depth grows
- *Optimization*: **What is the landscape of the empirical risk and how to optimize it efficiently?**
  - Wide networks may have simple landscape for GD/SGD algorithms …
- *Generalization:* **How can deep learning generalize well without overfitting the noise?**
  - Implicit regularization: SGD finds flat local maxima, Max-Margin classifier?
  - "Benign overfitting"?

Optimization vs. Generalization

- Consider the ***empirical risk*** *minimization* under i.i.d. samples

$$\hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i;\theta)) + \mathcal{R}(\theta)$$

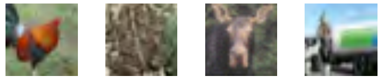
- The ***population risk*** with respect to unknown distribution

$$R(\theta) = \mathbf{E}_{x,y\sim P}\ell(y, f(x;\theta))$$

- Fundamental Theorem of Machine Learning (for 0-1 misclassification loss, called 'errors' below)
  - How to make training loss/error small? – Optimization issue
  - How to make generalization gap small? – Model Complexity issue

$$\underbrace{R(\theta)}_{\text{test/validation/generalization loss}} = \underbrace{\hat{R}_n(\theta)}_{\text{training loss}} + \underbrace{R(\theta) - \hat{R}_n(\theta)}_{\text{generalization gap}}$$

Why big models generalize well?



CIFAR10

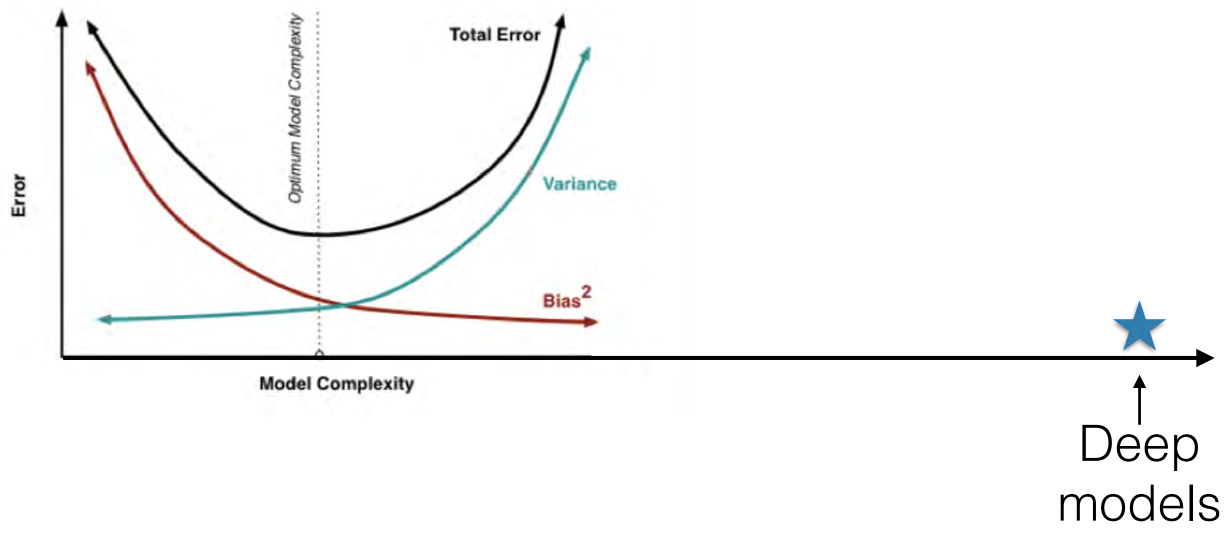
n=50,000
d=3,072
k=10

What happens when I turn off the regularizers?

Model	parameters	p/n	Train ***loss***	Test ***error***
CudaConvNet	145,578	2.9	0	23%
CudaConvNet (with regularization)	145,578	2.9	0.34	18%
MicroInception	1,649,402	33	0	14%
ResNet	2,401,440	48	0	13%

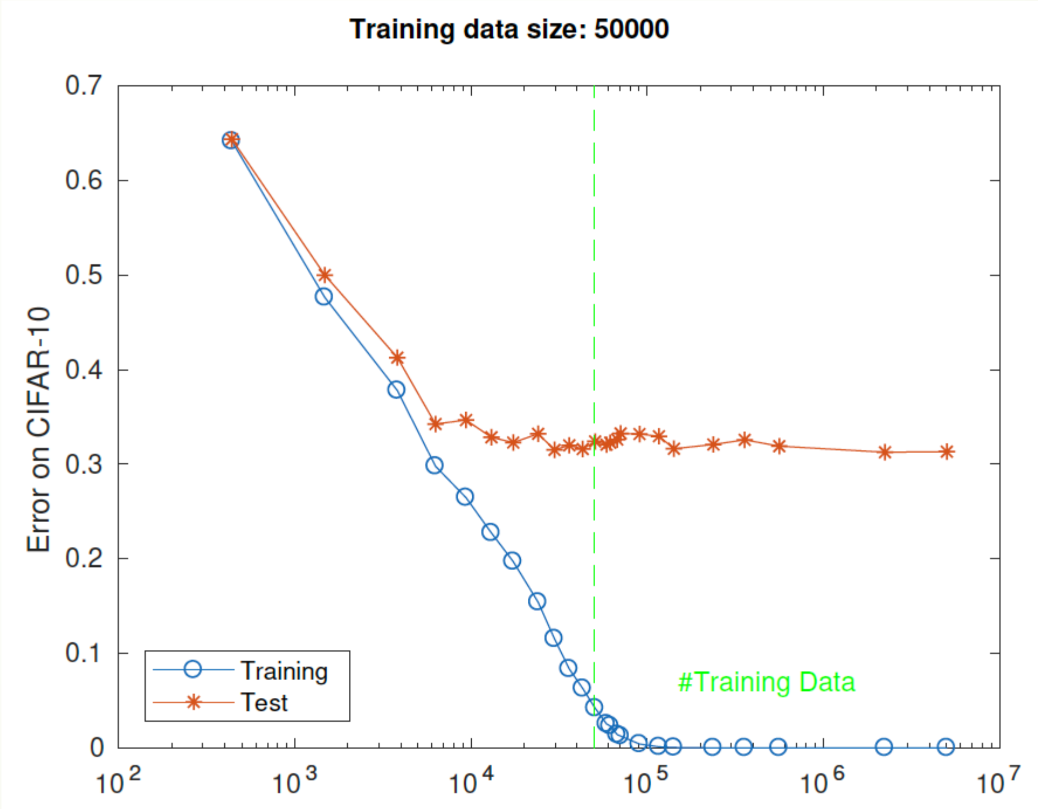
Ben Recht et al. 2016

The Bias-Variance Tradeoff?



Models where p>20n are common

Over-parameterized models



As model complexity grows (p>n), **training error goes down to zero**, but **test error does not increase**. *Why overparameterized models do not overfit here?* -- Tommy Poggio, 2018

- *Optimization:* how to achieve zero training loss/error in deep learning?
  - Overparameterized wide networks can do this via SGD
  - *Landscape of training loss* of such networks is simple! (**Joan Bruna**, **Rong Ge** et al.)
- *Generalization:* why overparameterized models do not overfit?
  - *Generalization gap* is determined by the Rademacher Complexity (Lipschitz) of networks, rather than number of parameters (**Peter Bartlett** et al.)
  - *Implicit regularization:* GD/SGD finds max margin classifiers (**Nati Srebro** et al.)
  - **Misha Belkin et al.**: Double descent for under-parameterized models vs. single descent for over-parameterized models

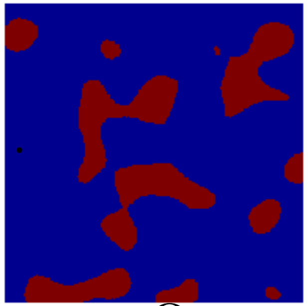
What’s the Landscape of Empirical Risks and How to optimize them efficiently?

Over-parameterized models lead to simple landscapes while SGD finds flat minina.

Sublevel sets and topology

- Given loss $E(\theta)$, $\theta \in \mathbb{R}^d$, we consider its representation in terms of level sets:

$$E(\theta) = \int_0^{\infty} \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{ y \in \mathbb{R}^d \ ; \ E(y) \leq u \} \ .$$



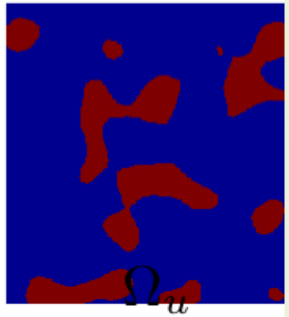
Ω_u

- A first notion we address is about the topology of the level sets .

- In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level $\boldsymbol{u}$?

Topology of Non-convex Risk Landscape

- A first notion we address is about the topology of the level sets .
  - In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u?

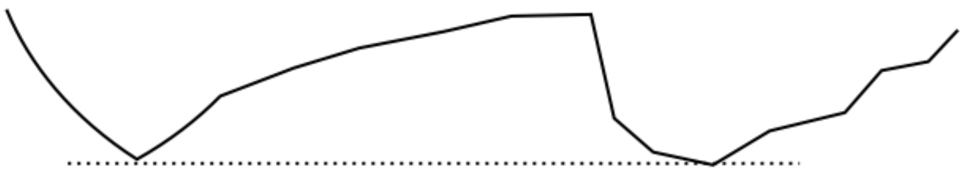


- This is directly related to the question of global minima:

Proposition: If $N_u = 1$ for all u then E has no poor local minima.

$$\text{(i.e. no local minima } y^* \text{ s.t. } E(y^*) > \min_y E(y))$$

- We say E is *simple* in that case.
- The converse is clearly not true.



Weaker: P.1, no spurious local valleys

Given a parameter space Θ and a loss function $L(\theta)$ as in (2), for all $c \in \mathbb{R}$ we define the sub-level set of L as

$$\Omega_L(c) = \{\theta \in \Theta : L(\theta) \leq c\}.$$

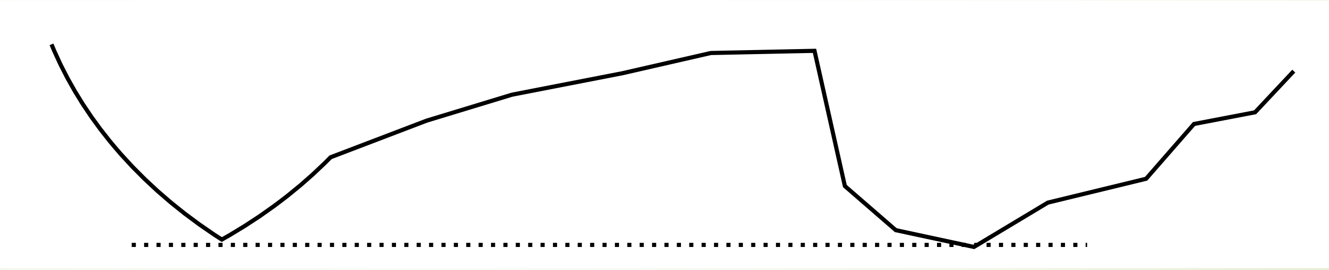
We consider two (related) properties of the optimization landscape. The first one is the following:

P.1 Given any *initial* parameter $\theta_0 \in \Theta$, there exists a continuous path $\theta : t \in [0, 1] \mapsto \theta(t) \in \Theta$ such that:

(a) $\theta(0) = \theta_0$

(b) $\theta(1) \in \arg\min_{\theta \in \Theta} L(\theta)$

(c) The function $t \in [0, 1] \mapsto L(\theta(t))$ is non-increasing.



The landscape has no spurious local valleys.

Overparameterized LN -> Single Basin

$$E(W_1, \ldots, W_K) = \mathbb{E}_{(X,Y)\sim P}\|W_K \ldots W_1 X - Y\|^2 \ .$$

Proposition: [BF'16]

1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.
2. (2-layer case, ridge regression) $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P}\|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$ satisfies $N_u = 1 \ \forall \ u$ if $n_1 > \min(n, m)$.

- We pay extra redundancy price to get simple topology.

Bruna, Freeman, 2016

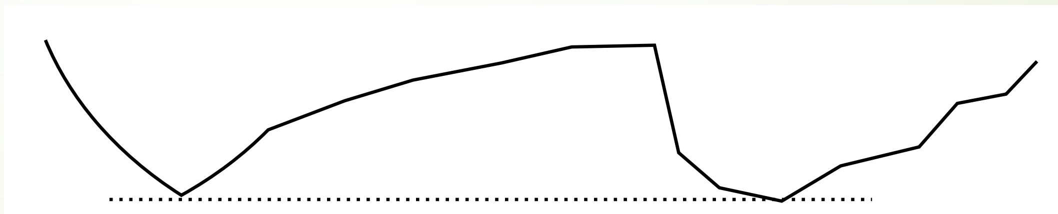
Venturi-Bandeira-Bruna'18

$$\Phi(x;\theta) = W_{K+1}\cdots W_1 x \ , \tag{13}$$

where $\theta = (W_{K+1}, W_K, \ldots, W_2, W_1) \in \mathbb{R}^{n\times p_{K+1}} \times \mathbb{R}^{p_{K+1}\times p_K} \times \cdots \mathbb{R}^{p_2\times p_1} \times \mathbb{R}^{p_1\times n}$.

Theorem 8 *For linear networks (13) of any depth $K \geq 1$ and of any layer widths $p_k \geq 1$, $k \in [1, K+1]$, and input-output dimensions n, m, the square loss function (2) admits no spurious valleys.*

Symmetry $f(W_i) = f(QW_i)$ $(Q \in GL(R^{n_l}))$ helps remove the network width constraint.



2-layer Neural Networks

[Venturi, Bandeira, Bruna, 2018]

Theorem 5 *The loss function*

$$L(\theta) = \mathbb{E}\|\Phi(X;\theta) - Y\|^2$$

of any network $\Phi(x;\theta) = U\rho W x$ *with effective intrinsic dimension* $q < \infty$ *admits no spurious valleys, in the over-parametrized regime* $p \geq q$*. Moreover, in the over-parametrized regime* $p \geq 2q$ *there is only one global valley.*

- Reproducing Kernel Hilbert Spaces (RKHS) are exploited in the proof!
- Matrix factorizations are of similar ideas.

Rong GE et al.

- For neural networks, not all local/global min are connected, even in the overparametrized setting.
- Solutions that satisfy **dropout stability** are connected.
- Possible to switch dropout stability with **noise stability** (used for proving generalization bounds for neural nets)

Thank you!