Towards Demystifying Overparameterization in Deep Learning

Mahdi Soltanolkotabi
Department of Electrical and Computer Engineering

USC University of Southern California

Mathematics of Imaging Workshop # 3
Henri Poincare Institute

Collaborators:

Samet Oymak and Mingchen Li

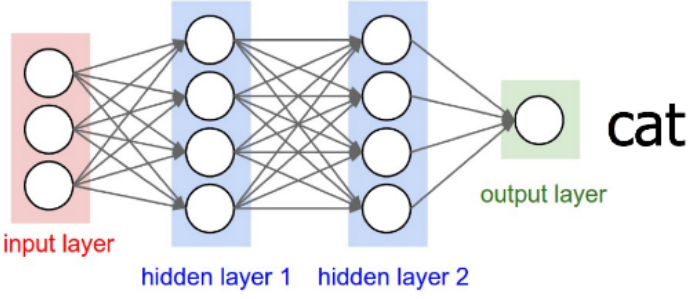
Motivation (Theory)

Many success stories

Neural networks very effective at learning from data

'It's able to create knowledge itself': Google unveils AI that learns on its own

In a major breakthrough for artificial intelligence, AlphaGo Zero took just three days to master the ancient Chinese board game of Go ... with no human help



HR Departments Turn to AI-Enabled Recruiting in Race for Talent

As the battle for talent becomes more competitive, companies are turning toward artificial intelligence to help with recruiting and other human resources tasks

A.I. Shows Promise Assisting Physicians

Doctors competed against A.I. computers to recognize illnesses on magnetic resonance images of a human brain during a competition in Beijing last year. The human doctors lost.

Mark Schiefelbein/Associated Press

Lots of hype

Andrew Ng ✓
@AndrewYNg

Following

Should radiologists be worried about their jobs? Breaking news: We can now diagnose pneumonia from chest X-rays better than radiologists.
stanfordmlgroup.github.io/projects/chexn…

Some failures

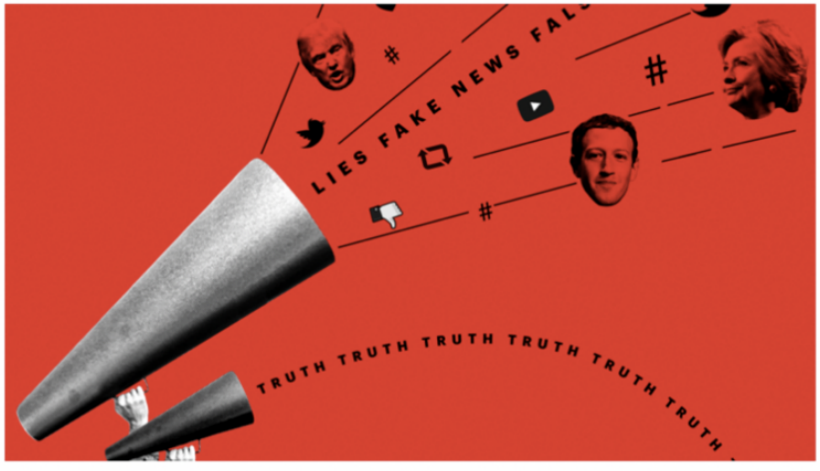
Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez @sarahintampa / 3 years ago Comment

The Grim Conclusions of the Largest-Ever Study of Fake News

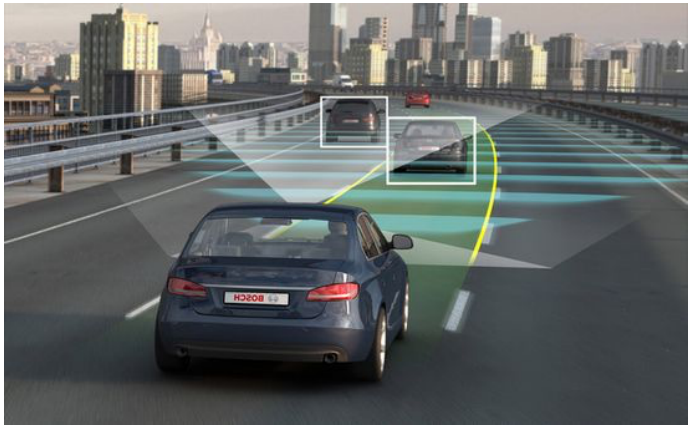
Falsehoods almost always beat out the truth on Twitter, penetrating further, faster, and deeper into the social network than accurate information.

ROBINSON MEYER MAR 8, 2018



Need more principled understanding

Deep learning-based AI increasingly used in human facing services



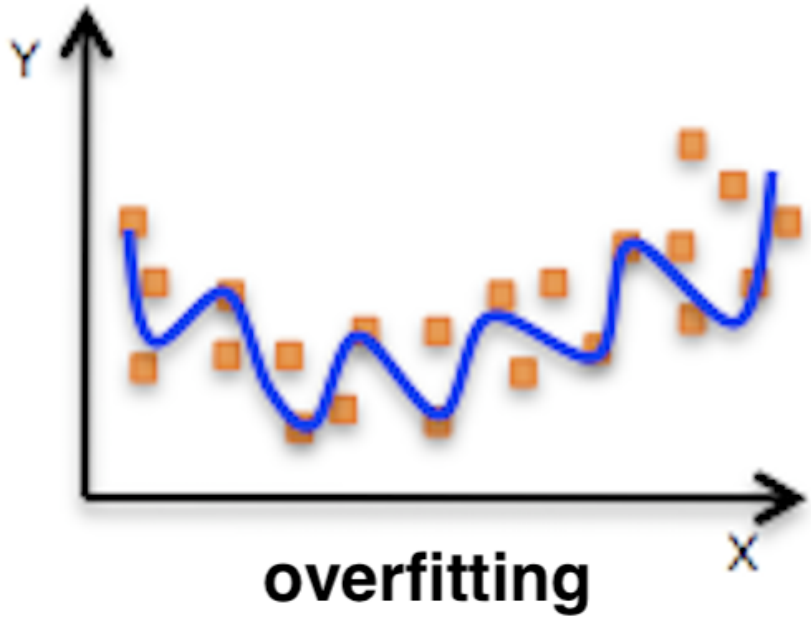
Challenges:

- *Optimization: Why can they fit?*
- *Generalization: Why can they predict?*
- *Achitecture: Which neural nets?*

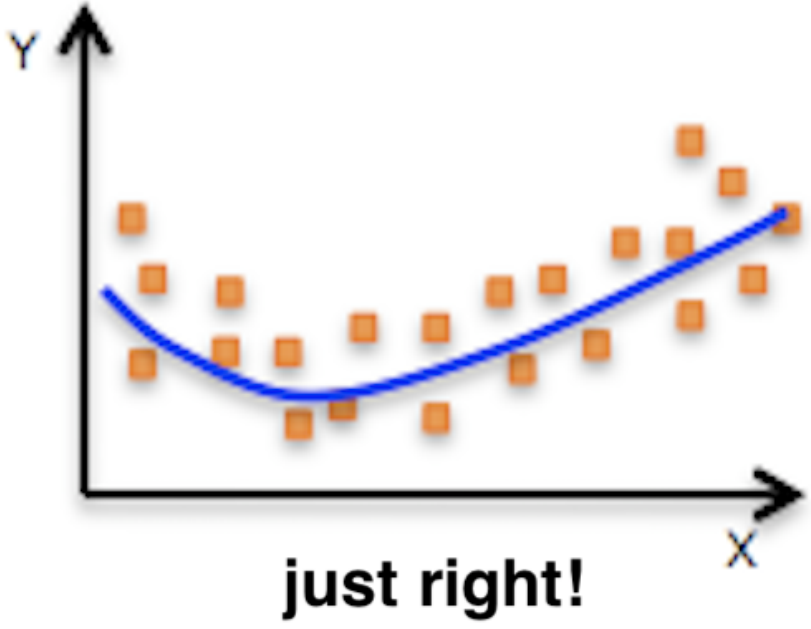
This talk: Overparameterization without overfitting

Mystery

# of parameters >> # training data

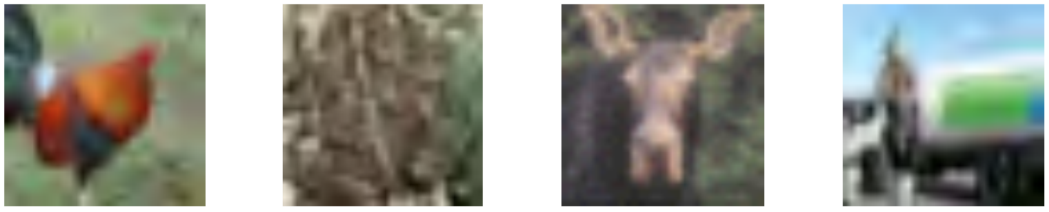


overfitting



just right!

Surprising experiment I (stolen from B. Recht)



CIFAR10

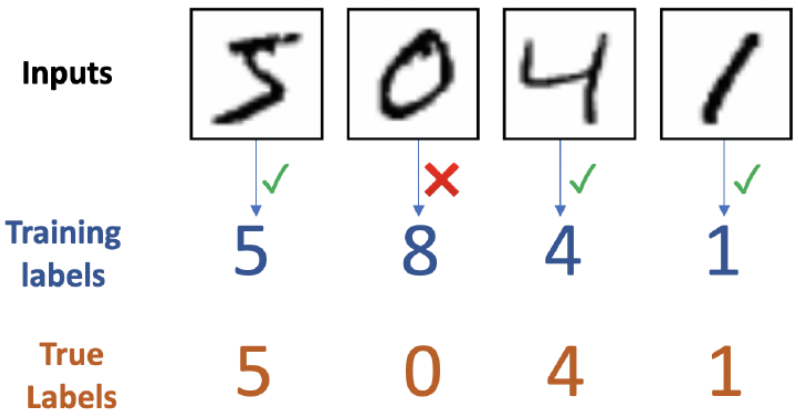
p parameters, $n = 50,000$ training samples, $d = 3072$ feature size, and 10 classes

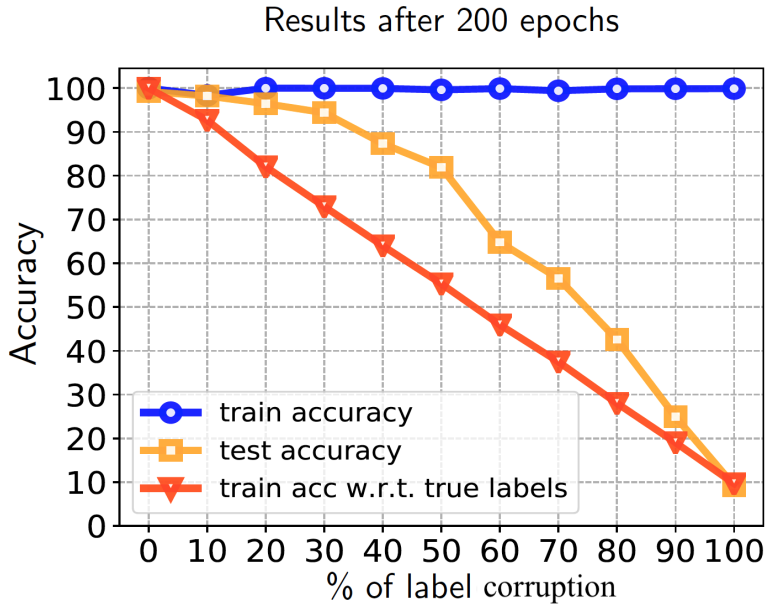
Model	parameters	p/n	Train ***loss***	Test ***error***
CudaConvNet	145,578	2.9	0	23%
CudaConvNet (with regularization)	145,578	2.9	0.34	18%
MicroInception	1,649,402	33	0	14%
ResNet	2,401,440	48	0	13%

Surprising experiment II-Overfitting to corruption

Add corruption

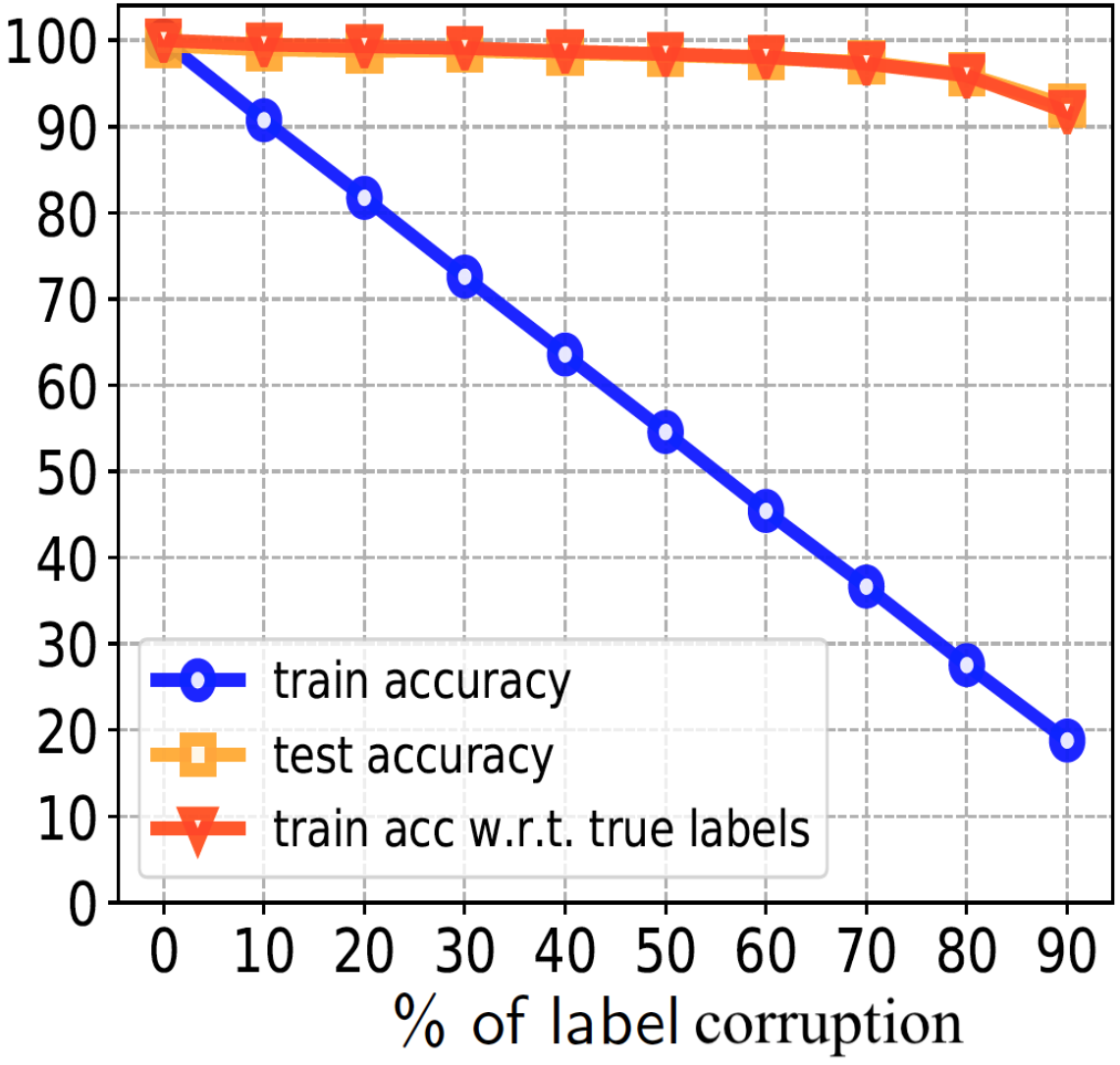
- Corrupt a fraction of training labels by replacing with another random label
- No corruption on test labels





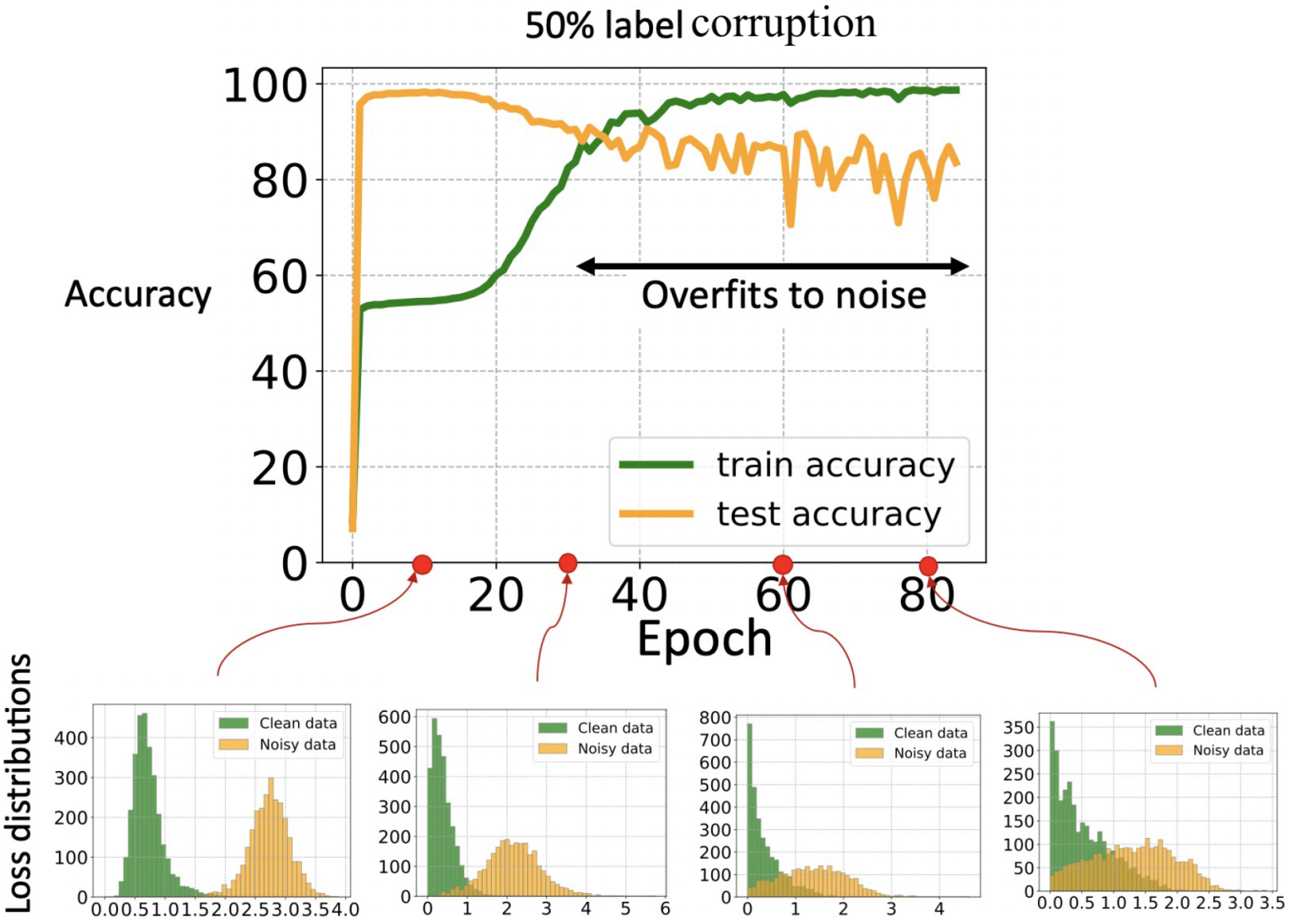
Surprising experiment III-Robustness

Repeat the same experiment but stop early



Surprising experiment III-Robustness

Repeat the same experiment but stop early



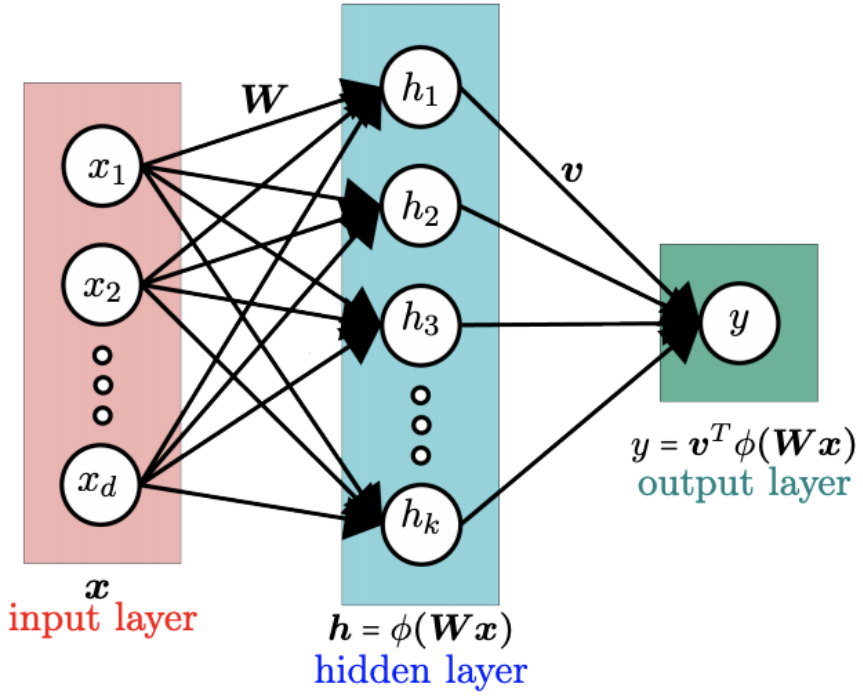
Benefits of overparameterization for neural networks

- *Benefit I: Tractable nonconvex optimization*

- *Benefit II: Robustness to corruption with early stopping*

Benefit I: Tractable nonconvex optimization

One-hidden layer

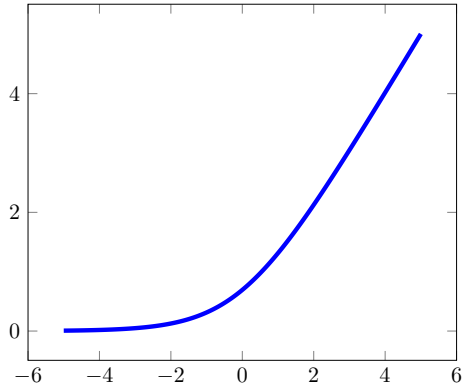


$$\boldsymbol{y}_i = \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i)$$

Theory for smooth activations

Data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\|\boldsymbol{x}_i\|_{\ell_2} = 1$

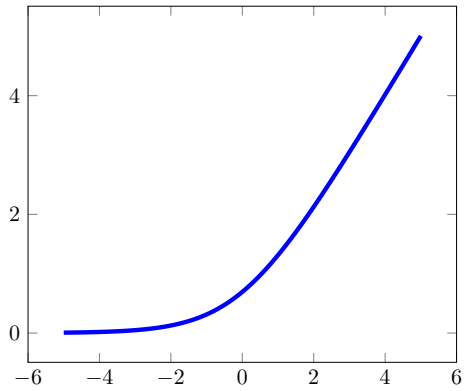
$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) := \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)^2$$



Theory for smooth activations

Data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\|\boldsymbol{x}_i\|_{\ell_2} = 1$

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) := \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)^2$$



- Set $\boldsymbol{v}$ at random or balanced (half $+$, half $-$)
- Run gradient descent $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \mu_\tau \nabla \mathcal{L}(\boldsymbol{W}_\tau)$ with random initialization

Theorem (Oymak and Soltanolkotabi 2019)

Assume

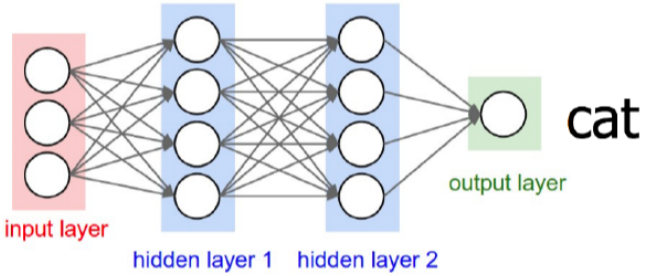
- *Smooth activations* $|\phi'(z)| \leq B$ *and* $|\phi''(z)| \leq B$
- *Overparameterization* $\sqrt{kd} \gtrsim \kappa(\boldsymbol{X})n$
- *Initialization with i.i.d.* $\mathcal{N}(0,1)$ *entries*

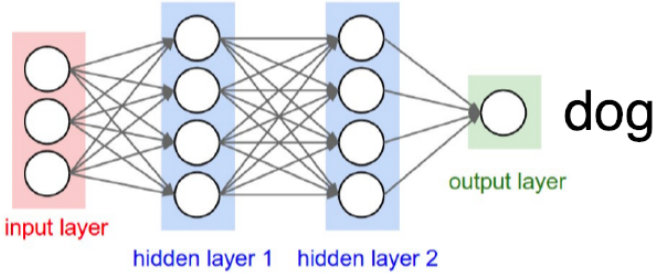
Then, with high probability

- *Zero training error:* $\mathcal{L}(\boldsymbol{W}_\tau) \leq \left(1 - c\frac{d}{n}\right)^{2\tau} \mathcal{L}(\boldsymbol{W}_0)$
- *Iterates remain close to initialization:* $\frac{\|\boldsymbol{W} - \boldsymbol{W}_0\|_F}{\|\boldsymbol{W}_0\|_F} \lesssim \frac{\sqrt{kd}}{\sqrt{n}}$

Dependence on data?

Diversity of input data is important...





$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

$$\kappa(\boldsymbol{X}) := \frac{\sqrt{\frac{d}{n}}\,\|\boldsymbol{X}\|}{\lambda(\boldsymbol{X})}$$

Definition (Neural network covariance matrix and eigenvalue)

- Neural net covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{X}) :=& \frac{1}{k}\,\mathbb{E}_{\boldsymbol{W}_0}\left[\mathcal{J}(\boldsymbol{W}_0)\mathcal{J}^T(\boldsymbol{W}_0)\right] \\ =& \mathbb{E}_{\boldsymbol{w}}\left[\left(\phi'(\boldsymbol{X}\boldsymbol{w})\phi'(\boldsymbol{X}\boldsymbol{w})^T\right) \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right)\right]. \end{aligned}$$

- Eigenvalue $\lambda(\boldsymbol{X}) := \lambda_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{X})\right)$

Hermite expansion

Lemma

Let $\{\mu_r(\phi')\}_{r=0}^{\infty}$ be the Hermite coefficients of ϕ'. Then,

$$\boldsymbol{\Sigma}(\boldsymbol{X}) = \sum_{r=0}^{+\infty} \mu_r^2(\phi') \underbrace{\left(\boldsymbol{X}\boldsymbol{X}^T\right) \odot \ldots \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right)}_{r+1} \succeq \underbrace{\mu_2^2(\phi')}_{(\mathbb{E}[\phi''(g)])^2} \left(\boldsymbol{X}\boldsymbol{X}^T\right) \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right)$$

arbitrary activation $\Leftrightarrow$ quadratic activation

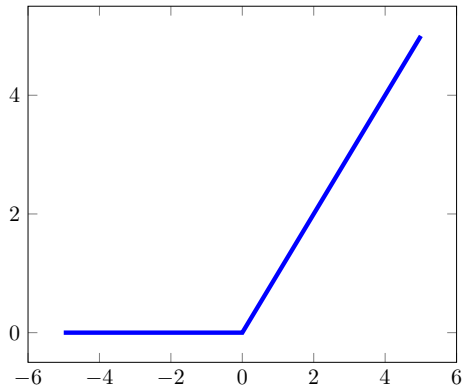
Conclusion

For generic data e.g. $\boldsymbol{x}_i$ i.i.d. uniform on the unit sphere $\kappa(\boldsymbol{X})$ scales like a constant

Theory for ReLU activations

Data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\|\boldsymbol{x}_i\|_{\ell_2} = 1$

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) := \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)^2$$



- Set $\boldsymbol{v}$ at random or balanced (half $+$, half $-$)
- Run gradient descent $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \mu_\tau \nabla \mathcal{L}(\boldsymbol{W}_\tau)$ with random initialization

Theorem (Oymak and Soltanolkotabi 2019)

Assume

- *ReLU activation* $\phi(z) = ReLU(z) = \max(0, z)$
- *Overparameterization* $\sqrt{kd} \gtrsim \kappa^3(\boldsymbol{X}) \frac{n}{d} \times n$
- *Initialization with i.i.d.* $\mathcal{N}(0,1)$ *entries*

Then, with high probability

- *Zero training error:* $\mathcal{L}(\boldsymbol{W}_\tau) \leq \left(1 - c\frac{d}{n}\right)^{2\tau} \mathcal{L}(\boldsymbol{W}_0)$
- *Iterates remain close to initialization:* $\frac{\|\boldsymbol{W} - \boldsymbol{W}_0\|_F}{\|\boldsymbol{W}_0\|_F} \lesssim \frac{\sqrt{kd}}{\sqrt{n}}$

Theory for SGD

Data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\|\boldsymbol{x}_i\|_{\ell_2} = 1$

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) := \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)^2$$

- Set $\boldsymbol{v}$ at random or balanced (half $+$, half $-$)
- Run gradient descent $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \mu_\tau \nabla \mathcal{L}(\boldsymbol{W}_\tau)$ with random initialization

Theorem (Oymak and Soltanolkotabi 2019)

Assume

- *Smooth activations* $|\phi'(z)| \leq B$ *and* $|\phi''(z)| \leq B$
- *Overparameterization* $\sqrt{kd} \gtrsim \kappa(\boldsymbol{X})n$
- *Initialization with i.i.d.* $\mathcal{N}(0,1)$ *entries*

Then, with high probability

- *Zero training error:* $\mathbb{E}[\mathcal{L}(\boldsymbol{W}_\tau)] \leq \left(1 - c\frac{d}{n^2}\right)^{2\tau} \mathcal{L}(\boldsymbol{W}_0)$
- *Iterates remain close to initialization:* $\frac{\|\boldsymbol{W} - \boldsymbol{W}_0\|_F}{\|\boldsymbol{W}_0\|_F} \lesssim \frac{\sqrt{kd}}{\sqrt{n}}$

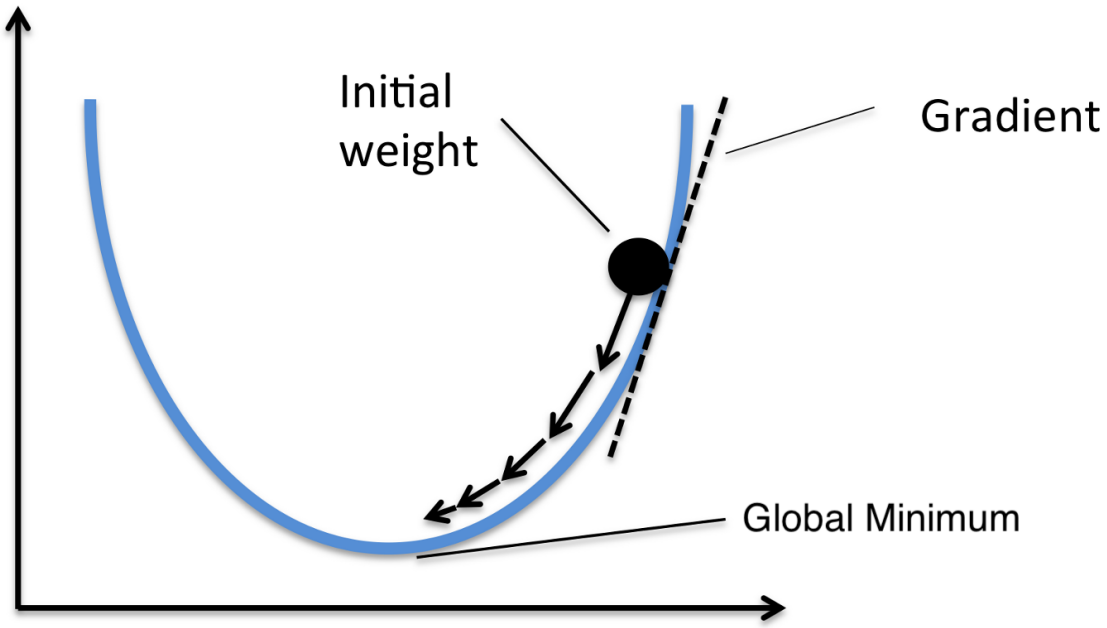
Proof Sketch

Prelude: over-parametrized linear least-squares

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_{\ell_2}^2 \quad \text{with} \quad \boldsymbol{X} \in \mathbb{R}^{n \times p} \quad \text{and} \quad n \leq p.$$

Gradient descent starting from $\boldsymbol{\theta}_0$ has three properties:

- Global convergence
- Converges to a global optimum which is closest to $\boldsymbol{\theta}_0$
- Total gradient path length is relatively short



Over-parametrized nonlinear least-squares

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\left\|f(\boldsymbol{\theta}) - \boldsymbol{y}\right\|_{\ell_2}^2,$$

where

$$\boldsymbol{y} := \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\boldsymbol{\theta}) := \begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ f(\boldsymbol{x}_2;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_n;\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \le p.$$

Gradient descent: start from some initial parameter $\boldsymbol{\theta}_0$ and run

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta_\tau \nabla\mathcal{L}(\boldsymbol{\theta}_\tau),$$

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})^T(f(\boldsymbol{\theta}) - \boldsymbol{y}).$$

Here, $\mathcal{J}(\boldsymbol{\theta}) \in \mathbb{R}^{n\times p}$ is the Jacobian matrix with entries $\mathcal{J}_{ij} = \frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$.

Key lemma

Lemma

Following assumptions on $\mathcal{B}(\boldsymbol{\theta}_0, R)$ *with* $R := \frac{4\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$

- *Jacobian at initialization:* $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \geq 2\alpha$
- *Bounded Jacobian spectrum:* $\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta$
- *Lipschitz Jacobian:* $\left\|\mathcal{J}(\widetilde{\theta}) - \mathcal{J}(\boldsymbol{\theta})\right\| \leq L\left\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_F$
- *Small initial residual:* $\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \leq \frac{\alpha^2}{4L}$

Then using step size $\eta \leq \frac{2}{\beta}$

- *Global geometric convergence:* $\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2}^2 \leq \left(1 - \frac{\eta\alpha^2}{2}\right)^\tau \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}^2$
- *iterates stay close to init.:* $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \leq \frac{4}{\alpha}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \leq 4\frac{\beta}{\alpha}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_{\ell_2}$
- *Total gradient path bounded:* $\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} \leq \frac{4}{\alpha}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}$

Key Ideal

Track dynamics of

$$\mathcal{V}_\tau := \|\boldsymbol{r}_\tau\|_{\ell_2} + \frac{1}{2}\left(1 - \eta\beta^2\right)\sum^{\tau-1}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_{\ell_2}.$$

Proof sketch (SGD)

Challenge: show that SGD remains in the local neighborhood

- Attempt I: Show $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2}$ is a super martingale (see also [Tan and Vershynin 2017])
- Attempt II: Show that $\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2} + \lambda \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2}$ is a super martingale
- Final attempt: Show that

$$\frac{1}{K}\sum_{j=1}^{K} \|\boldsymbol{\theta}_\tau - \boldsymbol{v}_i\|_{\ell_2} + \frac{3\eta}{n} \left\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\left(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\right)\right\|_{\ell_2}$$

is a super-martingale. Here, $\boldsymbol{v}_i$ is a very fine cover of $\mathcal{B}(\boldsymbol{\theta}_0, R)$

Over-parametrized nonlinear least-squares for neural nets

$$\min_{\boldsymbol{W}\in\mathbb{R}^{k\times d}} \mathcal{L}(\boldsymbol{W}) := \frac{1}{2}\left\|f(\boldsymbol{W})-\boldsymbol{y}\right\|_{\ell_2}^2,$$

where

$$\boldsymbol{y} := \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\boldsymbol{W}) := \begin{bmatrix} f(\boldsymbol{W},\boldsymbol{x}_1) \\ f(\boldsymbol{W},\boldsymbol{x}_2) \\ \vdots \\ f(\boldsymbol{W},\boldsymbol{x}_n) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \le kd.$$

Linearization via Jacobian

$$\mathcal{J}(\boldsymbol{W}) = \boldsymbol{X} * \left(\phi'\left(\boldsymbol{X}\boldsymbol{W}^T\right)\mathsf{diag}(\boldsymbol{v})\right)$$

Key Techniques

- Hadamard product

$$\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W}) = \left(\phi'(\boldsymbol{X}\boldsymbol{W}^T)\phi'(\boldsymbol{W}\boldsymbol{X}^T)\right) \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right)$$

Theorem (Schur 1913)

For two PSD $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$

$$\lambda_{\min}\left(\boldsymbol{A} \odot \boldsymbol{B}\right) \geq \left(\min_i \boldsymbol{B}_{ii}\right) \lambda_{\min}(\boldsymbol{A})$$

$$\lambda_{\max}\left(\boldsymbol{A} \odot \boldsymbol{B}\right) \leq \left(\max_i \boldsymbol{B}_{ii}\right) \lambda_{\max}(\boldsymbol{A})$$

- Random matrix theory

$$\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W}) = \sum_{\ell=1}^{k} \left(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)^T\right) \odot \left(\boldsymbol{X}\boldsymbol{X}^T\right)$$

Side corollary: Nonconvex matrix recovery

- Features: $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n \in \mathbb{R}^{d \times d}$.
- Labels: $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$
- Solve Nonconvex matrix factorization

$$\min_{\boldsymbol{U} \in \mathbb{R}^{d \times r}} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^T \rangle \right)^2$$

Theorem (Oymak and Soltanolkotabi 2018)

Assume

- *i.i.d. Gaussian* $\boldsymbol{A}_i$
- *any label* y_i
- *Initialization at well conditioned matrix* $\boldsymbol{U}_0$

Then, gradient descent iterations $\boldsymbol{U}_\tau$ *converge with a geometric rate to a close global optima as soon as* $n \leq dr$.

- Burer-Monteiro and many others $r \geq \sqrt{n}$
- For Gaussian $\boldsymbol{A}_i$ we allow $r \geq \frac{n}{d}$

when $n \approx dr_0$

- Burer-Monteiro: $r \gtrsim \sqrt{dr_0}$ Ours: $r \gtrsim r_0$

Previous work

- Unrealistic quadratic: [Soltanolkotabi, Javanmard, Lee 2018] and [Venturi, Bandeira, Bruna,...]
- Smooth activations: [Du, Lee, Li, Wang, Zhai 2018]

$$kd \gtrsim n^2 \quad \text{versus} \quad k \gtrsim n^4.$$

- ReLU activation: [Du et. al. 2018]

$$k \gtrsim \frac{n^4}{d^3} \quad \text{versus} \quad k \gtrsim n^6.$$

- Separation: [Li and Liang 2018] [Allen-Zhu, Li, Song 2018]

$$k \gtrsim \frac{n^{12}}{\delta^4} \quad \text{versus} \quad k \gtrsim n^{25}????.$$

- Begin to move beyond "lazy training" [Chizat & Bach, 2018];
- Faster convergence rate
- Deep: [Du, Lee, Li, Wang, Zhai 2018] and [Allen-Zhu, Li, Song 2018]
- Mean field analysis for infinitely wide: [Mei et al., 2018]; [Chizat & Bach, 2018]; [Sirignano & Spiliopoulos, 2018]; [Rotskoff & Vanden-Eijnden, 2018]; [Wei et al., 2018].

Related recent literature

- Approximation capability
  [Barron 1994], [Telgarsky 2016],[Bolcskei, Grohs, Kutyniok, and Petersen 2017]
- More over-parameterization ($n \leq ck$)
  [Poston, Lee, Choie, and Kwon 1991], [Haeffele and Vidal 2015], [Nguyen and Hein 2017]
- Under-parameterized with resampling
  [Oymak 2018], [Ge, Ma, Lee 2017], [Zhong, Song, Jain, Bartlett, and Dhillon 2017]
  [Brutzkus and Globerson 2017] and [Li and Yuan 2017]
- Other learning methods (Tensors, kernels, etc)
  [Janzamin, Sedghi, and Anandkumar 2015], [Goel and Klivans 2017]
- Generalization
  [Hardt, Benjamin Recht, Yoram Singer 2016],
  [Brutzkus, Globerson, Malach, and Shalev-Shwartz 2017],
  [Golowich, Rakhlin, Shamir 2017],
  [Dziugaite and Roy 2017], [Bartlett, Foster, Telgarsky 2017],
  [Neysahbur, Bhojanapalli, McAllester, Srebro 2017]
  [Arora, Ge, Neyshabur, and Zhang 2018]
  [Arora, Cohen, Hazan 2018], [Azzian, Hassibi 2018]
- Interface with statistical physics
  [Choromanskaya, Henaff, Mathieu, Arous, LeCun 2015],
  [Lee, Bahri, Novak, Schoenholz, Pennington, Sohl-dickstein 2018],
  [Novak, Bahri, Abolafia, Pennington, Sohl-Dickstein 2018],
- Many others...

The need for overparameterization beyond width

Simple exercise: initialize $\boldsymbol{W}$ at random and just fit output layer weights

$$\mathcal{L}(\boldsymbol{v}) := \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\phi\left(\boldsymbol{W}\boldsymbol{x}_i\right) - y_i\right)^2 = \frac{1}{2}\left\|\phi\left(\boldsymbol{X}\boldsymbol{W}^T\right)\boldsymbol{v} - \boldsymbol{y}\right\|_{\ell_2}^2,$$

Simple least-squares problem

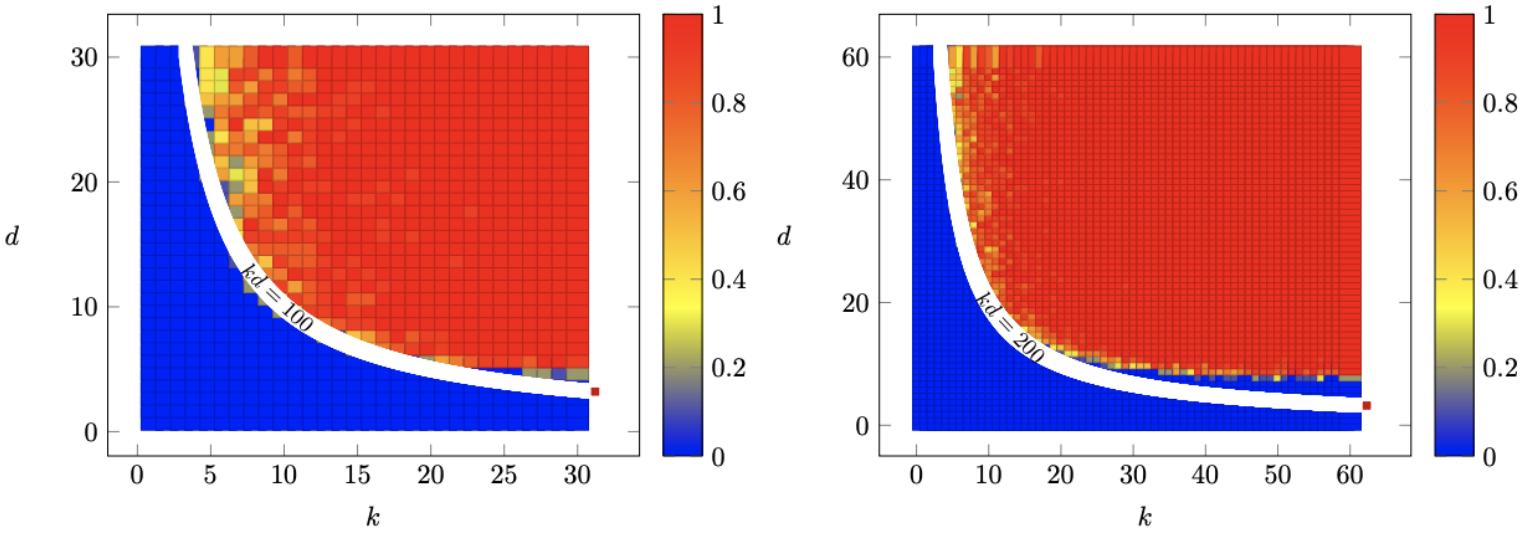
$$\hat{\boldsymbol{v}} := \boldsymbol{\Phi}^T\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)^{-1}\boldsymbol{y} \quad \text{where} \quad \boldsymbol{\Phi} := \phi\left(\boldsymbol{X}\boldsymbol{W}^T\right).$$

Theorem (Oymak and Soltanolkotabi 2019)

Fitting the output layer perfectly interpolates the data w.h.p. as soon as

$$k \gtrsim n$$

There is still a huge gap!



(a) softplus activation with $n = 100$

(b) softplus activation with $n = 200$

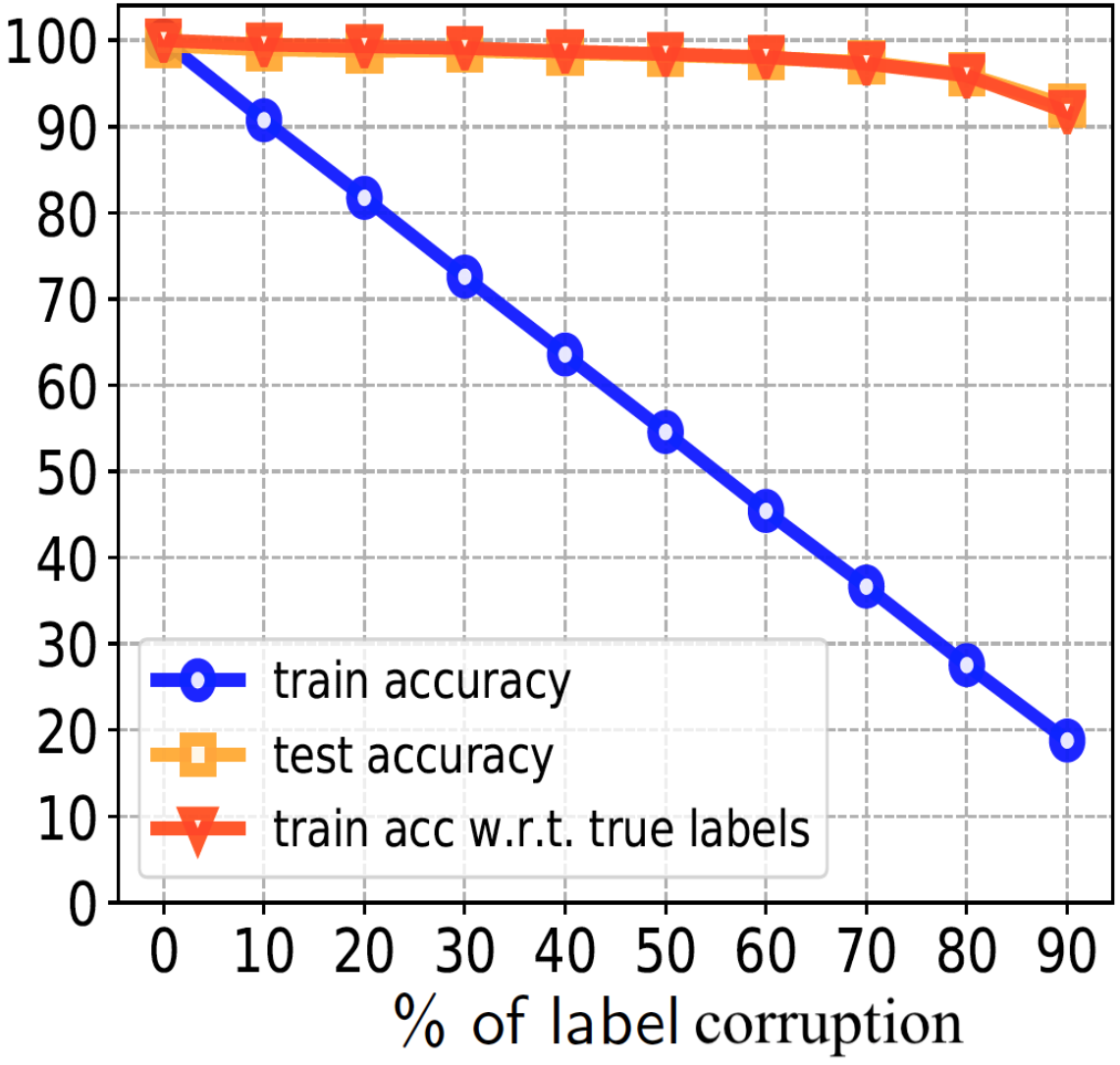
(c) ReLU activation with $n = 100$

(d) ReLU activation with $n = 200$

Benefit II: Robustness to corruption

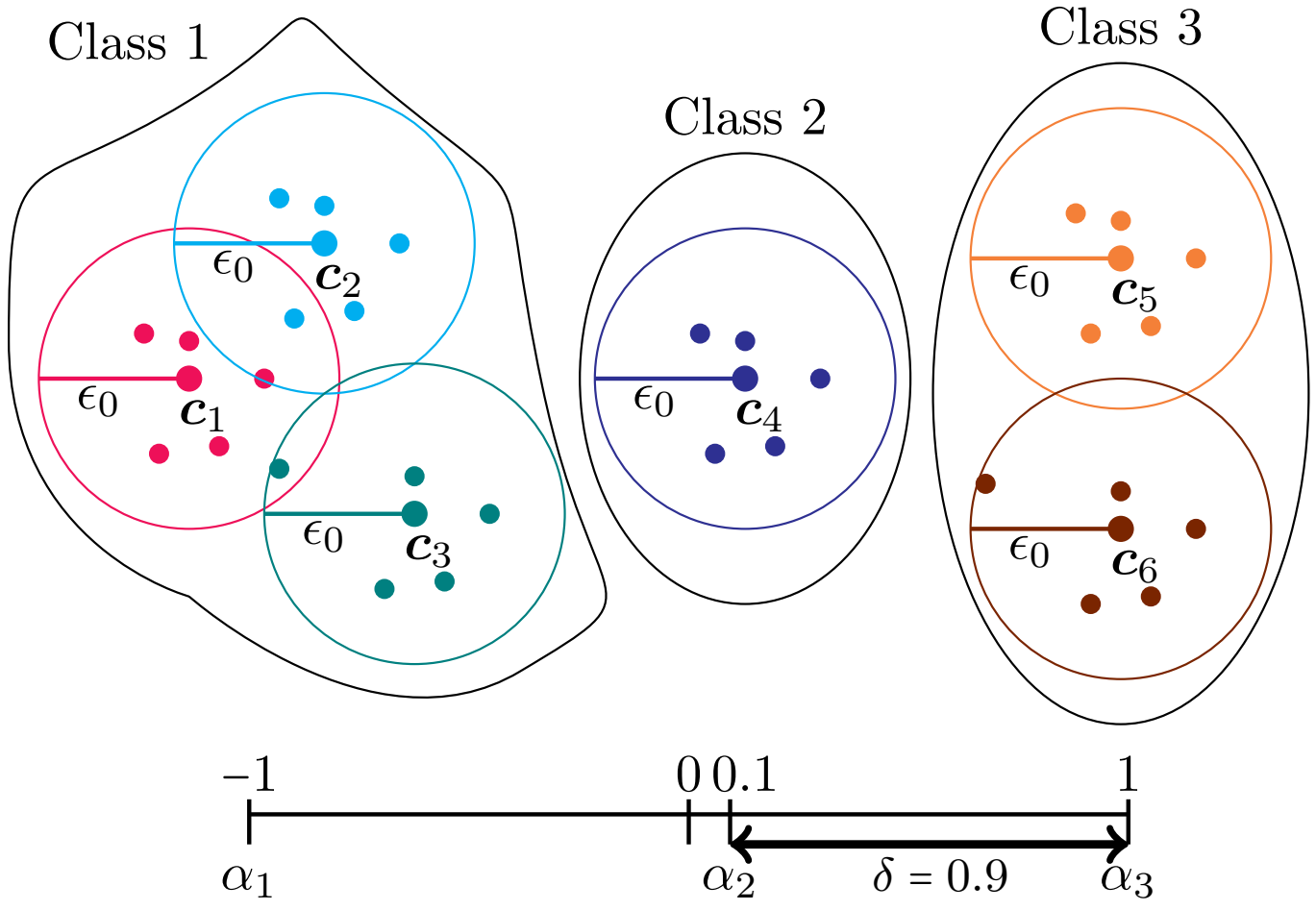
Surprising experiment III-Robustness

Repeat the same experiment but stop early



Model (without corruption)

clean data: (ϵ_0, δ)-clusterable data
input/label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times [-1, 1]$
L clusters and K classes



Robustness to corruption

Clean data points $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$, corrupt $s := \rho n$ to get corrupted data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.
Fit

$$\mathcal{L}(\boldsymbol{W}) := \frac{1}{2} \sum_{i=1}^{n} \left(f(\boldsymbol{W}, \boldsymbol{x}_i) - \boldsymbol{y}_i\right)^2$$

via gradient descent

Theorem (Oymak and Soltanolkotabi 2019)

Assume

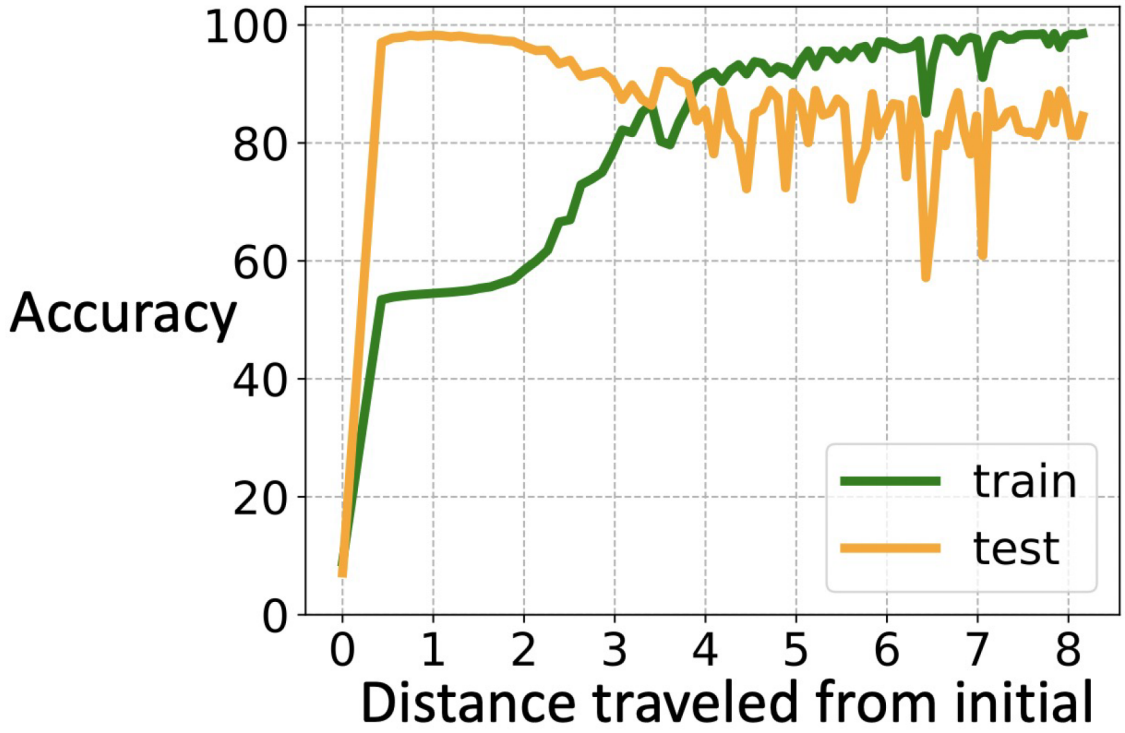
- *Corruption level* $\rho < \frac{1}{16}$
- *Cluster radius* $\epsilon \lesssim 1/L^2$
- *# Overparameterization* $k \times d \gtrsim \kappa^2(\boldsymbol{C}) L^4$

Starting from random initialization, after $\tau \sim L \log(1/\rho)$ *iterations, gradient descent finds a model with perfect accuracy, i.e.*

$$\textit{closest label to } f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) = \textit{true label } \bar{y}_i$$

Learning versus overfitting

Key insight: distance from initialization



Theorem (Oymak and Soltanolkotabi 2019)

- *With early stopping* ($\tau \sim L \log(\rho)$) *distance is bounded* $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \lesssim \sqrt{L}$
- *To overfit to the corruption you have to travel far* $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \propto \sqrt{s}$

Proof Sketch

High-level intuition

- Intuition I: Network should learn when there is no corruption
- Intuition II: Network should not fit to the corruption

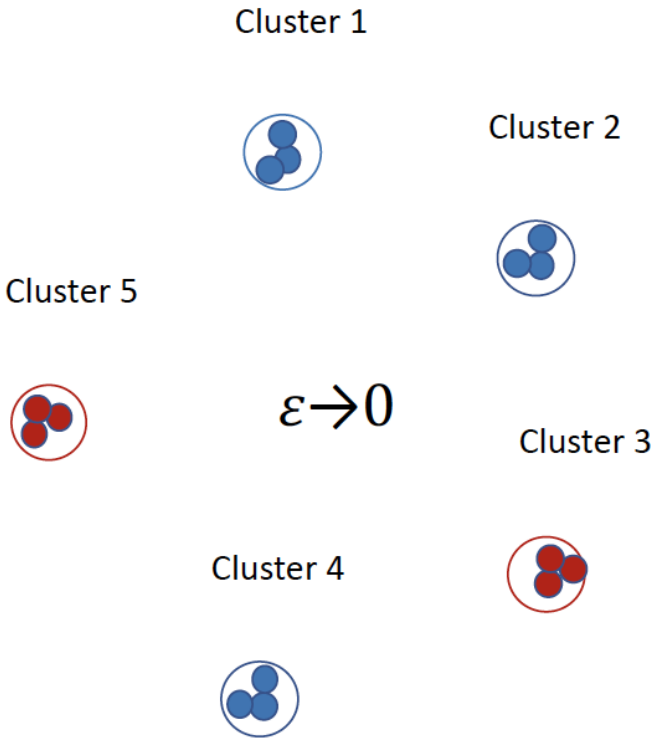
Key Idea I

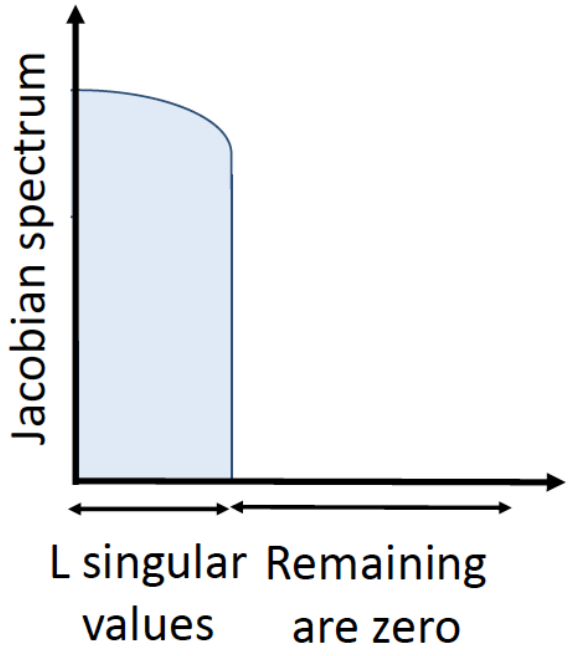
Reminder

- Gradient $\nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta}) \left(f(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{y} \right)$
- Jacobian $\mathcal{J}(\boldsymbol{\theta}) = \left[\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_1)}{\partial \boldsymbol{\theta}} \quad \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_2)}{\partial \boldsymbol{\theta}} \quad \dots \quad \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_n)}{\partial \boldsymbol{\theta}} \right] \in \mathbb{R}^{p \times n}$

Key Ideal I

If $\epsilon = 0$, there are only L distinct inputs. $\mathcal{J}$ has exactly rank L.





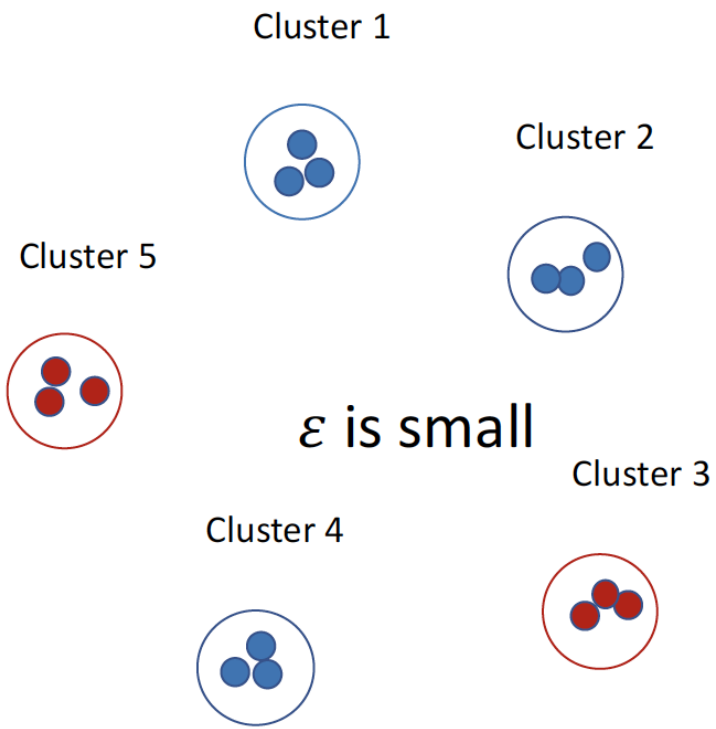
Key Idea I

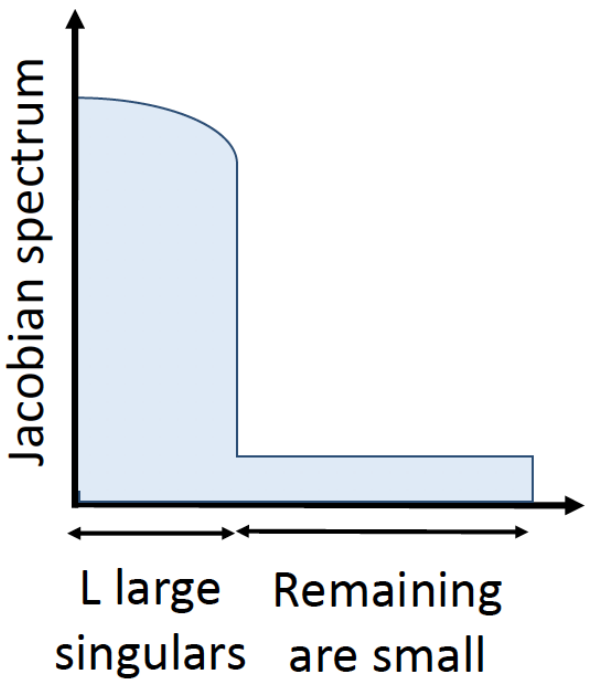
Reminder

- Gradient $\nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})\left(f(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{y}\right)$
- Jacobian $\mathcal{J}(\boldsymbol{\theta}) = \left[\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_1)}{\partial \boldsymbol{\theta}} \quad \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_2)}{\partial \boldsymbol{\theta}} \quad \cdots \quad \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_n)}{\partial \boldsymbol{\theta}}\right] \in \mathbb{R}^{p \times n}$

Key Ideal I

If ϵ is small, there are only L distinct inputs. $\mathcal{J}$ has approximately rank L.



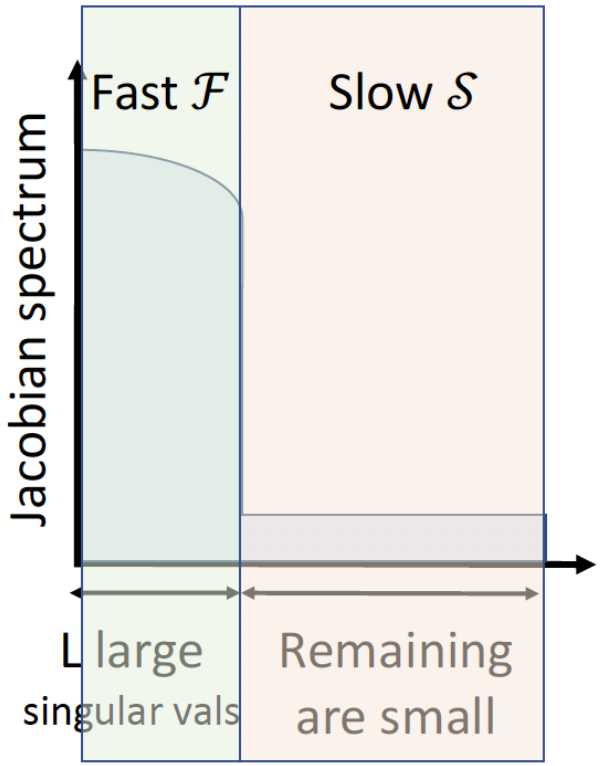


Key Idea II

Key Ideal II

Two complementary subspaces

- *Fast (data) subspace $\mathcal{F}$: Subspace associated with top L right singular vectors of $\mathcal{J}$*
- *slow (noise) subspace $\mathcal{S}$: Complement of $\mathcal{F}$*



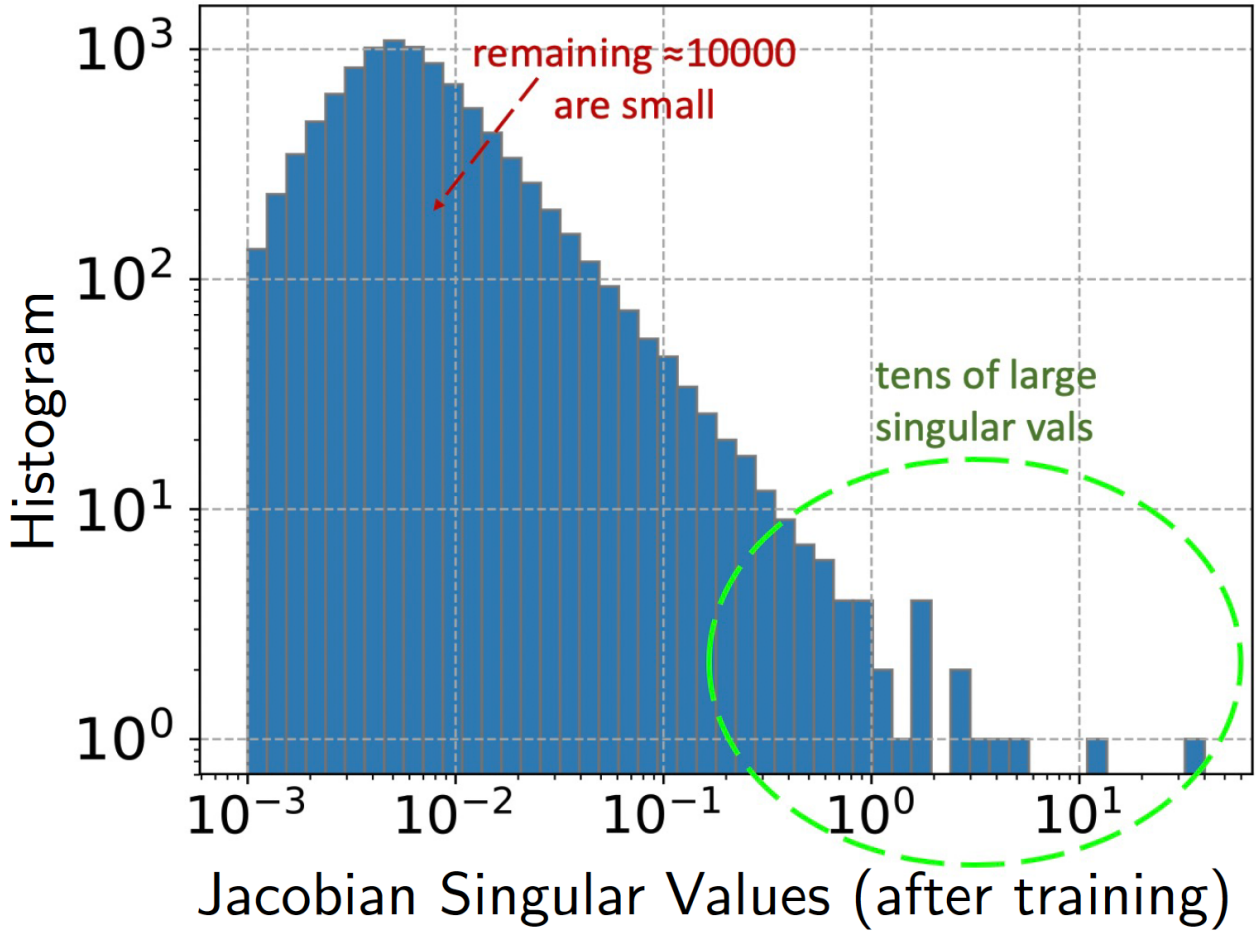
Interaction of Jacobian and residual in the gradient $\nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})\left(f(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{y}\right)$
Residual can be decomposed into two terms

$$r(\boldsymbol{\theta}) := f(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{y} = \underbrace{f(\boldsymbol{\theta}, \boldsymbol{X}) - \bar{\boldsymbol{y}}}_{\text{Residual w.r.t. true labels}} + \underbrace{\bar{\boldsymbol{y}} - \boldsymbol{y}}_{corruption}$$

- Residual w.r.t. true labels falls mostly onto $\mathcal{F}$ and quickly goes to zero
- Corruption $\boldsymbol{y} - \bar{\boldsymbol{y}}$ falls mostly onto $\mathcal{S}$ and goes very slowly to zero

What about real data?

- Dataset: CIFAR10
- Model: ResNET20
- Task: Binary classification (airplane vs truck)
- $n = 10,000$ and $p = 270,000$



Conclusion

Provable benefits of overparameterization

- More tractable optimization
- Robustness to corruption

Mandatory Postdoc Announcement

References

- Theoretical insights into the Optimization Landscape of Over-parameterized Shallow Neural Nets M. Soltanolkotabi, A. Javanmard, and J. D. Lee 2017.
- Over-parametrized nonlinear learning: Gradient descent follows the shortest path? S. Oymak and M. Soltanolkotabi
- Gradient Descent is Provably Robust to Label Noise for Overparameterized Neural Networks. S. Oymak and M. Soltanolkotabi
- Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. S. Oymak and M. Soltanolkotabi
- Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks. S. Oymak and M. Soltanolkotabi

Thanks!

Funding acknowledgment