What is Machine Learning

(and what we don't understand about it)

Andrea Montanari

Stanford University

July 21, 2020

Happy to be here

My Ph.D. was in Physics!

...ICTP contributed to broaden my interests

INTERNATIONAL SUMMER SCHOOL on

STATISTICAL PHYSICS AND PROBABILISTIC METHODS IN COMPUTER SCIENCE:
A Primer for Physicists, Mathematicians & Computer Scientists

23 August - 3 September 1999
Miramare, Trieste, Italy

The Abdus Salam International Centre for Theoretical Physics will organize a School on "Statistical Physics and Probabilistic Methods in Computer Science", from 23 August to 3 September 1999. The School will be followed by a Topical Conference on "NP-HARDNESS AND PHASE TRANSITIONS", from 6 to 10 September 1999 (please see separate announcement).

Members of the Steering Committee Board include Professors B.BOLLOBAS (Univ. of Memphis, U.S.A. and Cambridge Univ., U.K.); C. BORGS (Microsoft, Seattle); J. CHAYES (Microsoft, Seattle); S. KIRKPATRICK (IBM, NY); R. MONASSON (ENS, Paris); B. SELMAN (Cornell, NY); J. SPENCER (NY Univ) and R. ZECCHINA (The Abdus Salam ICTP, Trieste).

I. PURPOSE AND NATURE:

The aim of the School is to encourage young, qualified Mathematicians Computer Scientists and Theoretical Physicists to broaden their horizons, learn new subjects and apply the sophisticated tools developed in mathematics and theoretical physics to the field of computer science. The two week School will be devoted to introductory and tutorial lectures, where the multidisciplinary nature of the subjects and methods will be emphasized.

II. LIST OF TOPICS:

- Complexity Theory;
- Analysis of Algorithms;
- Statistical Mechanics Approach to Phase Transitions in Random Combinatorics;
- Heuristics Optimization and Simulation of Hard Physical/Combinatorial Models;
- Random Graphs and Hypergraphs;

Back to today's topic

What is Machine Learning?

Machine Learning (AI) in the news

What is machine learning?

Machine-learning algorithms find and apply patterns in data. And they pretty much run the world.

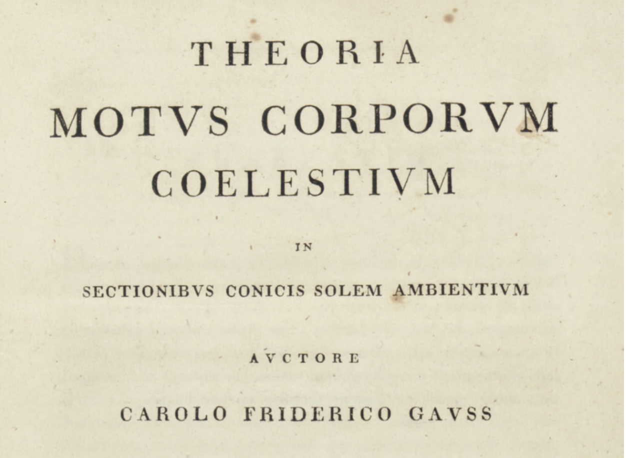
by **Karen Hao** November 17, 2018

Machine-learning algorithms are responsible for the vast majority of the artificial intelligence advancements and applications you hear about. (For

'. . . find . . . patterns in data. . . '

- Isn't this all of science?
- Haven't we been doing this for a while?

'. . . find . . . patterns in data. . . '



- Gauss, 1809: First use of least squares fitting

Outline

- **Three ways to think about 'patterns from data':**
  - Classical statistics
    [Fisher, Pearson, Wald,. . . 1920—. . .]
  - Statistical learning / Nonparametric estimation
    [Vapnik / Fix, Hodges, . . . 1970/1950—. . .]
  - Deep learning
    [In progress. . . 2010—. . .]
- **Some recent mathematical developments**

Canonical setting ('regression', 'supervised learning')

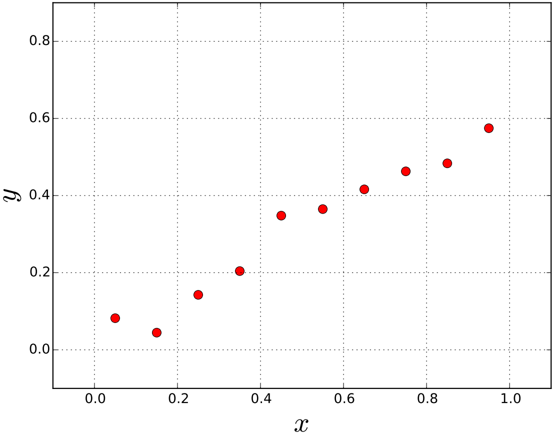
- **Data**

$$(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$$
$$y_i \in \mathbb{R} : \quad \text{'label', 'response',}$$
$$\boldsymbol{x}_i \in \mathbb{R}^d \quad \text{'features' vector', 'covariates'.}$$

- **Want** to predict new labels

$$f : \mathbb{R}^d \to \mathbb{R}$$

Example (low-dimensional)



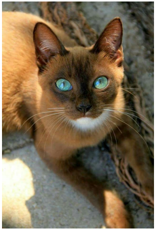
Example (high-dimensional)

$\Big(x_1 =$  $, y_1 = +1\Big)$

$\Big(x_2 =$  $, y_2 = -1\Big)$

$\Big(x_3 =$ 

$, y_3 = +1\Big)$

$\Big(x_4 =$  $, y_4 = -1\Big)$

Classical statistics

Mathematical model

- **Statistical model**

$$\{\mathrm{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$$

- **Data** $(y, X) := \{(y_i, x_i)\}_{i \le n}$

$$\{(y_i, x_i)\}_{i \le n} \sim_{iid} \mathrm{P}_{\boldsymbol{\theta}_0} \qquad \boldsymbol{\theta}_0 \in \Theta\,.$$

- **Estimator**

$$\hat{\theta} : (y, X) \to \hat{\theta}(y, X)$$

- **Predictive model**

$$f_{\boldsymbol{\theta}}(x) = \mathbb{E}_{\boldsymbol{\theta}}(y|x)\,.$$

Mathematical model

- **Statistical model**

$$\{\mathrm{P}_{\boldsymbol{\theta}} : \ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$$

- **Data** $(\boldsymbol{y}, \boldsymbol{X}) := \{(y_i, \boldsymbol{x}_i)\}_{i \le n}$

$$\{(y_i, \boldsymbol{x}_i)\}_{i \le n} \sim_{iid} \mathrm{P}_{\boldsymbol{\theta}_0} \qquad \boldsymbol{\theta}_0 \in \Theta\,.$$

- **Estimator**

$$\hat{\theta} : (y, X) \to \hat{\theta}(y, X)$$

- **Predictive model**

$$f_{\theta}(x) = \mathbb{E}_{\theta}(y|x)\,.$$

Mathematical model

- **Statistical model**

$$\{\mathrm{P}_{\boldsymbol{\theta}}:\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$$

- **Data** $(\boldsymbol{y}, \boldsymbol{X}) := \{(y_i, \boldsymbol{x}_i)\}_{i \le n}$

$$\{(y_i, \boldsymbol{x}_i)\}_{i \le n} \sim_{iid} \mathrm{P}_{\boldsymbol{\theta}_0} \qquad \boldsymbol{\theta}_0 \in \Theta\,.$$

- **Estimator**

$$\hat{\boldsymbol{\theta}} : (\boldsymbol{y}, \boldsymbol{X}) \to \hat{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{X})$$

- **Predictive model**

$$f_{\theta}(x) = \mathbb{E}_{\theta}(y|x)\,.$$

Mathematical model

- **Statistical model**

$$\{\mathrm{P}_{\boldsymbol{\theta}} : \; \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$$

- **Data** $(\boldsymbol{y}, \boldsymbol{X}) := \{(y_i, \boldsymbol{x}_i)\}_{i \le n}$

$$\{(y_i, \boldsymbol{x}_i)\}_{i \le n} \sim_{iid} \mathrm{P}_{\boldsymbol{\theta}_0} \qquad \boldsymbol{\theta}_0 \in \Theta\,.$$

- **Estimator**

$$\hat{\boldsymbol{\theta}} : (\boldsymbol{y}, \boldsymbol{X}) \to \hat{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{X})$$

- **Predictive model**

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}}(y|\boldsymbol{x})\,.$$

Example

Logistic model $y_i \in \{+1, -1\}, \boldsymbol{x}_i \in \mathbb{R}^d$

$$\mathrm{P}_{\boldsymbol{\theta}}(y = +1|\boldsymbol{x}) = \frac{e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}}{1 + e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}} \, ,$$

Estimation

Empirical risk (here $\boldsymbol{z}_i := (y_i, \boldsymbol{x}_i)$)

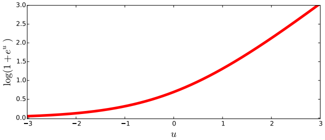
$$\text{minimize} \quad \widehat{R}_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}(\boldsymbol{\theta}; \boldsymbol{z}_i)$$

Loss: $\boldsymbol{\ell} : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$

Rationale (population risk)

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \mathbb{E}\{\boldsymbol{\ell}(\boldsymbol{\theta}; \boldsymbol{Z})\}$$

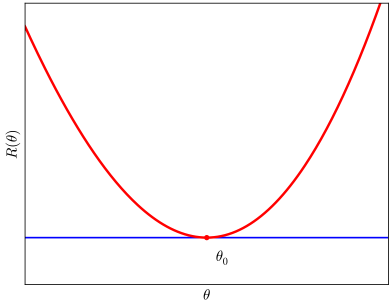
Example

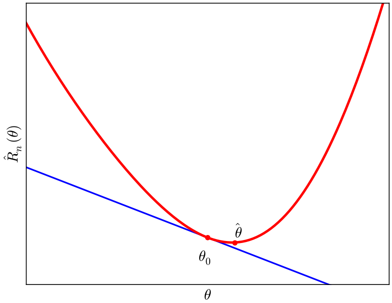


Logistic regression $\boldsymbol{z} = (y, \boldsymbol{x}) \in \{0, 1\} \times \mathbb{R}^p$

$$\ell(\boldsymbol{\theta}; y, \boldsymbol{x}) = -y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle + \log(1 + e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle})$$

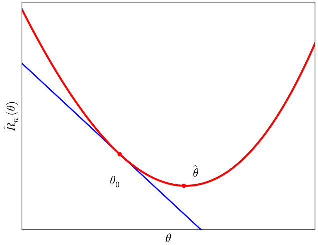
Why does this work?





$$\nabla R(\boldsymbol{\theta}_0) = \mathbf{0} \qquad \nabla \widehat{R}_n(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

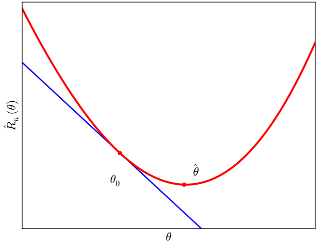
Why does this work?



$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \frac{1}{\boldsymbol{\kappa}} \|\nabla \widehat{R}_n(\boldsymbol{\theta}_0)\|_2 = \frac{1}{\boldsymbol{\kappa}} \Big\| \frac{1}{n} \sum_{i=1}^{n} \nabla \boldsymbol{\ell}(\boldsymbol{\theta}_0; \boldsymbol{z}_i) \Big\|_2 = O\left(\sqrt{\frac{p}{n}}\right)$$

Need $n \gg p$!

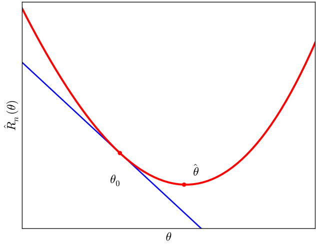
Why does this work?



$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \frac{1}{\boldsymbol{\kappa}}\|\nabla \widehat{R}_n(\boldsymbol{\theta}_0)\|_2 = \frac{1}{\boldsymbol{\kappa}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla \boldsymbol{\ell}(\boldsymbol{\theta}_0; z_i)\right\|_2 = O\left(\sqrt{\frac{p}{n}}\right)$$

Need $n \gg p$!

Why does this work?



$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \frac{1}{\boldsymbol{\kappa}} \|\nabla \widehat{R}_n(\boldsymbol{\theta}_0)\|_2 = \frac{1}{\boldsymbol{\kappa}} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \boldsymbol{\ell}(\boldsymbol{\theta}_0; z_i) \right\|_2 = O\left(\sqrt{\frac{p}{n}}\right)$$

Need $n \gg p$!

Statistical Learning

Is this realistic?

$\Big(x_1 =$  $, y_1 = +1\Big)$ $\Big(x_2 =$  $, y_2 = -1\Big)$

$$\{\mathrm{P}_{\boldsymbol{\theta}} : \ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$$

$$\{(y_i, \boldsymbol{x}_i)\}_{i \le n} \sim_{iid} \mathrm{P}_{\boldsymbol{\theta}_0} \qquad \boldsymbol{\theta}_0 \in \Theta\,.$$

Idea

$\Big(x_1 =$  $, y_1 = +1 \Big)$ $\Big(x_2 =$  $, y_2 = -1 \Big)$

Try to optimize over 'all' functions

Three pillars

1. Empirical Risk Minimization
2. Uniform convergence
3. Convex optimization

Textbook: Ben-David and Shalev-Shwartz, *Understanding Machine Learning*

Pillar #1: Empirical Risk Minimization

Objective:

$$\text{minimize} \quad R(f) := \mathbb{E}\{\mathsf{dist}(y_{\text{new}}, f(\boldsymbol{x}_{\text{new}}))\}, \quad (y_{\text{new}}, \boldsymbol{x}_{\text{new}}) \sim \mathbb{P}.$$

Problem: We do not know $\mathbb{P}$!

Idea:

$$\begin{aligned} \text{minimize} \quad & \widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{dist}(y_i, f(\boldsymbol{x}_i)), \\ \text{subj. to} \quad & f \in \mathcal{F} \end{aligned}$$

Pillar #1: Empirical Risk Minimization

Objective:

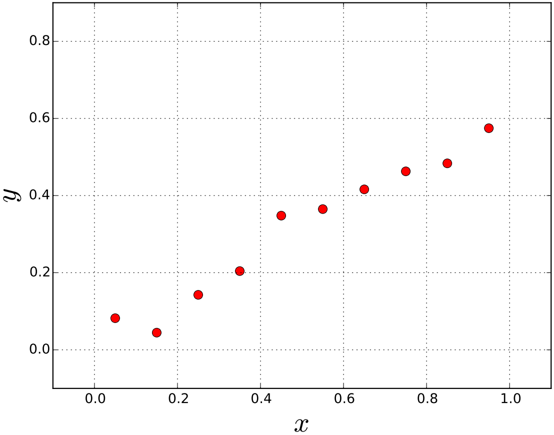
$$\text{minimize} \quad R(f) := \mathbb{E}\{\mathsf{dist}(y_{\rm new}, f(\boldsymbol{x}_{\rm new}))\}\,,, \qquad (y_{\rm new}, \boldsymbol{x}_{\rm new}) \sim \mathbb{P}\,.$$

Problem: We do not know $\mathbb{P}$!

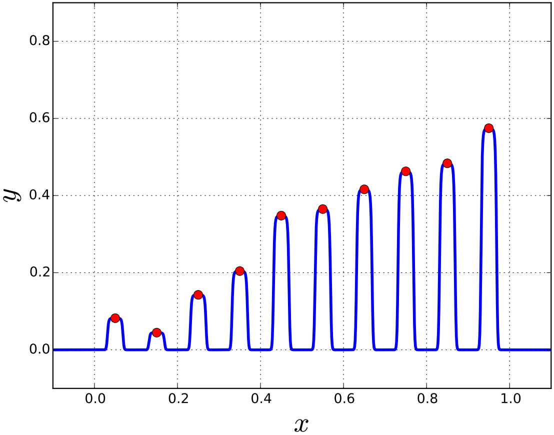
Idea

$$\begin{aligned} &\text{minimize} \quad \widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{dist}(y_i, f(\boldsymbol{x}_i))\,, \\ &\text{subj. to} \quad f \in \mathcal{F} \end{aligned}$$

Why constrain $f \in \mathcal{F}$? Baby example



Why constrain $f \in \mathcal{F}$? Baby example



Quiz

- Can you give an example of $\mathcal{F}$?
- Can you give an example of loss dist?

Pillar #2: Uniform convergence

Cartoon statement

$$\sup_{f \in \mathcal{F}} \left| \widehat{R}_n(f) - R(f) \right| \lesssim \varepsilon(n, \mathcal{F})$$

$$\varepsilon(n, \mathcal{F}) \ll 1 \quad \Leftrightarrow \quad n \gg \mathrm{Cplx}(\mathcal{F})$$

If the sample size is larger than the model complexity, then test error $\approx$ training error

Pillar #2: Uniform convergence

Cartoon statement

$$\sup_{f \in \mathcal{F}} \left| \widehat{R}_n(f) - R(f) \right| \lesssim \varepsilon(n, \mathcal{F})$$

$$\varepsilon(n, \mathcal{F}) \ll 1 \quad \Leftrightarrow \quad n \gg \mathrm{Cplx}(\mathcal{F})$$

If the sample size is larger than the model complexity, then test error $\approx$ training error

Pillar #2: Uniform convergence

Cartoon statement

$$\sup_{f \in \mathcal{F}} \left| \widehat{R}_n(f) - R(f) \right| \lesssim \varepsilon(n, \mathcal{F})$$

$$\varepsilon(n, \mathcal{F}) \ll 1 \quad \Leftrightarrow \quad n \gg \mathrm{Cplx}(\mathcal{F})$$

If the sample size is larger than the model complexity, then test error $\approx$ *training error*

Pillar #3: Convex optimization

$$\begin{aligned} \text{minimize} \quad & \widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{dist}(y_i, f(\boldsymbol{x}_i))\,, \\ \text{subj. to} \quad & f \in \mathcal{F} = \{f(\,\cdot\,;\boldsymbol{\theta}) : \ \ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\} \end{aligned}$$

Pillar #3: Convex optimization

$$\begin{aligned} \text{minimize} \quad & \widehat{R}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \mathsf{dist}(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta})), \\ \text{subj.to} \quad & \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p \end{aligned}$$

Pillar #3: Convex optimization

$$\begin{aligned} \text{minimize} \quad & \widehat{R}_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{dist}(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))\,, \\ \text{subj.to} \quad & \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p \end{aligned}$$

Idea: *Choose* Θ, dist, f *such that* $\theta \mapsto \mathsf{dist}(y_i, f(x_i; \theta))$ *is convex.*

Pillar #3: Convex optimization

$$\begin{aligned} \text{minimize} \quad & \widehat{R}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \mathsf{dist}(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta})), \\ \text{subj.to} \quad & \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p \end{aligned}$$

Idea: *Choose* Θ, dist, f *such that* $\boldsymbol{\theta} \mapsto \mathsf{dist}(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))$ *is convex.*

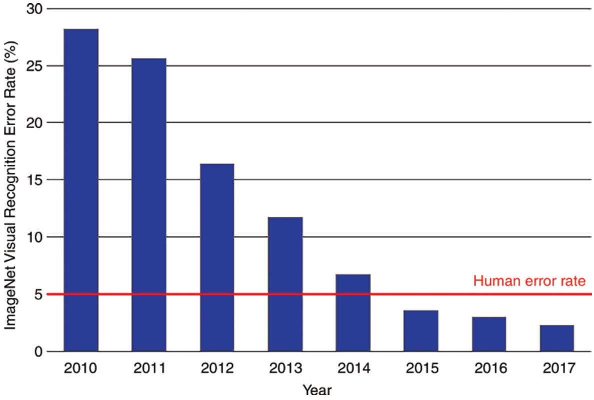
Deep Learning

Since ~2010, none of the three pillars seem to hold anymore[1].

[1] For many applications

ImageNet challenge

$$[n = 14 \cdot 10^6, y_i \in \{1, 2, \ldots, 2 \cdot 10^4\}]$$



↑

'Deep learning' revolution

Multi-layer (fully connected) neural network

$$\boldsymbol{\theta} = (\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_L)$$
$$\boldsymbol{\theta} \in \boldsymbol{\Theta} := \mathbb{R}^{N_1 \times N_0} \times \cdots \times \mathbb{R}^{N_L \times N_{L-1}}, \quad N_0 = d, N_L = 1,$$

$$f(\cdot; \boldsymbol{\theta}) := \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_{L-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1.$$

where

$$\boldsymbol{W}_\ell(x) := \boldsymbol{W}_\ell x,$$

$$\sigma(x) := (\sigma(x_1), \ldots, \sigma(x_N)),$$

Examples: $\sigma(x) = \tanh(x)$, $\sigma(x) = \max(x, 0)$,...

Multi-layer (fully connected) neural network

$$\boldsymbol{\theta} = (\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_L)$$
$$\boldsymbol{\theta} \in \boldsymbol{\Theta} := \mathbb{R}^{N_1 \times N_0} \times \cdots \times \mathbb{R}^{N_L \times N_{L-1}}, \quad N_0 = d, N_L = 1\,,$$

$$f(\,\cdot\,;\boldsymbol{\theta}) := \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_{L-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1\,.$$

where

$$\boldsymbol{W}_\ell(\boldsymbol{x}) := \boldsymbol{W}_\ell \boldsymbol{x}\,,$$

$$\boldsymbol{\sigma}(\boldsymbol{x}) := (\sigma(x_1), \ldots, \sigma(x_N))\,,$$

Examples: $\sigma(x) = \mathtt{tanh}(x)$, $\sigma(x) = \mathtt{max}(x, 0)$,...

Pillar #3: Convex optimization

$$f(\cdot\,;\boldsymbol{\theta}) := \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_{L-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1\,,$$

Example $\ell(y, f) = (y - f)^2$

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta})\big)^2\,.$$

Highly nonconvex!

Pillar #3: Convex optimization

$$f(\,\cdot\,;\boldsymbol{\theta}) := \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_{L-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1\,,$$

Example $\ell(y, f) = (y - f)^2$

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta})\big)^2\,.$$

Highly nonconvex!

Pillar #3: Convex optimization

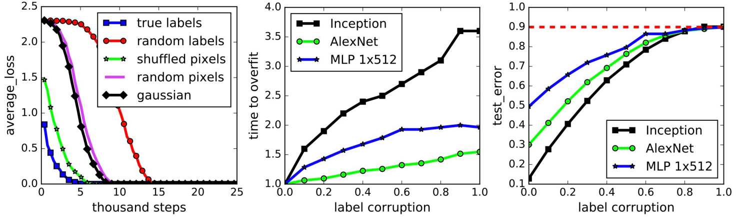
$$f(\,\cdot\,;\boldsymbol{\theta}) := \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_{L-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1\,,$$

Example $\ell(y, f) = (y - f)^2$

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta})\big)^2\,.$$

Highly nonconvex!

Pillar #2: Uniform convergence



(a) learning curves (b) convergence slowdown (c) generalization error growth

[Zhang, Bengio, Hardt, Recht, Vinyals, 2016]

Remarks

- $\mathcal{F}$ rich enough to 'interpolate' data points
- Test error $\gg$ Train error ≈ 0
- Outside uniform convergence regime

Pillar #1: Empirical Risk Minimization

$$\text{minimize} \qquad \widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

Gradient Descent (GD)

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \varepsilon_t \nabla_{\boldsymbol{\theta}} \widehat{R}_n(\boldsymbol{\theta}^{t})$$

Stochastic Gradient Descent (SGD)

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \varepsilon_t \nabla_{\boldsymbol{\theta}} \ell(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}^{t}))$$

$\widehat{R}_n(\boldsymbol{\theta})$ does not tell the full story!

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \varepsilon_t \nabla_{\boldsymbol{\theta}} \widehat{R}_n(\boldsymbol{\theta}^t)$$

- Nonconvex optimization
- Many global optima ($\widehat{R}_n(\boldsymbol{\theta}) \approx 0$)
- Output depends on
  - Initialization
  - Step-size schedule ε_t
  - Other algorithm details (batch size, dropout, ...)

$\widehat{R}_n(\boldsymbol{\theta})$ does not tell the full story!

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \boldsymbol{\varepsilon}_t \nabla_{\boldsymbol{\theta}} \widehat{R}_n(\boldsymbol{\theta}^t)$$

- Nonconvex optimization
- Many global optima ($\widehat{R}_n(\boldsymbol{\theta}) \approx 0$)
- Output depends on
  - Initialization
  - Step-size schedule ε_t
  - Other algorithm details (batch size, dropout, ...)

$\widehat{R}_n(\boldsymbol{\theta})$ does not tell the full story!

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \varepsilon_t \nabla_{\boldsymbol{\theta}} \widehat{R}_n(\boldsymbol{\theta}^t)$$

- Nonconvex optimization
- Many global optima ($\widehat{R}_n(\boldsymbol{\theta}) \approx 0$)
- Output depends on
  - Initialization
  - Step-size schedule ε_t
  - Other algorithm details (batch size, dropout, ...)

$\widehat{R}_n(\boldsymbol{\theta})$ does not tell the full story!

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \varepsilon_t \nabla_{\boldsymbol{\theta}} \widehat{R}_n(\boldsymbol{\theta}^t)$$

- Nonconvex optimization
- Many global optima ($\widehat{R}_n(\boldsymbol{\theta}) \approx 0$)
- Output depends on
  - Initialization
  - Step-size schedule ε_t
  - Other algorithm details (batch size, dropout, . . .)

Some recent mathematical developments

A simple example: Random features ridge regression

Ghorbani, Mei, Misiakiewicz, M, arXiv:1904.12191, 1906.08899

Mei, M, arXiv:1908.05355

Related work

- Belkin, Rakhlin, Tsybakov, 2018
- Liang, Rakhlin, 2018
- Hastie, Montanari, Rosset, Tibshirani, 2019
- Belkin, Hsu, Xu,2019
- Bartlett, Long, Lugosi, Tsigler, 2019
- Muthukumar, Vodrahalli, Sahai, 2019
- Many papers in 2020

This work: Exact asymptotics in a very simple neural net
(random feature models)

Related work

- Belkin, Rakhlin, Tsybakov, 2018
- Liang, Rakhlin, 2018
- Hastie, Montanari, Rosset, Tibshirani, 2019
- Belkin, Hsu, Xu,2019
- Bartlett, Long, Lugosi, Tsigler, 2019
- Muthukumar, Vodrahalli, Sahai, 2019
- Many papers in 2020

This work: Exact asymptotics in a very simple neural net
(random feature models)

Data

- $\{(y_i, \boldsymbol{x}_i)\}_{i \le n}$ iid
- $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d})), \quad d \gg 1$
- **Response**

$$y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i\,, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)\,.$$

Random features model

$$\mathcal{F}_{\mathsf{RF}}^{N}(\boldsymbol{W}) \equiv \Big\{ \hat{f}(\boldsymbol{x};\boldsymbol{a}) = \sum_{i=1}^{N} a_i\, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \; : \quad a_i \in \mathbb{R}\; \forall i \le N \Big\},$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N] \quad \boldsymbol{w}_i \sim_{iid} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$$

- Two-layers, fully connected
- Train only second layer
- Model is liear in the parameters!

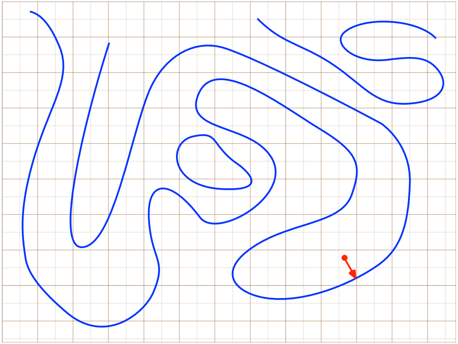
Neal, 1996; Balcan, Blum, Vempala 2006; Rahimi, Recht; 2008; Bach, 2016

Training

- **Ridge regression**

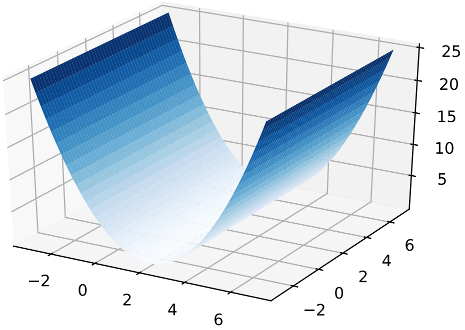
$$\widehat{\boldsymbol{a}}_{\mathrm{RR}}(\lambda) := \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i, \boldsymbol{a}))^2 + \frac{N\lambda}{d}\|\boldsymbol{a}\|_2^2 \right\} .$$

Why $\mathcal{F}_{\mathsf{RF}}$? Lazy regime



Jacot, Gabriel, Hongler, 2018; Du, Zhai, Poczos, Singh 2018; Allen-Zhu, Li, Song 2018; Chizat, Bach, 2019; Ghorbani, Mei, Misiakiewicz, M, 2019; Arora, Du, Hu, Li, Salakhutdinov, Wang, 2019; Oymak, Soltanolkotabi, 2019; . . .

Why ridge regression?



Results: Wide limit, polynomial asymptotics

Polynomial asymptotics

- $N = \infty; \quad n \asymp d^\alpha$
- Same paper: $n = \infty$, $N \asymp d^\alpha$

Ghorbani, Mei, Misiakiewicz,

Prediction error of Kernel Ridge Regression

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$. *Let* $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2),$$

$$\mathsf{P}_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg. } > \ell \textit{ polynomials}$$

Further, no kernel method can do better.

- Optimal error → interpolants ($\lambda = 0$)
- Staircase phenomenon.

Prediction error of Kernel Ridge Regression

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$. *Let* $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

$$\mathsf{P}_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg.} > \ell \textit{ polynomials}$$

Further, no kernel method can do better.

- Optimal error → interpolants ($\lambda = 0$)
- Staircase phenomenon.

Prediction error of Kernel Ridge Regression

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$*. Let* $\ell \in \mathbb{Z}$*, and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$*,* $\varepsilon > 0$*. Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$*,*

$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

$$\mathsf{P}_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg. } > \ell \textit{ polynomials}$$

Further, no kernel method can do better.

- Optimal error $\to$ interpolants ($\lambda = 0$)
- Staircase phenomenon.

Prediction error of Kernel Ridge Regression

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1 |x|)$. *Let* $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

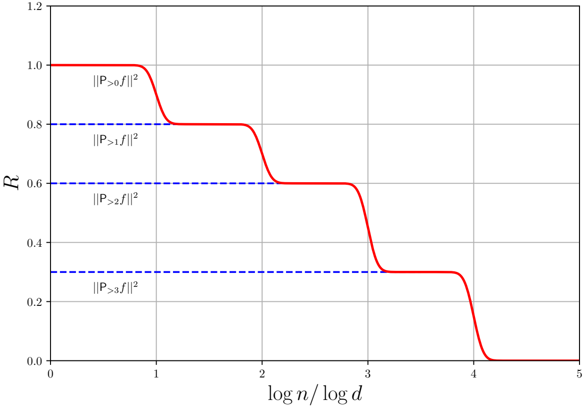
$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

$$\mathsf{P}_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg.} > \ell \textit{ polynomials}$$

Further, no kernel method can do better.

- Optimal error $\to$ interpolants ($\lambda = 0$)
- Staircase phenomenon.

Prediction error of Kernel Ridge Regression



Proportional asymptotics

Proportional asymptotics

- $n \asymp d$
- $N \asymp d$

Setting

- True function

$$f_*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle + f_*^{\mathrm{NL}}(\boldsymbol{x})$$

f_*^{NL} non-linear isotropic.

- $\|\boldsymbol{\beta}_0\|_2 = F_1$, $\|f_*^{\mathrm{NL}}\|_{L^2} = F_*$

- $n, N, d \to \infty$: $N/d \to \psi_1$, $n/d \to \psi_2$.

- $R(\hat{f}_\lambda) \equiv$ prediction error

Precise asymptotics

Theorem (Mei, M. 2019)

Decomposr $\sigma(x) = \sigma_0 + \sigma_1 x + \sigma^{\mathrm{NL}}(x)$ *where (for* $G \sim \mathsf{N}(0,1)$*)*

$$\mathbb{E}[G\sigma^{\mathrm{NL}}(G)] = \mathbb{E}[\sigma^{\mathrm{NL}}(G)] = 0, \qquad \zeta^2 := \frac{\sigma_1^2}{\mathbb{E}[\sigma^{\mathrm{NL}}(G)^2]}\,.$$

Then, for any $\overline{\lambda} = \lambda/\overline{b}_*^2 > 0$

$$R(\hat{f}_\lambda) = F_1^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + (\tau^2 + F_*^2)\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + F_*^2 + o_d(1)\,,$$

where $\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda})$, $\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda})$ *are explicitly given below.*

Variance computed in [Hastie, M, Rosset, Tibshirani, 2019]

Precise asymptotics

Theorem (Mei, M. 2019)

Decomposr $\sigma(x) = \sigma_0 + \sigma_1 x + \sigma^{\rm NL}(x)$ *where (for* $G \sim \mathsf{N}(0,1)$*)*

$$\mathbb{E}[G\sigma^{\rm NL}(G)] = \mathbb{E}[\sigma^{\rm NL}(G)] = 0, \qquad \zeta^2 := \frac{\sigma_1^2}{\mathbb{E}[\sigma^{\rm NL}(G)^2]}\,.$$

Then, for any $\overline{\lambda} = \lambda/\overline{b}_*^2 > 0$

$$R(\hat{f}_\lambda) = F_1^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + (\tau^2 + F_*^2)\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + F_*^2 + o_d(1)\,,$$

where $\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda})$, $\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda})$ *are explicitly given below.*

Variance computed in [Hastie, M, Rosset, Tibshirani, 2019]

Explicit formulae

Let $(\nu_1(\xi), \nu_2(\xi))$ be the unique solution of

$$\nu_1 = \psi_1 \Big(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2} \Big)^{-1} ,$$
$$\nu_2 = \psi_2 \Big(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2} \Big)^{-1} ;$$

Let

$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}) \cdot \nu_2(\boldsymbol{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}),$$

and

$$\begin{aligned}
\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv & -\chi^5 \zeta^6 + 3\chi^4 \zeta^4 + (\psi_1 \psi_2 - \psi_2 - \psi_1 + 1)\chi^3 \zeta^6 - 2\chi^3 \zeta^4 - 3\chi^3 \zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1 \psi_2 + 1)\chi^2 \zeta^4 + 2\chi^2 \zeta^2 + \chi^2 + 3\psi_1 \psi_2 \chi \zeta^2 - \psi_1 \psi_2 , \\
\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv & \; \psi_2 \chi^3 \zeta^4 - \psi_2 \chi^2 \zeta^2 + \psi_1 \psi_2 \chi \zeta^2 - \psi_1 \psi_2 , \\
\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv & \; \chi^5 \zeta^6 - 3\chi^4 \zeta^4 + (\psi_1 - 1)\chi^3 \zeta^6 + 2\chi^3 \zeta^4 + 3\chi^3 \zeta^2 + (-\psi_1 - 1)\chi^2 \zeta^4 - 2\chi^2 \zeta^2 - \chi^2 .
\end{aligned}$$

We then have

$$\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})} , \qquad \mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})} .$$

- *Kernel inner product random matrices*

Cheng, Singer, 2016; Do, Vu, 2017; Fan, M, 2017; Pennington Wohra, 2018;. . .

An interpretation

'Noisy linear features model'

Nonlinear features

$$\hat{f}(\boldsymbol{x}_i; \boldsymbol{a}) = \langle \boldsymbol{a}, \boldsymbol{x}_i \rangle \,,$$
$$u_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle) = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma^{\mathrm{NL}}(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle)$$

Noisy linear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \langle \boldsymbol{a}, \tilde{\boldsymbol{u}} \rangle \,,$$
$$\tilde{u}_{ij} = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma_* \, z_{ij} \,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$
$$\sigma_* := \|\sigma^{\mathrm{NL}}\|_{L^2}$$

Gaussian, correlated

‘Noisy linear features model’

Nonlinear features

$$\hat{f}(\boldsymbol{x}_i; \boldsymbol{a}) = \langle \boldsymbol{a}, \boldsymbol{x}_i \rangle \,,$$
$$u_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle) = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma^{\mathrm{NL}}(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle)$$

Noisy linear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \langle \boldsymbol{a}, \tilde{\boldsymbol{u}} \rangle \,,$$
$$\tilde{u}_{ij} = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma_* \, z_{ij} \,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0, 1)$$
$$\sigma_* := \|\sigma^{\mathrm{NL}}\|_{L^2}$$

Gaussian, correlated

Conceptual version of our theorem

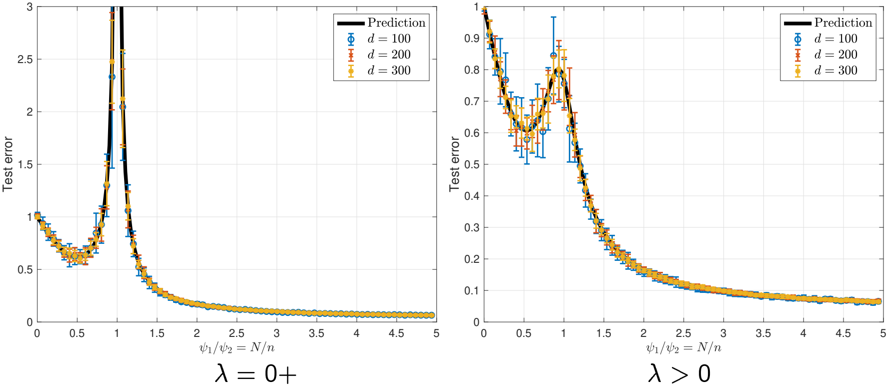
Theorem (Mei, M, 2019)

Consider random-features ridge regression in the proportional asymptotics

$$d \to \infty, \qquad N/d \to \psi_1, \qquad n/d \to \psi_2.$$

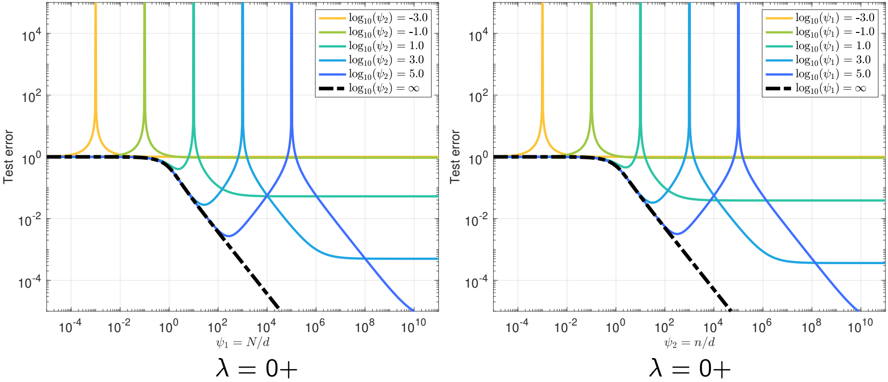
Then the nonlinear features model and noisy linear features model are 'asymptotically equivalent.'

Simulations vs theory

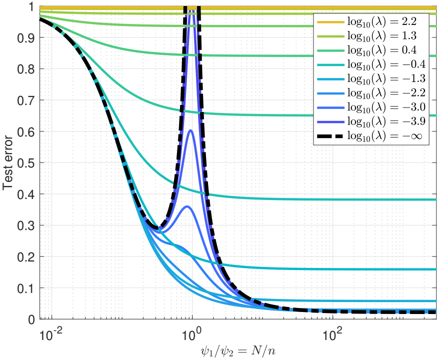


Insigths

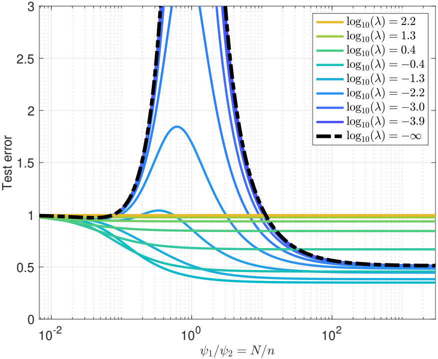
Insight #1: Optimum at $N/n \to \infty$



Insight #2: No double descent for optimal λ

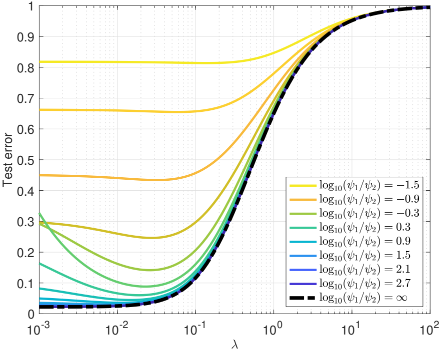


SNR$= 5$

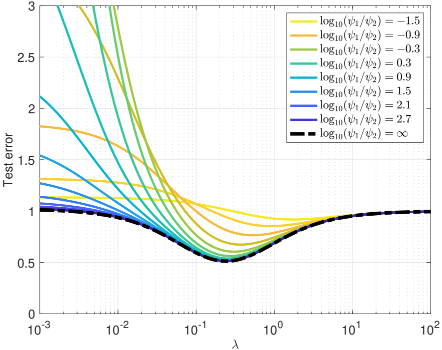


SNR$= 1/5$

Insight #3: $\lambda = 0+$ optimal at high SNR



SNR= 5



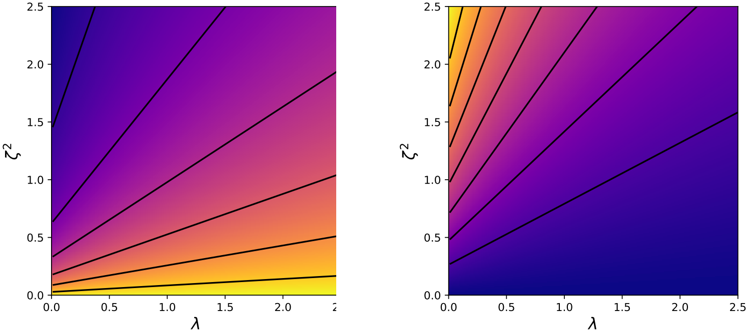
SNR= 1/10

- High SNR: Minimum at $\lambda = 0+$.
- Low SNR: Minimum at $\lambda > 0$.

Insight #4: Nonlinearity *is* regularization

- **Wide limit** $\psi_1 = N/d \to \infty$, $\psi_2 = n/d < \infty$

Insight #4: Nonlinearity *is* regularization



Bias Variance

Decreasing the $\zeta^2 := \frac{\mathbb{E}\{\sigma(G)F\}^2}{\mathbb{E}[\sigma^{\mathrm{NL}}(G)^2]} \Leftrightarrow$ Increasing λ

Extensions

- Anisotropic distributions

  [Ghorbani, Mei, Misiakiewicz, M, 2020]

- Binary classification

  [M, Ruan, Sohn, Yan, 2019]

- Neural tangent features

  [M, Zhong, 2020]

Conclusion

Conclusion

- Lots of new phenomena to be understood
- Mathematics/Theoretical Physics can play a useful role
- Machine Learning is a useful new tool
- We just have to understand for what :)

Thanks!

Conclusion

- Lots of new phenomena to be understood
- Mathematics/Theoretical Physics can play a useful role
- Machine Learning is a useful new tool
- We just have to understand for what :)

Thanks!

Conclusion

- Lots of new phenomena to be understood
- Mathematics/Theoretical Physics can play a useful role
- Machine Learning is a useful new tool
- We just have to understand for what :)

Thanks!