Introduction

Theoretical Deep Learning #2: generalization ability

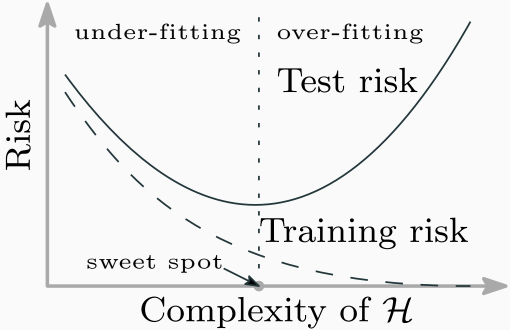
Eugene Golikov

MIPT, fall 2019

Neural Networks and Deep Learning Lab., MIPT

Complexity-risk curve

Classic "bias-variance trade-off" curve:

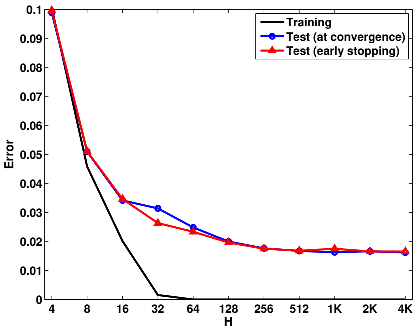


The figure is borrowed from Belkin et al. (2018)[1].

[1] `https://arxiv.org/abs/1812.11118`

Complexity-risk curve

The same curve for neural networks:

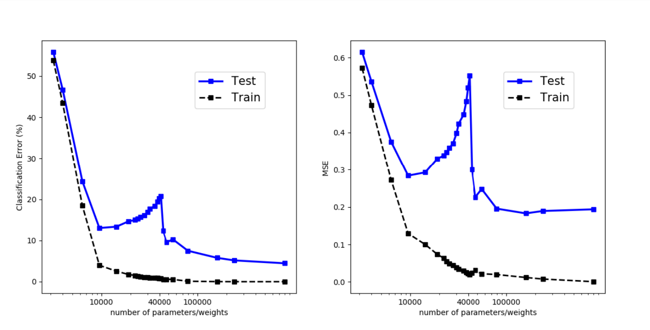


The figure is borrowed from Neyshabur et al. (2014)[2].

[2] https://arxiv.org/abs/1412.6614

Complexity-risk curve

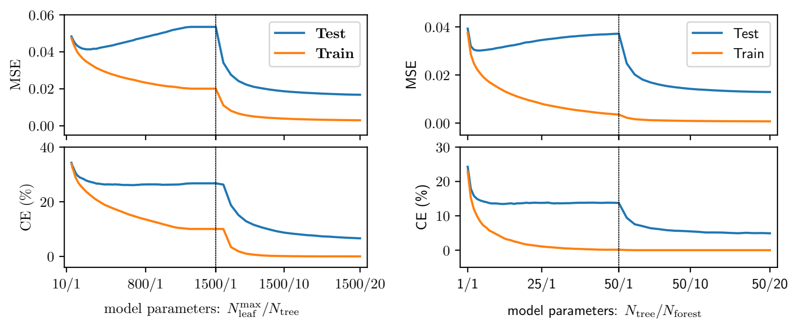
Where is the sweet spot?



The figure is borrowed from Belkin et al. (2018).

Complexity-risk curve

Similar curves for random forest and boosting:

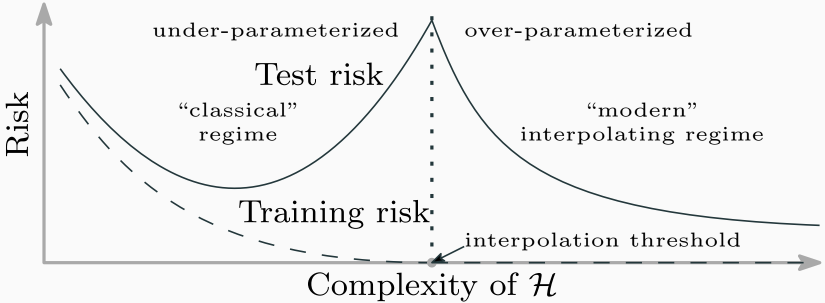


Left: random forest, right: boosting on decision trees.

Figures are borrowed from Belkin et al. (2018).

Complexity-risk curve

General "double descent" curve:



The figure is borrowed from Belkin et al. (2018).

Complexity-risk curve

Neural networks appear to be in interpolating regime; what phenomena do we observe there?

1. SGD achieves zero training risk (TDL #1);
2. Test risk decreases as complexity grows (this course).

TDL #1 recap

Assume fully-connected feed-forward network with single output neuron and l_2 loss.

Define:

- $n = \#$training samples;
- $H = \#$hidden layers;
- $m =$ width of the widest hidden layer;
- $d =$ total number of parameters;
- $\sigma(\cdot)$ – activation function.

TDL #1 recap

- **Observation:** SGD achieves zero training risk.
- **Natural hypothesis:** All local minima of training risk are global.
- **Results:**
  1. **Kawaguchi (2016)[3]:** True, if σ is identity map.
  2. **Yu & Chen (1995)[4]:** True, if $H = 1$, $m \geq n$ and if σ is real analytic.
  3. **Nguyen & Hein (2017)[5]:** (Almost) true for $H > 1$, if $m \geq n$, if net contracts after the widest layer and if σ is real analytic.
  4. **Nguyen (2019)[6]:** True for any $H \geq 1$, if $m \geq n$, if net contracts after the widest layer and if σ is leaky ReLU.

[3] http://www.mit.edu/~kawaguch/publications/kawaguchi-nips16.pdf

[4] https://ieeexplore.ieee.org/document/410380/

[5] https://arxiv.org/abs/1704.08045

[6] http://proceedings.mlr.press/v97/nguyen19a/nguyen19a.pdf

TDL #1 recap

- **Problem:** Globality of local minima is not sufficient for SGD to converge fast!
- **Hypothesis:** SGD converges to a global minimum in linear time whp over initialization.
- **Results for $H = 1$:**
  1. **Du et al. (2018)[7]:** True wp $\geq 1 - \delta$, if $m = \Omega(n^6/\delta^3)$ and σ is ReLU.
  2. **Song & Yang (2019)[8]:** True wp $\geq 1 - \delta$, if $m = \Omega(n^4 \operatorname{poly}\log(n/\delta))$ and σ is ReLU.
  3. **Kawaguchi & Huang (2019)[9]:** True wp $\geq 1 - \delta$, if $d = \Omega(n \log(n/\delta))$ and σ is real analytic.

[7] https://openreview.net/forum?id=S1eK3i09YQ
[8] https://arxiv.org/abs/1906.03593
[9] https://arxiv.org/abs/1908.02419

TDL #2 announcement

- **Observation:** Test risk of networks found by SGD decreases as width grows.
- **Hypothesis:** There is a network complexity measure with following properties:
  1. It correlates with test risk;
  2. It is implicitly minimized by SGD.
- **Possible candidates** are specific parameter norms.

TDL #2 announcement

Our goal is to bound the risk difference:

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \mathrm{bound}(N(\hat{f}_n), n, \delta) \quad \text{w.p. } \geq 1 - \delta \text{ over dataset } S_n,$$

where

- $R(f)$ — risk of predictor f,
- $\hat{f}_n$ — solution found by SGD,
- $N(f)$ is the complexity measure of predictor f.

Usual form of bound:

$$\mathrm{bound}(N, n, \delta) = \tilde{O}\left(\sqrt{\frac{N + \log(1/\delta)}{n}}\right).$$

TDL #2 announcement

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \mathrm{bound}(N(\hat{f}_n), n, \delta) \quad \text{w.p. } \geq 1 - \delta \text{ over dataset } S_n.$$

Worst-case bounds:

$$\mathrm{bound} = \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)).$$

Lead to complexity measures that depend on $\mathcal{F}$ — **dimension** of space of predictors.

TDL #2 announcement

$$R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \mathrm{bound}(N(\hat{f}_n), n, \delta) \quad \text{w.p. } \geq 1 - \delta \text{ over dataset } S_n.$$

PAC-Bayes bounds:

$$N(\hat{f}_n) = KL(Q(S_n) \| P), \quad \hat{f}_n \sim Q(S_n).$$

Here $Q(S_n)$ — "posterior" over predictors, P — "prior" over predictors.

TDL #2 announcement

Hypothesis:

SGD prefers predictors that minimize some complexity measure $N(f)$:

$$\hat{f}_n = \mathrm{SGD}(S_n) \in \operatorname*{Arg\,min}_{f:\ \hat{R}_n(f)=0} N(f).$$

Results:

- Linear regression: for zero init GD chooses minimum l_2-norm solution.
- Neural network: depends on magnitude of init (Woodworth et al., 2019)[10].

[10] `https://arxiv.org/abs/1906.05827`

Organization

Lectures:

1 per week, ~ 8 lectures total.

Lab(s) ($\sim 40\%$ of final grade):

We use pytorch; GPU is desirable.

Theoretical assignments:

Possibly, no theoretical assignments this time.

Oral exam ($\sim 60\%$ of final grade):

In the form of interview.

Main resource: `https://github.com/deepmipt/tdl2`

Link to **telegram chat** and **homework submission rules** are on github page.