

Why it's not overfitting

Thursday, March 26, 2020 2:54 PM

3/25

Why it's not overfitting? -- From Classical Statistics to Modern ML: the Lessons of Deep Learning -
Mikhail Belkin, From Classical Statistics to Modern ML: the Lessons of Deep Learning, From Classical
Statistics to Modern ML: the Lessons of Deep Learning, <https://www.youtube.com/watch?v=5-Kqb80h9rk&feature=youtu.be>

Mikhail (Misha) Belkin, https://scholar.google.com/citations?hl=en&user=lwd9DdkAAAAJ&view_op=list_works&sortby=pubdate

- VC dimension: A measure of the capacity of a space of function that can be learned by a statistical classification algorithm. It's defined as the cardinality of the largest set of points that the algorithm can shatter, which was originally defined by Vladimir Vapnik.

The ERM/SRM theory of learning

- Goal of ML: $f^* = \operatorname{argmin}_f E_{\text{unseen data}} L(f(x), y)$
- Goal of ERM: $f_{\text{ERM}}^* = \operatorname{argmin}_{f_w \in \mathcal{H}} \frac{1}{n} \sum_{\text{training data}} L(f_w(x_i), y_i)$
- ML: the goal of Machine Learning is to find a function that minimize the loss in the future (Or the expected loss over prob dist under some assumption).
- ERM: the goal of Empirical risk minimization is to find a function over a class of function that minimize the loss over the training data
- How should we compare them in the classical approach?
 - According to Vapnik: 1) Uniform Law of Large number. 2) Capacity control
 - uniform Law of large number** says that the ERM is approximately equal to the expected loss, so it allow us to compare the empirical loss to the expected loss
 - Capacity control:** say there are H (A set of function) contains a function that approximate the goal of Machine Learning function (The target funct f^*)

- 1. The theory of induction is based on the uniform law of large numbers.*
 - 2. Effective methods of inference must include capacity control.*

V. Vapnik, Statistical Learning Theory, 1998

- You can see your ERM solution is nearly optimal (Close to expected loss)
 - (1)+(2) $\Rightarrow E_{\text{unseen data}} L(f_{\text{ERM}}^*(x), y) \approx E_{\text{unseen data}} L(f^*(x), y)$
- Uniform laws of large numbers (Aka WYSIWG bounds), e.g. vc-dim, fat shattering, rademacher, covering number, margin...

Model or function complexity, e.g., VC, margin or $\|f\|_{\mathcal{H}}$

- Expected risk:
what you get

$$E(L(f_{ERM}^*, y)) \leq \frac{1}{n} \sum L(f_{ERM}^*(x_i), y_i) + O^*\left(\sqrt{\frac{C}{n}}\right)$$

- Empirical risk:
what you see

- In the left: the expected risk, what you get, the future
- In the right: the Empirical risk (What you see when in your training data) + capacity term (e.g. VC, margin)

- Capacity control

6.1 THE SCHEME OF THE STRUCTURAL RISK MINIMIZATION INDUCTION PRINCIPLE 223

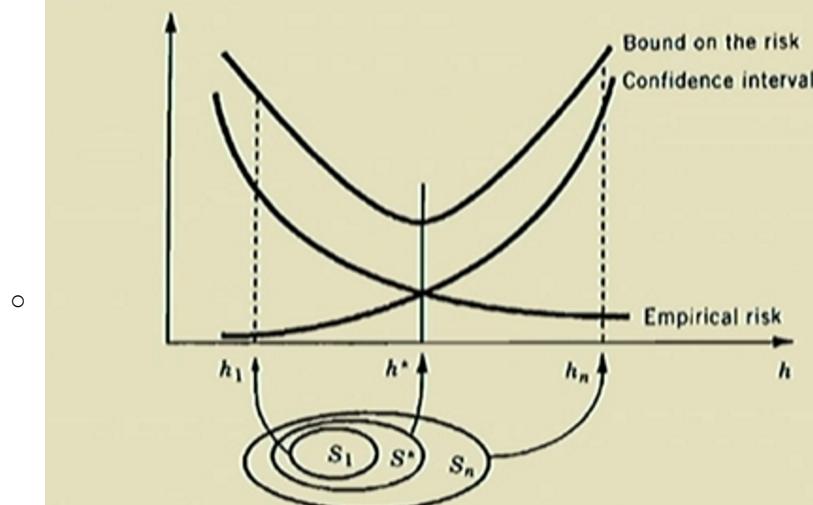
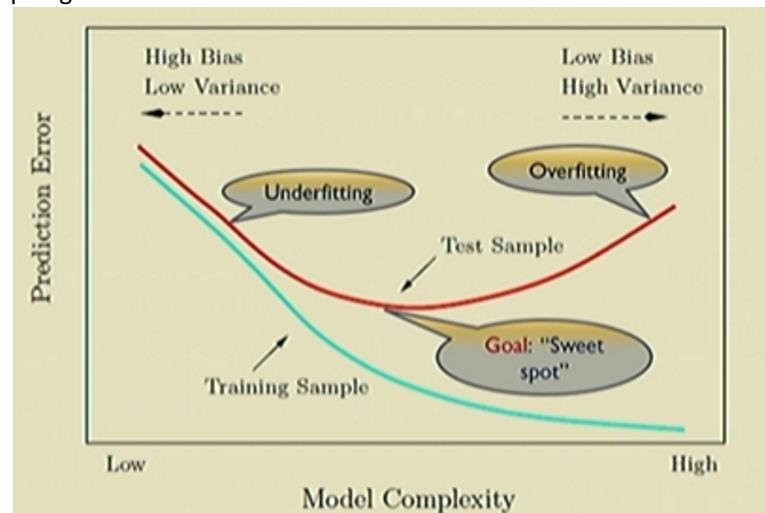


FIGURE 6.2. The bound on the risk is the sum of the empirical risk and of the confidence interval. The empirical risk is decreased with the index of element of the structure, while the confidence interval is increased. The smallest bound of the risk is achieved on some appropriate element of the structure.

V.Vapnik, Statistical Learning Theory, 1998

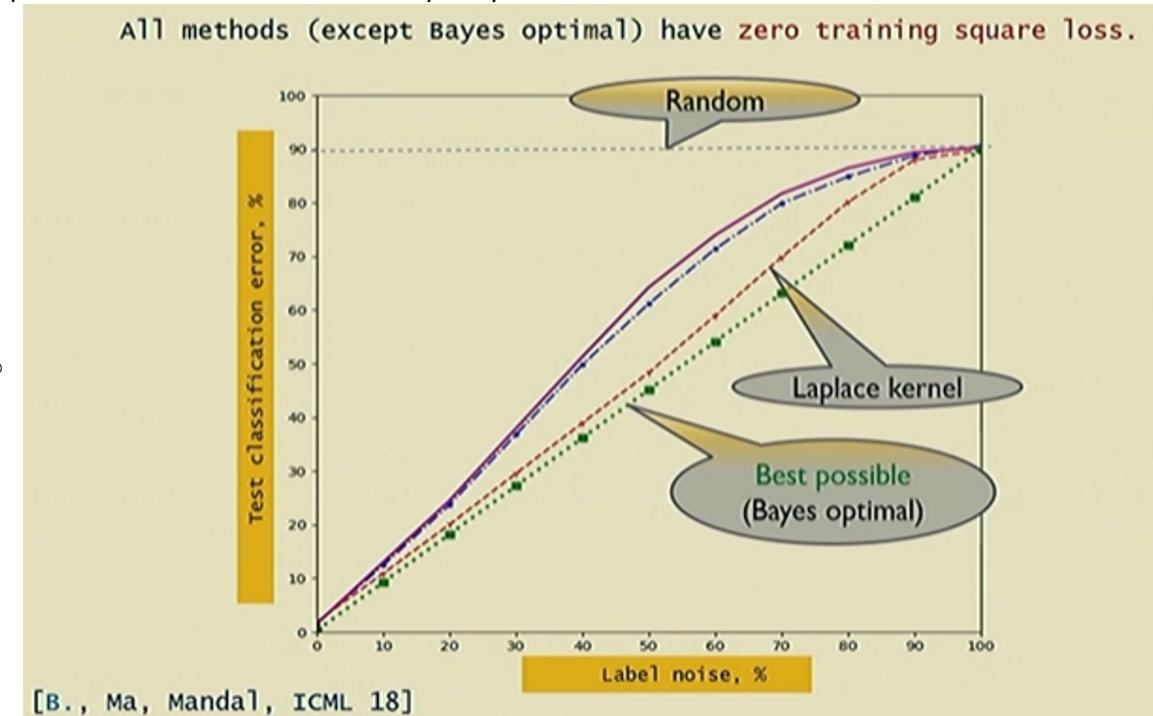
- U-shaped generalization curve



- In the left side is underfitting, and right side is overfitting, and our goal is to find this "sweet spot"
- However, a model with **zero training error** is overfitting to the training data and will typically generalize poorly.
- zero training error is also known as Interpolation, in mathematical term!

- Does interpolation overfit?

- This is a paper, prove that you can interpolation and still have a good test result!
- Interpolation doesn't overfit even for very noisy data:



- Randomization: take 10% of the data and assign random labels
- Green line: 50%, the best you can possibly do, based on Bayes optimal classifier.
- Random guess: is 90%, because the dataset has 10 classes.
- Red line: a kernel machine train to have zero loss on the noisy data. However, the result doesn't seem to overfit.

- why bounds fail?

correct

$$0.7 < O^* \left(\sqrt{\frac{c(n)}{n}} \right) < 0.9 \quad n \rightarrow \infty$$

- There are two problem:

1. The constant in O^* needs to be exact. There are no known bounds like that.

□

2. Conceptually, how would the quantity $c(n)$ "know" about the Bayes risk?

- Interpolation is best practice for deep learning:
 - The best way to solve the problem from practical standpoint is you build a very big system
.... Basically you want to make sure you hit the zero training error (Interpolation)

Yann Lecun (IPAM talk, 2018):

Deep learning breaks some basic rules of statistics.

- The modern ML: The key lesson

The new theory of induction cannot be based on uniform laws of large numbers with capacity control.

- So the generalization theory will be helpful for understanding interpolation?

- What theoretical analyses do we have?

➤ VC-dimension/Rademacher complexity/covering/Pac-Bayes/margin bounds.	Uniform bounds: training loss expected loss
➤ Cannot deal with interpolated classifiers when Bayes risk is non-zero.	
➤ Generalization gap cannot be bound when empirical risk is zero.	
➤ Algorithmic stability.	Typically Diverge
➤ Does not apply when empirical risk is zero, expected risk nonzero.	
➤ Regularization-type analyses (Tikhonov, early stopping/SGD, etc.)	oracle bounds expected loss optimal loss
➤ Diverge as $\lambda \rightarrow 0$ for fixed n .	
➤ Classical smoothing methods (nearest neighbors, Nadaraya-Watson).	
➤ Most classical analyses do not support interpolation.	
➤ But 1-NN! (Also Hilbert regression Scheme, [Devroye, et al. 98])	

- 1-NN

- Analysis doesn't based on complexity bounds
 - Estimating expected loss, not the generalization gap

This talk so far:

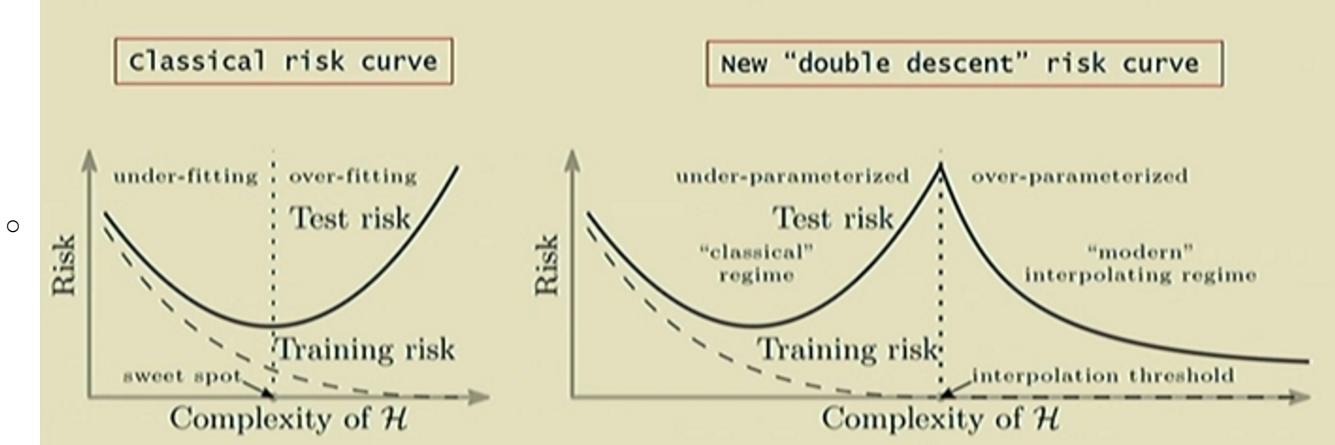
- Effectiveness of interpolation.
- Theory of interpolation cannot be based on uniform bounds.
- Statistical validity of interpolating nearest neighbor methods.

- Yet, there is a mismatch between A and C. Methods we considered theoretically seem quite different from those used in practice.

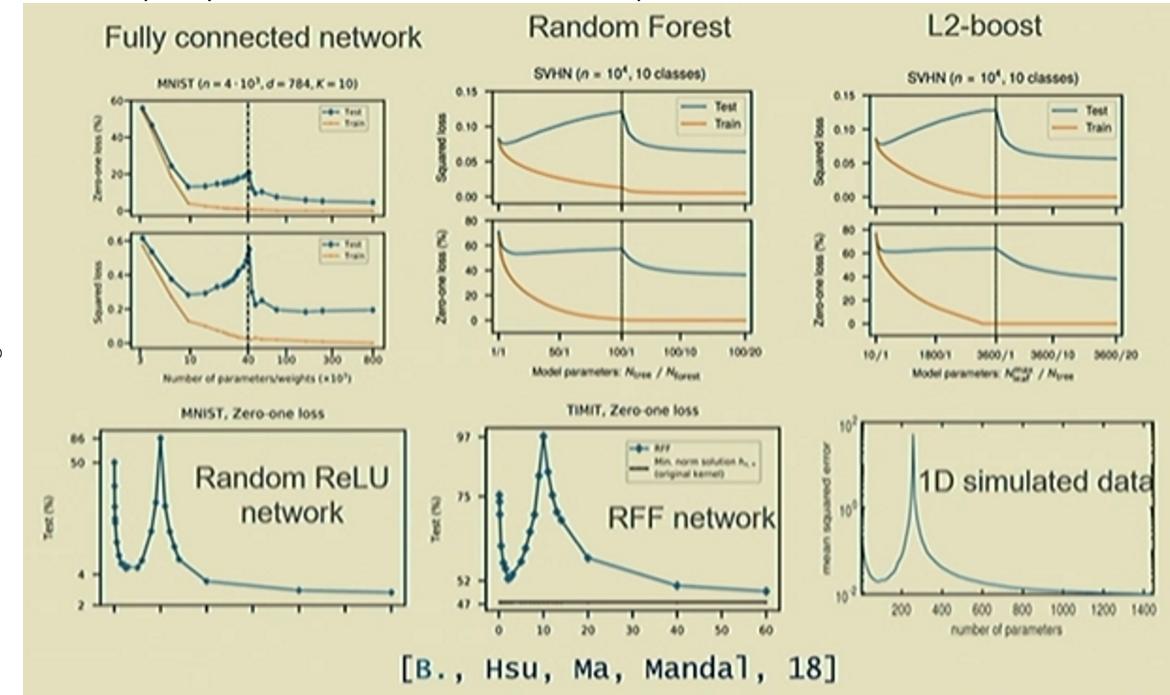
Key questions:

- How do classical analyses relate to interpolation?
- Dependence of generalization on model complexity?
- What is the role of optimization?

- "Double descent" risk curve: Reconciling modern machine-learning practice and the classical bias-variance trade-off, <https://www.pnas.org/content/pnas/116/32/15849.full.pdf>



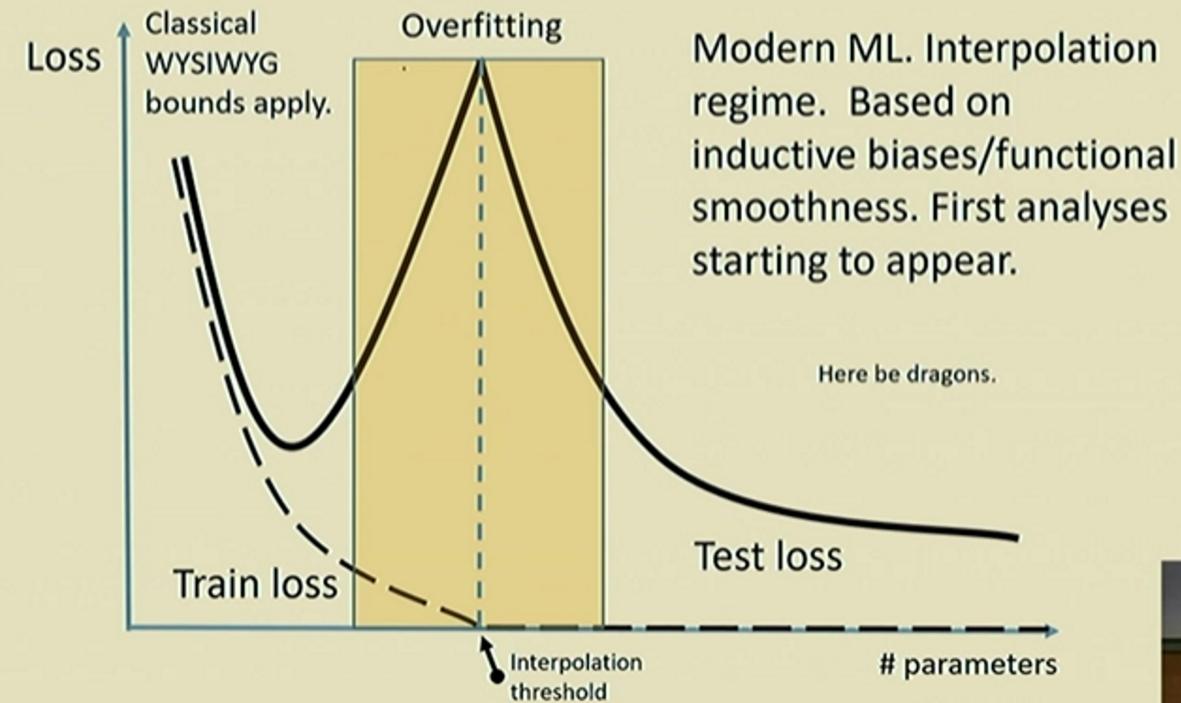
- Complexity of H : can be thought as the number of parameters



[B., Hsu, Ma, Mandal, 18]

- More parameters are better: an example
- Random Fourier network
- Why is training better as I am increasing the number of features?
 - the space of solution, the interpolated data
- New understanding of overfitting
 - Previous: while the training loss is low, we must be overfitting, and therefore we should decrease the number of parameter, such introduce regularization,
 - Now: there are two way to avoid the overfitting. 1)The classical way is to reduce the number of parameter. 2)The modern ML is to increase the number of parameter, moving to the right. It's counterintuitive, but

The landscape of generalization



- Classical Optimization(Under-parametrized)
 - Many local minima
 - SGD(Fixed step size) doesn't converge
- Modern Optimization(**Interpolation**/over-parametrized)
 1. Every local minimum is global (for networks wide enough)
[Li, Ding, Sun, 18], [Yu, Chen, 95]
 2. Local methods converge to global optima
 - [Kawaguchi, 16] [Soheil, et al, 16] [Bartlett, et al, 17]
[Soltanolkotabi, et al, 17, 18] [Du, et al, 19] ...
 3. Small batch SGD (fixed step size) converges as fast as GD per iteration.
[Ma, Bassily, B., ICML 18] [Bassily, Ma, B., 18]