Coffee Talk #8

January 11, 2023

Ozgur Taylan Turan

Benign, Tempered or Catastrophic: A Taxonomy of Overfitting[1]

[1] N. Mallinar, J. B. Simon, A. Abedsoltan, P. Pandit, M. Belkin, and P. Nakkiran (July 13, 2022). *Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting*. arXiv: 2207.06569 [cs, stat]

Why this paper?

- Discussion revolving around overfitting in one of the PR-Lab meetings.

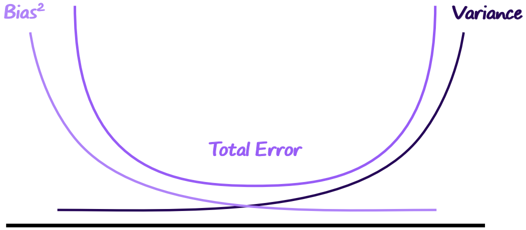
Aim

- To create a taxonomy of overfitting for interpolating methods!
- To show *overfitting* does not necessarily a good or a bad thing!

Introduction-A

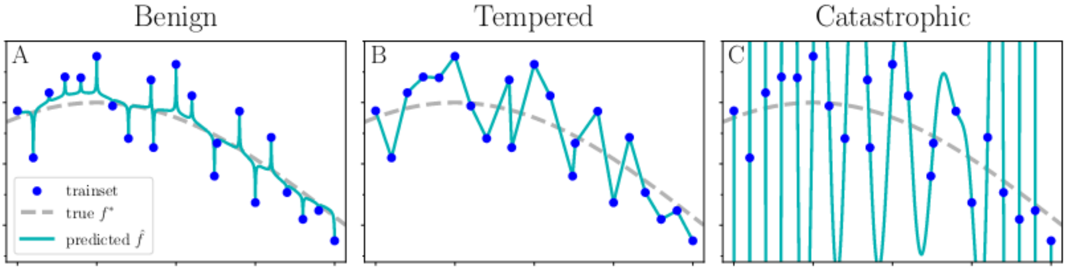
Classical statistical learning theory vs Double Descent

- Double Descent and Benign overfitting



Introduction-B

Overfitting: fitting exactly to your training data.
Interpolation $\sim$ Overfitting (in Belkin's works!)



Problem Setting

- Interpolation $:=$ Zero Training Error
- $\hat{f} : \mathcal{X} \to \mathbb{R}$ from $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^{n}$
- Var$[y_i | x_i] > 0$ Non zero target noise!

- Generalization performance $\mathcal{R}(\hat{f}) := \mathbb{E}_{\mathcal{D}}[(\hat{f}(x) - y)^2]$
- Bayes optimal solution: $f^* = \arg\min_f R(f)$ gives the risk R^*

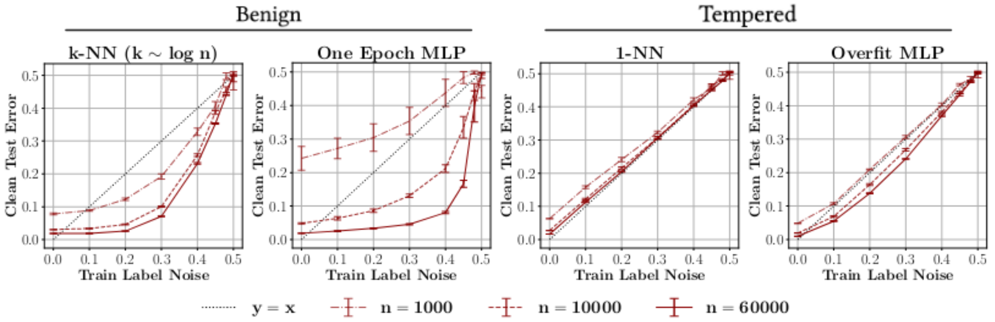
Problem Setting-Learning Procedures

Learning Procedure

- $\mathcal{A} := \{\mathcal{A}_n\}_n$ are potentially stochastic functions
- $\mathcal{A}_n : \mathcal{D}_n \to \hat{f}$ are potentially stochastic functions
- Risk of a Learning Procdure:

$$R_n := \mathbb{E}_{\mathcal{A}_n, \mathcal{D}_n}[R(\mathcal{A}_n(\mathcal{D}_n))]$$

Benign	$\lim_{n\to\infty} R_n = R^*$
Tempered	$\lim_{n\to\infty} R_n \in (R^*, \infty)$
Catastrophic	$\lim_{n\to\infty} R_n = \infty$



Benign	$\lim_{n\to\infty} R_n = R^*$
Tempered	$\lim_{n\to\infty} R_n \in (R^*, \infty)$
Catastrophic	$\lim_{n\to\infty} R_n = \infty$

Overfitting-Kernel Regression-A

Neural Tangent Kernel (NTK)[1]: infinite width and time limit NNs with converge to "Ridless" ($\lambda = 0$) Kernel Ridge Regression for Least Squares problem! NTK opens up a path for understanding the generalization of NNs.

$$\hat{f}(x) = K(x, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \delta \mathbf{I})^{-1}\mathbf{Y}$$

[1] A. Jacot, F. Gabriel, and C. Hongler (Feb. 10, 2020). *Neural Tangent Kernel: Convergence and Generalization in Neural Networks*. arXiv: 1806.07572 [cs, math, stat]

Overfitting-Kernel Regression-B

- Exploiting $\mathbb{E}_{x'}[K(x,x')\phi_i(x')] = \lambda_i\phi_i(x)$ and $\mathbb{E}_x[\phi_i(x)\phi_j(x)] = \delta_{ij}$

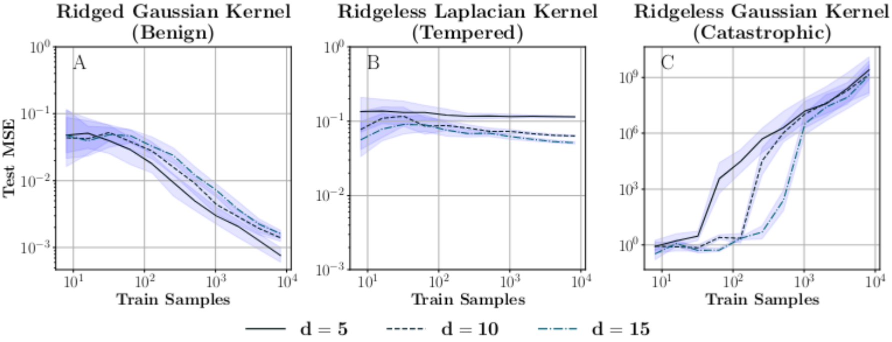
After some math[1] R_n can be approximated by *Modewise Learnabilities*:

- A positive Ridge parameter or slow eigendecay $\rightarrow$ **Benign Overfitting**
- Powerlaw decay of eigenvalues $\rightarrow$ **Tempered Overfitting**
- Any decay faster than power law $\rightarrow$ **Catastrophic Overfitting**

[1] J. B. Simon, M. Dickens, D. Karkada, and M. R. DeWeese (Oct. 12, 2022). *The Eigenlearning Framework: A Conservation Law Perspective on Kernel Regression and Wide Neural Networks*. arXiv: 2110.03922 [cs, stat]

Overfitting-Kernel Regression-B

- A positive Ridge parameter or slow eigendecay $\rightarrow$ **Benign Overfitting**
- Powerlaw decay of eigenvalues $\rightarrow$ **Tempered Overfitting**
- Any decay faster than power law $\rightarrow$ **Catastrophic Overfitting**



What do they do more?

- Computer vision examples showing tempered overfitting for interpolating ANNs.
- MLP on synthetic data: first benign, then tempered overfitting.

Conclusions

- Chill-out it! There is small chance that overfitting will hurt in modern applications of interpolating models?!

Benign	Tempered	Catastrophic
• Early-stopped DNNs	• **Interpolating DNNs**	• **Gaussian KR**
• KR with ridge	• **Laplace KR**	• Critically-parameterized regression
• k-NN ($k \sim \log n$)	• **ReLU NTKs**	
• Nadaraya-Watson kernel smoothing with Hilbert kernel	• k-NN (constant k)	
	• Simplicial interpolation	

THANKS!