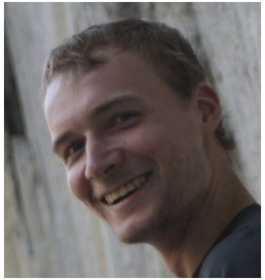
Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting

`paper at tinyurl.com/TemperedOverfitting`

In review at NeurIPS 2022

Neil Mallinar*
UC San Diego



James B. Simon*
UC Berkeley

Amirhesam Abedsoltan
UC San Diego

Parthe Pandit
UC San Diego

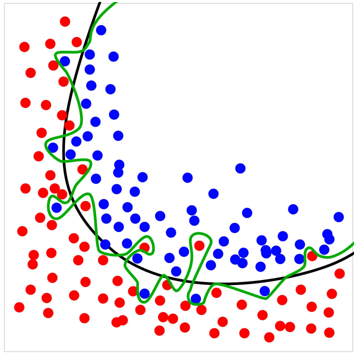
Mikhail Belkin
UC San Diego

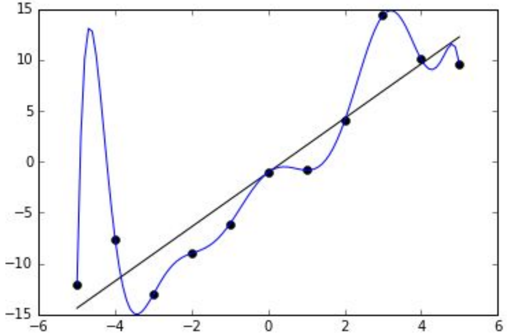
Preetum Nakkiran
Apple

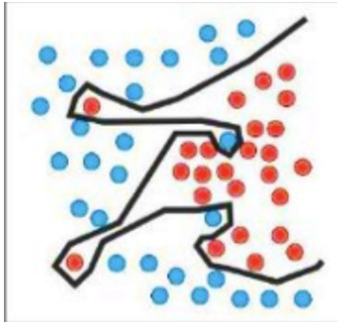
*Co-first author

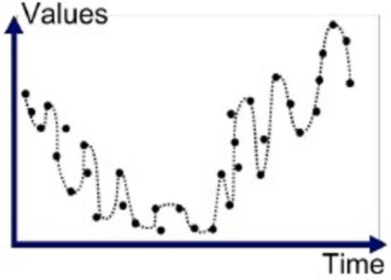
paper at tinyurl.com/TemperedOverfitting

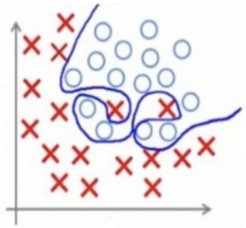
The classical story of overfitting

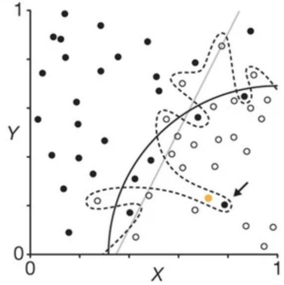


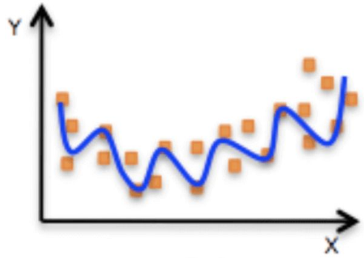




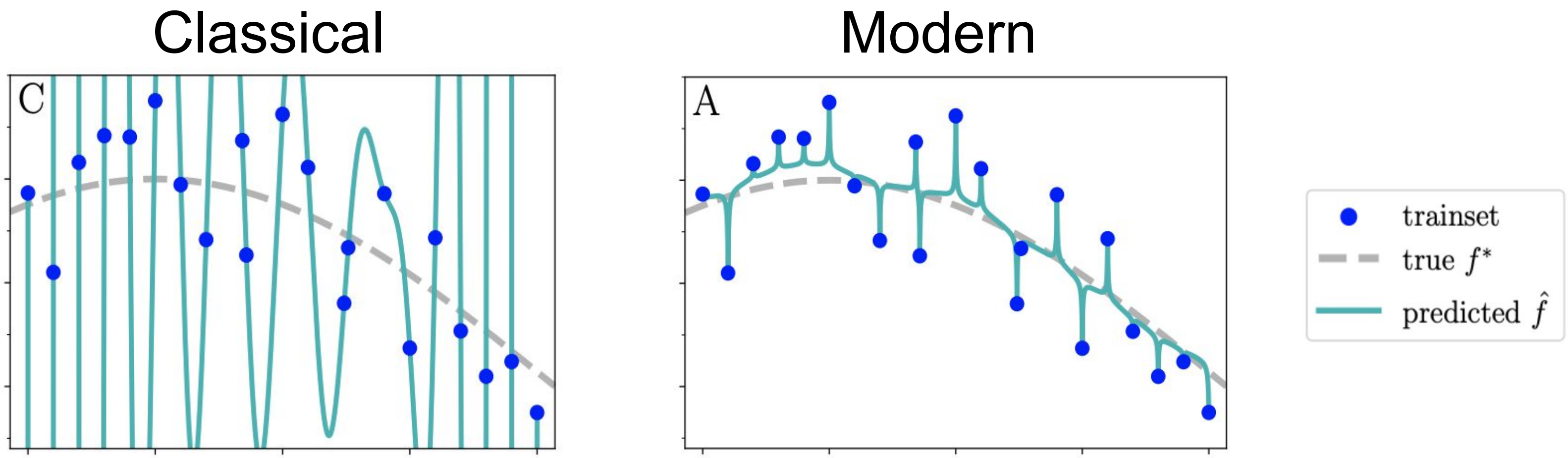








Overfitting can be benign



Just Interpolate: Kernel "Ridgeless" Regression Can Generalize

Tengyuan Liang*[1] and Alexander Rakhlin†[2]

[1]University of Chicago, Booth School of Business
[2]Massachusetts Institute of Technology

How harmful is overfitting for standard deep neural networks (DNNs)?

Our Setting

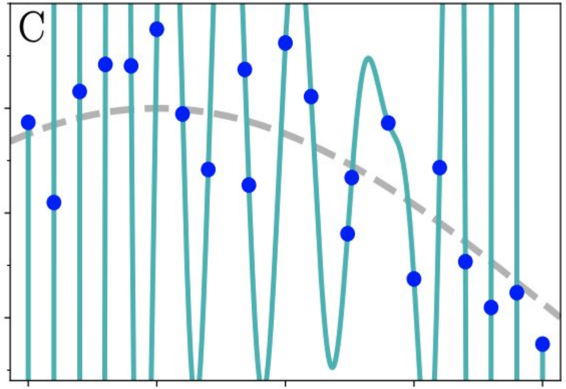
$\hat{f}_n : \mathcal{X} \to \mathbb{R}$ - estimator trained on n samples
$R(\hat{f}) = \mathbb{E}[(\hat{f}(x) - y)^2]$ - population risk (MSE)
$R^* = \min_f R(f)$ - optimal risk
$R_n = \mathbb{E}[R(\hat{f}_n))]$

Want to estimate:

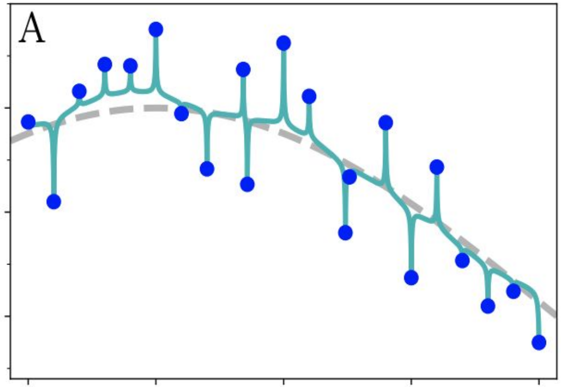
$\lim_{n \to \infty} R_n = ?$

A taxonomy of overfitting

Classical

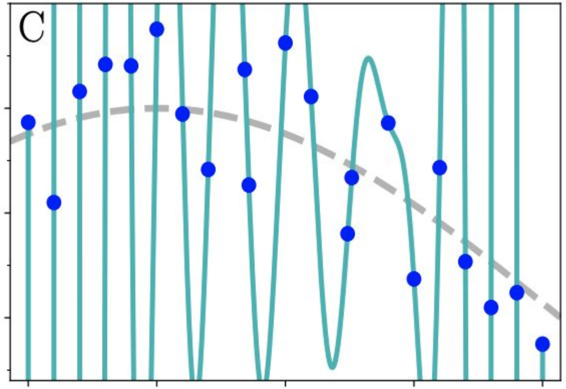


Modern



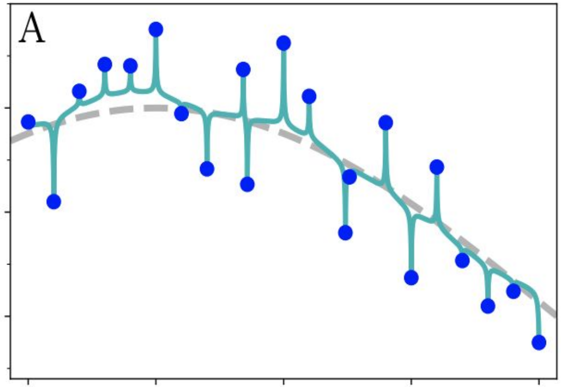
A taxonomy of overfitting

Catastrophic



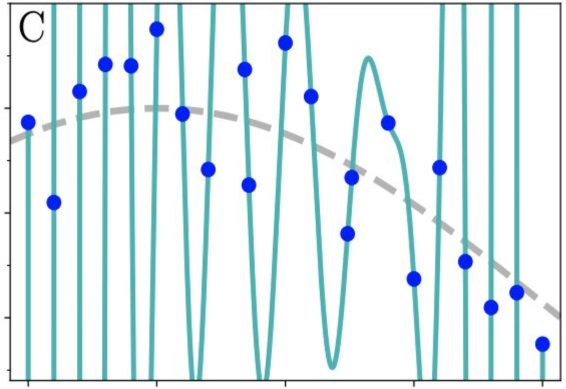
$\lim_{n\to\infty} R_n = \infty$

Modern



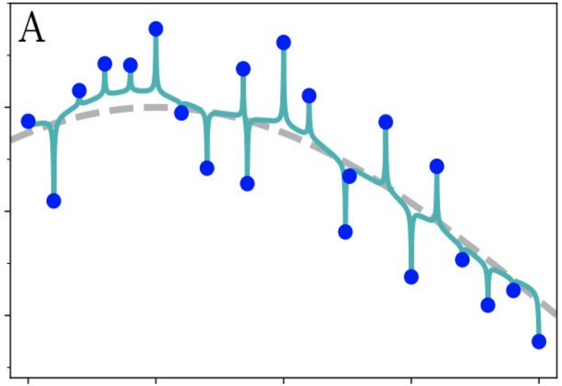
A taxonomy of overfitting

Catastrophic



$\lim_{n\to\infty} R_n = \infty$

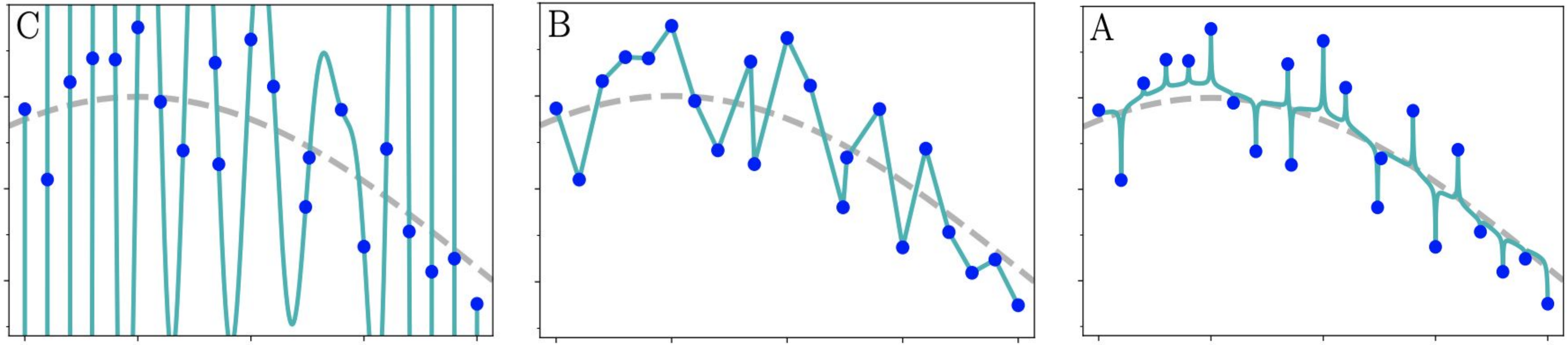
Benign



$\lim_{n\to\infty} R_n = R^*$

A taxonomy of overfitting

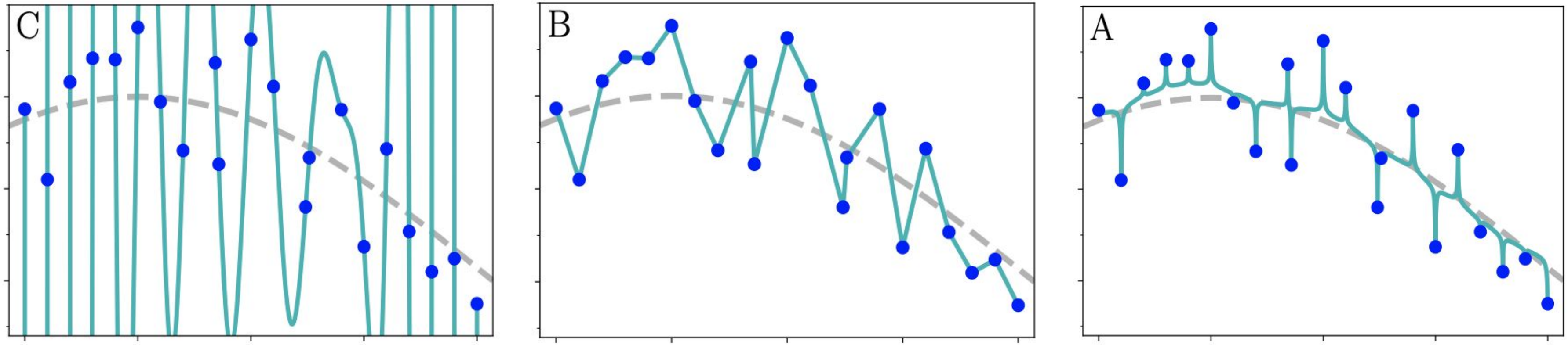
Catastrophic | Tempered | Benign



$\lim_{n\to\infty} R_n = \infty$ | $\lim_{n\to\infty} R_n = (R^*, \infty)$ | $\lim_{n\to\infty} R_n = R^*$

A taxonomy of overfitting

Catastrophic | **Tempered** | **Benign**



$\lim_{n\to\infty} R_n = \infty$

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

$\lim_{n\to\infty} R_n = R^*$

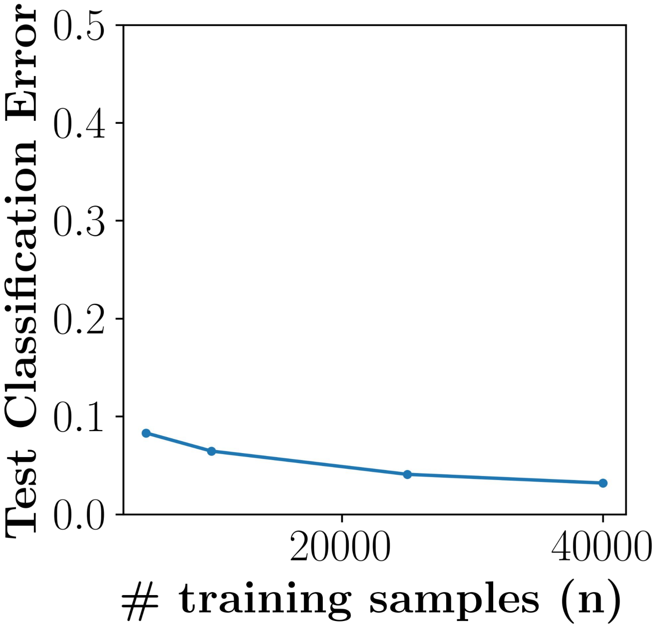
random classifier

How harmful is overfitting for standard deep neural networks (DNNs)?

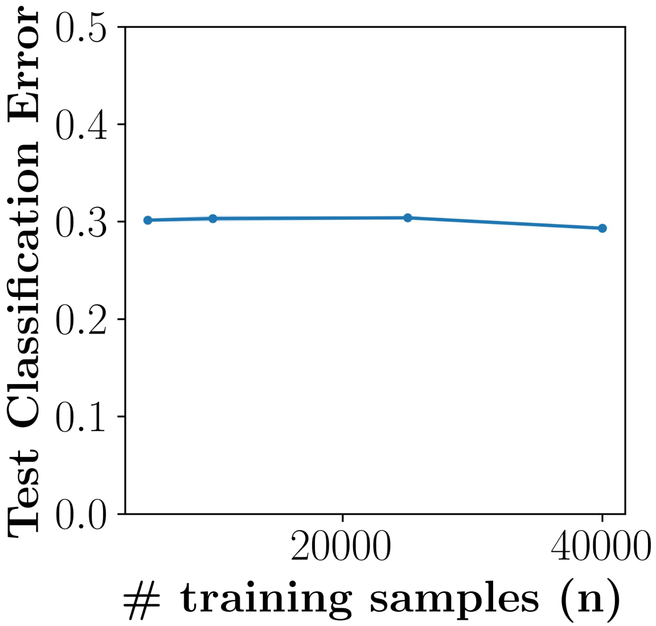
A simple experiment

Binary CIFAR-10, WideResNets interpolating training data

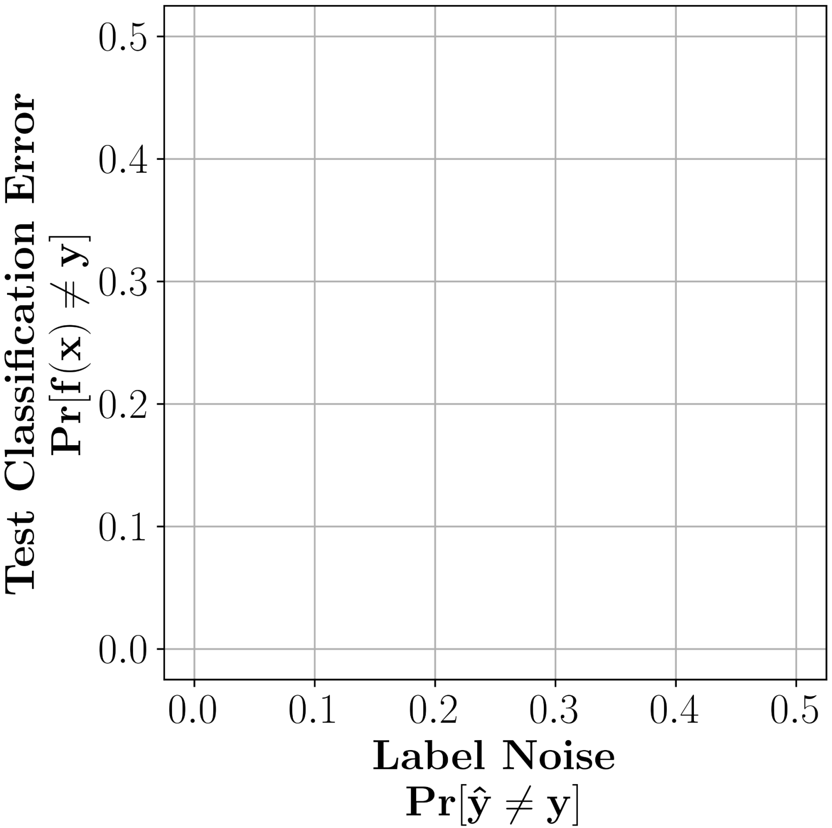
No added label noise



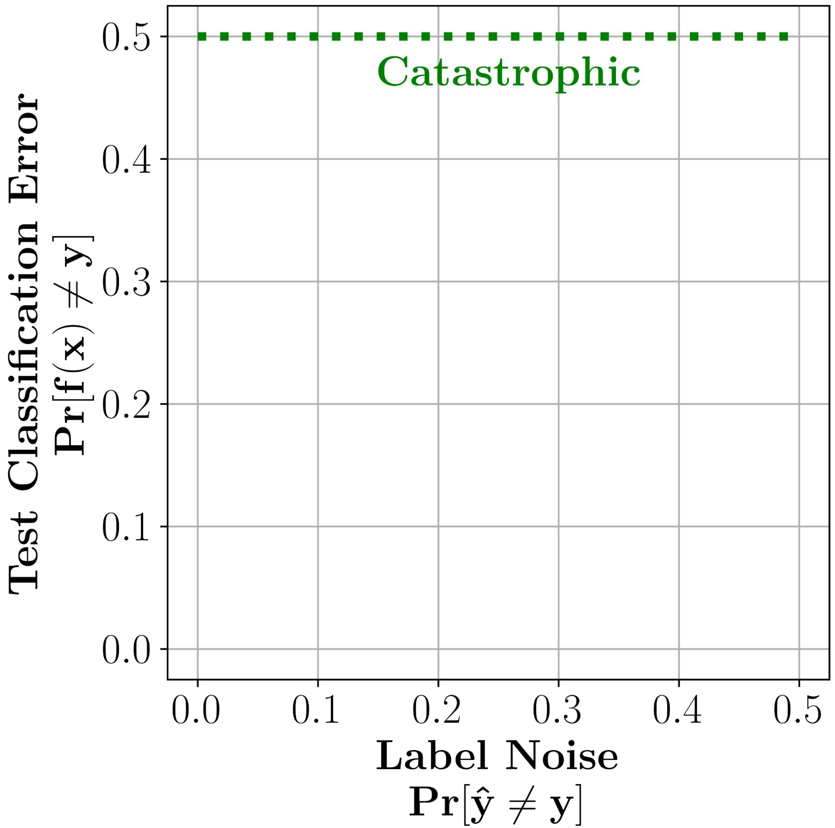
Flip 30% of labels



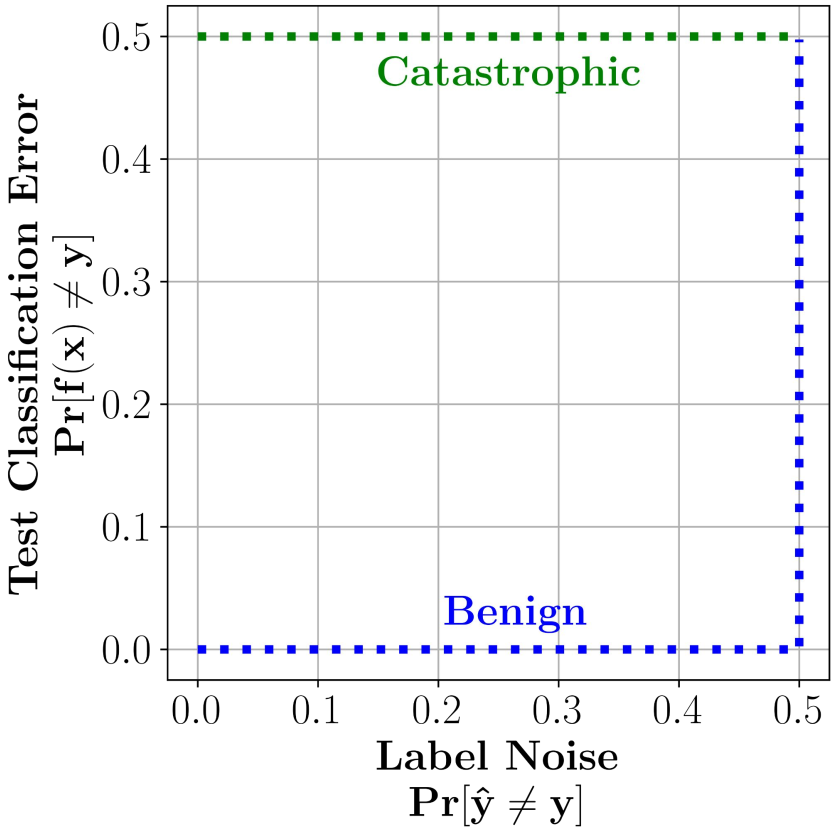
A simple experiment



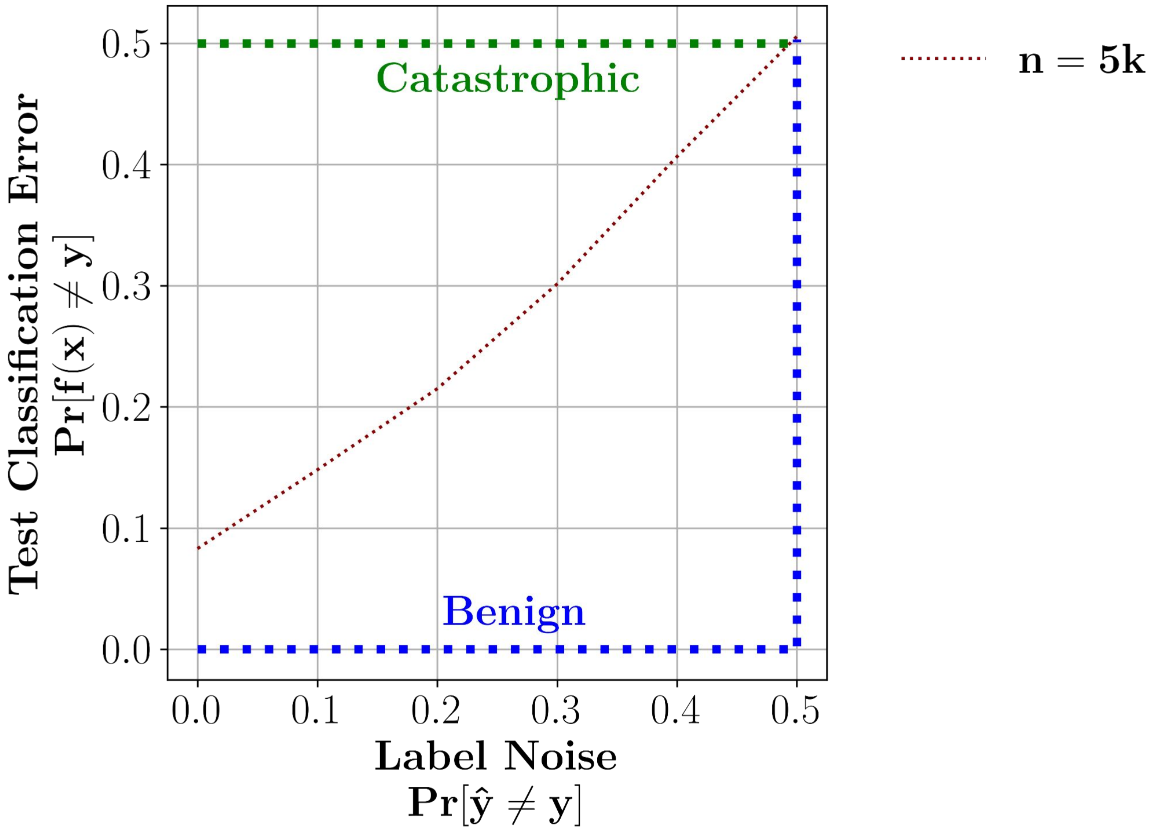
A simple experiment



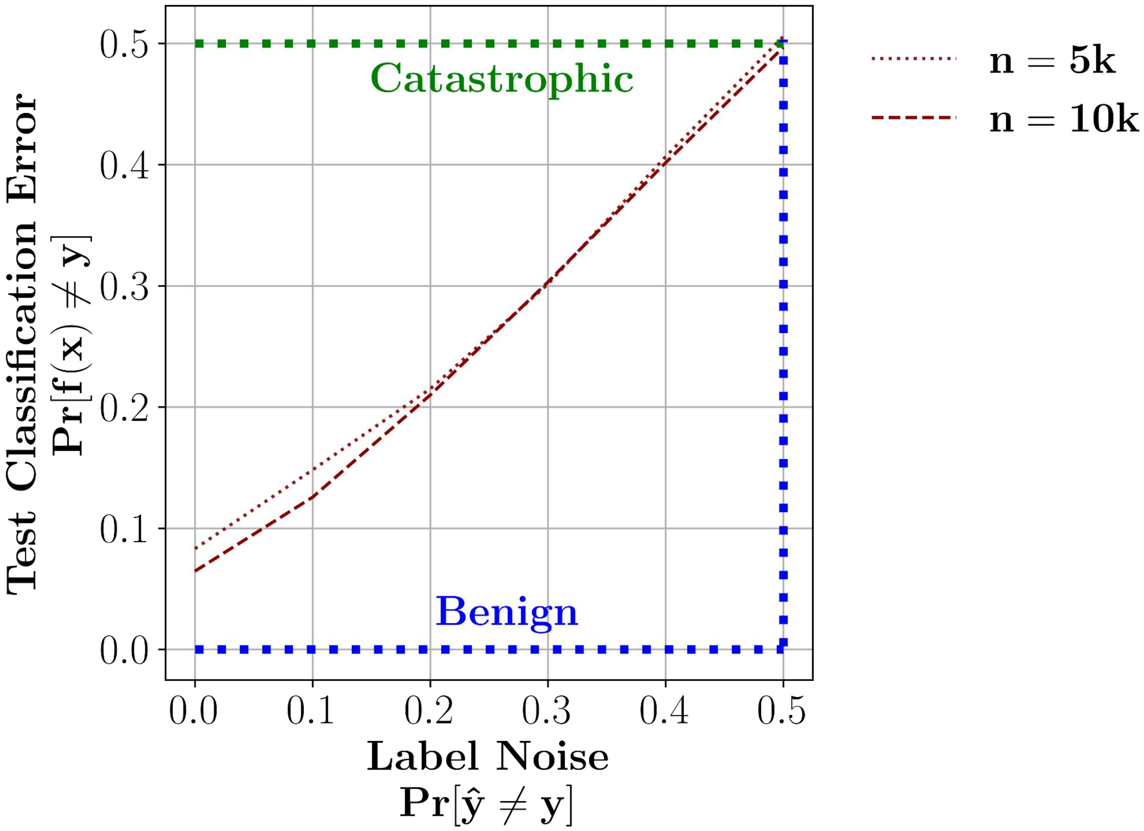
A simple experiment



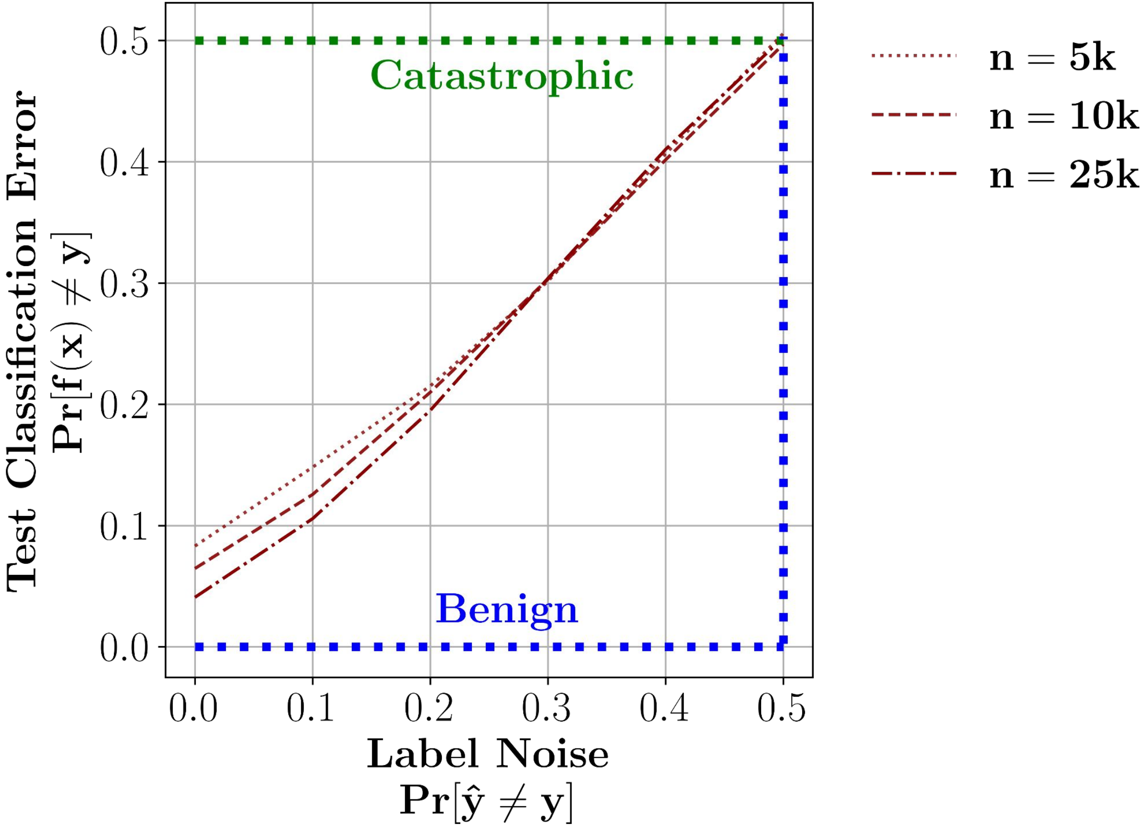
A simple experiment



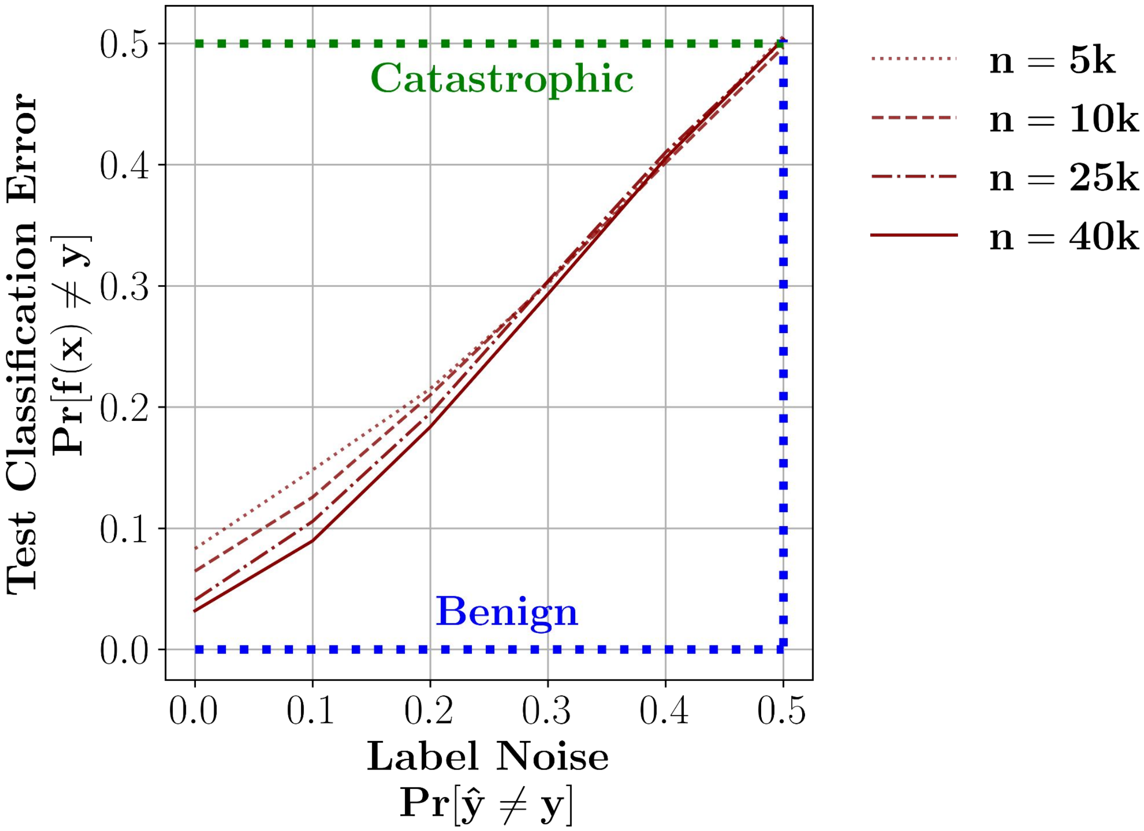
A simple experiment



A simple experiment



A simple experiment



Outline

1. Using the taxonomy

2. Empirical results on deep neural networks (DNNs)

3. Overfitting in kernel regression (KR)

Outline

1. Using the taxonomy

2. Empirical results on deep neural networks (DNNs)

3. Overfitting in kernel regression (KR)

(Some) prior works

Bartlett, Long, Lugosi, Tsigler (2020)	Linear regression
Liang, Rakhlin (2018)	Kernel ridgeless regression
Mei, Montanari (2019)	Random feature regression (ridgeless limits)
Belkin, Hsu, Mitra (2018)	Kernel smoothers / nearest neighbors
Rakhlin, Zhai (2019)	Laplace kernel interpolation
Koehler, Zhou, Sutherland, Srebro (2021)	High-dim linear regression
Ji, Li, Telgarsky (2021)	Early-stopped neural networks
Beaglehole, Belkin, Pandit (2022)	Shift-invariant kernel interpolators

d - input (ambient) dimension, n - number of training samples
Benign overfitting commonly shown for d > n or (d, n) scale jointly
Generalization error bounds in d, n

Motivation for a taxonomy

We consider: fixed input dimension (d), take n $\rightarrow \infty$

In this setting, *benign* = consistent

Prior works show inconsistency of interpolators on noisy data in low / fixed dimension (Rakhlin & Zhai '19; Beaglehole, Belkin, Pandit '22)

Two ways to be inconsistent when interpolating:

1. *tempered* (bounded risk as a function of label noise)
2. *catastrophic* (unbounded risk)

Example methods in the taxonomy

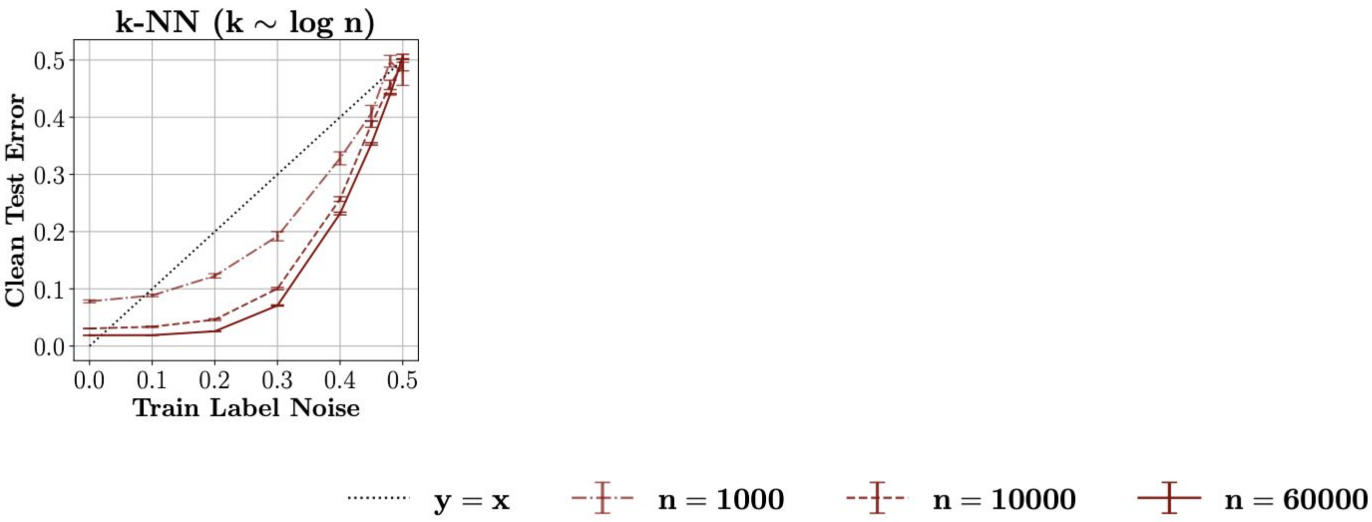
Benign (Consistent)	**Tempered** (Inconsistent)	Catastrophic (Inconsistent)
$\lim_{n\to\infty} \mathcal{R}_n = R^*$	$\lim_{n\to\infty} \mathcal{R}_n \in (R^*, \infty)$	$\lim_{n\to\infty} \mathcal{R}_n = \infty$
• Ridged kernel regression (KR) • k-NN, k ~ log n • Nadaraya-Watson estimator with singular kernel	• **Interpolating DNNs** • **Laplacian KR** • k-NN, constant k	• Models at double descent peak • Polynomial regression w/ degree = n • **Gaussian KR**

Example methods in the taxonomy

Classifying Binary MNIST (even/odd)

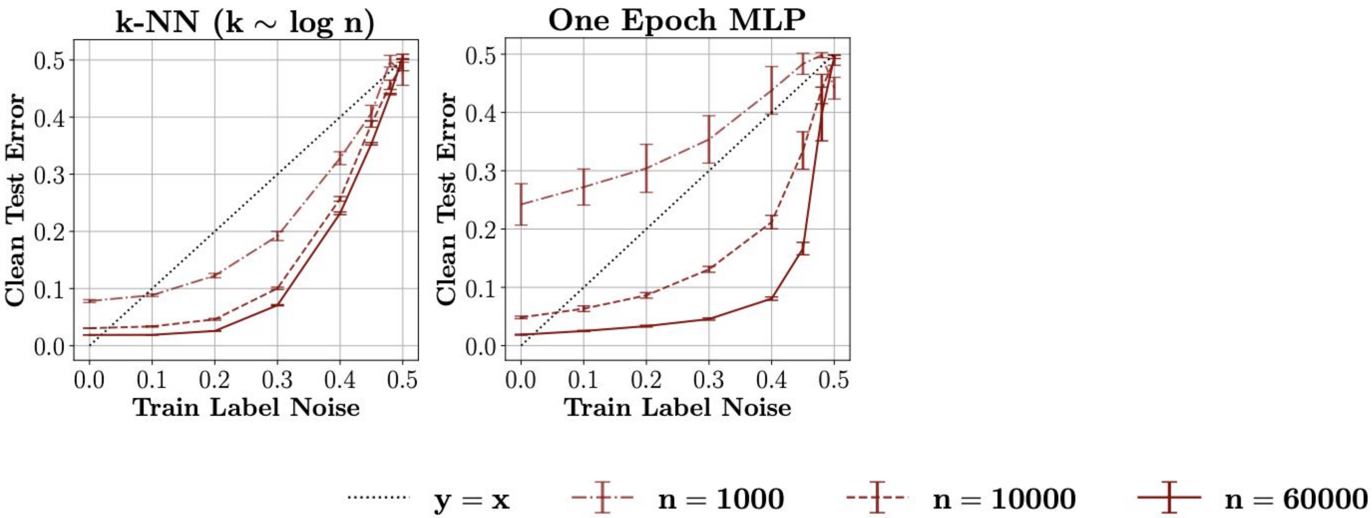
Example methods in the taxonomy

Classifying Binary MNIST (even/odd)



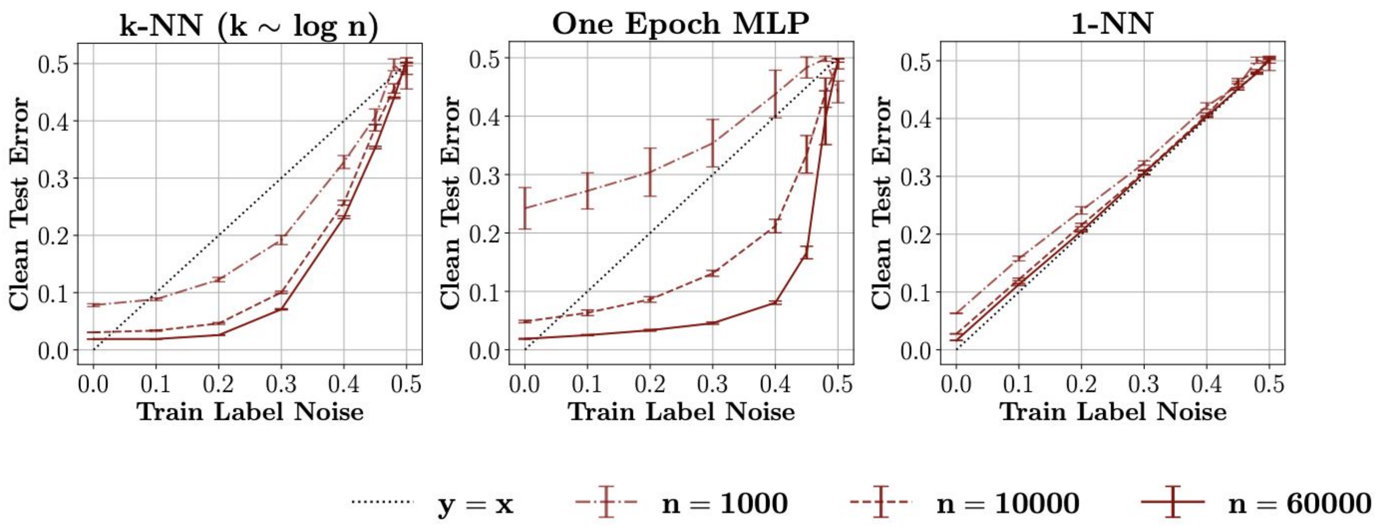
Example methods in the taxonomy

Classifying Binary MNIST (even/odd)



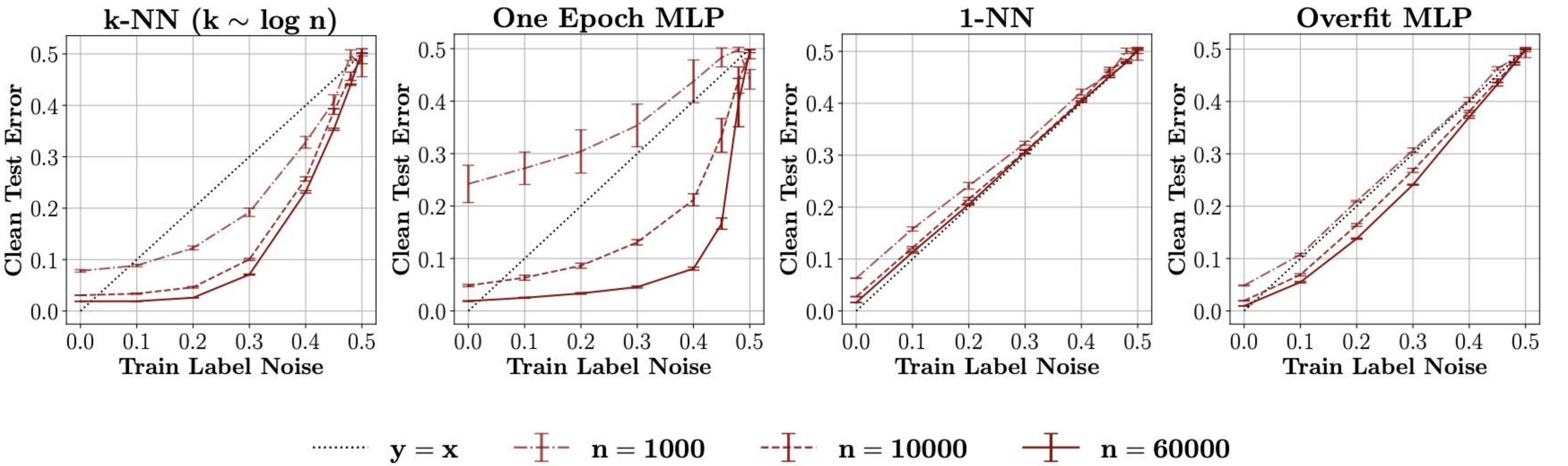
Example methods in the taxonomy

Classifying Binary MNIST (even/odd)



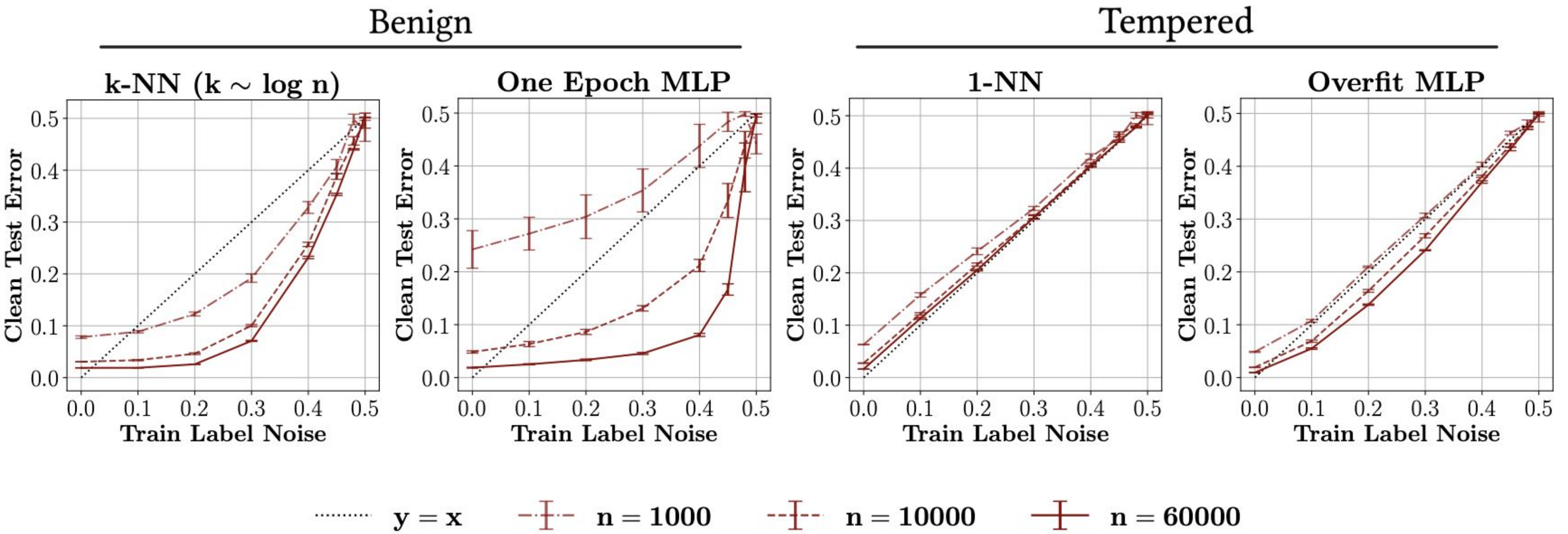
Example methods in the taxonomy

Classifying Binary MNIST (even/odd)



Example methods in the taxonomy

Classifying Binary MNIST (even/odd)

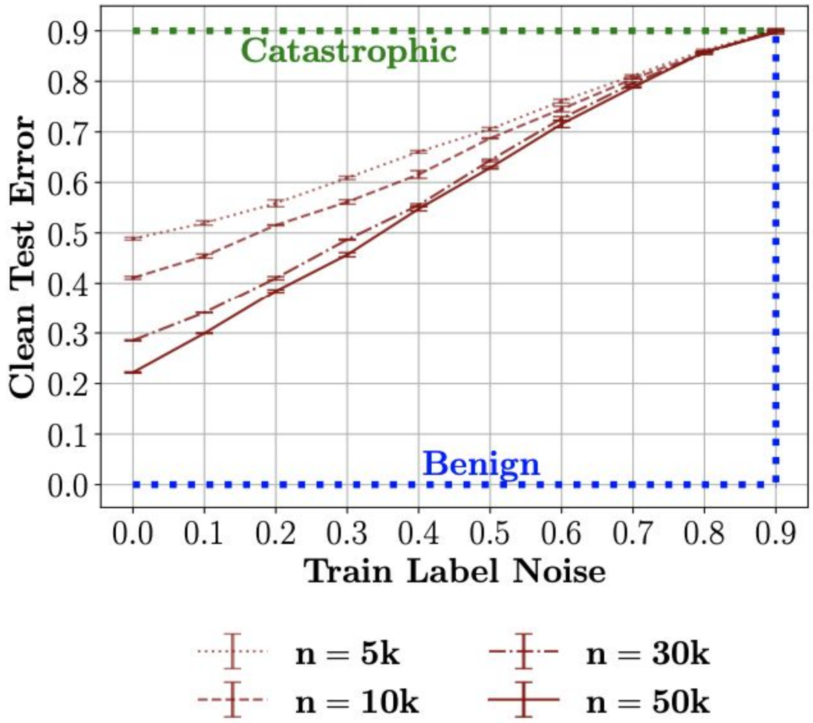


Outline

1. Using the taxonomy
2. Empirical results on deep neural networks (DNNs)
3. Overfitting in kernel regression (KR)

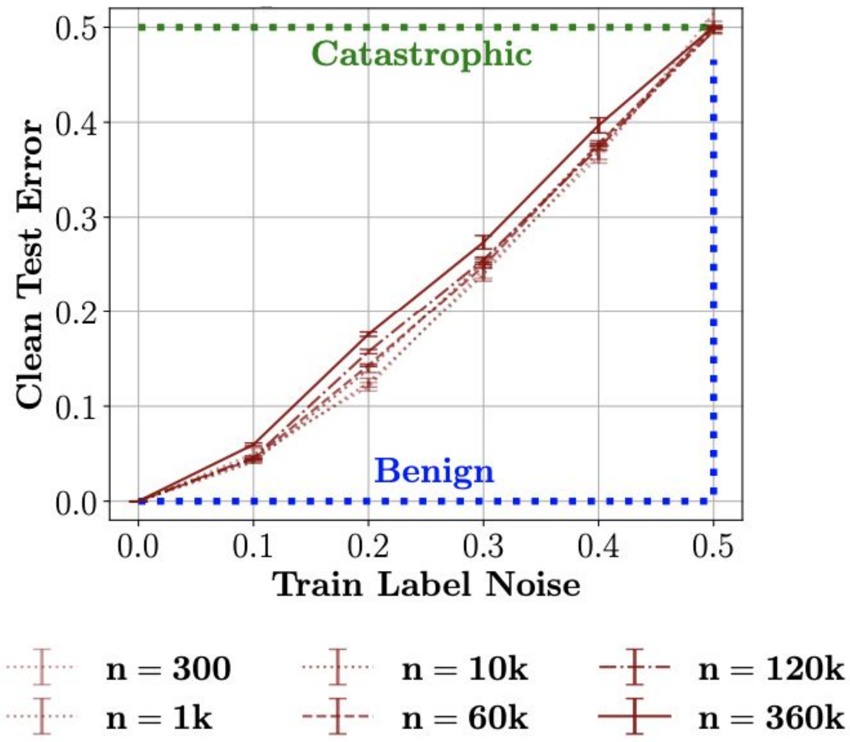
Interpolating DNNs are tempered

- Multi-class classification, CIFAR-10, WideResNet



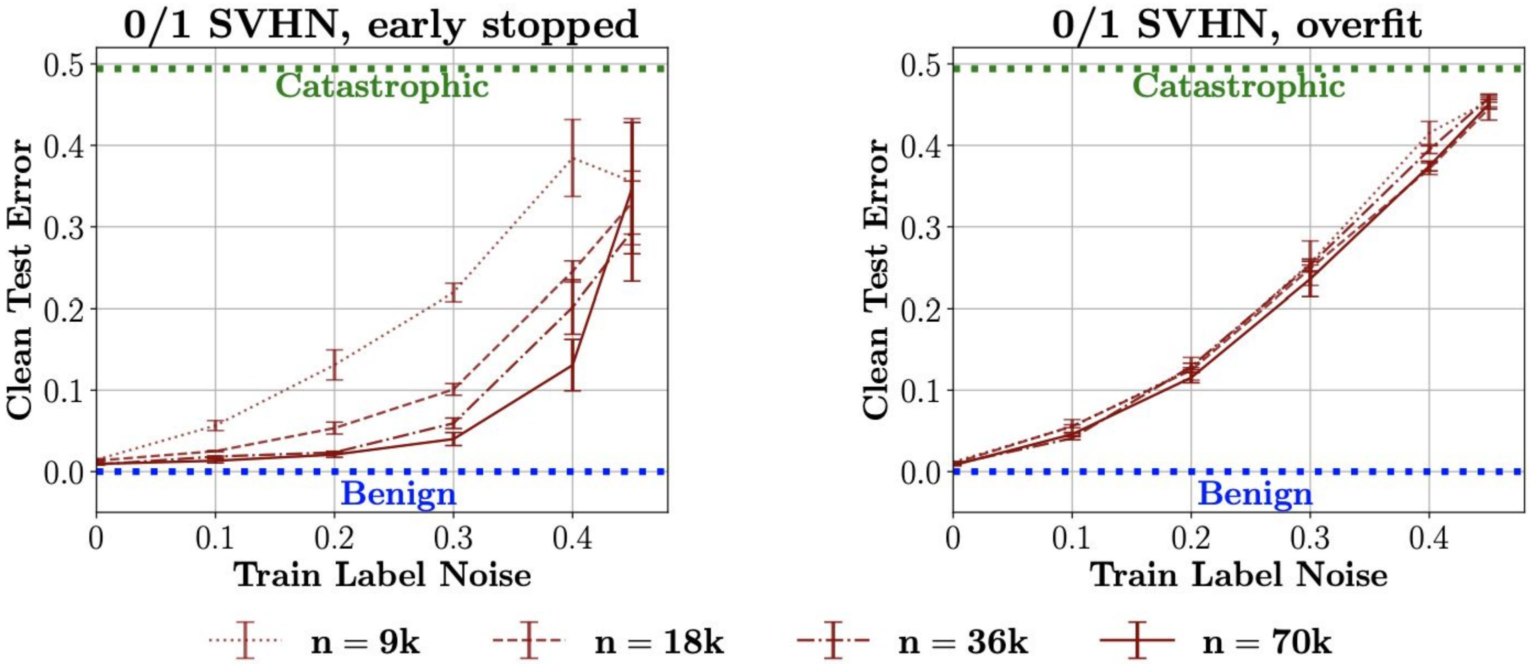
Interpolating DNNs are tempered

- Binary classification, synthetic data on 10-dim hypersphere, y = 1, MLP



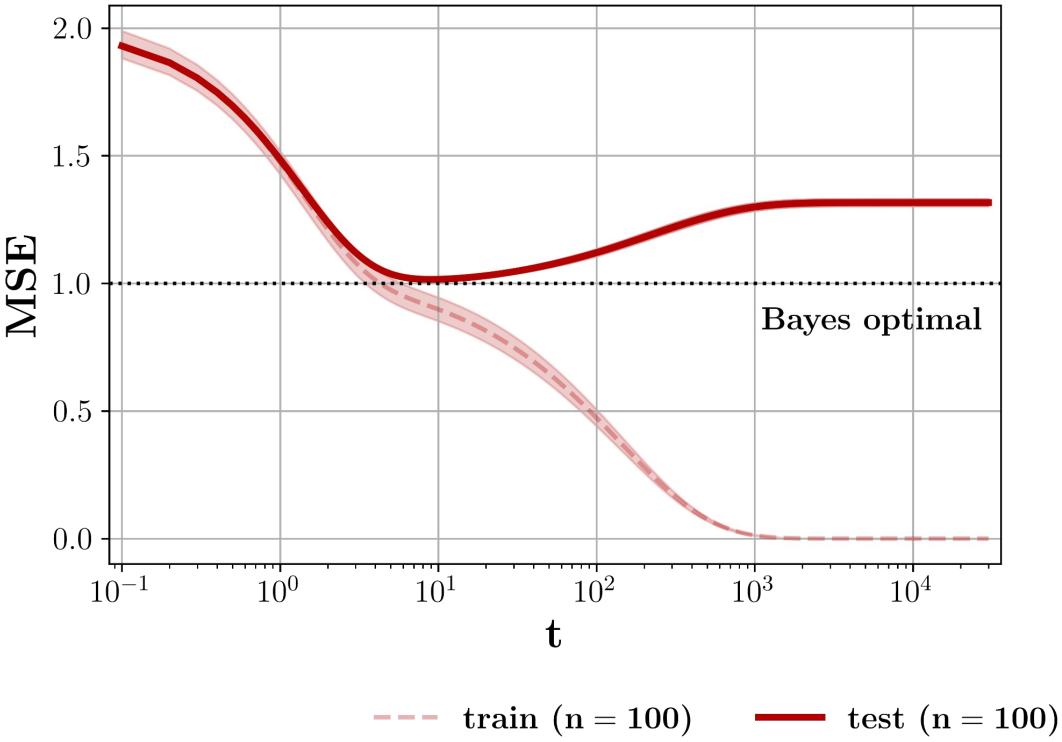
Early-stopped DNNs are benign

Shallow & wide ReLU nets are consistent w/ early stopping (Ji, Li, Telgarsky, 2021)



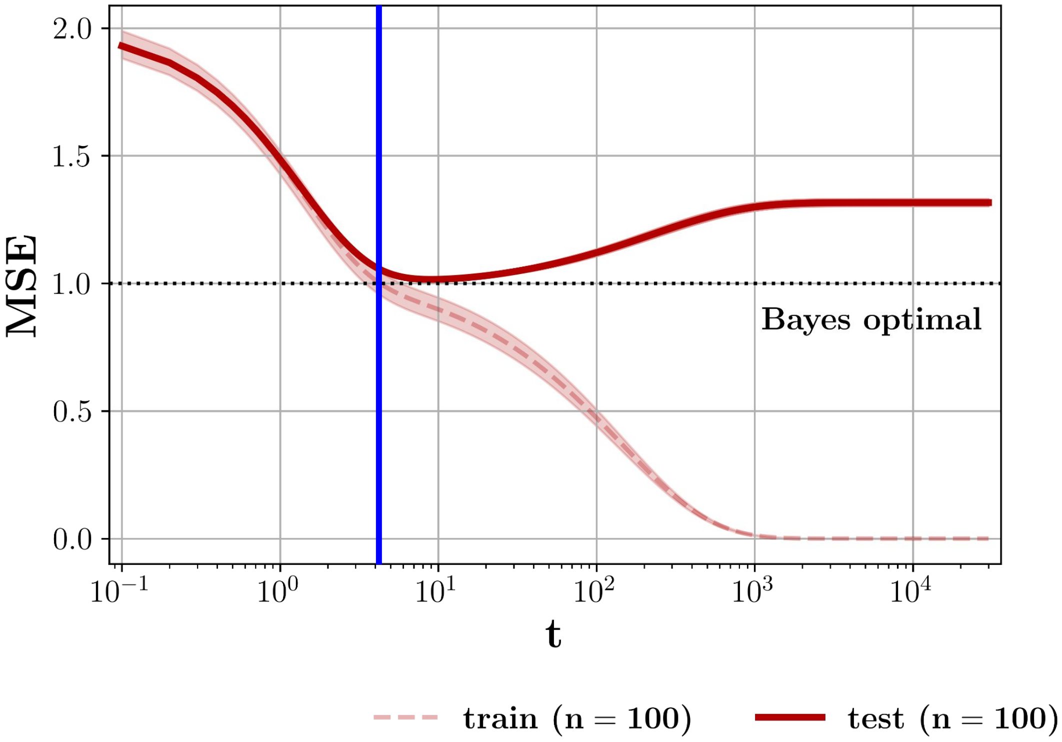
Time dynamics of MLP regression

ReLU MLP
Full batch GD
X = 5-dim sphere
y = 1 + N(0,1)



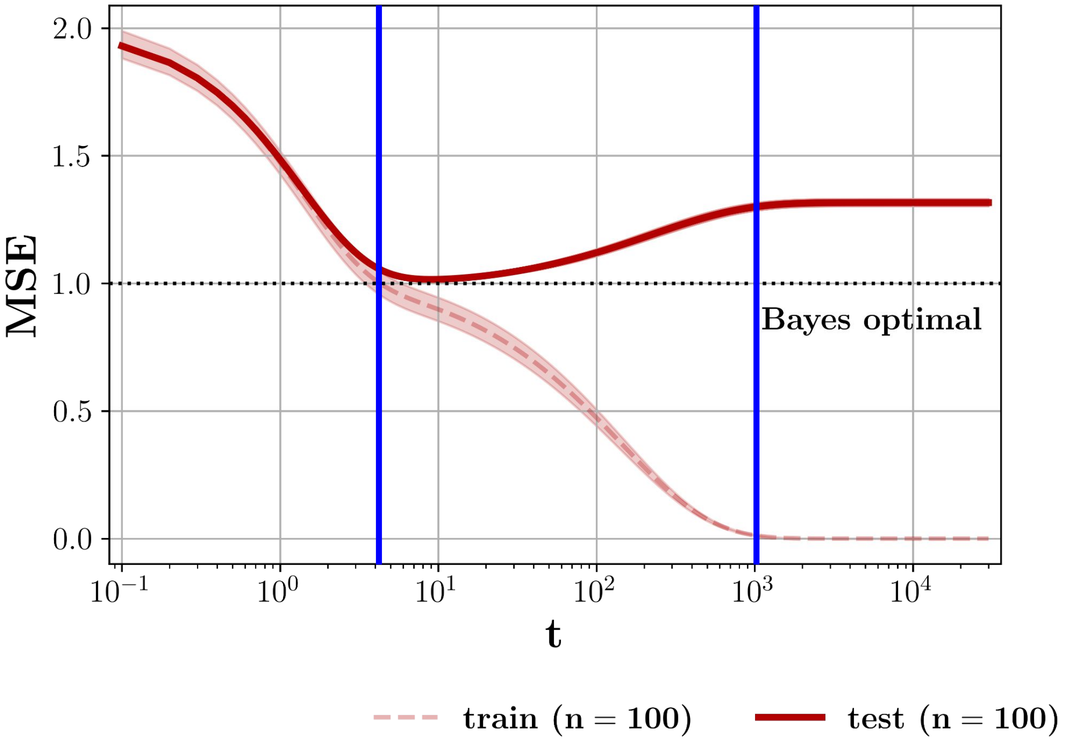
Time dynamics of MLP regression

ReLU MLP
Full batch GD
X = 5-dim sphere
y = 1 + N(0,1)



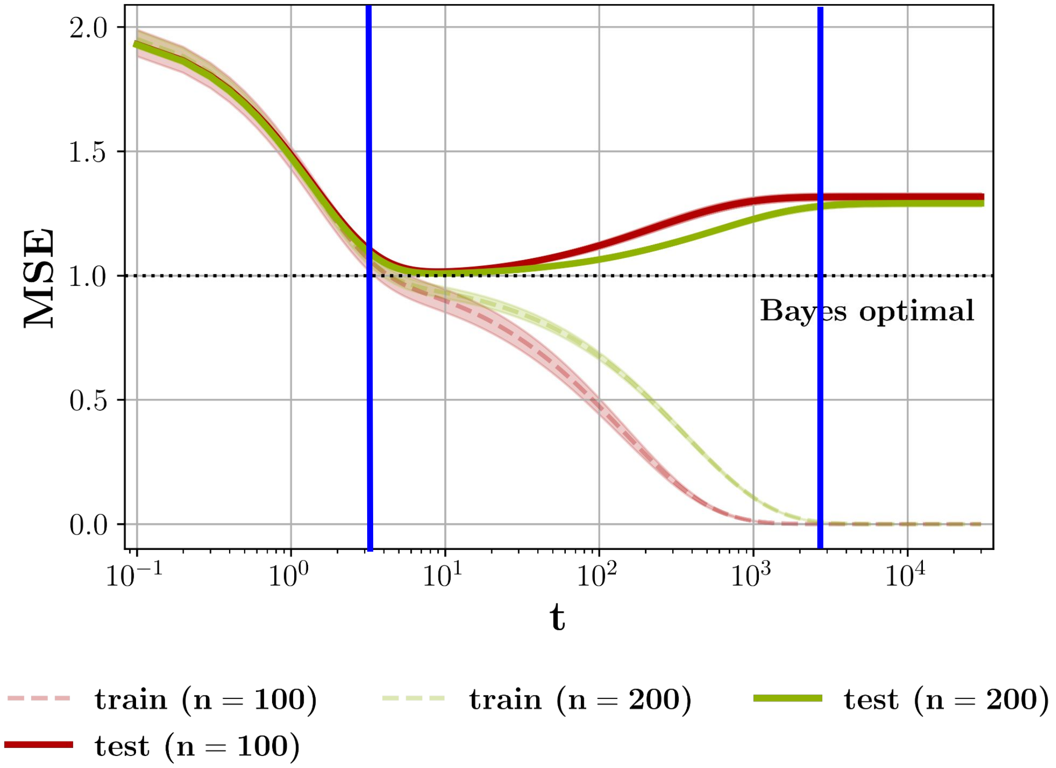
Time dynamics of MLP regression

ReLU MLP
Full batch GD
X = 5-dim sphere
y = 1 + N(0,1)



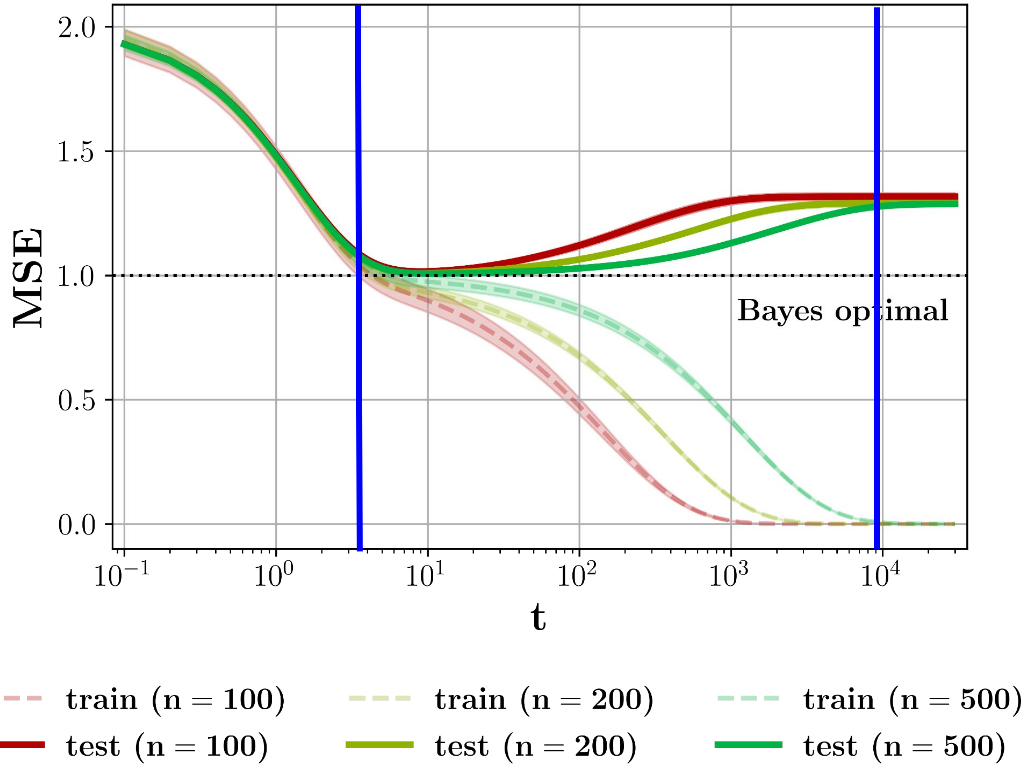
Time dynamics of MLP regression

ReLU MLP
Full batch GD
X = 5-dim sphere
y = 1 + N(0,1)



Time dynamics of MLP regression

ReLU MLP
Full batch GD
X = 5-dim sphere
y = 1 + N(0,1)



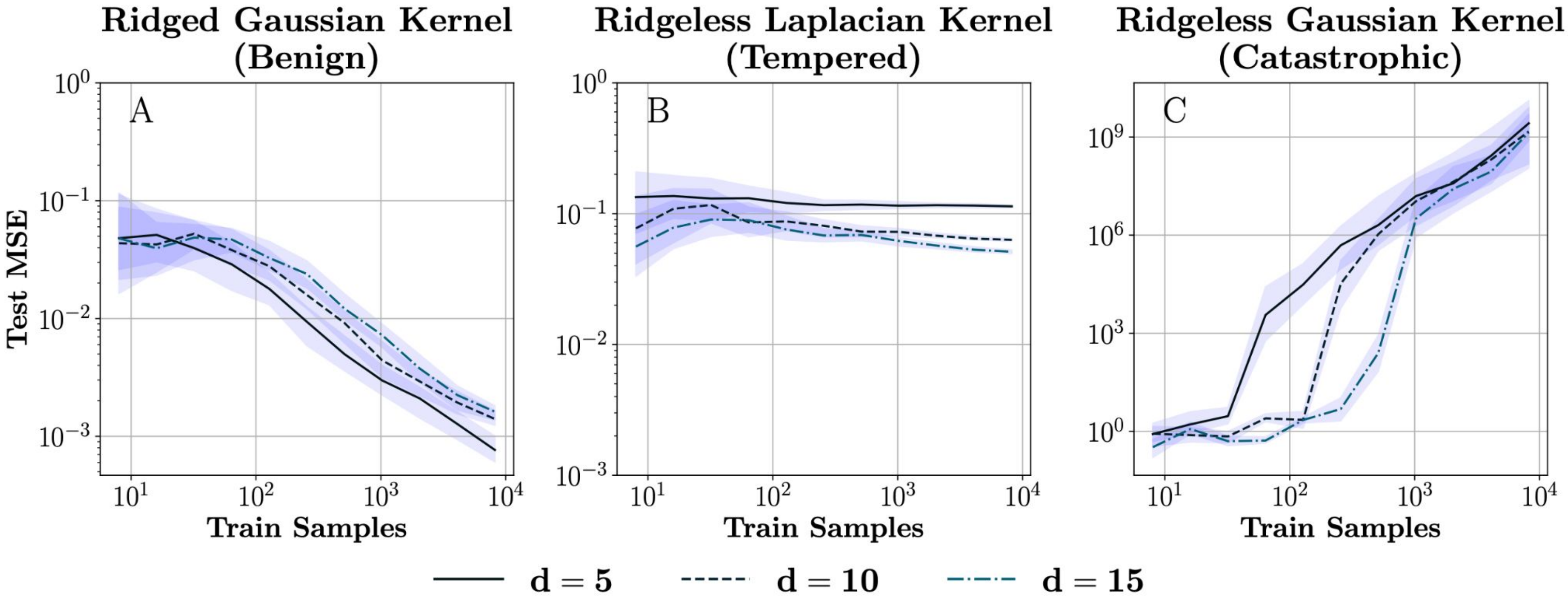
Outline

1. Using the taxonomy

2. Empirical results on deep neural networks (DNNs)

3. Overfitting in kernel regression (KR)

Kernel regression can exhibit all three types of fitting



Kernel regression

$x_i \overset{\text{iid}}{\sim} p, \quad p \text{ is a measure over } \mathbb{R}^d \qquad \mathcal{D} = \{x_i\}_{i=1}^n$

Goal: fit $f(x) = f^*(x) + \eta, \quad f^* : \mathbb{R}^d \to \mathbb{R}, \quad \eta \sim \mathcal{N}(0, \epsilon^2)$

Kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (no special assumptions)

$$\boxed{\hat{f}(x) = K(x, \mathcal{D})(K(\mathcal{D}, \mathcal{D}) + \delta \mathbf{I}_n)^{-1} f(\mathcal{D})}$$

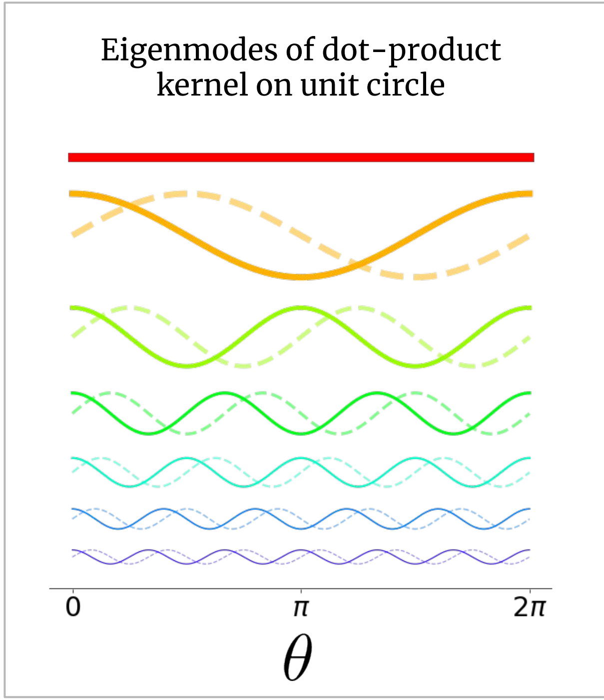
The kernel eigensystem

$$\langle g, h \rangle \equiv \mathbb{E}_x \left[g(x) h(x) \right]$$

Eigensystem:

$$\boxed{\begin{array}{c} \mathbb{E}_{x'} [K(x, x') \phi_i(x')] = \lambda_i \phi_i(x) \\ \langle \phi_i, \phi_j \rangle = \delta_{ij} \qquad \lambda_i \geq 0 \end{array}}$$

Target function: $$f^*(x) = \sum_i \mathbf{v}_i \phi_i(x),$$



KR = linear regression with eigenfunction features

$$K(x_1, x_2) = \sum_i \lambda_i \phi_i(x_1)\phi_i(x_2)$$

With feature map $\boldsymbol{\psi}$ ...

$$\boldsymbol{\psi}(x) \equiv \begin{bmatrix} \lambda_1^{1/2}\phi_1(x) \\ \lambda_2^{1/2}\phi_2(x) \\ \vdots \\ \lambda_i^{1/2}\phi_i(x) \\ \vdots \end{bmatrix}$$

... the feature-feature inner product is $\boldsymbol{K}$ ...

$$\boldsymbol{\psi}(x_1)^\top \boldsymbol{\psi}(x_2) = K(x_1, x_2)$$

... and LR with features $\boldsymbol{\psi}$ is equivalent to KR with kernel $\boldsymbol{K}$.

$$\hat{f}_{\mathrm{LR}}(x) = \boldsymbol{\psi}(x)^\top \left(\boldsymbol{\psi}(\mathcal{D})\boldsymbol{\psi}(\mathcal{D})^\top\right)^+ \boldsymbol{\psi}(\mathcal{D}) f(\mathcal{D})$$

$$\hat{f}_{\mathrm{KR}}(x) = K(x, \mathcal{D}) K(\mathcal{D}, \mathcal{D})^{-1} f(\mathcal{D})$$

Approximation: features are Gaussian and uncorrelated

$$\boldsymbol{\psi}(x) \equiv \begin{bmatrix} \lambda_1^{1/2}\phi_1(x) \\ \lambda_2^{1/2}\phi_2(x) \\ \vdots \\ \lambda_i^{1/2}\phi_i(x) \\ \vdots \end{bmatrix}$$

$$\mathbb{E}_x[\phi_i(x)\phi_j(x)] = \delta_{ij}$$

$$\Rightarrow \mathbb{E}_x[\psi_i(x)\psi_j(x)] = \lambda_i\delta_{ij}$$

"Universality" assumption:

$$\boldsymbol{\psi}(x) \sim \mathcal{N}(0, \mathrm{diag}(\lambda_1, \ldots))$$

Closed-form expression for generalization of LR with Gaussian covariates?

The "eigenlearning" equations

[Hastie et al. '19], [Bordelon et al. '20], [Jacot et al. '20], [Bartlett et al. '21], [Loureiro et al. '21], [Simon et al. '21]

test MSE

noise-fitting MSE

$$\mathcal{R}_n \approx \mathcal{E}_n \equiv \mathcal{E}_0 \left(\sum_i (1 - \mathcal{L}_i)^2 v_i^2 + \sigma^2 \right), \quad \text{where} \quad \mathcal{E}_0 \equiv \frac{n}{n - \sum_j \mathcal{L}_j^2},$$

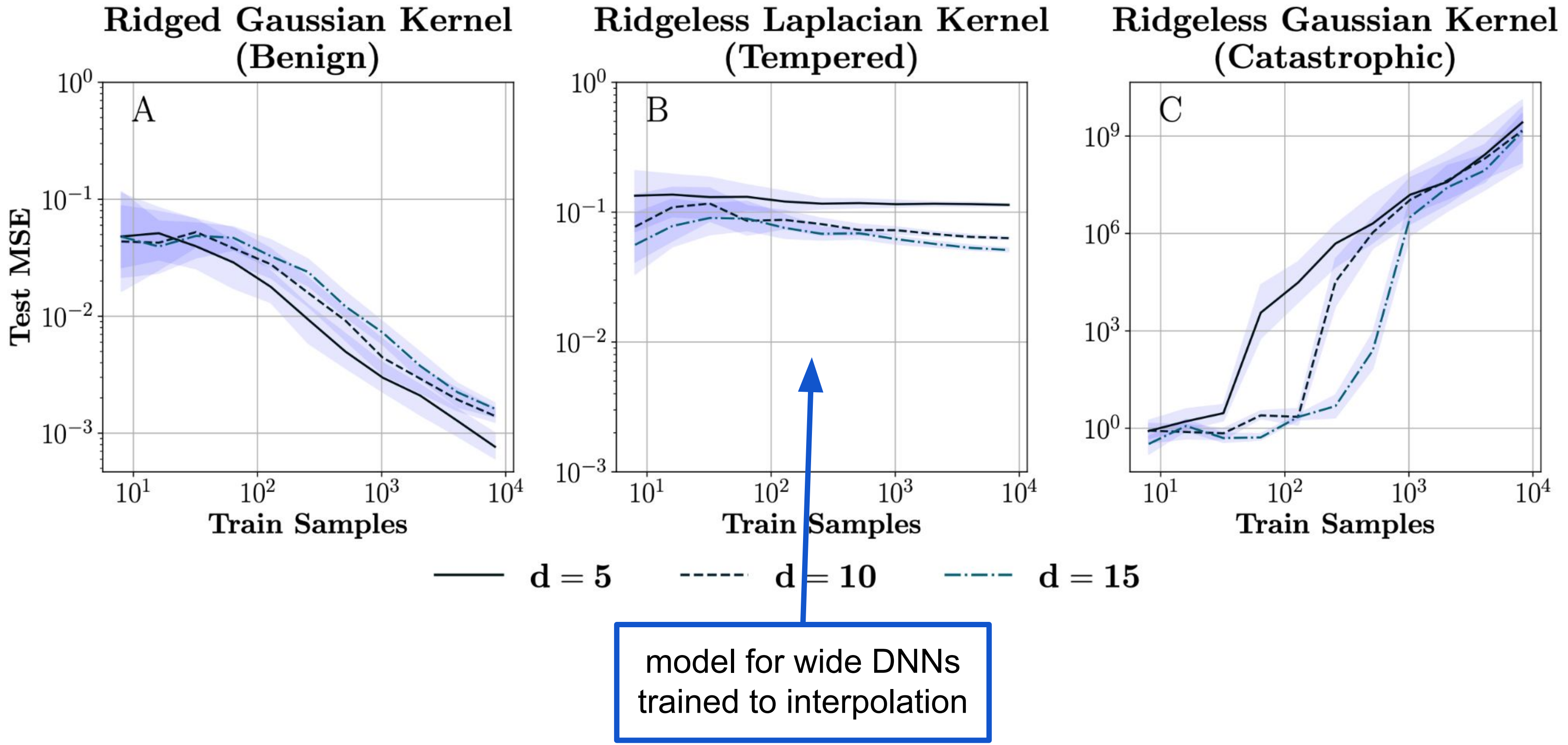
$$\mathcal{L}_i \equiv \frac{\lambda_i}{\lambda_i + C}, \quad \text{and} \quad \underbrace{C \geq 0 \text{ satisfies } \sum_i \frac{\lambda_i}{\lambda_i + C} + \frac{\delta}{C} = n.}_{\text{eigenvalue threshold}}$$

"eigenmode learnability" $\in [0,1]$

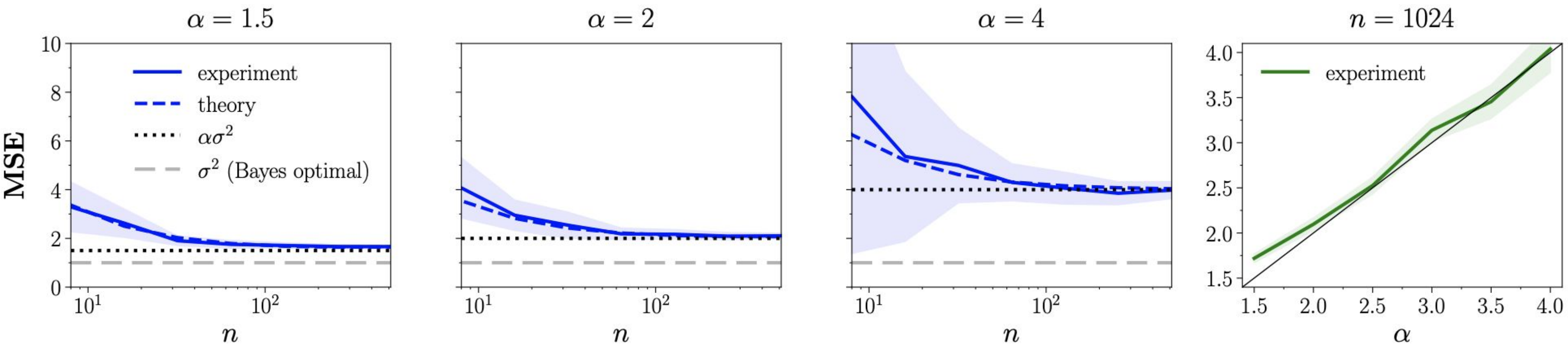
eigenvalue threshold

The Trichotomy Theorem

Spectrum	Limiting risk	
$\delta > 0$ $\delta = 0 \; and \; \lambda_i = i^{-1} \log^{-\alpha} i \; for \; some \; \alpha > 1$	$\lim\limits_{n \to \infty} \mathcal{E}_n = \sigma^2$	**BENIGN**


Ridged Gaussian Kernel
(Benign)
Ridgeless Laplacian Kernel
(Tempered)
Ridgeless Gaussian Kernel
(Catastrophic)
A
B
C
Test MSE
Train Samples
d = 5
d = 10
d = 15
model for wide DNNs
trained to interpolation

Linear regression with $\lambda_i \propto i^{-\alpha}$



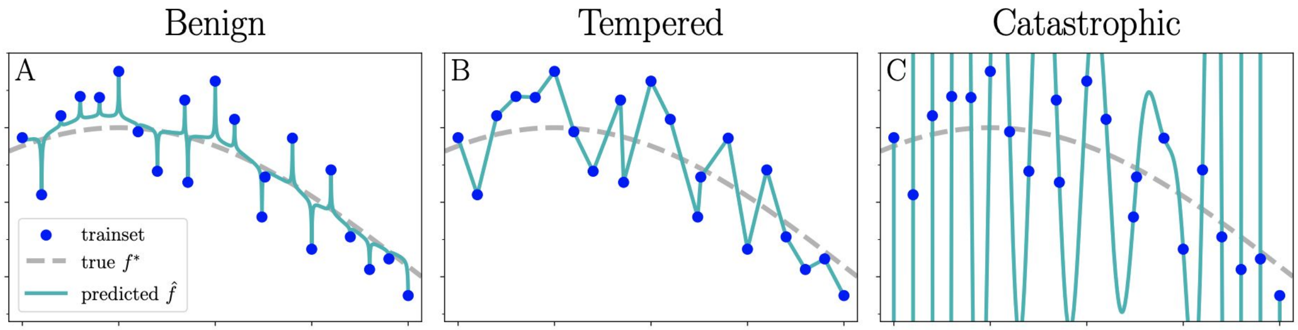
asymptotically "overfitting by a factor of the exponent"!

Implications of the Trichotomy Theorem

- Laplace kernels are **tempered**, ridgeless Gaussian kernels are **catastrophic**
- NTKs' fitting depends on activation function:
  - ReLU -> powerlaw spectrum -> **tempered**
  - other choices (e.g. erf) -> superpowerlaw spectrum -> **catastrophic**
- Ridge parameter ≈ early stopping, so early-stopped DNNs are **benign**

Conclusions

- There are three ways to overfit
- Common interpolating methods fall into the intermediate regime
- For KR, ridge + kernel spectrum determine the regime



Open Questions

1. How do input / manifold dimensionality affect overfitting?
2. Theoretical results "beyond kernels"?
3. Trichotomy theorem for classification?
4. Trichotomy theorem with exhaustive conditions?
5. Do any closed-form kernels give benign overfitting?

References

Bartlett, P., Long, P., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 2020.

Belkin, M., Hsu, D.J., & Mitra, P.P. (2018). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. ArXiv, abs/1806.05161.

Ji, Z., Li, J., and Telgarsky, M. Early-stopped neural networks are consistent. Advances in Neural Information Processing Systems, 34, 2021

Koehler, F., Zhou, L., Sutherland, D., and Srebro, N. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overtting. Advances in Neural Information Processing Systems, 34, 2021.

Liang, T. & Rakhlin, A. Just interpolate: Kernel" ridgeless" regression can generalize. arXiv preprint arXiv:1808.00387, 2018. 3, 1.2

Mei, S. & Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. arXiv preprint arXiv:1908.05355, 2019. 3, 1.2, 2.4

Rakhlin, A. & Zhai, X.. (2019). Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. <i>Proceedings of the Thirty-Second Conference on Learning Theory</i>, in <i>Proceedings of Machine Learning Research</i> 99:2595-2623 Available from https://proceedings.mlr.press/v99/rakhlin19a.html.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. ArXiv, abs/1611.03530.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64, 3 (March 2021), 107–115. https://doi.org/10.1145/3446776

References (cont.)

Bordelon, Blake, Abdulkadir Canatar, and Cengiz Pehlevan. "Spectrum dependent learning curves in kernel regression and wide neural networks." International Conference on Machine Learning. PMLR, 2020.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nature Communications, 12(1):1–12, 2021.

Bartlett, Peter L., Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint." Acta numerica 30 (2021): 87-201.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. In Advances in Neural Information Processing Systems 33, 2020b.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. arXiv, abs/1903.08560, 2020.

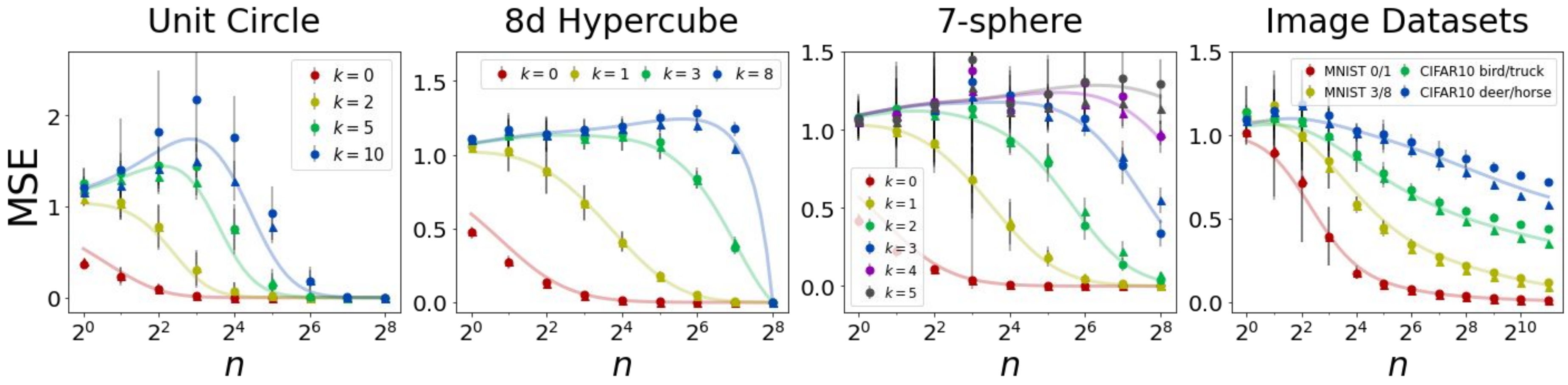
Simon, J. B., Dickens, M., and DeWeese, M. R. Neural tangent kernel eigenvalues accurately predict generalization. arXiv, abs/2110.03922, 2021.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborova, L. Learning curves of generic features maps for realistic datasets with a teacher student model. In Advances in Neural Information Processing Systems 34, 2021.

Beaglehole, D., Belkin, M., Pandit, P. Kernel Ridgeless Regression is Inconsistent in Low Dimensions. arXiv, abs/2205.13525, 2022.

Eigenlearning theory closely matches experiment in many real and synthetic tasks

width 500 FCNs (circles), NTK regression (triangles), theoretical "eigenlearning" predictions (solid curves)



Details on Formalism

Data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathbb{R}$
Estimator $f : \mathcal{X} \to \mathbb{R}$
Pop. Risk (MSE): $R(f) = \mathbb{E}_{\mathcal{D}}[(f(x) - y)^2]$
Optimal regressor: $f^* = \arg\min_f R(f) = \mathbb{E}_{\mathcal{D}}[y|x]$
Bayes risk: $R^* = R(f^*)$

$$R(f) = \mathbb{E}_{\mathcal{D}}[(f(x) - f^*(x))^2] + R^*$$

Details on Formalism

$\mathcal{D}_n \sim \mathcal{D}$ - n iid samples
$\hat{f}_n : \mathcal{X} \to \mathbb{R}$ - estimator trained on $\mathcal{D}_n$
Expected regressor: $\bar{f}_n = \mathbb{E}_{\hat{f}_n}[\hat{f}_n]$

$$\begin{aligned} R_n &= \mathbb{E}_{\hat{f}_n}[R(\hat{f}_n)] \\ &= \mathbb{E}_{\hat{f}_n}[\mathbb{E}_{\mathcal{D}}[(\hat{f}_n(x) - f^*(x))^2]] + R^* \\ &= \mathbb{E}_{\mathcal{D}}[(\bar{f}_n(x) - f^*(x))^2] + \mathbb{E}_{\hat{f}_n,\mathcal{D}}[(\bar{f}_n(x) - \hat{f}_n(x))^2] + R^* \\ &= B_n^2 + V_n + R^* \end{aligned}$$

Consistency: $B_n, V_n \to 0, n \to \infty$
Inconsistent: $B_n \not\to 0$ or $V_n \not\to 0$ or both

Details on Formalism

$$R_n = B_n^2 + V_n + R^*$$

Benign / Consistent: $B_n, V_n \to 0, n \to \infty$
Tempered / Inconsistent: $B_n \to O(1)$ or $V_n \to O(1)$ or both, $n \to \infty$

- both $B_n, V_n \not\to \infty, n \to \infty$
- both $B_n, V_n \not\to 0, n \to \infty$

Catastrophic / Inconsistent: $B_n \to \infty$ or $V_n \to \infty$ or both, $n \to \infty$

Details on Formalism

$R_n = B_n^2 + V_n + R^*$

Benign / Consistent: $B_n, V_n \to 0, n \to \infty$
Tempered / Inconsistent: $B_n \to O(1)$ or $V_n \to O(1)$ or both, $n \to \infty$

- both $B_n, V_n \not\to \infty, n \to \infty$
- both $B_n, V_n \not\to 0, n \to \infty$
- conjecture: $B_n \to 0, V_n \to O(1), n \to \infty$?

Catastrophic / Inconsistent: $B_n \to \infty$ or $V_n \to \infty$ or both, $n \to \infty$