Reconsidering Overfitting in the Age of Overparameterized Models

slides & refs

NeurIPS 2023 Tutorial, New Orleans

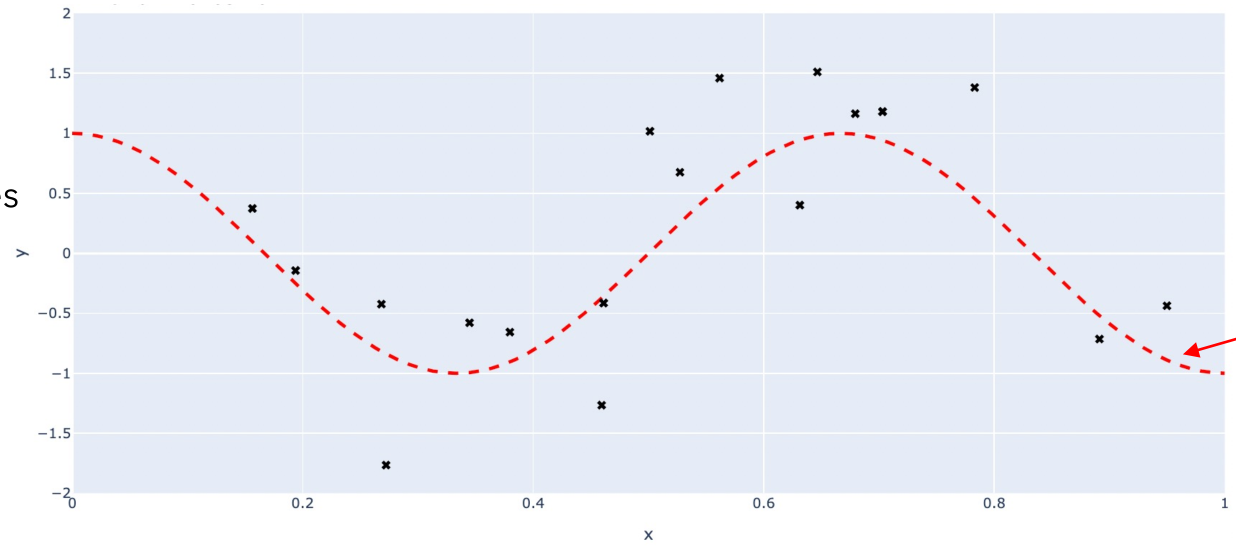
Speakers: Spencer Frei, Vidya Muthukumar, Fanny Yang, Moderator: Daniel Hsu




UC DAVIS UNIVERSITY OF CALIFORNIA
Georgia Tech
ETH zürich
COLUMBIA UNIVERSITY

Textbook wisdom on overfitting

n = 20 samples



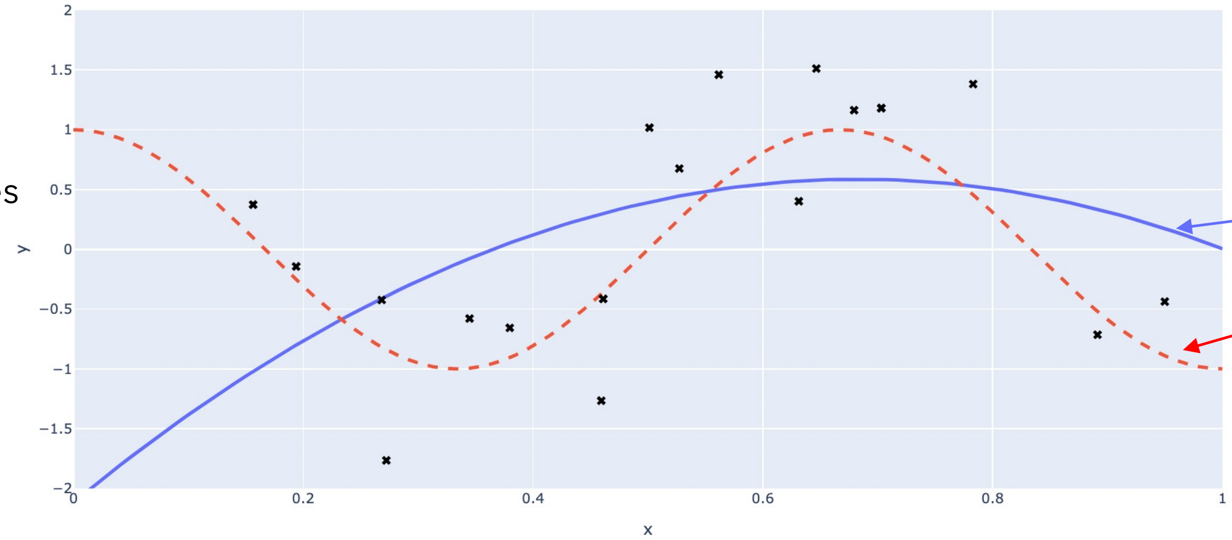
✖ random x_i, y_i noisy version of $f^\star(x_i)$

true $f^\star(x)$

Textbook wisdom on overfitting

n = 20 samples

polynomial fit
degree d = 2



✖ random x_i,
y_i noisy version of $f^\star(x_i)$

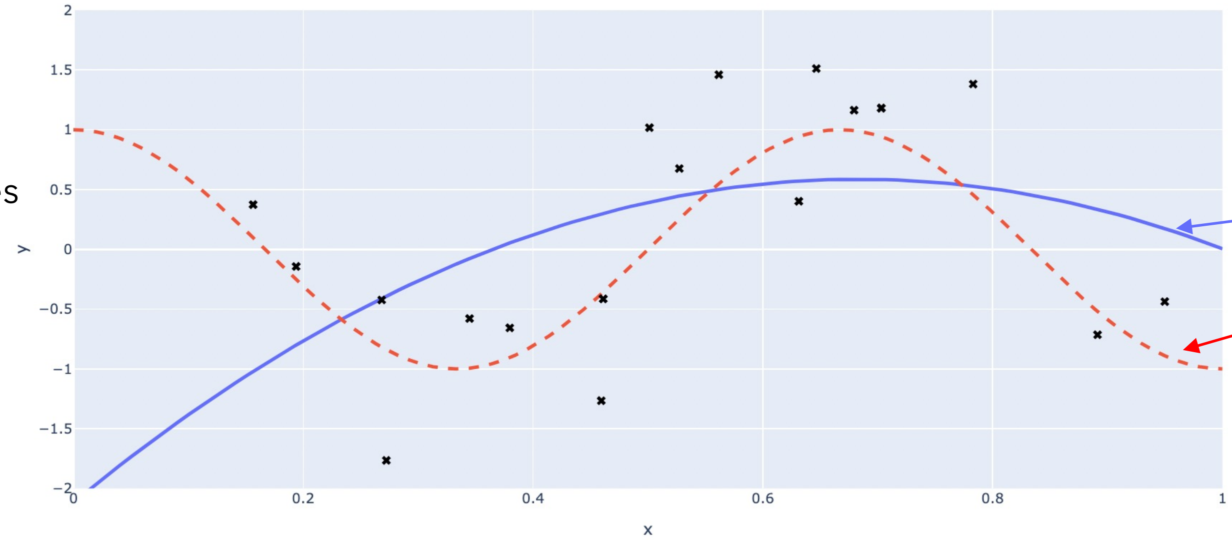
predicted $\hat{f}(x)$

true $f^\star(x)$

Textbook wisdom on overfitting

n = 20 samples

polynomial fit
degree d = 2



✖ random x_i, y_i noisy version of $f^\star(x_i)$

predicted $\hat{f}(x)$

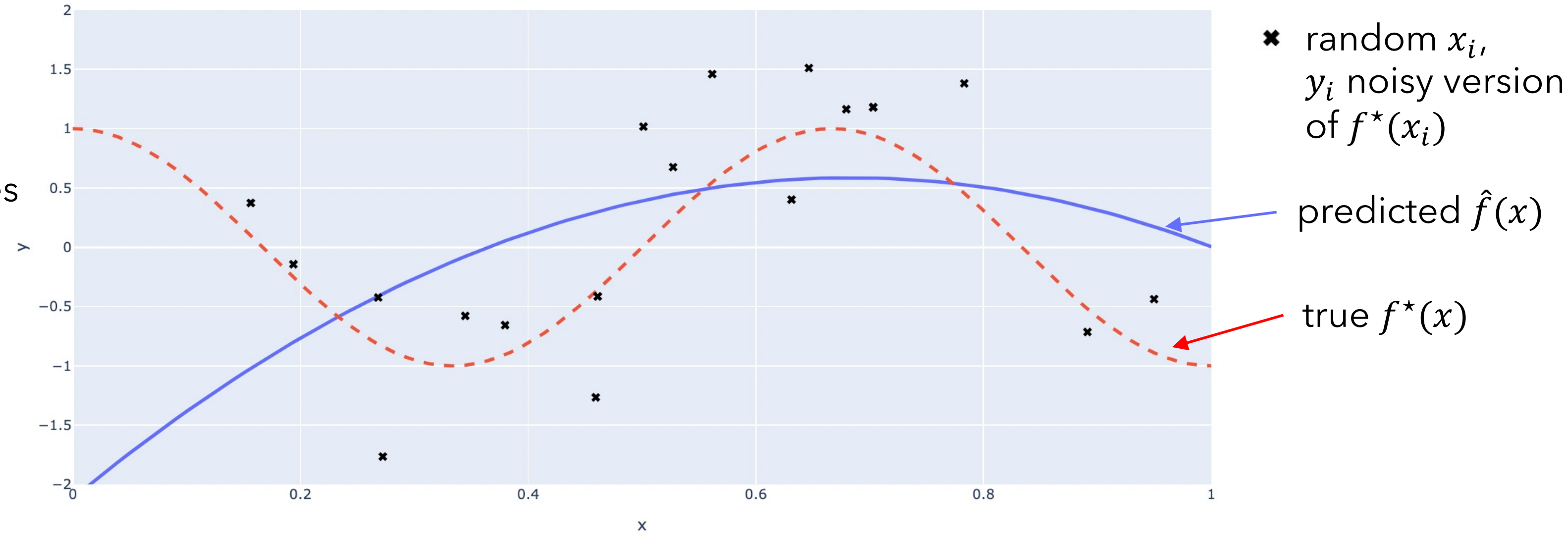
true $f^\star(x)$

Small models cannot fit perfectly:
- cannot express function of interest (high statistical bias)

Textbook wisdom on overfitting

n = 20 samples

polynomial fit
degree d = 2



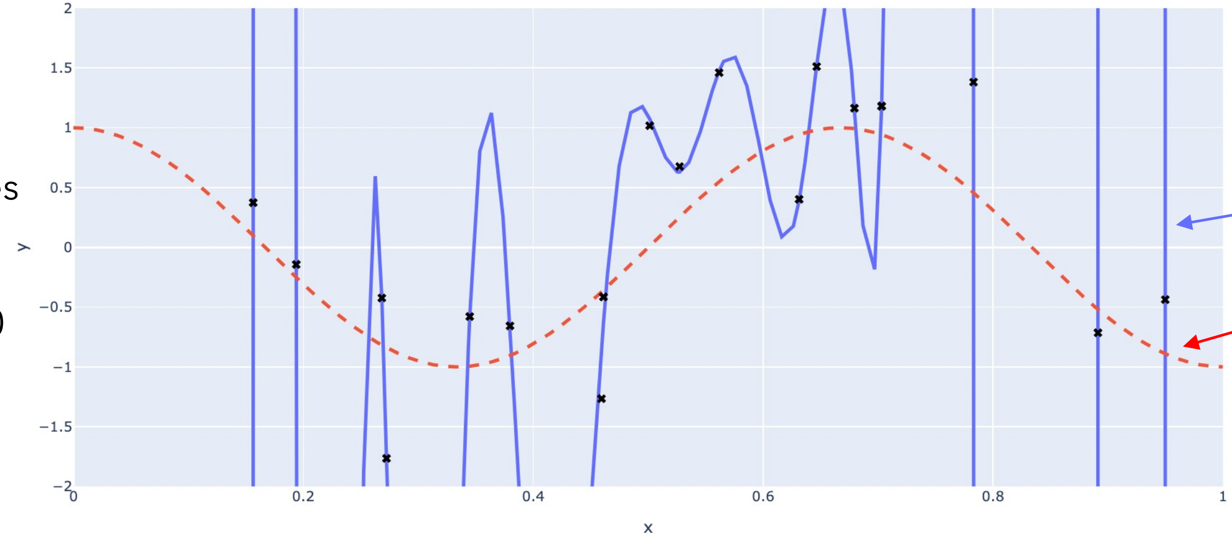
Small models cannot fit perfectly:

- cannot express function of interest (high statistical bias)
- largely ignores noise → does not fluctuate a lot (small variance)

Textbook wisdom on overfitting

n = 20 samples

polynomial fit
degree d = 20



✖ random x_i, y_i noisy version of $f^\star(x_i)$

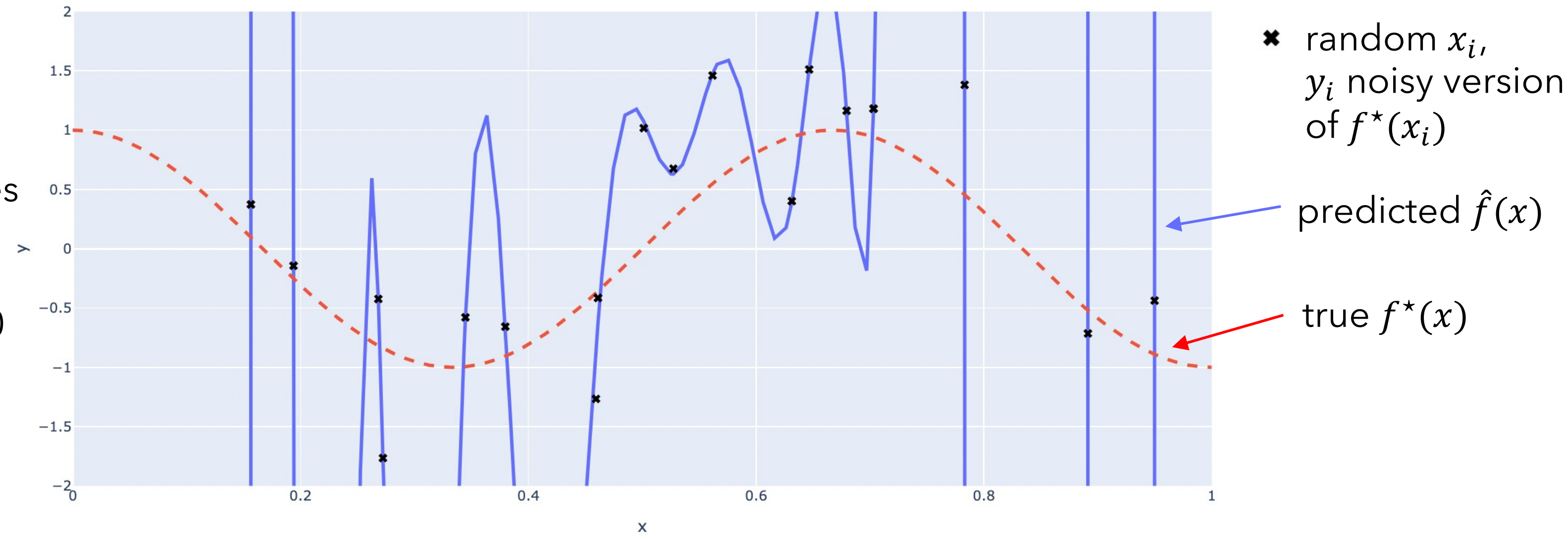
predicted $\hat{f}(x)$

true $f^\star(x)$

Textbook wisdom on overfitting

n = 20 samples

polynomial fit
degree d = 20



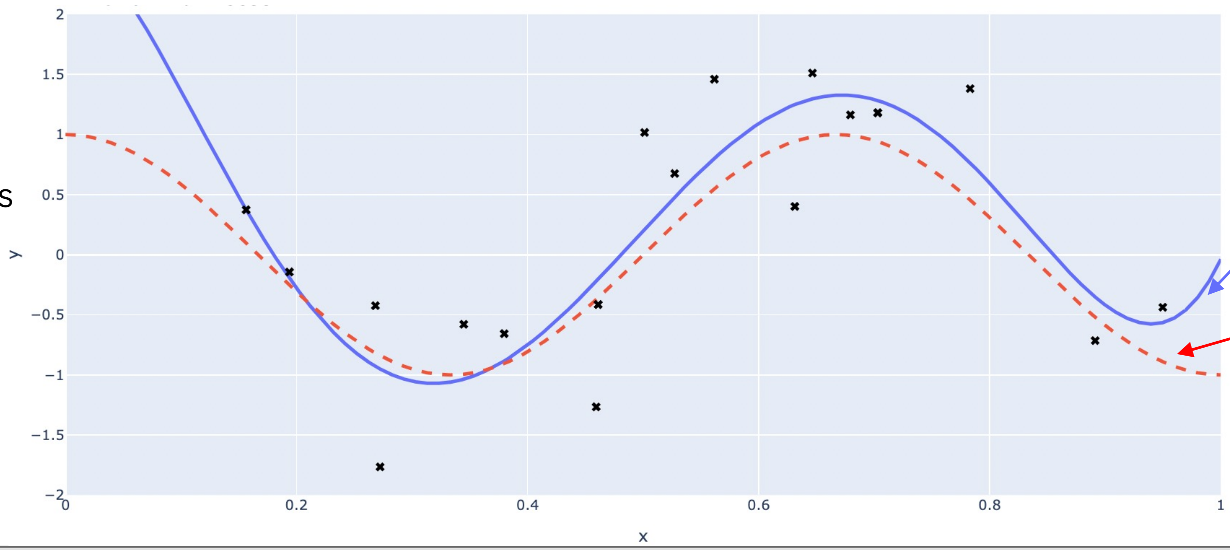
Large models fit perfectly (overfit):

- flexible and can express function of interest (small bias)
- fits too much of the noise (overfit) → fluctuates a lot (high variance)

Textbook wisdom: Avoid fitting noise

n = 20 samples

polynomial fit
degree d = 5



✖ random x_i, y_i noisy version of $f^\star(x_i)$

predicted $\hat{f}(x)$

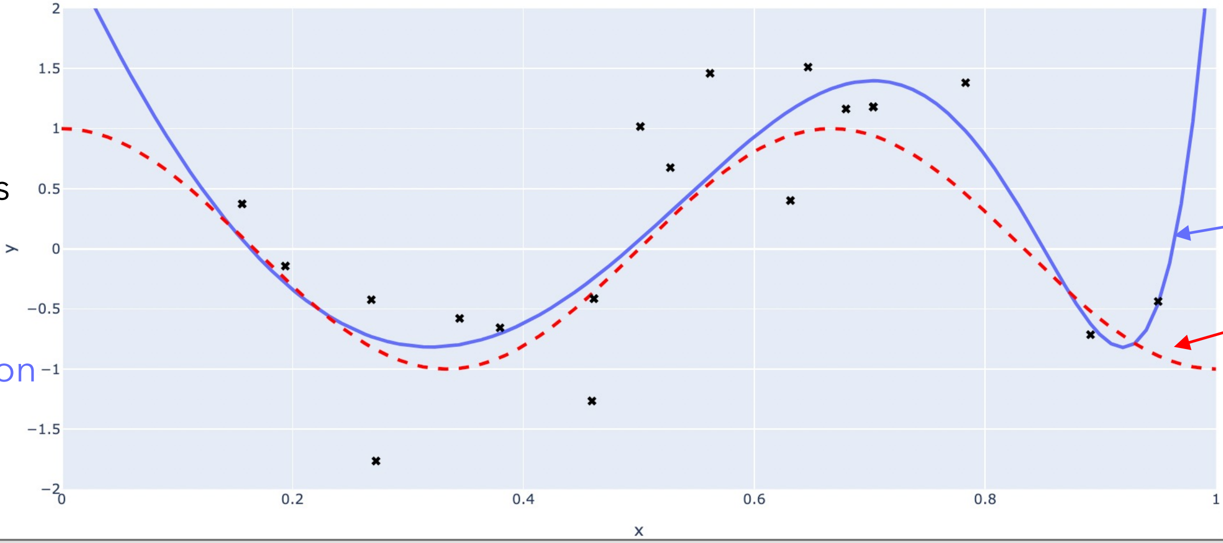
true $f^\star(x)$

Classical theory: Improve generalization by optimizing expressivity via bias-variance trade-off

Textbook wisdom: Avoid fitting noise

n = 20 samples

polynomial fit
degree d = 20
w/ regularization



random x_i, y_i noisy version of $f^\star(x_i)$

predicted $\hat{f}(x)$

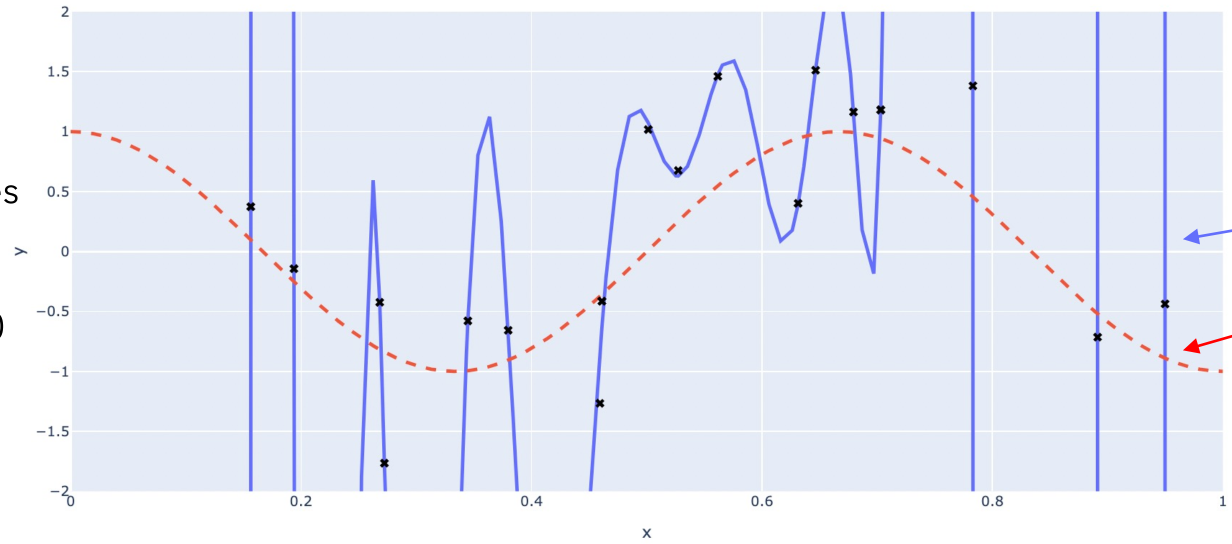
true $f^\star(x)$

Classical theory: Improve generalization by optimizing expressivity via bias-variance trade-off

Textbook wisdom on overfitting

n = 20 samples

polynomial fit
degree d = 20



✖ random x_i, y_i noisy version of $f^\star(x_i)$

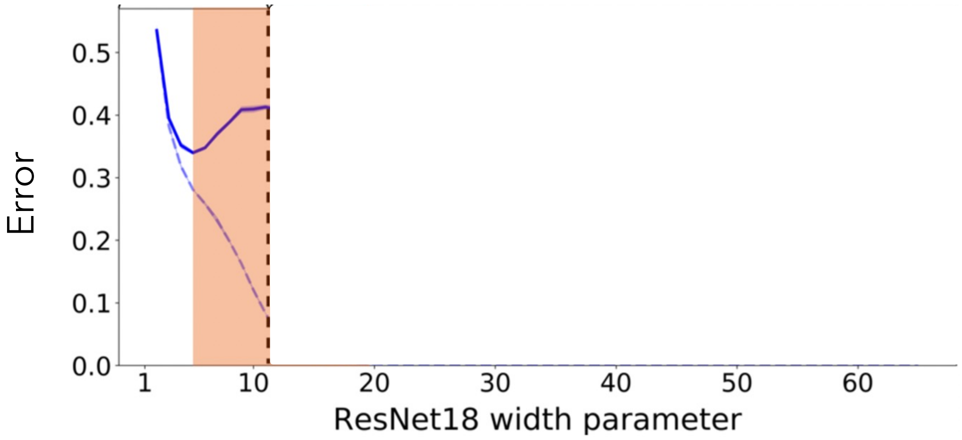
predicted $\hat{f}(x)$

true $f^\star(x)$

What happens if we increase the polynomial degree even further without regularizing?

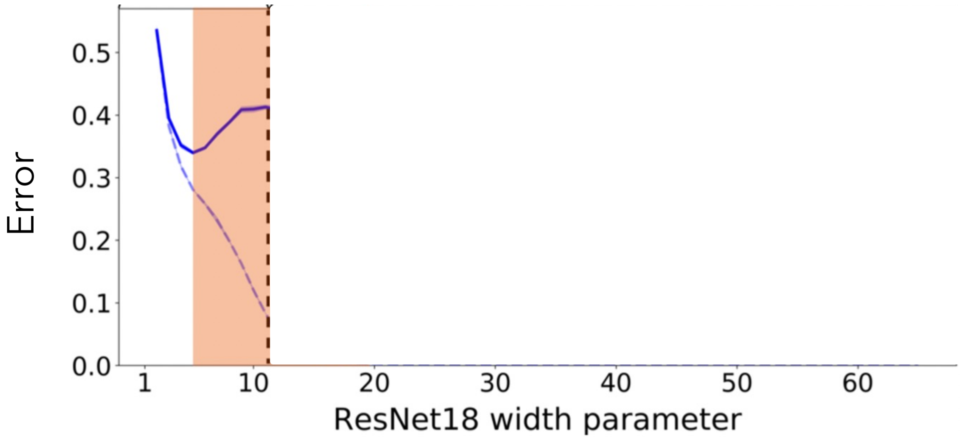
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*

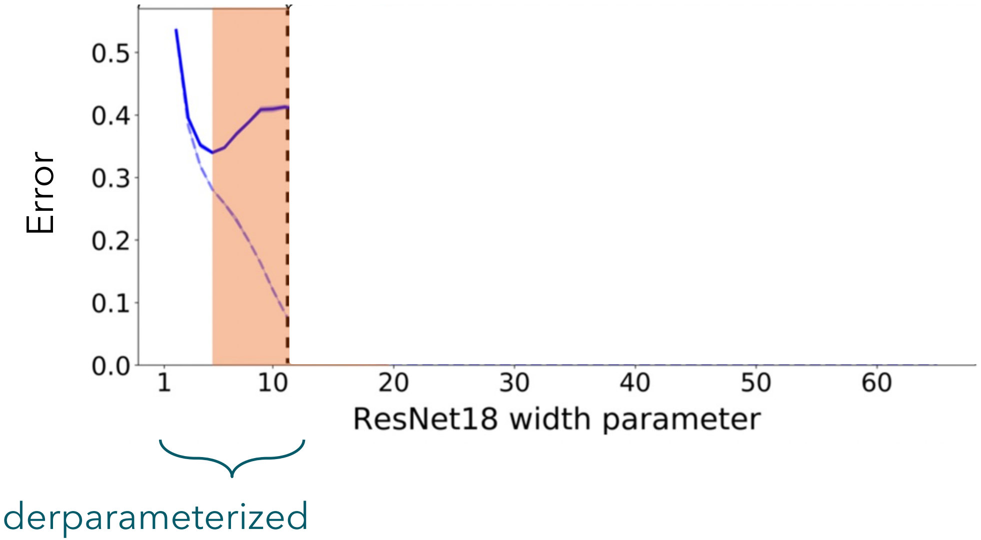


underparameterized

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Double descent on neural networks

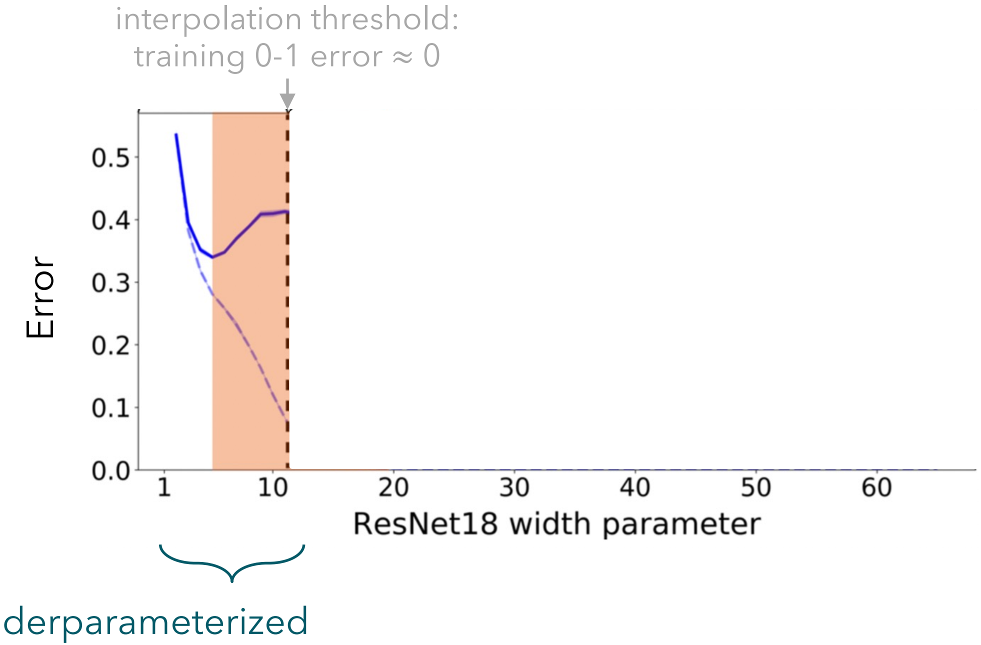
Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Double descent on neural networks

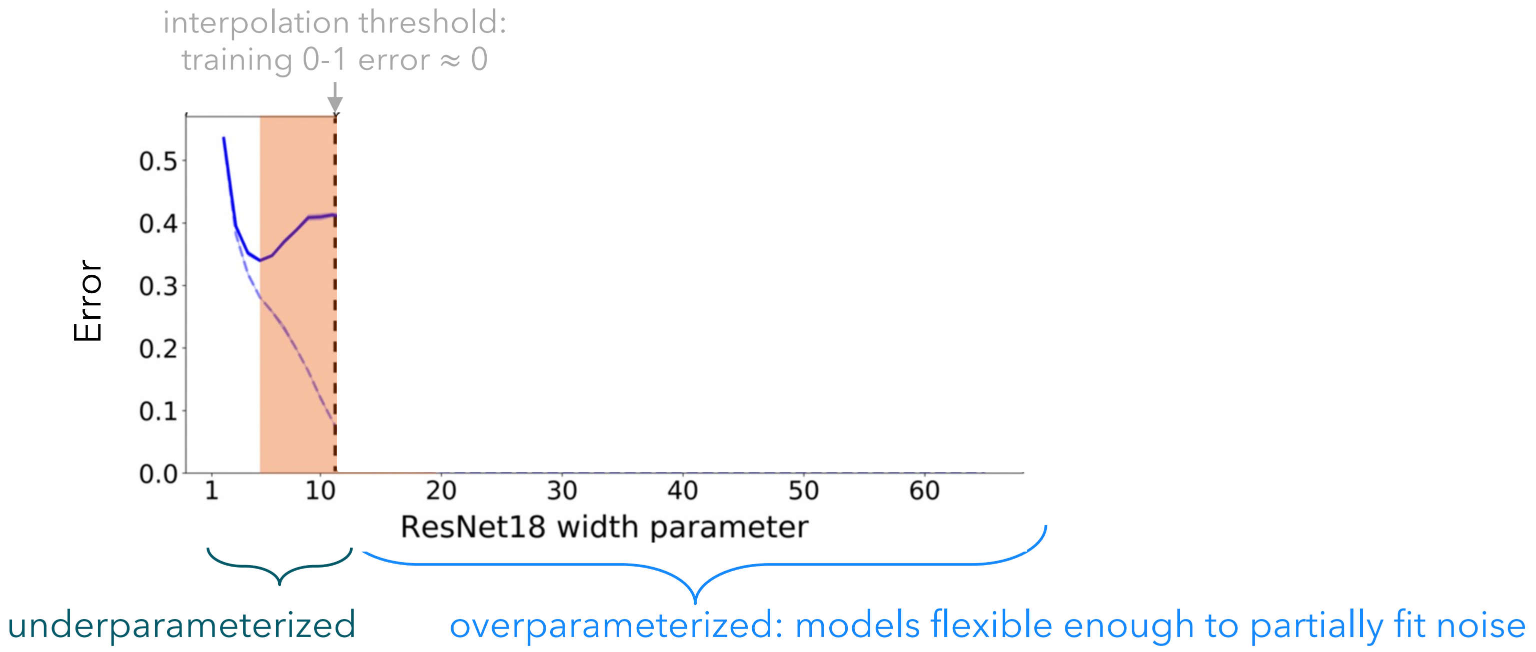
Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*

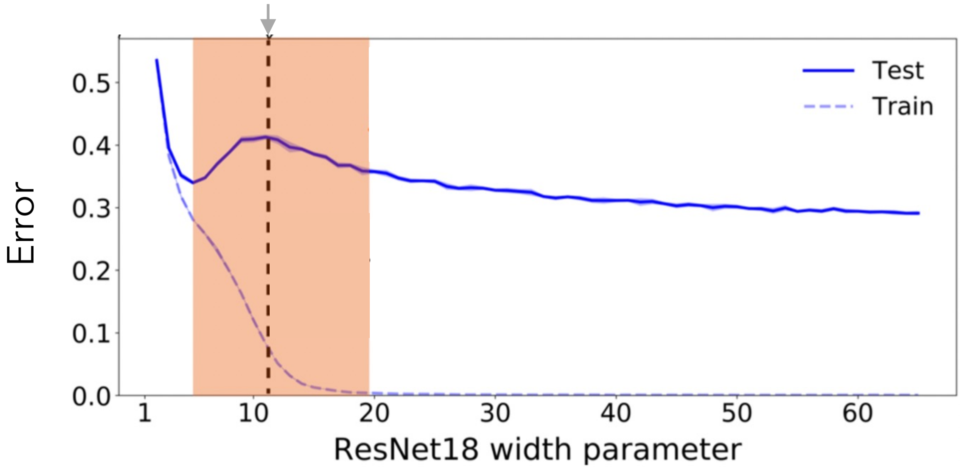


[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*

interpolation threshold:
training 0-1 error ≈ 0



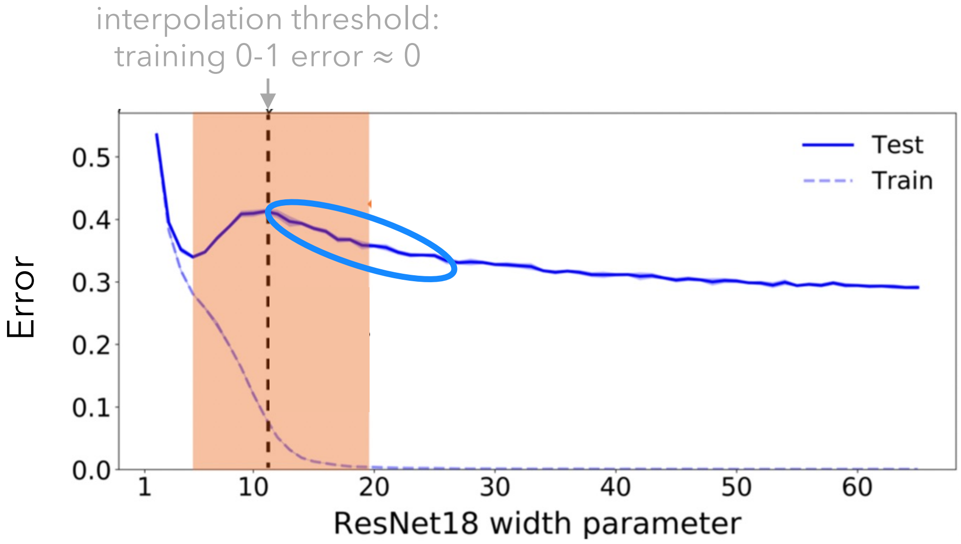
underparameterized

overparameterized: models flexible enough to partially fit noise

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. I: Second descent beyond interpolation

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*

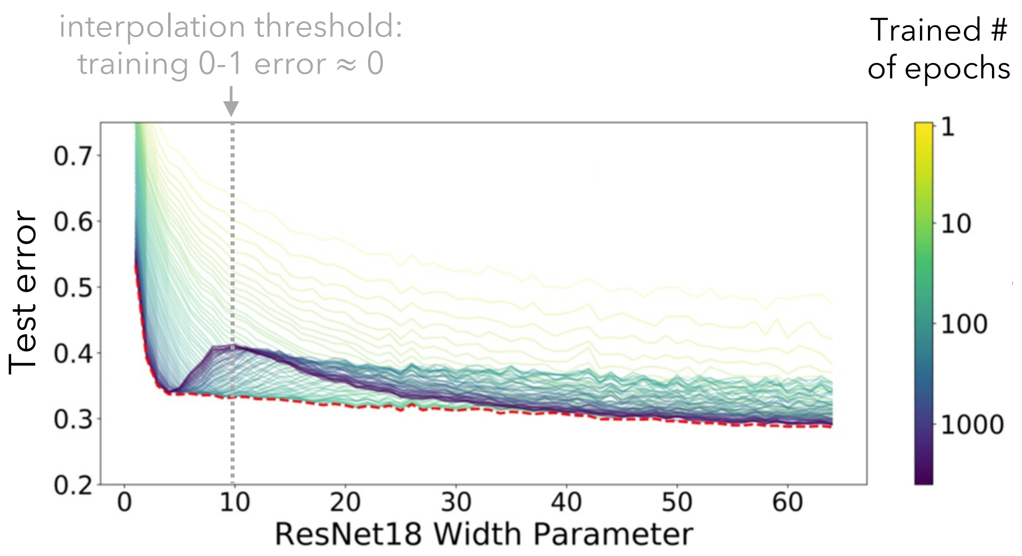


1 After interpolation threshold, we have a second "descent" (double descent) for interpolators

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. II: Harmless interpolation for large models

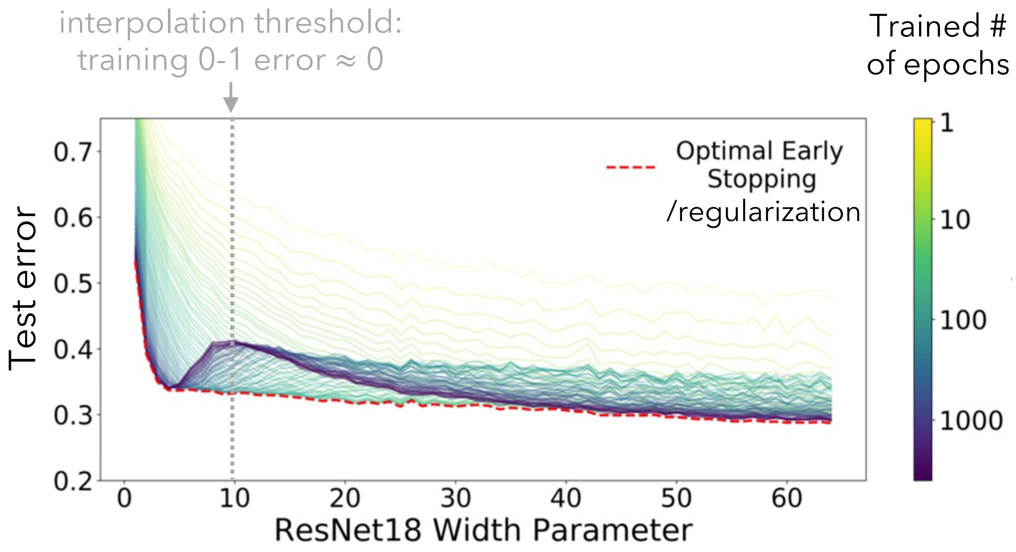
Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. II: Harmless interpolation for large models

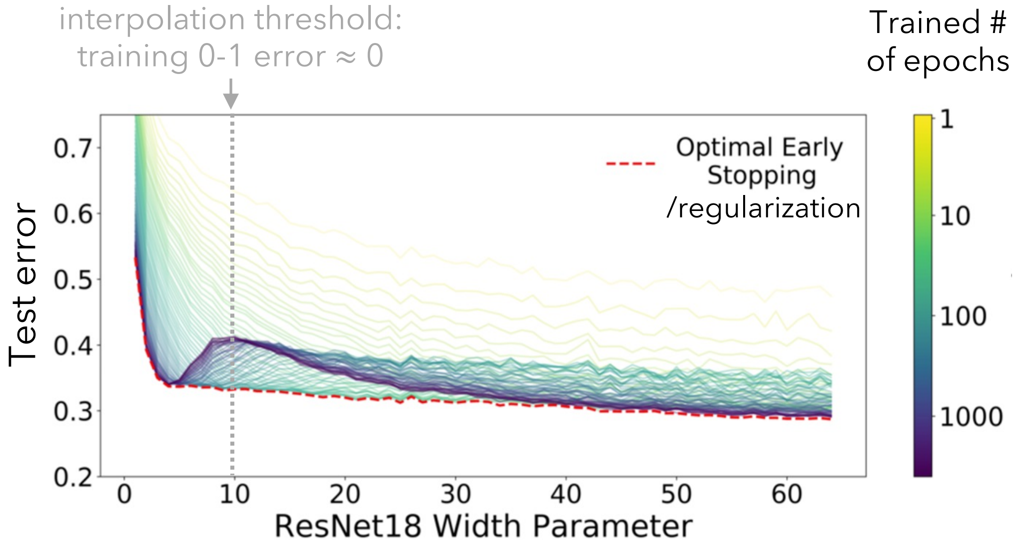
Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*

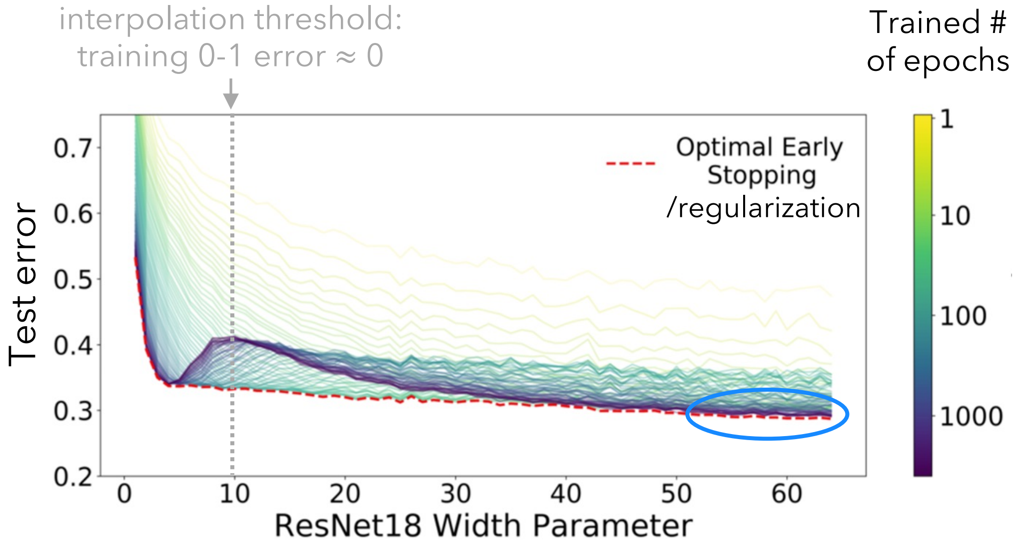


compare dark blue (at convergence) with red dashed (best stopping time)

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*

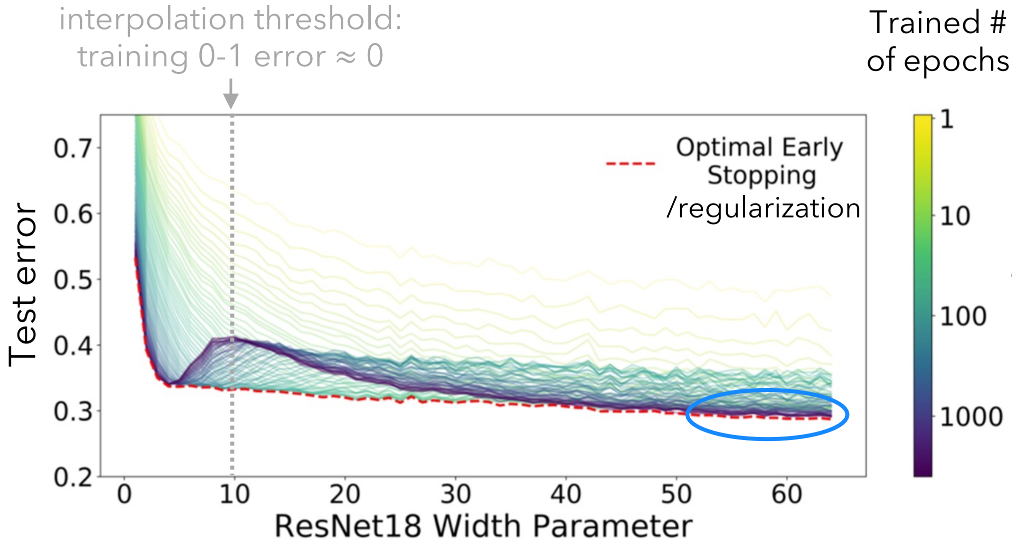


compare dark blue
(at convergence)
with red dashed
(best stopping time)

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



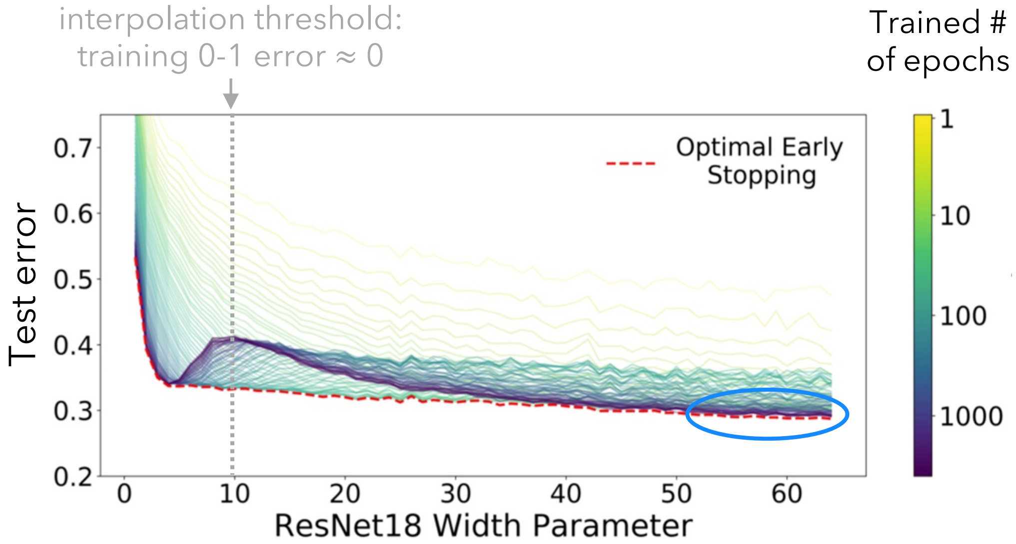
compare dark blue
(at convergence)
with red dashed
(best stopping time)

2 For large models, interpolation is not worse than regularization (harmless interpolation)

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Obs. III: Good generalization for large models

Classification using neural networks and Adam on CIFAR-10 *with 15% additional label noise*



3

For large models, we achieve reasonably good test accuracy

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '20]

Textbooks need an update…

uploaded 2016

DOI:10.1145/3446776

Understanding Deep Learning (Still) Requires Rethinking Generalization

By **Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht,** and **Oriol Vinyals**

Communications of the ACM, 2021

panelist today

*and many more papers that expressed the need for “rethinking”

Textbooks need an update…

uploaded 2016

DOI:10.1145/3446776

Understanding Deep Learning (Still) Requires Rethinking Generalization

By Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals

added 2021

panelist today

Communications of the ACM, 2021

*and many more papers that expressed the need for "rethinking"

What the field set out to understand…

Try to understand when the following happens:

1. Second "descent" as model size grows grows beyond interpolation threshold

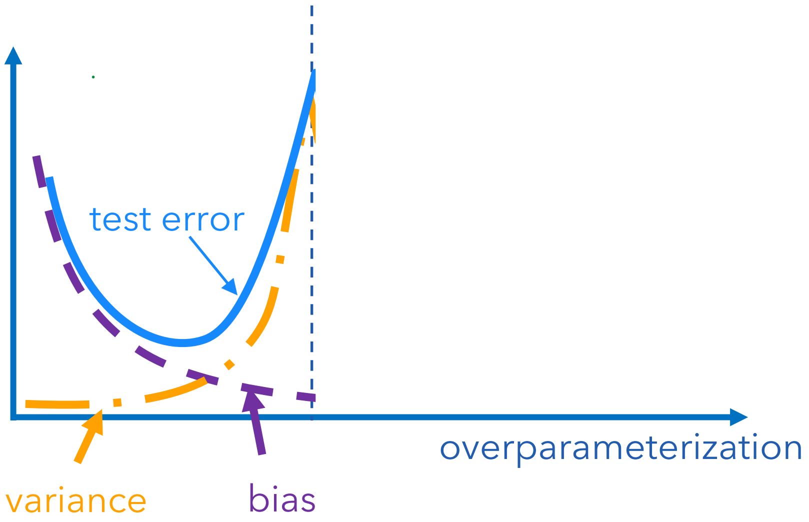
2. Harmless interpolation for large models, i.e. interpolation ~ opt. regularization

3. Good test performance for large models, close to best possible prediction error

What the field set out to understand…

Try to understand when the following happens:

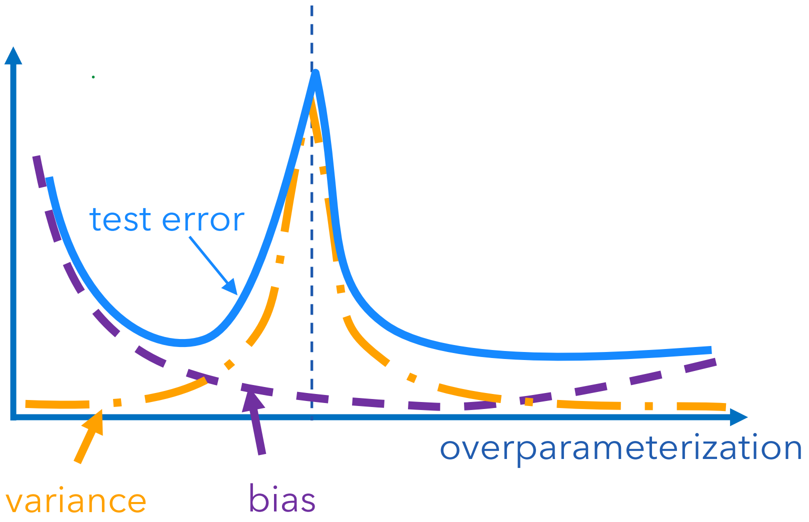
1. Second "descent" as model size grows grows beyond interpolation threshold
2. Harmless interpolation for large models, i.e. interpolation ~ opt. regularization
3. Good test performance for large models, close to best possible prediction error



What the field set out to understand…

Try to understand when the following happens:

1. Second "descent" as model size grows grows beyond interpolation threshold
2. Harmless interpolation for large models, i.e. interpolation ~ opt. regularization
3. Good test performance for large models, close to best possible prediction error



What the field set out to understand…

Try to understand when the following happens:

As overparameterization ↑:

1. Second "descent" as model size grows grows beyond interpolation threshold
2. Harmless interpolation for large models, i.e. interpolation ~ opt. regularization
3. Good test performance for large models, close to best possible prediction error

variance decays → (1), (2)

bias stays low → (3)

What the field set out to understand…

Try to understand when the following happens:

1. Second "descent" as model size grows grows beyond interpolation threshold
2. Harmless interpolation for large models, i.e. interpolation ~ opt. regularization
3. Good test performance for large models, close to best possible prediction error

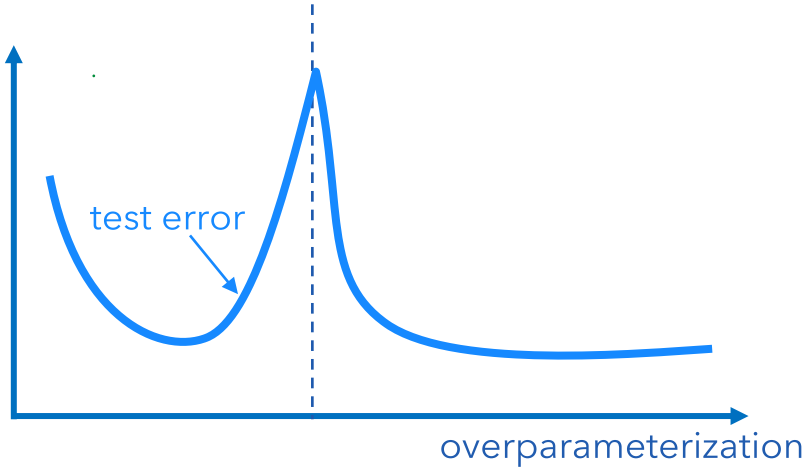
As overparameterization ↑:

- variance decays
- bias stays low

when is this the case?

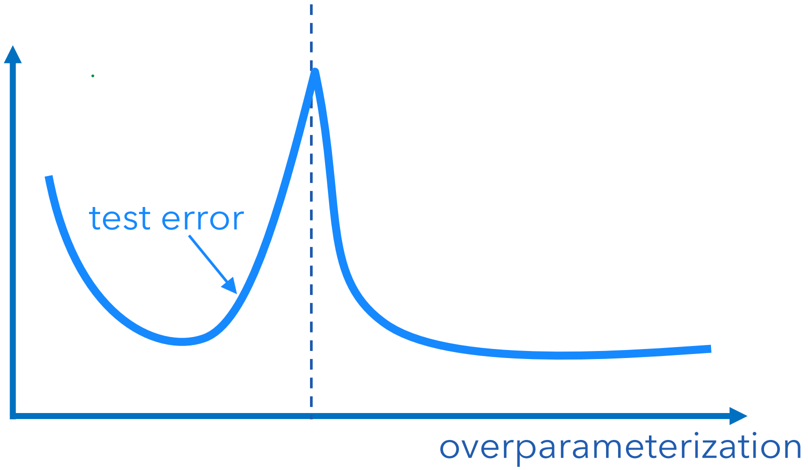
Which factors govern…

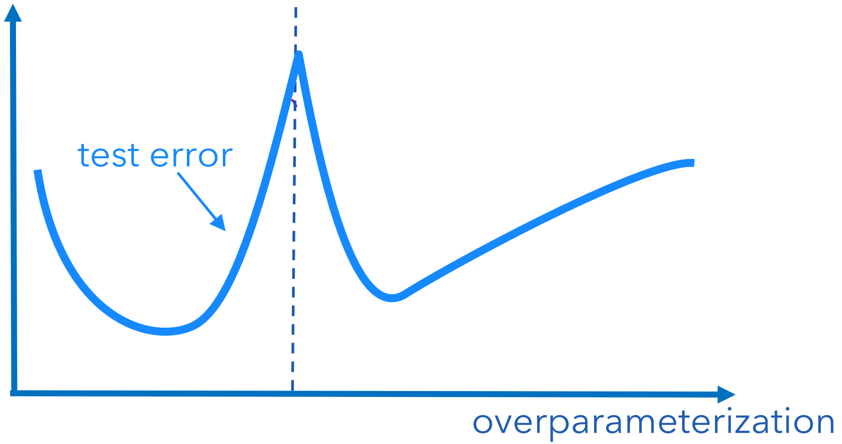
when we have this picture…



Which factors govern…

when we have this picture…





…rather than this picture

Seeking answers using theoretical analysis...

Neural network interpolators

- feature learning with
  overparameterization ≜
  e.g. width of hidden layers

- found w/ 1st order methods to
  minimize **non-convex losses**

← complexity to analyze model

Seeking answers using theoretical analysis...

Neural network interpolators → Kernel / random features

- feature learning with overparameterization ≜ e.g. width of hidden layers
- found w/ 1st order methods to minimize **non-convex losses**

- using p nonlinear features w/ overparameterization ≜ number of features $p \gg n$
- found w/ 1st order methods to minimize a **convex loss**

← complexity to analyze model

Seeking answers using theoretical analysis…

Neural network interpolators → Kernel / random features → Linear interpolators

Neural network interpolators

- feature learning with overparameterization ≜ e.g. width of hidden layers
- found w/ 1st order methods to minimize **non-convex losses**

Kernel / random features

- using p nonlinear features w/ overparameterization ≜ number of features $p \gg n$
- found w/ 1st order methods to minimize a **convex loss**

Linear interpolators

- using d input features with overparameterization ≜ dimension $d \gg n$
- found w/ 1st order methods to minimize a **convex loss**

← complexity to analyze model

Seeking answers using theoretical analysis…

Neural network interpolators → Kernel / random features → Linear interpolators

Neural network interpolators

- feature learning with overparameterization ≜ e.g. width of hidden layers
- found w/ 1st order methods to minimize **non-convex losses**

Kernel / random features

- using p nonlinear features w/ overparameterization ≜ number of features $p \gg n$
- found w/ 1st order methods to minimize a **convex loss**

Linear interpolators

- using d input features with overparameterization ≜ dimension $d \gg n$
- found w/ 1st order methods to minimize a **convex loss**

← complexity to analyze model

Seeking answers using theoretical analysis…

Neural network interpolators

- feature learning with overparameterization ≜ e.g. width of hidden layers
- found w/ 1st order methods to minimize **non-convex losses**

Kernel / random features

- using p nonlinear features w/ overparameterization ≜ number of features $p \gg n$
- found w/ 1st order methods to minimize a **convex loss**

Linear interpolators

- using d input features with overparameterization ≜ dimension $d \gg n$
- found w/ 1st order methods to minimize a **convex loss**

complexity to analyze model

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
    - For fixed interpolator, certain problem instances/distributions are more benign
    - For fixed problem instance, certain interpolators generalize better

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

Goal is ***not to find*** *better interpolators in practice*

but ***to understand when*** *interpolation is benign*

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

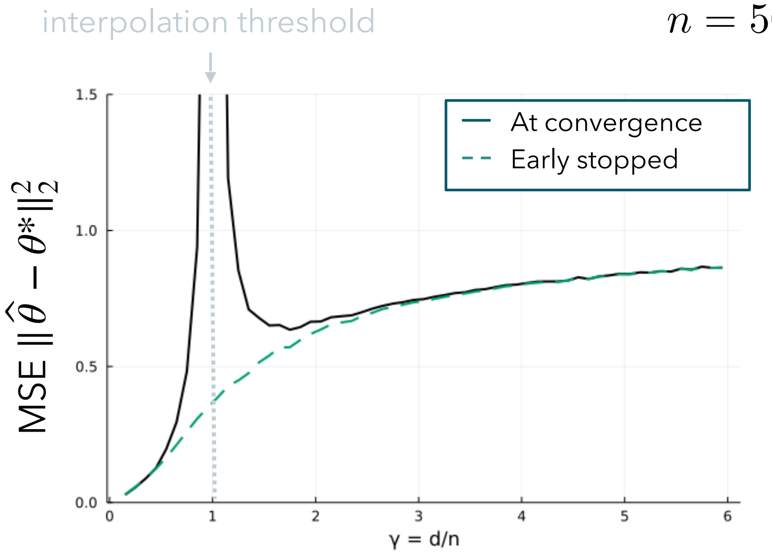
Benefits of overparameterization and interpolation in **linear models**

We run gradient descent on $\|\mathbf{Y} - \mathbf{X}\theta\|_2^2$ at $\theta_0 = 0$ for $\mathbf{Y} = \mathbf{X}\theta^* + \mathbf{W}$
(where $\mathbf{X}, \mathbf{W}$ are comprised of iid standard Gaussian entries)

Benefits of overparameterization and interpolation in **linear models**

We run gradient descent on $\|\mathbf{Y} - \mathbf{X}\theta\|_2^2$ at $\theta_0 = 0$ for $\mathbf{Y} = \mathbf{X}\theta^* + \mathbf{W}$
(where $\mathbf{X}, \mathbf{W}$ are comprised of iid standard Gaussian entries)

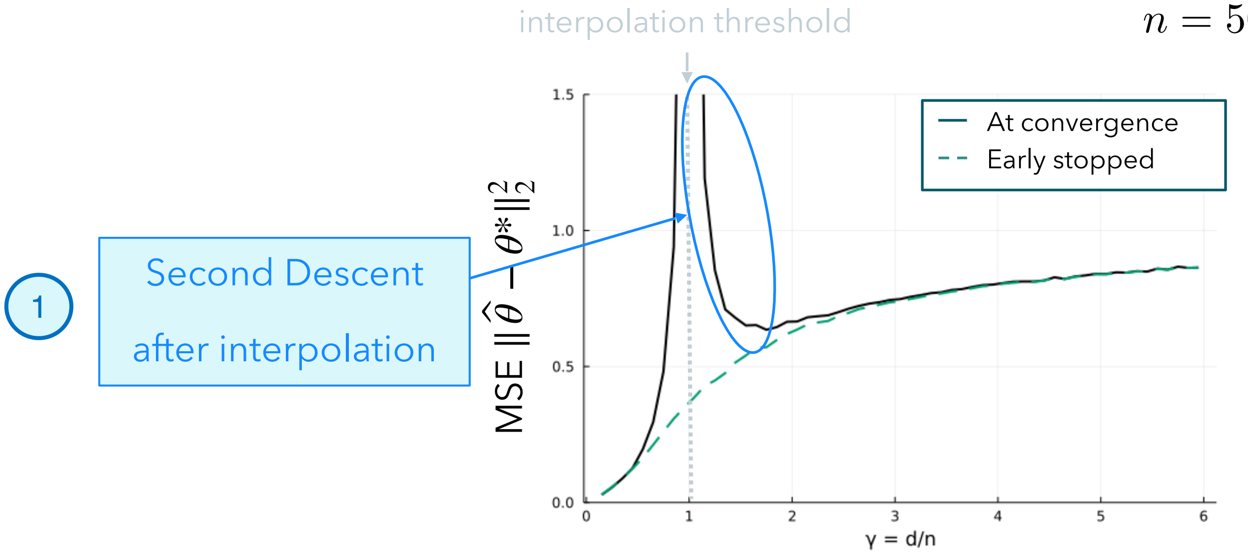
$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$



Benefits of overparameterization and interpolation in **linear models**

We run gradient descent on $\|\mathbf{Y} - \mathbf{X}\theta\|_2^2$ at $\theta_0 = 0$ for $\mathbf{Y} = \mathbf{X}\theta^* + \mathbf{W}$
(where $\mathbf{X}, \mathbf{W}$ are comprised of iid standard Gaussian entries)

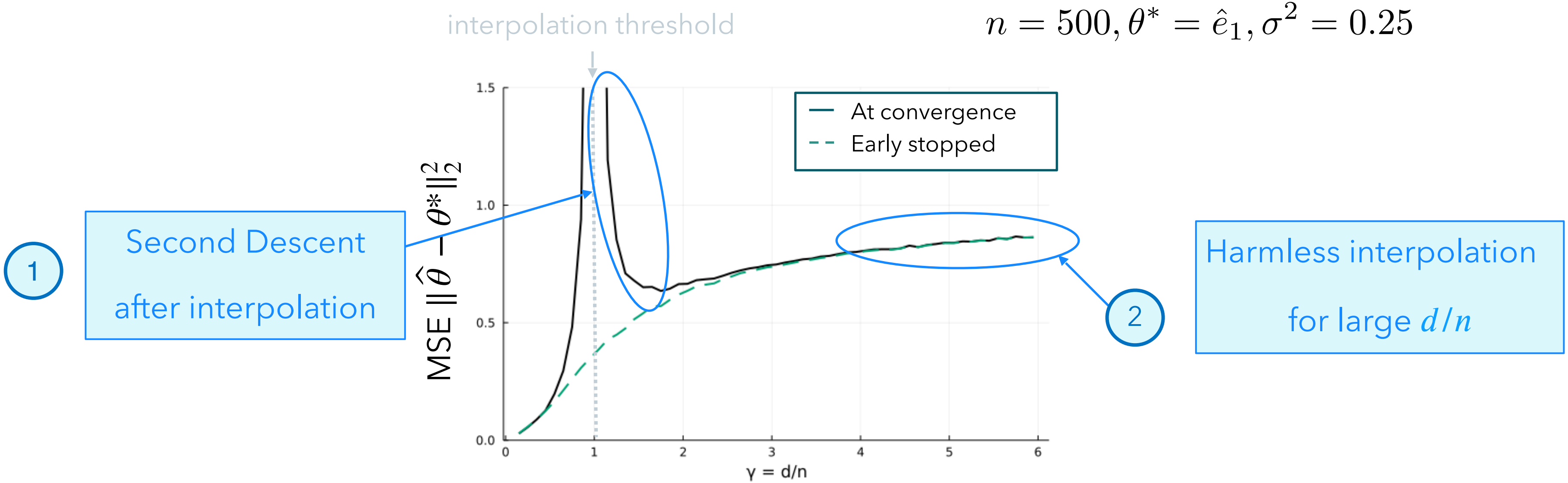
$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$



Benefits of overparameterization and interpolation in **linear models**

We run gradient descent on $\|\mathbf{Y} - \mathbf{X}\theta\|_2^2$ at $\theta_0 = 0$ for $\mathbf{Y} = \mathbf{X}\theta^* + \mathbf{W}$
(where $\mathbf{X}, \mathbf{W}$ are comprised of iid standard Gaussian entries)

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$



Formal setup: overparameterized linear regression

Formal setup: overparameterized linear regression

true parameter/signal (unknown)

$$Y = X^\top \theta^* + W$$

output

input features, dimension = d

noise, variance = σ^2

Formal setup: overparameterized linear regression

true parameter/signal (unknown)

$$Y = X^\top \theta^* + W$$

output

input features, dimension = d

noise, variance = σ^2

$$\mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma$$

(data covariance)

e.g. “isotropic covariance” means $\Sigma = I$

Formal setup: overparameterized linear regression

true parameter/signal (unknown)

$$Y = X^\top \theta^* + W$$

output

input features, dimension = d

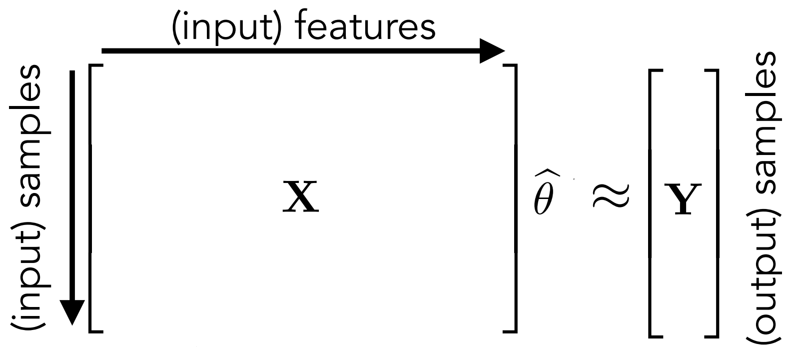
noise, variance = σ^2

$$\mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma$$

(data covariance)

e.g. "isotropic covariance" means $\Sigma = I$

(no. of features) $d > n$ **(no. of samples)**



$\mathbf{X}\hat{\theta} = \mathbf{Y}$ has infinitely many **interpolating solutions!**

Formal setup: overparameterized linear regression

true parameter/signal (unknown)

$$Y = X^\top \theta^* + W$$

output

input features, dimension = d

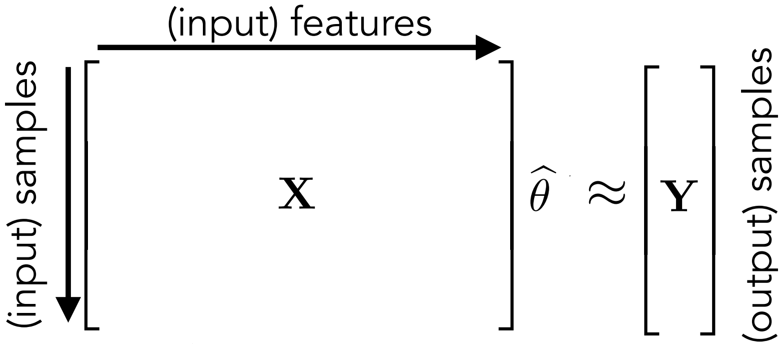
noise, variance = σ^2

$$\mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma$$

(data covariance)

e.g. “isotropic covariance” means $\Sigma = I$

(no. of features) $d > n$ **(no. of samples)**



$\mathbf{X}\widehat{\theta} = \mathbf{Y}$ has infinitely many **interpolating solutions!**

Solutions of study today:

The minimum-lp-norm interpolator

$$\widehat{\theta}_p = \arg\min \|\theta\|_p \text{ subject to } \mathbf{X}\theta = \mathbf{Y}.$$

(beginning with p = 2)

Formal setup: overparameterized linear regression

true parameter/signal (unknown)

$$Y = X^\top \theta^* + W$$

output

input features, dimension = d

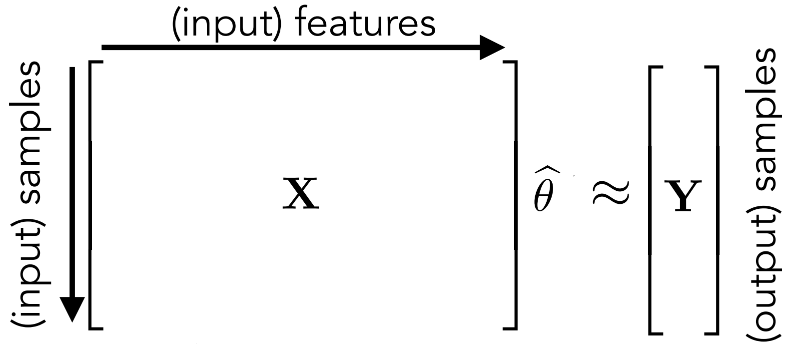
noise, variance = σ^2

$$\mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma$$

(data covariance)

e.g. "isotropic covariance" means $\Sigma = I$

(no. of features) $d > n$ **(no. of samples)**



$\mathbf{X}\widehat{\theta} = \mathbf{Y}$ has infinitely many **interpolating solutions!**

Solutions of study today:

The minimum-lp-norm interpolator

$$\widehat{\theta}_p = \arg\min \|\theta\|_p \text{ subject to } \mathbf{X}\theta = \mathbf{Y}.$$

(beginning with p = 2)

Error metric is **mean-squared-error:** $\mathscr{E}_{\text{MSE}} := \mathbb{E}\left[(X^\top(\widehat{\theta} - \theta^*))^2\right]$

Analysis framework

Non-asymptotic: we consider $d = n^\beta, \beta > 1$ (or $d \gg n$) and state results as:

- **Consistency:** goal is to have $\mathcal{E}_{\text{MSE}} \to 0$ as $n \to \infty$
- **Rates:** upper and lower bounds on $\mathcal{E}_{\text{MSE}}$ as a function of n that match up to universal constants (not depending on n, d, θ^*, Σ)

Analysis framework

Non-asymptotic: we consider $d = n^{\beta}, \beta > 1$ (or $d \gg n$) and state results as:

- **Consistency:** goal is to have $\mathcal{E}_{\text{MSE}} \to 0$ as $n \to \infty$
- **Rates:** upper and lower bounds on $\mathcal{E}_{\text{MSE}}$ as a function of n that match up to universal constants (not depending on n, d, θ^*, Σ)

An alternative asymptotic analysis framework (not the focus of this tutorial):

Considers $d \propto n, \dfrac{d}{n} = \gamma$.

Exact error expressions derived as a function of γ as $n, d \to \infty$ together.

Why these types of "low-norm" interpolators?

Popular optimization algorithms converge to "low-norm" solutions!

Gradient descent on squared loss

(Folklore, see e.g. Engl et al 1996)

Minimum-**l2**-norm interpolation

$$\widehat{\theta}_2 = \arg\min \|\theta\|_2 \text{ subject to } X_i^\top \theta = Y_i, i \in [n].$$

Why these types of "low-norm" interpolators?

Popular optimization algorithms converge to "low-norm" solutions!

Mirror descent on squared loss, Potential = $||\cdot||_p$

(Gunasekar et al, 2018)

Minimum-**lp**-norm interpolation

$$\widehat{\theta}_p = \arg\min \|\theta\|_p$$
subject to
$$X_i^\top \theta = Y_i, i \in [n].$$

Coordinate descent/least-angle regression

(Efron et al, 2004)

Minimum-**l1**-norm interpolation

$$\widehat{\theta}_1 = \arg\min \|\theta\|_1$$
subject to
$$X_i^\top \theta = Y_i, i \in [n].$$

Why these types of "low-norm" interpolators?

Popular optimization algorithms converge to "low-norm" solutions!

Mirror descent on squared loss, Potential = $\|\cdot\|_p$

(Gunasekar et al, 2018)

Minimum-**lp**-norm interpolation

$$\widehat{\theta}_p = \arg\min \|\theta\|_p \text{ subject to } X_i^\top \theta = Y_i, i \in [n].$$

Coordinate descent/least-angle regression

(Efron et al, 2004)

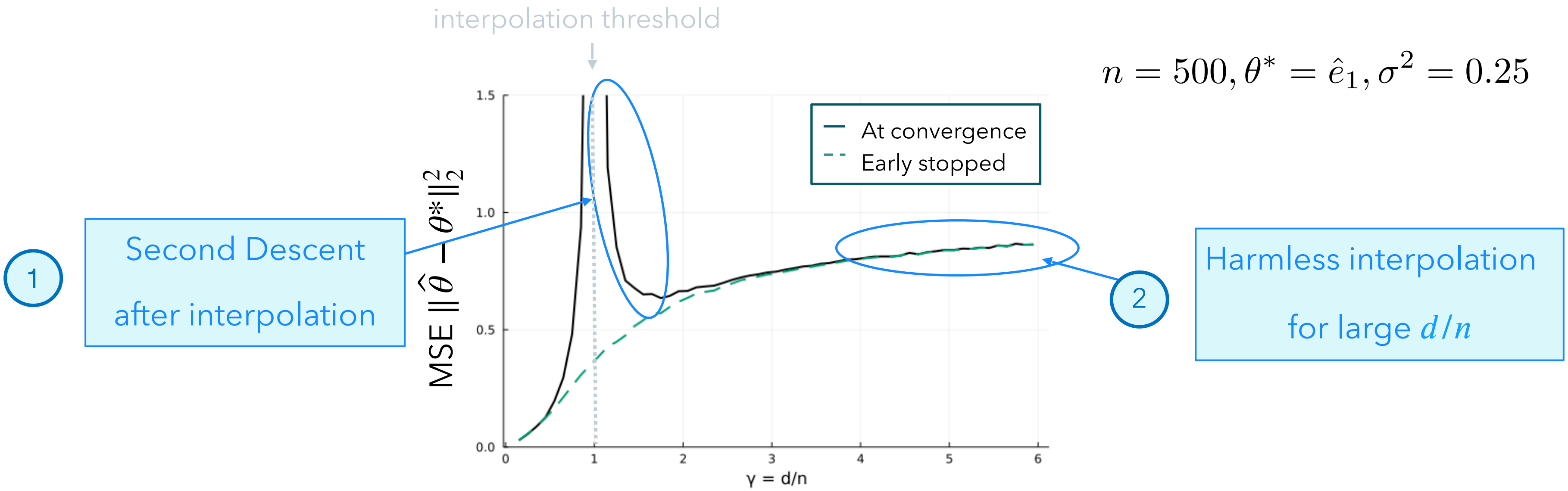
Minimum-**l1**-norm interpolation

$$\widehat{\theta}_1 = \arg\min \|\theta\|_1 \text{ subject to } X_i^\top \theta = Y_i, i \in [n].$$

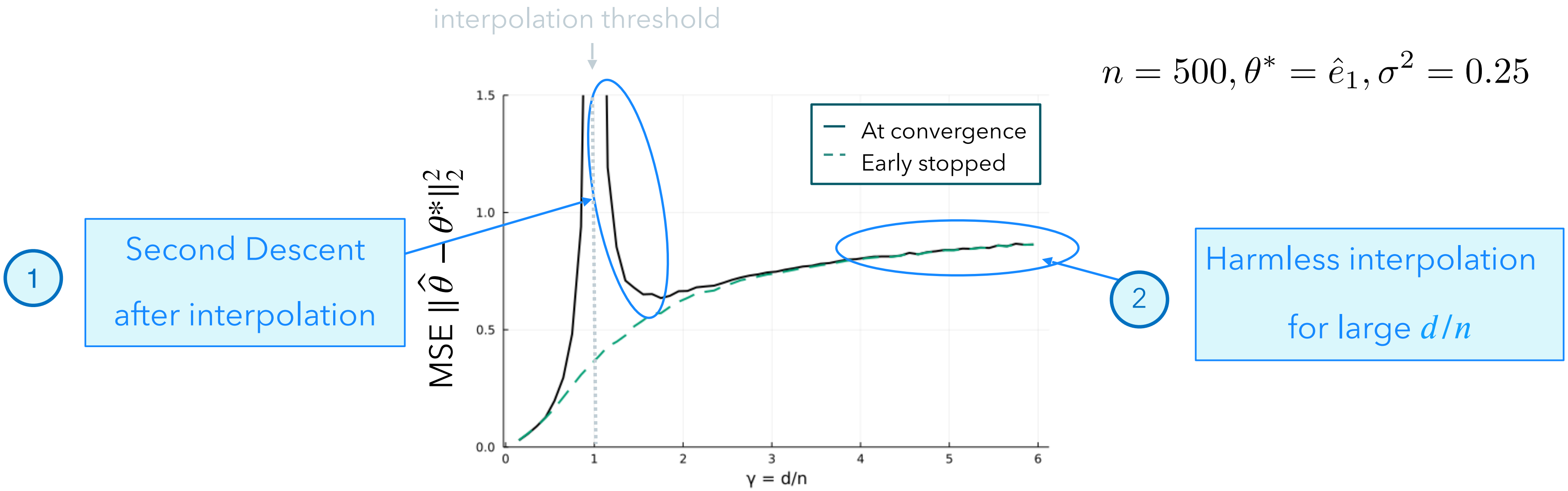
Implicit bias theory is a useful "sanity check" but not the full picture: do these solutions always generalize well?

Recall: what was observed for min-l2-norm interpolator

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$



Recall: what was observed for min-l2-norm interpolator



(1) and (2) are implied by **variance reduction with increased overparameterization!**

Theorem (isotropic covariance)*: $\text{Variance term} \asymp \frac{\sigma^2 n}{d}.$

*included in results of Hastie et al (2022), Bartlett et al (2020), Muthukumar et al (2020)

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

Variance reduction: main proof ideas

- **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^* + \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Variance reduction: main proof ideas

- **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Ideally: have this be
close to θ^* (error = bias)

Variance reduction: main proof ideas

- **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}}$$

Ideally: have this be close to θ^* (error = bias)

Ideally: have this be close to 0 (error = **variance**)

Variance reduction: main proof ideas

- **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}}$$

Ideally: have this be close to θ^* (error = bias)

Ideally: have this be close to 0 (error = **variance**)

- **Step 2:** variance term can also be expressed in closed form

$$\text{Variance} = \|\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}\|_2^2 = \mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Note: this calculation is simplified for isotropic data covariance, but works more generally (Bartlett et al, 2020)

Variance reduction: main proof ideas

- **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}}$$

Ideally: have this be close to θ^* (error = bias)

Ideally: have this be close to 0 (error = **variance**)

- **Step 2:** variance term can also be expressed in closed form

$$\text{Variance} = \|\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}\|_2^2 = \mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

$$= \mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Note: this calculation is simplified for isotropic data covariance, but works more generally (Bartlett et al, 2020)

Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)

$$\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d$$

Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)

$$\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d$$

$$\implies \mathbf{X}\mathbf{X}^\top \approx d\mathbf{I}$$

Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)

$$\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d$$

$$\implies \mathbf{X}\mathbf{X}^\top \approx d\mathbf{I}$$

$$\implies \text{Variance} = \mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W} \approx \frac{\|\mathbf{W}\|_2^2}{d}$$

Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)

$$\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d$$

$$\implies \mathbf{X}\mathbf{X}^\top \approx d\mathbf{I}$$

$$\implies \text{Variance} = \mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W} \approx \frac{\|\mathbf{W}\|_2^2}{d}$$

Total "noise energy"

$$\approx \frac{n\sigma^2}{d}$$

Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)

$$\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d$$

$$\implies \mathbf{X}\mathbf{X}^\top \approx d\mathbf{I}$$

$$\implies \text{Variance} = \mathbf{W}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{W} \approx \frac{\|\mathbf{W}\|_2^2}{d}$$

Total "noise energy"

$$\approx \frac{n\sigma^2}{d}$$

Intuition: noise energy is "spread out" along d feature dimensions, contributes more harmlessly as d increases

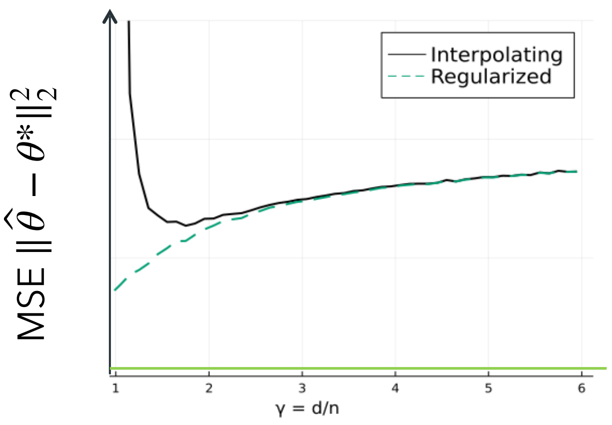
Note: can show corresponding **precise** results when $d \propto n$, $d, n \to \infty$ (Hastie et al, 2022)

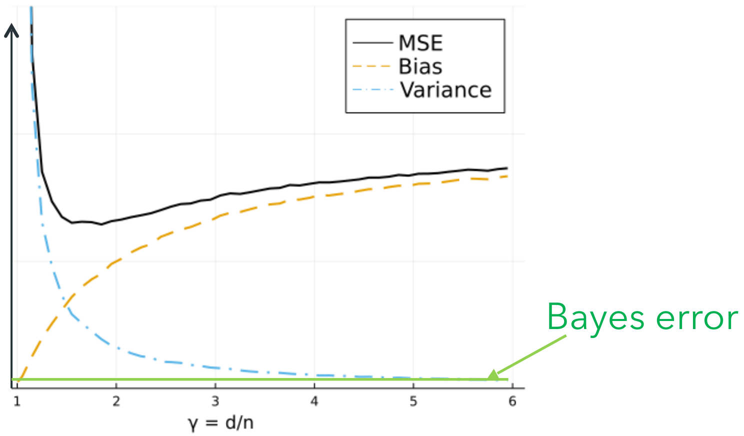
So is min-l2-norm interpolation *always* a good idea?

Interpolator $\widehat{\theta}_2 = \arg\min \|\theta\|_2$ subject to $\mathbf{X}\theta = \mathbf{Y}$ vs.

regularized estimator: $\arg\min \|\mathbf{X}\theta - \mathbf{Y}\|_2^2 + \lambda \|\theta\|_2^2$

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$



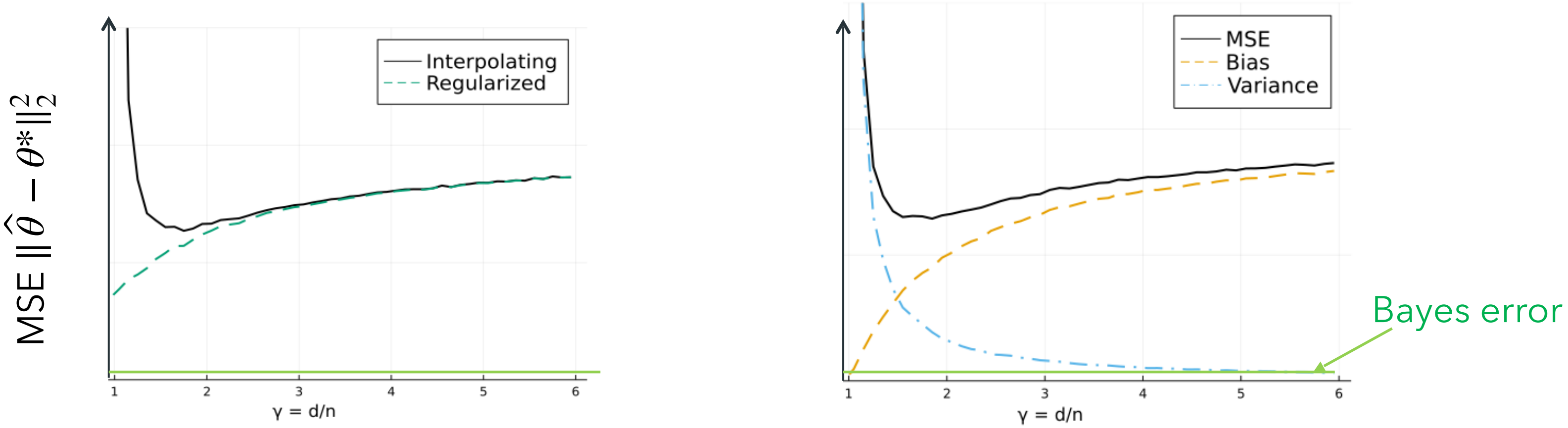


So is min-l2-norm interpolation *always* a good idea?

Interpolator $\widehat{\theta}_2 = \arg\min \|\theta\|_2$ subject to $\mathbf{X}\theta = \mathbf{Y}$ vs.

regularized estimator: $\arg\min \|\mathbf{X}\theta - \mathbf{Y}\|_2^2 + \lambda\|\theta\|_2^2$

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$



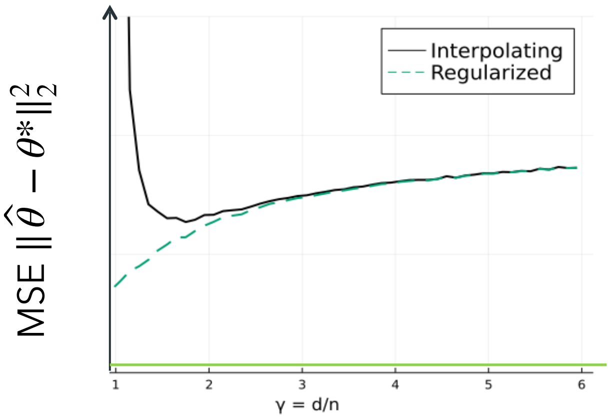
1 second descent ✓ 2 harmless interpolation ✓ 3 good generalization ✗

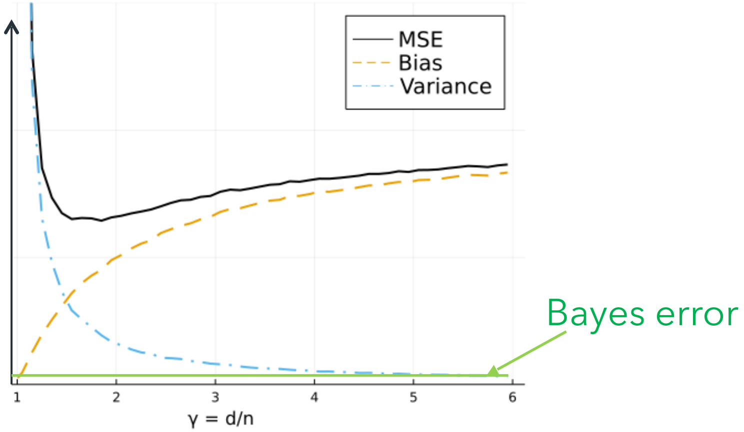
So is min-l2-norm interpolation *always* a good idea?

Interpolator $\widehat{\theta}_2 = \arg\min \|\theta\|_2$ subject to $\mathbf{X}\theta = \mathbf{Y}$ vs.

regularized estimator: $\arg\min \|\mathbf{X}\theta - \mathbf{Y}\|_2^2 + \lambda\|\theta\|_2^2$

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$





1 second descent ✓ 2 harmless interpolation ✓ 3 good generalization ✗

Core issue: **bias increases with d,** eventually dominates

Issues with isotropy and min-l2 inductive bias

Recall: minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Ideally: have this be
close to θ^* (error = bias)

Issues with isotropy and min-l2 inductive bias

Recall: minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Ideally: have this be
close to θ^* (error = bias)

$$\text{Bias} = \|(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X} - \mathbf{I})\theta^*\|_2^2$$

Issues with isotropy and min-l2 inductive bias

Recall: minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Ideally: have this be
close to θ^* (error = bias)

$$\text{Bias} = \|(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X} - \mathbf{I})\theta^*\|_2^2$$

Theorem*: $\text{Bias} \asymp \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

*included in results of Hastie et al (2022), Bartlett et al (2020)

Issues with isotropy and min-l2 inductive bias

Recall: minimum-l2-norm interpolator can be expressed in closed form

$$\widehat{\theta}_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y} = \boxed{\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\theta^*} + \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{W}$$

Ideally: have this be
close to θ^* (error = bias)

$$\text{Bias} = \|(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X} - \mathbf{I})\theta^*\|_2^2$$

Theorem*: $$\text{Bias} \asymp \left(1 - \frac{n}{d}\right)\|\theta^*\|_2^2$$

Intuition: under isotropy, **true parameter energy** also spread out across d features!

*included in results of Hastie et al (2022), Bartlett et al (2020)

Isotropy and min-l2-norm bias visualized at feature-by-feature level

Theorem: $\mathrm{Bias} \asymp \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

Intuition: under isotropy, **true parameter energy** also spread out across d features

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)

Isotropy and min-l2-norm bias visualized at feature-by-feature level

Theorem: $\text{Bias} \asymp \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

Intuition: under isotropy, **true parameter energy** also spread out across d features

Canonical setting: k-sparse signal

$$Y = X^\top \theta^* + W$$

$$\theta_j^* \neq 0 \text{ for } j \in [k], 0 \text{ otherwise}$$

$$k \ll n$$

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)

Isotropy and min-l2-norm bias visualized at feature-by-feature level

Theorem: $\text{Bias} \asymp \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

Intuition: under isotropy, **true parameter energy** also spread out across d features

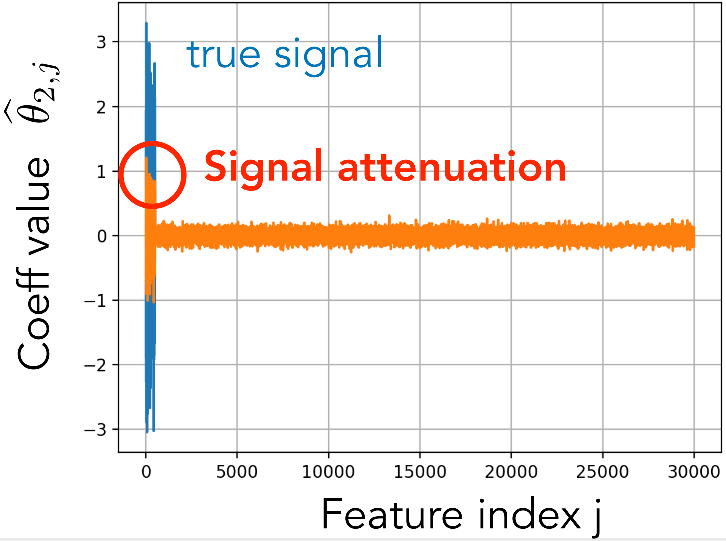
Canonical setting: k-sparse signal

$$Y = X^\top \theta^* + W$$

$$\theta^*_j \neq 0 \text{ for } j \in [k], 0 \text{ otherwise}$$

$$k \ll n$$

(k = 500, n = 5000, d = 30000)



This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)

Isotropy and min-l2-norm bias visualized at feature-by-feature level

Theorem: $\text{Bias} \asymp \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

Intuition: under isotropy, **true parameter energy** also spread out across d features

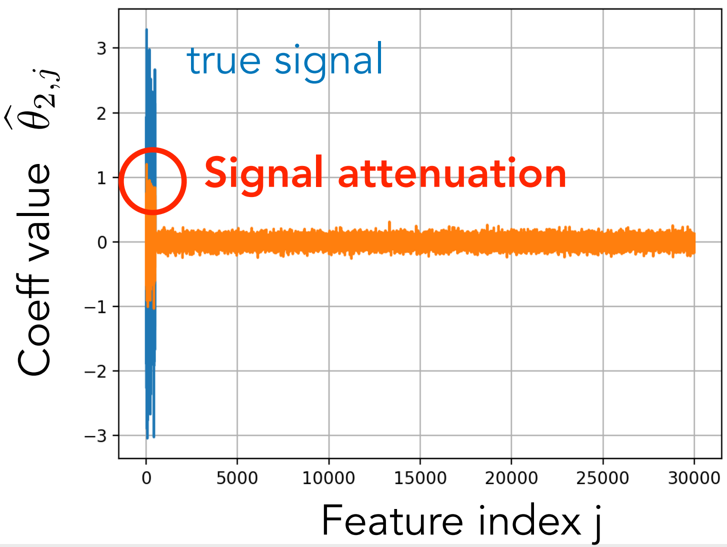
Canonical setting: k-sparse signal

$$Y = X^\top \theta^* + W$$

$$\theta_j^* \neq 0 \text{ for } j \in [k], 0 \text{ otherwise}$$

$$k \ll n$$

(k = 500, n = 5000, d = 30000)



Core issue for bias: $|\widehat{\theta}_j| \ll |\theta_j^*|$ **for all** $j \in [k]$!

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

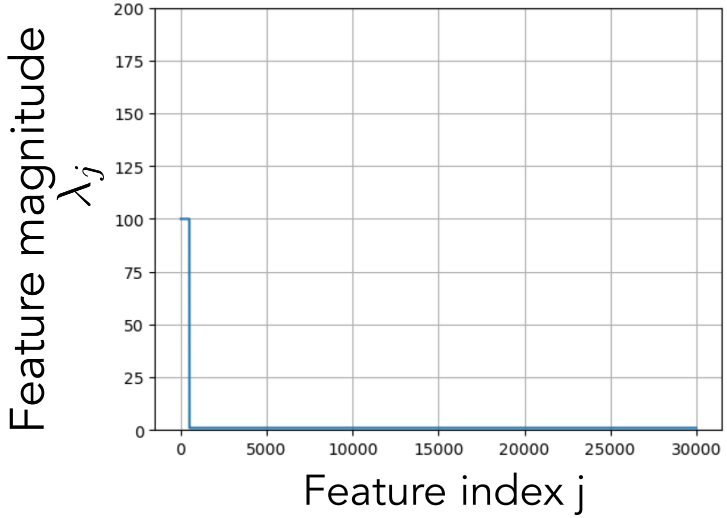
Anisotropy to the rescue: "upweighting" features aligned with signal

- A special case $\Sigma = \begin{bmatrix} R\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k} \end{bmatrix}, R \gg 1$ (spiked-covariance)

Anisotropy to the rescue: "upweighting" features aligned with signal

- A special case $\Sigma = \begin{bmatrix} R\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k} \end{bmatrix}, R \gg 1$ (spiked-covariance)

Effective "upweighting" on top k features

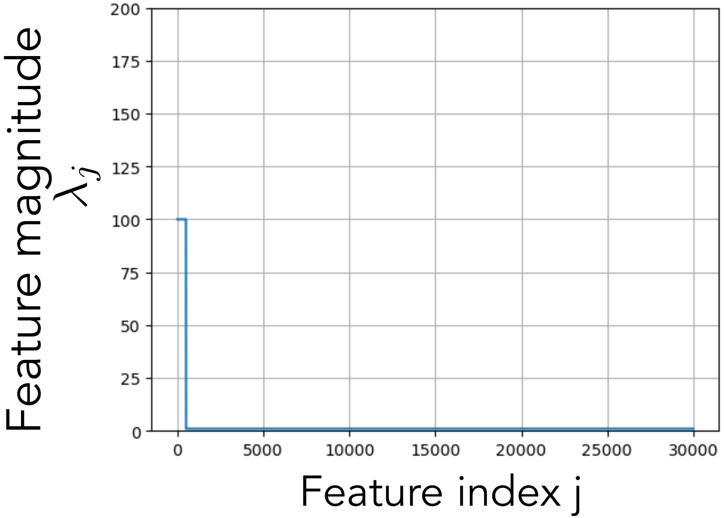


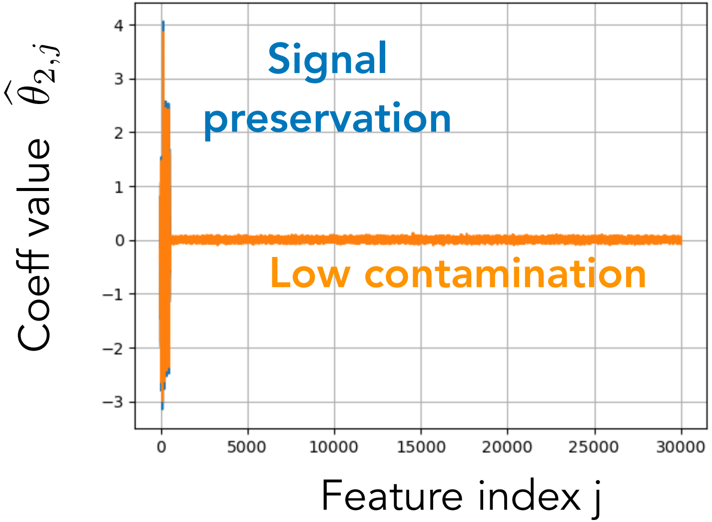
Anisotropy to the rescue: "upweighting" features aligned with signal

- A special case $\Sigma = \begin{bmatrix} R\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k} \end{bmatrix}, R \gg 1$ (spiked-covariance)

Effective "upweighting" on top k features

(k = 500, n = 5000, d = 30000, R = 100)

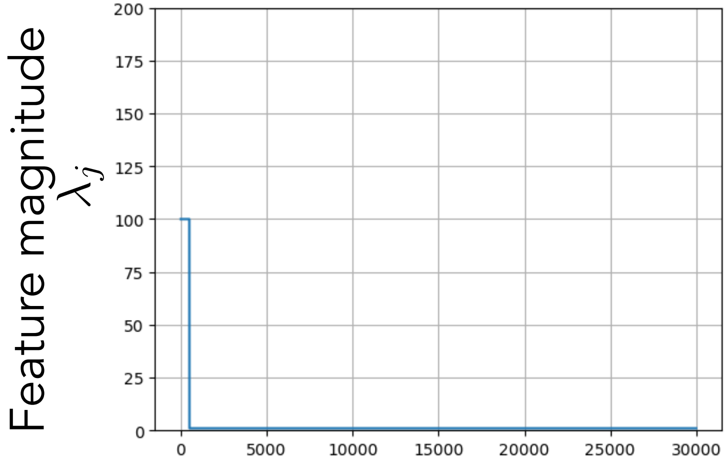




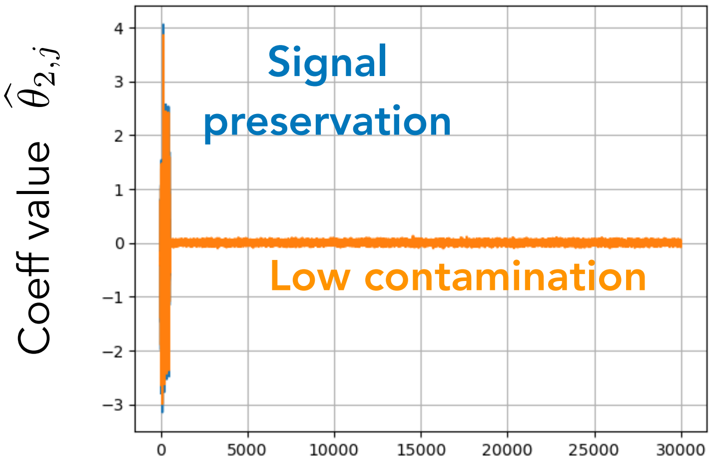
Anisotropy to the rescue: "upweighting" features aligned with signal

- A special case $\Sigma = \begin{bmatrix} R\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k} \end{bmatrix}, R \gg 1$ (spiked-covariance)

Effective "upweighting" on top k features



(k = 500, n = 5000, d = 30000, R = 100)

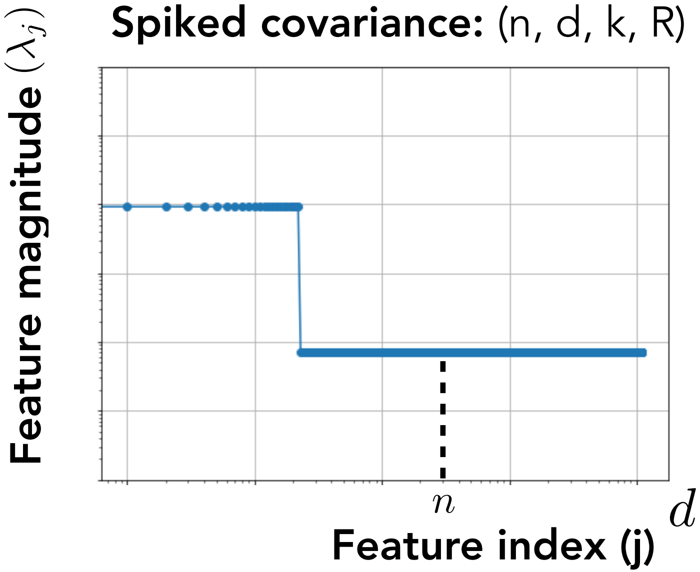


Low bias *iff* $\widehat{\theta}_j \approx \theta_j^*$ **for all** $j \in [k]$

Intuition: under near-orthogonality, $\widehat{\theta}_j \propto \sum_{i=1}^{n} y_i x_{i,j}$ - attenuation mitigated for larger R as $x_{i,j} \sim \mathcal{N}(0,R)$ for $j \in [k]$

A sensible model for l2: the **spiked-covariance** ensemble

$$\Sigma = \mathrm{diag}(\Lambda) =$$

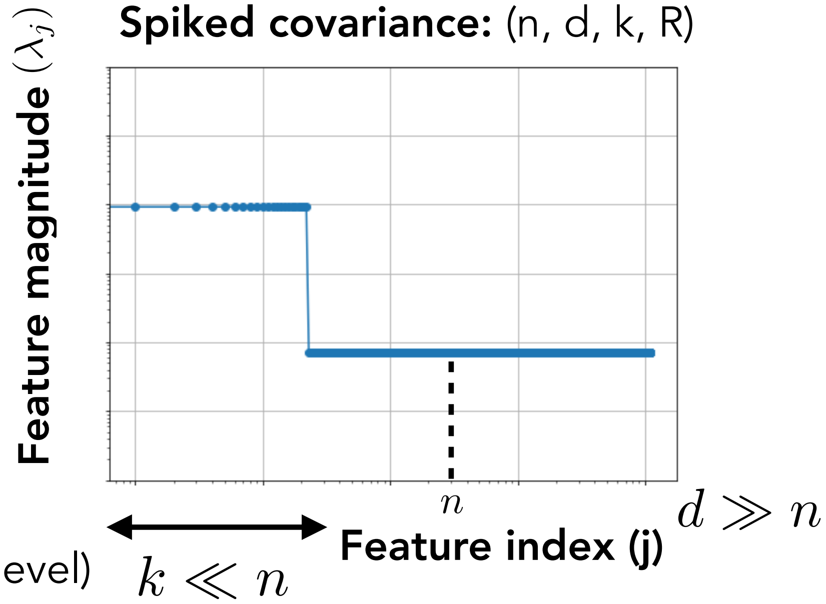


$d \gg n$

Conditions for **general anisotropic covariances** in terms of "effective ranks" by Bartlett et al (2020)

A sensible model for l2: the **spiked-covariance** ensemble

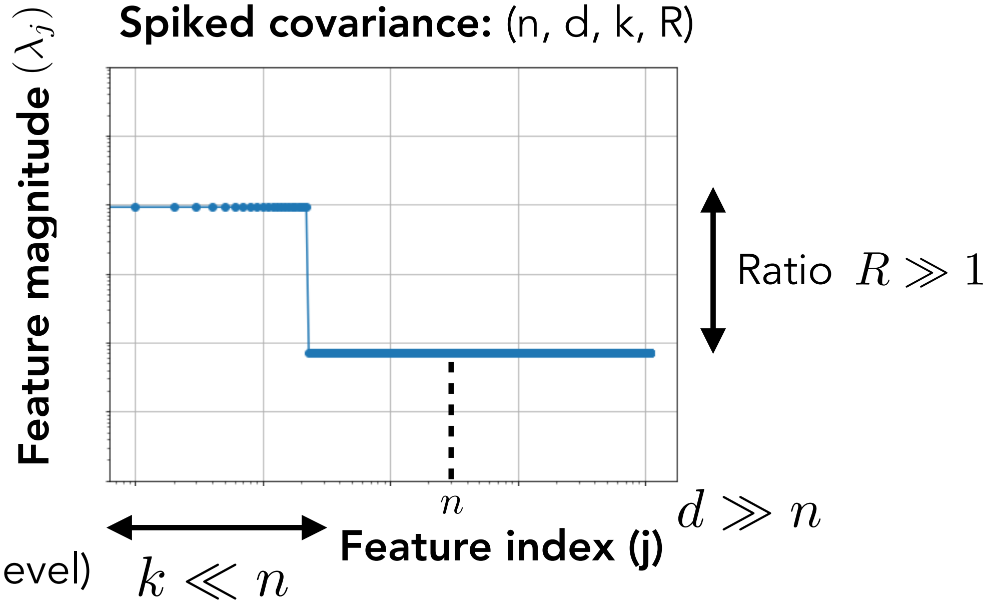
$$\Sigma = \mathrm{diag}(\Lambda) =$$



Conditions for **general anisotropic covariances** in terms of "effective ranks" by Bartlett et al (2020)

A sensible model for l2: the **spiked-covariance** ensemble

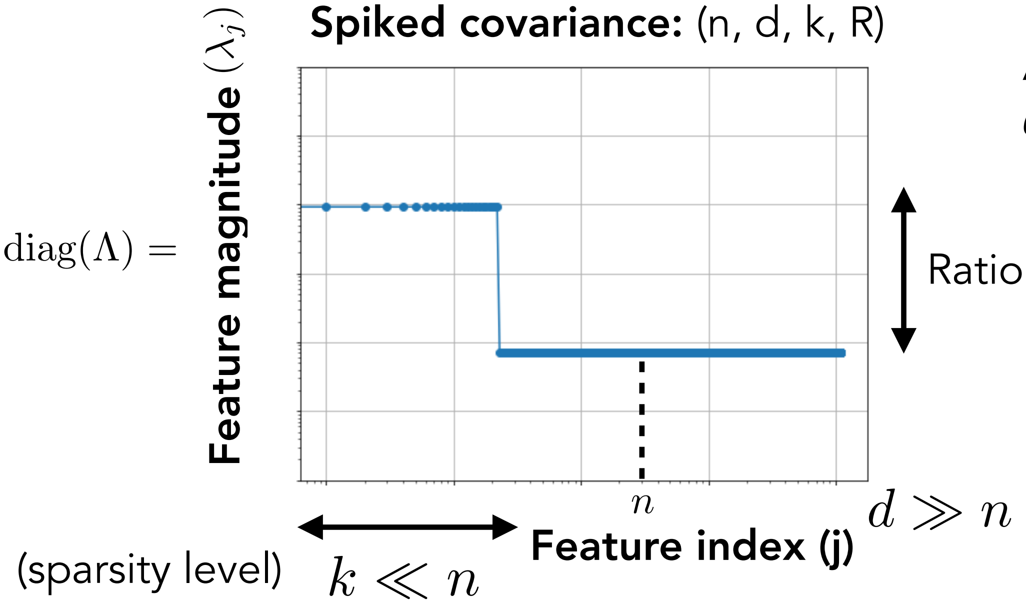
$\Sigma = \mathrm{diag}(\Lambda) =$



Conditions for **general anisotropic covariances** in terms of "effective ranks" by Bartlett et al (2020)

A sensible model for l2: the **spiked-covariance** ensemble

$\Sigma = \mathrm{diag}(\Lambda) =$



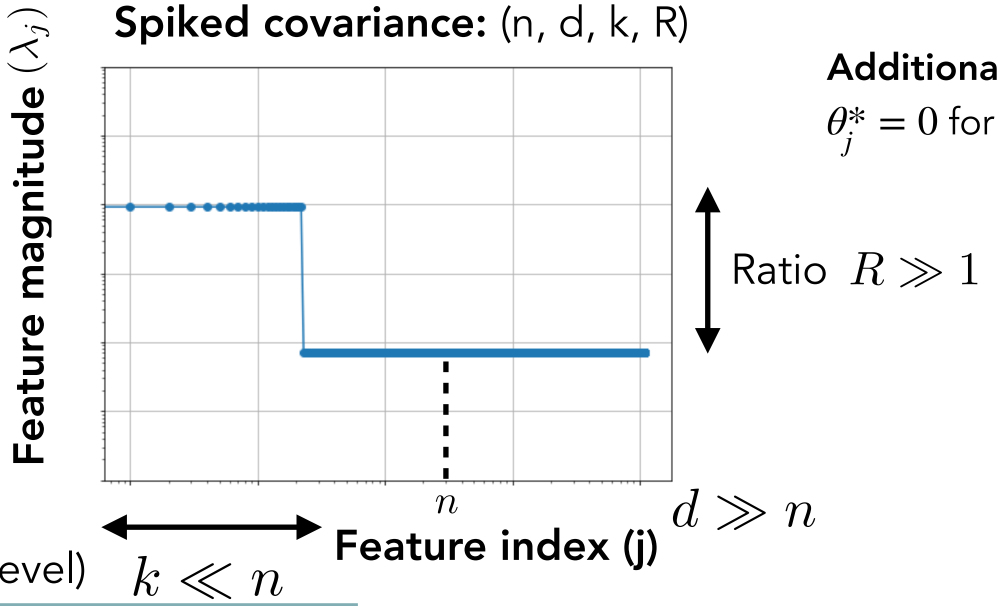
Additionally assume

$\theta^*_j = 0$ for all $j = k+1, \ldots, d$

Conditions for **general anisotropic covariances** in terms of "effective ranks" by Bartlett et al (2020)

A sensible model for l2: the **spiked-covariance** ensemble

$$\Sigma = \mathrm{diag}(\Lambda) =$$



(sparsity level)

Additionally assume

$\theta_j^* = 0$ for all $j = k+1, \ldots, d$

Will **always** achieve

Variance $\rightarrow 0$ as $n, d \rightarrow \infty$:

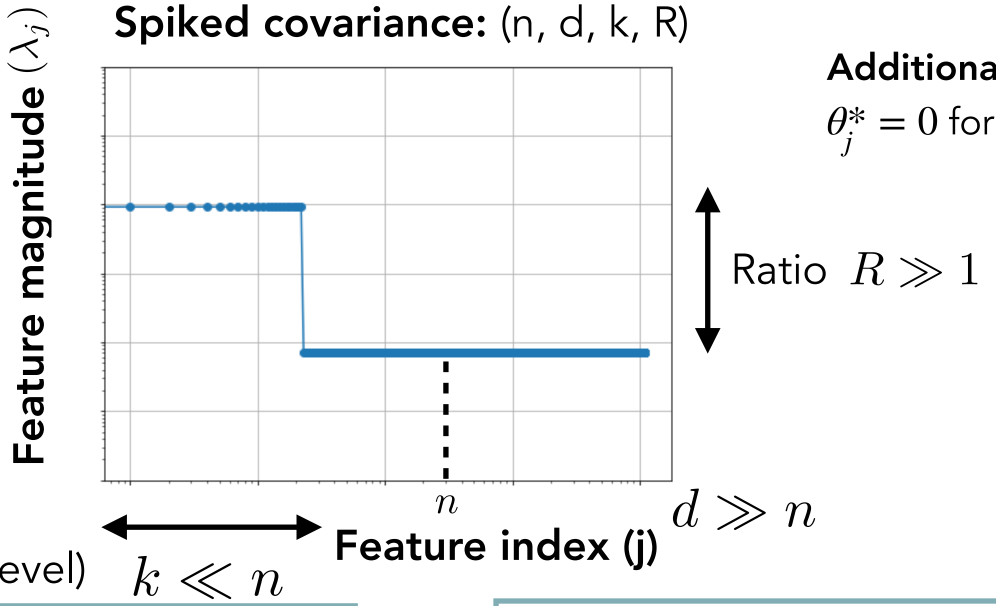
Noise hidden along (d-k) directions!

Conditions for **general anisotropic covariances** in terms of "effective ranks" by Bartlett et al (2020)

A sensible model for l2: the **spiked-covariance** ensemble

Spiked covariance: (n, d, k, R)

$$\Sigma = \mathrm{diag}(\Lambda) =$$



Additionally assume
$\theta^*_j = 0$ for all $j = k+1, \ldots, d$

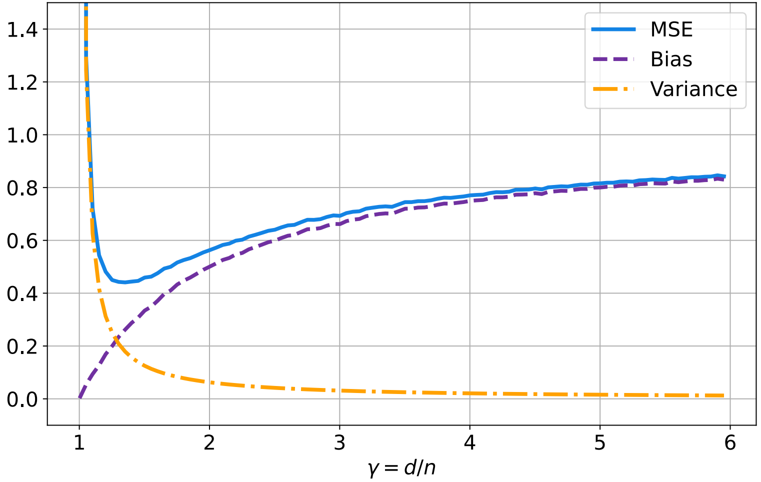
Will **always** achieve
Variance $\to 0$ as $n, d \to \infty$:
Noise hidden along (d-k) directions!

Also achieves Bias $\to 0$ as $n, d \to \infty$
provided that $R \gg \frac{d}{n}$

Conditions for **general anisotropic covariances** in terms of "effective ranks" by Bartlett et al (2020)

Summary: Uniform benefits of overparameterization with spiked covariance

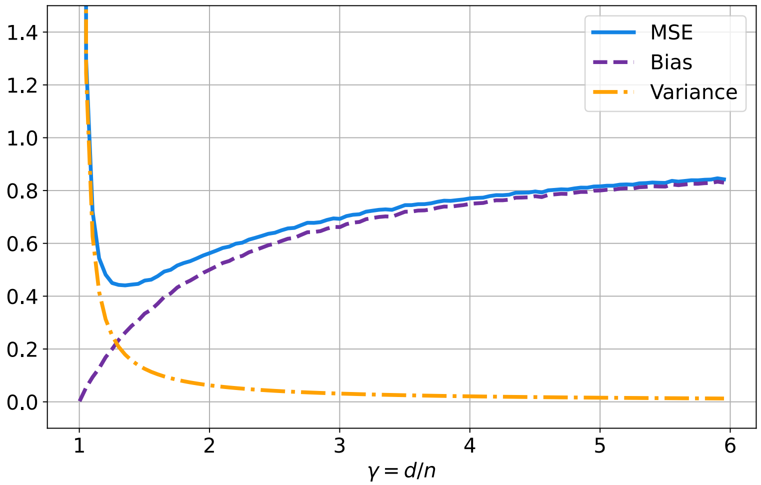
$$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$$



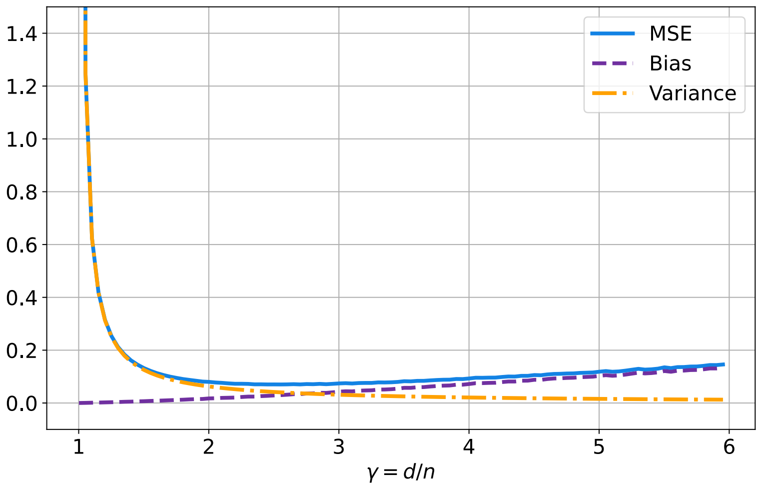
Isotropic covariance

Summary: Uniform benefits of overparameterization with spiked covariance

$$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$$



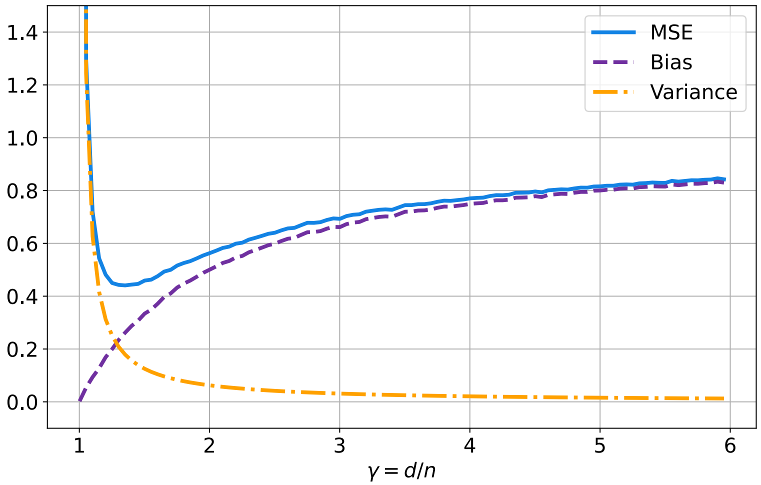
Isotropic covariance



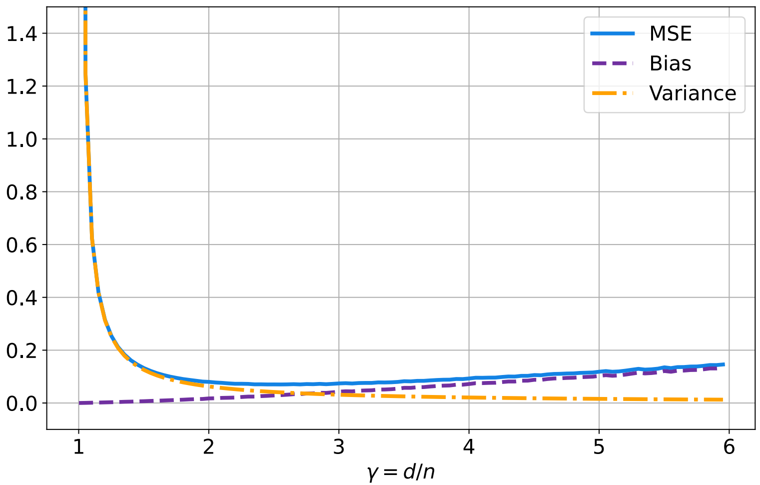
Spiked covariance, $R = 10$

Summary: Uniform benefits of overparameterization with spiked covariance

$$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$$



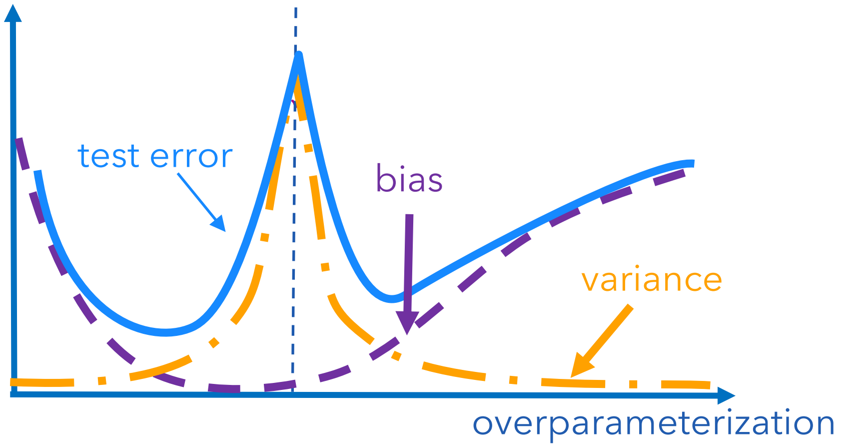
Isotropic covariance



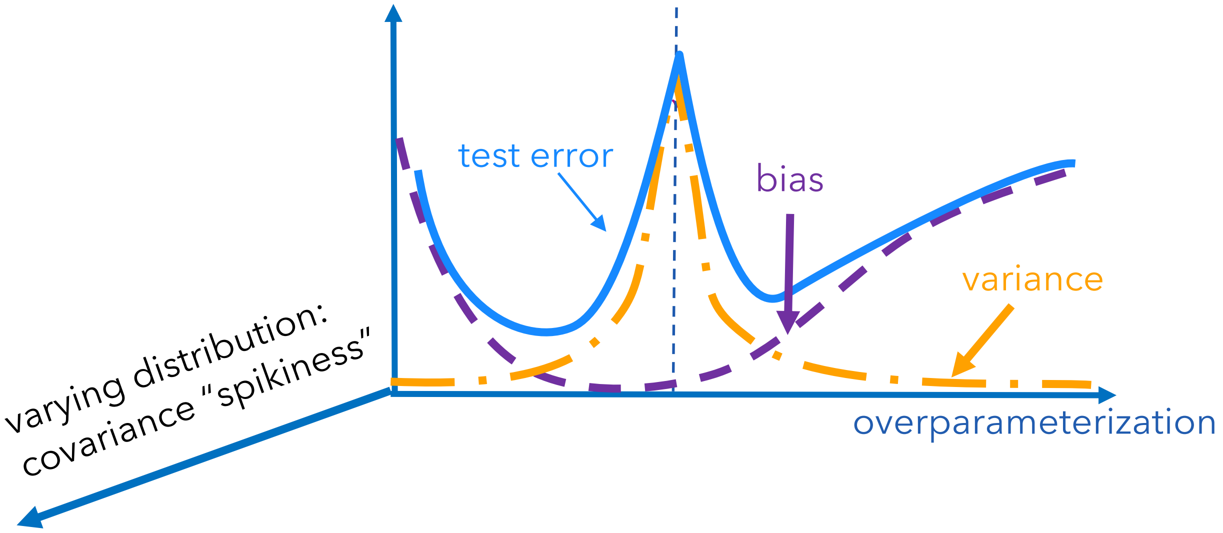
Spiked covariance, $R = 10$

For spiked covariance: (1) second descent ✓ (2) harmless interpolation ✓ (3) good generalization ✓

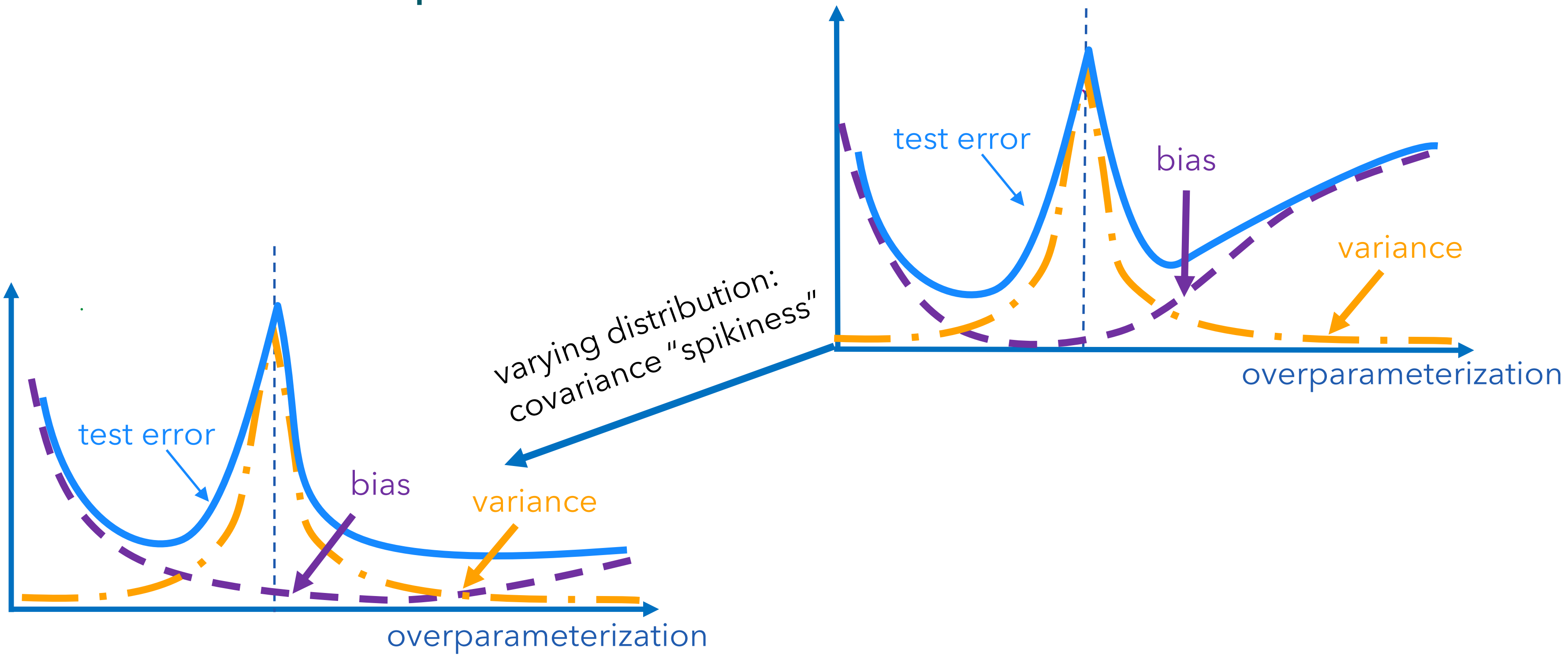
For fixed interpolator…



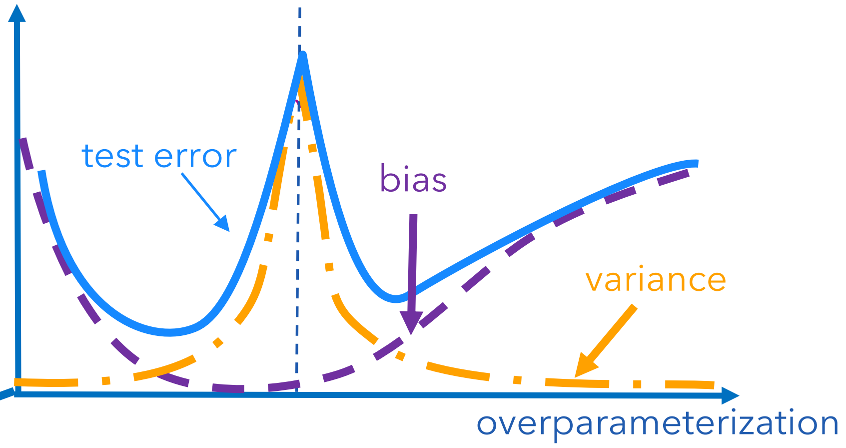
For fixed interpolator…



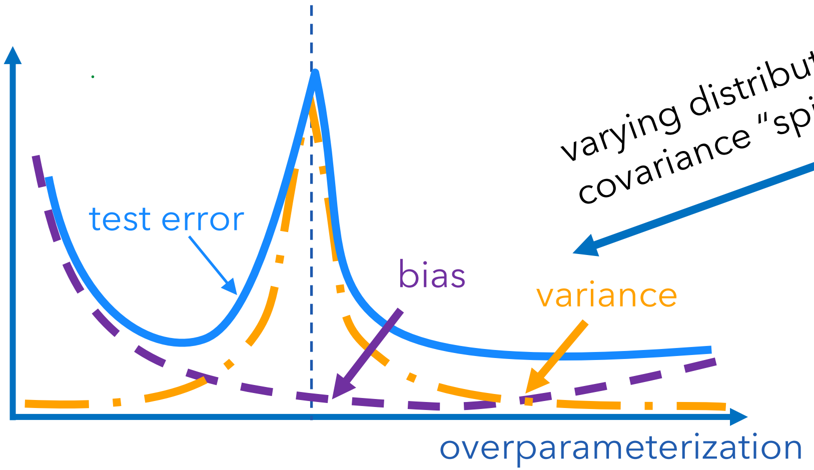
For fixed interpolator…



For fixed interpolator…



varying distribution:
covariance "spikiness"



For a fixed distribution (e.g. isotropic),

different algorithms → different interpolators

how do bias and variance behave?

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

Implicit bias → inductive bias

opt. algorithm
minimizing loss

Implicit bias → inductive bias

opt. algorithm minimizing loss —has implicit bias towards→ certain interpolator

Implicit bias → inductive bias

opt. algorithm
minimizing loss

has
implicit bias
towards

→

certain interpolator

e.g. 1st order method
on $||y - Xw||_2^2$

Implicit bias → inductive bias

opt. algorithm
minimizing loss

has
implicit bias
towards →

certain interpolator

e.g. 1st order method
on $||y - Xw||_2^2$

e.g. for $p \in [1,2]$
$\hat{w}_p = \text{argmin}_w ||w||_p$
$s.t. y = Xw$

Implicit bias → inductive bias

opt. algorithm minimizing loss —has implicit bias towards→ certain interpolator —has certain strength of inductive bias towards→ certain structure

e.g. 1st order method on $||y - Xw||_2^2$

e.g. for $p \in [1,2]$

$$\hat{w}_p = \text{argmin}_w \, ||w||_p \quad s.t. \, y = Xw$$

Implicit bias → inductive bias

opt. algorithm minimizing loss — has implicit bias towards → certain interpolator — has certain strength of inductive bias towards → certain structure

e.g. 1st order method on $||y - Xw||_2^2$

e.g. for $p \in [1,2]$

$$\hat{w}_p = \operatorname{argmin}_w ||w||_p \quad s.t.\ y = Xw$$

e.g. sparsity, invariances

Implicit bias → inductive bias

opt. algorithm
minimizing loss

has
implicit bias
towards

certain interpolator

has certain strength of
inductive bias
towards

certain structure

e.g. 1st order method
on $||y - Xw||_2^2$

e.g. for $p \in [1,2]$

$$\hat{w}_p = \text{argmin}_w \, ||w||_p \quad s.t. \, y = Xw$$

e.g. sparsity,
invariances

Implicit bias → inductive bias

opt. algorithm minimizing loss — has implicit bias towards → certain interpolator — has certain strength of inductive bias towards → certain structure

e.g. 1st order method on $||y - Xw||_2^2$

e.g. for $p \in [1,2]$

$$\hat{w}_p = \operatorname{argmin}_w ||w||_p \quad s.t.\ y = Xw$$

e.g. sparsity, invariances

Next: Recall how as $p \to 1$ has an inductive bias towards sparse solutions

Recall: Inductive bias for sparse linear models

isotropic

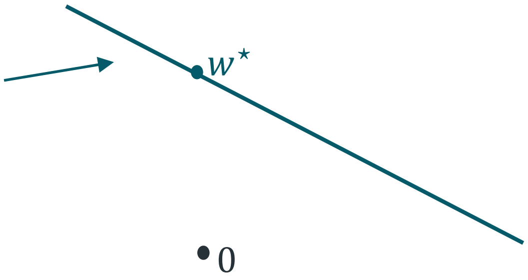
Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Recall: Inductive bias for sparse linear models

isotropic

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

subspace of all
linear interpolators
$\{w\colon Xw = y = Xw^\star\}$



for noiseless $\xi_i = 0$

Recall: Inductive bias for sparse linear models

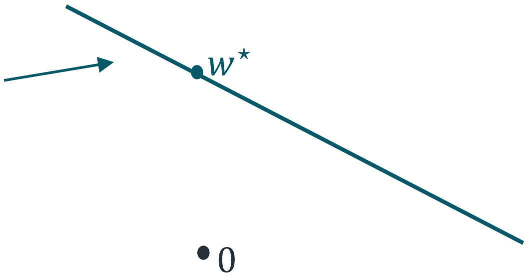
isotropic

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

subspace of all linear interpolators $\{w \colon Xw = y = Xw^\star\}$

for noiseless $\xi_i = 0$



Recall: Inductive bias for sparse linear models

isotropic

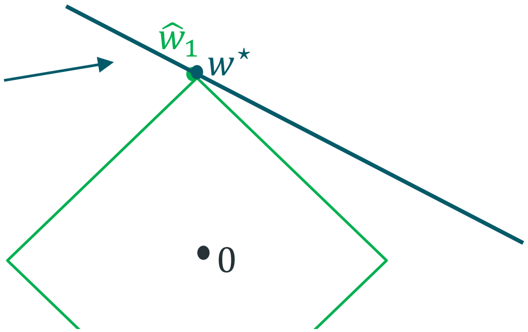
Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

- small $||w||_1$-norm encourages sparsity → aligns with $w^\star$ structure **(strong inductive bias)**

subspace of all linear interpolators $\{w\colon Xw = y = Xw^\star\}$

for noiseless $\xi_i = 0$



Recall: Inductive bias for sparse linear models

isotropic

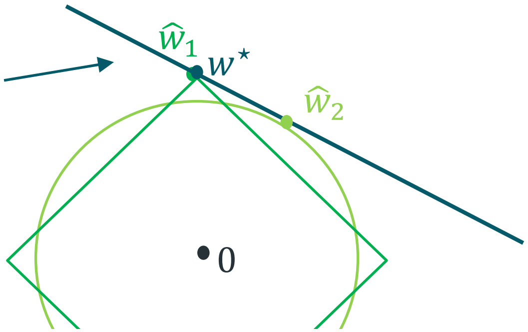
Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

- small $||w||_1$-norm encourages sparsity → aligns with $w^\star$ structure **(strong inductive bias)**
- small $||w||_2$-norm → does not restrict search space in right way! **(weak inductive bias)**

subspace of all linear interpolators $\{w: Xw = y = Xw^\star\}$

for noiseless $\xi_i = 0$



Recall: small ℓ_1-norm → small statistical bias

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$

Recall: small ℓ_1-norm → small statistical bias

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$

Noiseless
$y = Xw^\star$

Basis pursuit: $\hat{w}_1 = \text{argmin}_w \, ||w||_1 \; s.t. \; y = Xw$

Perfect recovery
w.h.p. for $n \sim k \log d$

Recall: small ℓ_1-norm → small statistical bias

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

for $i = 1, \dots, n$ samples and input and parameter dimension $d \gg n$

Noiseless $y = Xw^\star$	Basis pursuit: $\hat{w}_1 = \mathrm{argmin}_w \; ||w||_1 \; s.t. \; y = Xw$	Perfect recovery w.h.p. for $n \sim k \log d$
	⬇ when observations are noisy	
Noisy $y = Xw^\star + \xi$	Lasso: $\hat{w}_\lambda = \mathrm{argmin}_w \; ||y - Xw||_2^2 + \lambda ||w||_1$	Estimation error achieves minimax optimal rate $O\left(\frac{k \log d}{n}\right)$ for best λ

Recall: small ℓ_1-norm → small statistical bias

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

for $i = 1, \dots, n$ samples and input and parameter dimension $d \gg n$

Noiseless $y = Xw^\star$	Basis pursuit: $\hat{w}_1 = \mathrm{argmin}_w \, ||w||_1 \; s.t. \; y = Xw$	Perfect recovery w.h.p. for $n \sim k \log d$
	↓ when observations are noisy	
Noisy $y = Xw^\star + \xi$	Lasso: $\hat{w}_\lambda = \mathrm{argmin}_w \, ||y - Xw||_2^2 + \lambda ||w||_1$	Estimation error achieves minimax optimal rate $O\left(\frac{k \log d}{n}\right)$ for best λ

➡ $p = 1$ has a strong inductive bias towards sparse solutions → small *statistical* bias!

e.g. BP: [Candes, Tao ‘05, Donoho ‘06], Lasso: [Bunea, Tsybakov, Wegkamp ‘07, vandeGeer ‘08], [Wainwright ‘09]

Recall: small ℓ_1-norm → small statistical bias

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

for $i = 1, \dots, n$ samples and input and parameter dimension $d \gg n$

Noiseless
$y = Xw^\star$

Basis pursuit: $\hat{w}_1 = \mathrm{argmin}_w \, ||w||_1 \; s.t. \; y = Xw$

Perfect recovery
w.h.p. for $n \sim k \log d$

when observations are noisy

Noisy
$y = Xw^\star + \xi$

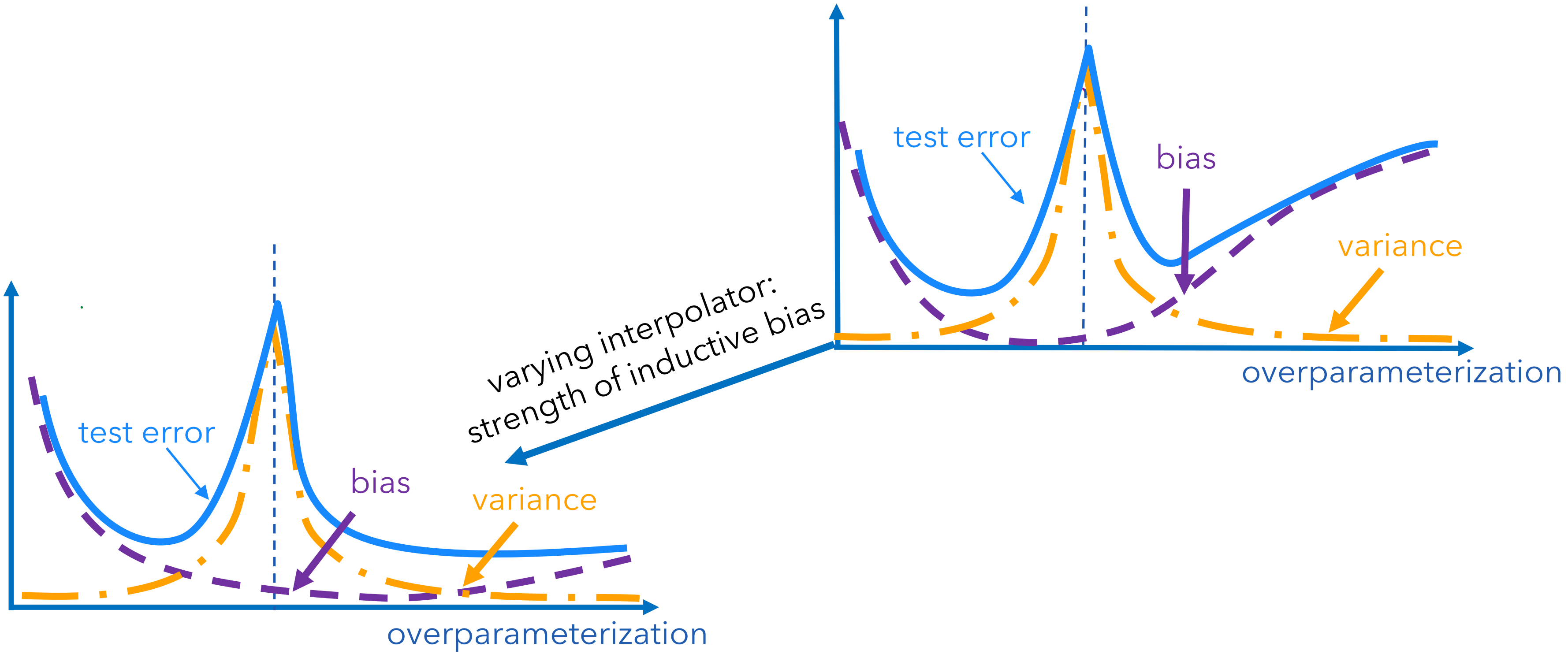
Lasso: $\hat{w}_\lambda = \mathrm{argmin}_w \, ||y - Xw||_2^2 + \lambda ||w||_1$

Estimation error achieves
optimal minimax rate
$O\left(\frac{k \log d}{n}\right)$ for best λ

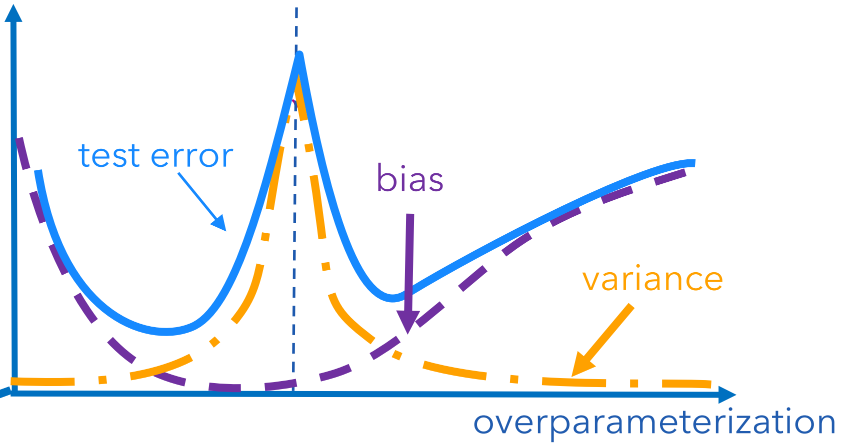
Previously unknown: prediction/estimation error of min-ℓ_1 interpolation for **noisy data**

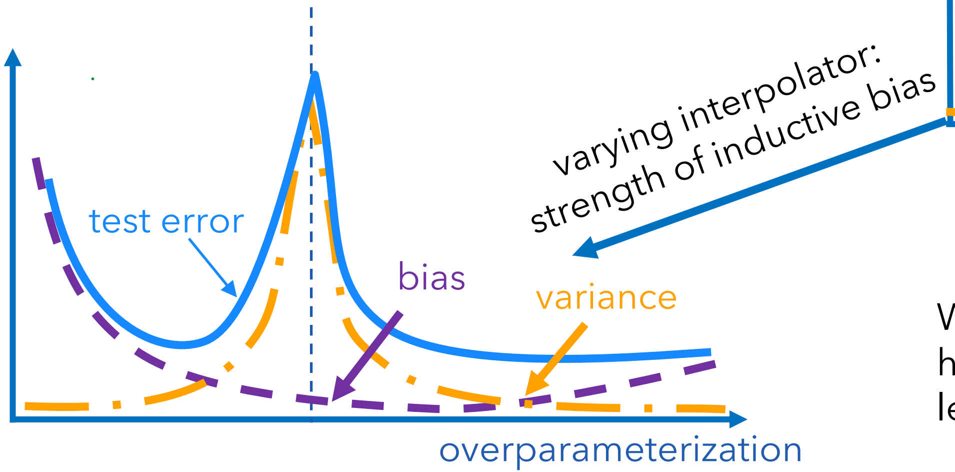
e.g. BP: [Candes, Tao '05, Donoho '06], Lasso: [Bunea, Tsybakov, Wegkamp '07, vandeGeer '08], [Wainwright '09]

For fixed distribution…



For fixed distribution…





When interpolating noise,
how strong of an inductive bias
leads to good generalization

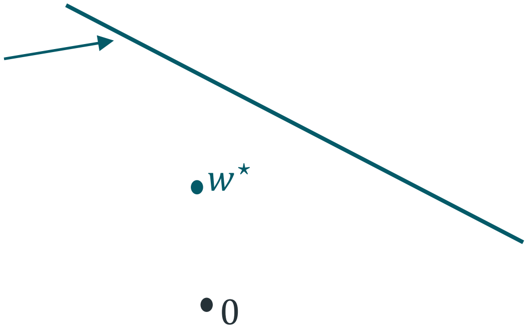
Inductive bias for noisy sparse linear models

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w ||w||_p \; s.t. \, y = Xw$

subspace of all linear interpolators $\{w\colon Xw = y = Xw^\star + \xi\}$

for i.i.d noise ξ_i



Inductive bias for noisy sparse linear models

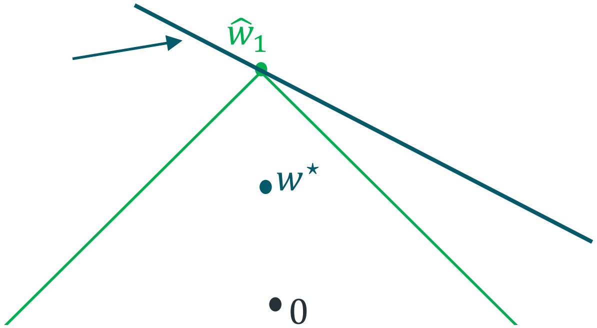
Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

- small $||w||_1$-norm encourages sparsity → aligns with $w^\star$ structure **(strong inductive bias)**

subspace of all
linear interpolators
$\{w\colon Xw = y = Xw^\star + \xi\}$

for i.i.d noise ξ_i



Inductive bias for noisy sparse linear models

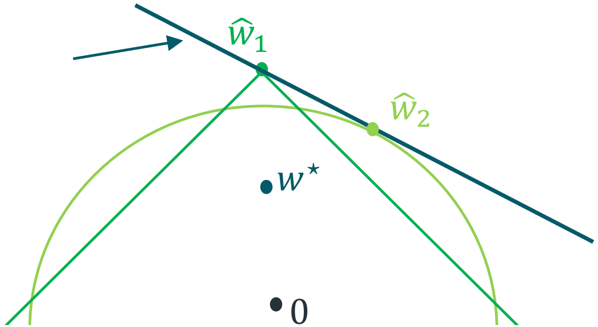
Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \mathrm{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

- small $||w||_1$-norm encourages sparsity → aligns with $w^\star$ structure **(strong inductive bias)**
- small $||w||_2$-norm → does not restrict search space in right way! **(weak inductive bias)**

subspace of all
linear interpolators
$\{w\colon Xw = y = Xw^\star + \xi\}$

for i.i.d noise ξ_i



Inductive bias for noisy sparse linear models

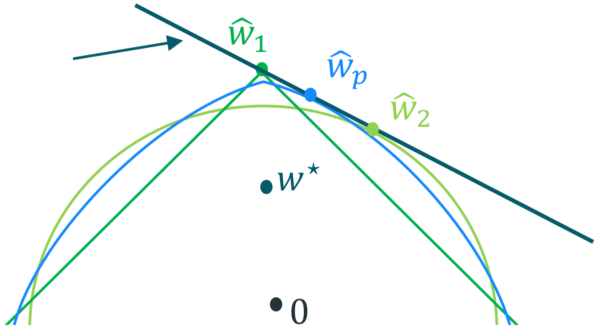
Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \mathrm{argmin}_w ||w||_p \; s.t. \; y = Xw$

- small $||w||_1$-norm encourages sparsity → aligns with $w^\star$ structure **(strong inductive bias)**
- small $||w||_2$-norm → does not restrict search space in right way! **(weak inductive bias)**

subspace of all
linear interpolators
$\{w : Xw = y = Xw^\star + \xi\}$

for i.i.d noise ξ_i



Varying inductive bias via $p \in [1,2]$

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w ||w||_p \; s.t. \; y = Xw$

- Consider overparameterized regime $d \gg n$, think of $d \propto n^\beta$ with $\beta > 1$ (high-dimensional)

Varying inductive bias via $p \in [1,2]$

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

- Consider overparameterized regime $d \gg n$, think of $d \propto n^\beta$ with $\beta > 1$ (high-dimensional)
- Compare estimators using tight, high-probability, non-asymptotic *statistical rates* of prediction error

$$\mathbb{E}_{x \sim N(0,I)} \left(x^\top \hat{w}_p - x^\top w^\star\right)^2 = \left|\left|\hat{w}_p - w^\star\right|\right|^2 = \Theta(h(n,d)) \text{ as } n \to \infty \text{ for some function } h \downarrow$$

Varying inductive bias via $p \in [1,2]$

Fixed distribution: $y_i = \langle w^\star, x_i \rangle + \xi_i$ with **sparse** $w^\star$, i.e. $||w||_0 = k \ll d$, i.i.d. noise ξ_i and $x_i \sim N(0, I)$

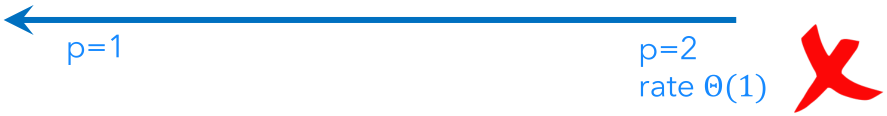
Min-ℓ_p-norm interpolation $\hat{w}_p = \mathrm{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

- Consider overparameterized regime $d \gg n$, think of $d \propto n^\beta$ with $\beta > 1$ (high-dimensional)
- Compare estimators using tight, high-probability, non-asymptotic *statistical rates* of prediction error

$$\mathbb{E}_{x \sim N(0,I)} \left(x^\top \hat{w}_p - x^\top w^\star\right)^2 = \left|\left|\hat{w}_p - w^\star\right|\right|^2 = \Theta(h(n,d)) \text{ as } n \to \infty \text{ for some function } h \downarrow$$

strong inductive bias towards sparsity ← p=1 —— p=2 no inductive bias towards sparsity

Strong inductive bias: $p = 1$



Inconsistent

but harmless interpolation

Strong inductive bias: $p = 1$

decreasing statistical bias

p=1

p=2
rate $\Theta(1)$

Inconsistent

but harmless
interpolation

Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors ξ but $O(\sigma^2)$ for ξ_i i.i.d. with variance σ^2

  [Chinot, Loeffler, vandeGeer '20], [Wojtaszczyk '10]

Inconsistent

decreasing statistical bias

p=1

p=2
rate $\Theta(1)$

but harmless interpolation

Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors ξ but $O(\sigma^2)$ for ξ_i i.i.d. with variance σ^2
  [Chinot, Loeffler, vandeGeer '20], [Wojtaszczyk '10]
- Lower bound for i.i.d. noise for sub-Gaussians $\Omega\left(\frac{\sigma^2}{\log\left(\frac{d}{n}\right)}\right)$ [Muthukumar, Vodrahalli, Subramanian, and Sahai '20]

Inconsistent

decreasing statistical bias

p=1

p=2
rate $\Theta(1)$

but harmless interpolation

Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors ξ but $O(\sigma^2)$ for ξ_i i.i.d. with variance σ^2
  [Chinot, Loeffler, vandeGeer '20], [Wojtaszczyk '10]
- Lower bound for i.i.d. noise for sub-Gaussians $\Omega\left(\frac{\sigma^2}{\log\left(\frac{d}{n}\right)}\right)$ [Muthukumar, Vodrahalli, Subramanian, and Sahai '20]
- Tight bounds for i.i.d. noise for Gaussian covariates $\frac{\sigma^2}{\log(d/n)} + O\left(\frac{\sigma^2}{\log^{3/2}(d/n)}\right)$ [Wang, Donhauser, Yang '22]

for $d \asymp n^\beta$ with $\beta > 1$ we obtain the rate $\Theta\left(\frac{1}{(\beta-1)\log n}\right)$

decreasing statistical bias

p=1

p=2
rate $\Theta(1)$

Inconsistent

but harmless interpolation

Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors ξ but $O(\sigma^2)$ for ξ_i i.i.d. with variance σ^2

  [Chinot, Loeffler, vandeGeer '20], [Wojtaszczyk '10]

- Lower bound for i.i.d. noise for sub-Gaussians $\Omega\left(\frac{\sigma^2}{\log(\frac{d}{n})}\right)$ [Muthukumar, Vodrahalli, Subramanian, and Sahai '20]

- Tight bounds for i.i.d. noise for Gaussian covariates $\frac{\sigma^2}{\log(d/n)} + O\left(\frac{\sigma^2}{\log^{3/2}(d/n)}\right)$ [Wang, Donhauser, Yang '22]

  for $d \asymp n^\beta$ with $\beta > 1$ we obtain the rate $\Theta\left(\frac{1}{(\beta-1)\log n}\right)$

Consistent

but harmful interpolation:

opt. regularized $O\left(\frac{k \log n}{n}\right)$

decreasing statistical bias

p=1

rate $\Theta\left(\frac{1}{\log n}\right) = \tilde{\Theta}(1)$

p=2

rate $\Theta(1)$

Inconsistent

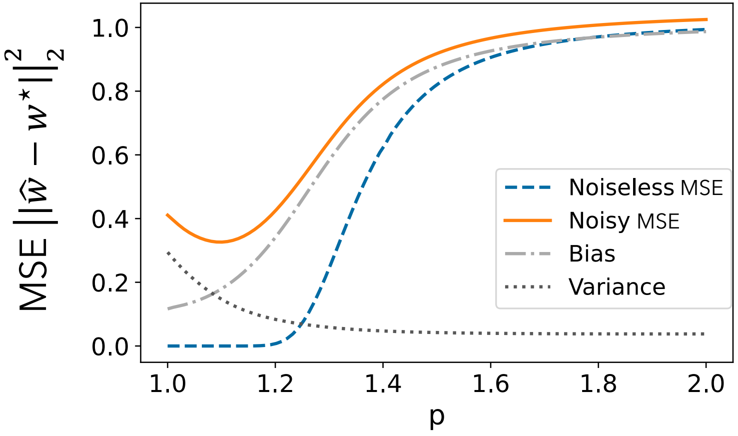
but harmless interpolation

The problem of $p = 1$ lies in the variance…

For $p = 1$ and $k = 1$, “sensitivity to noise” and variance larger than for $p = 2$

The problem of $p = 1$ lies in the variance…

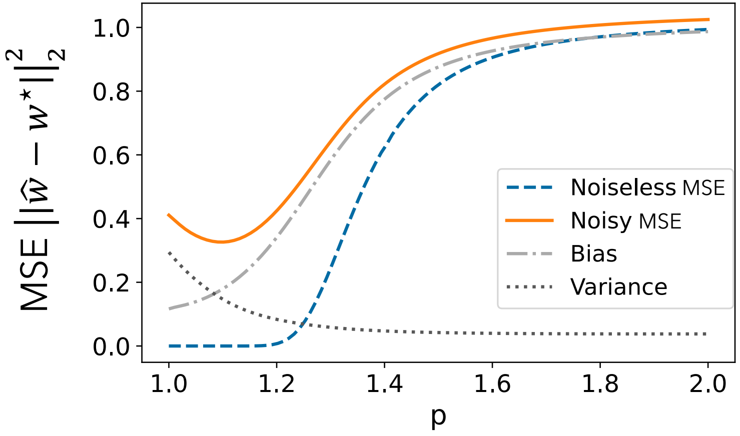
For $p = 1$ and $k = 1$, "sensitivity to noise" and variance larger than for $p = 2$



for $d = 5000, n = 100$

The problem of $p = 1$ lies in the variance…

For $p = 1$ and $k = 1$, "sensitivity to noise" and variance larger than for $p = 2$

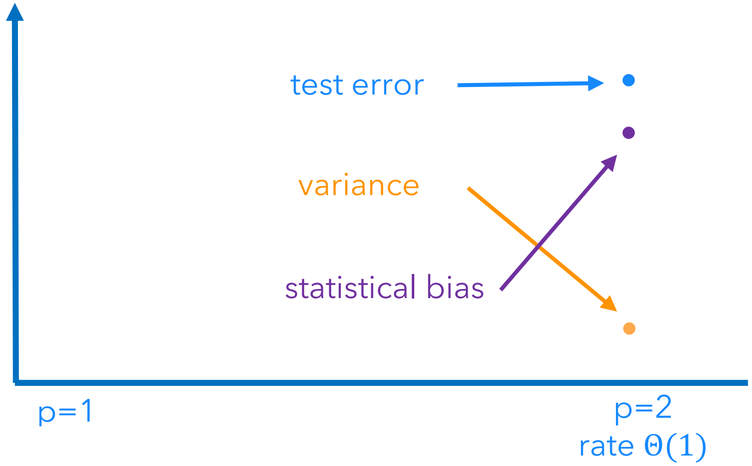


for $d = 5000, n = 100$

as overparameterization increases, variance decay is slower for $p = 1$ than for $p = 2$!

A bias-variance trade-off for $p \in [1,2]$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \;\; s.t. \, y = Xw$

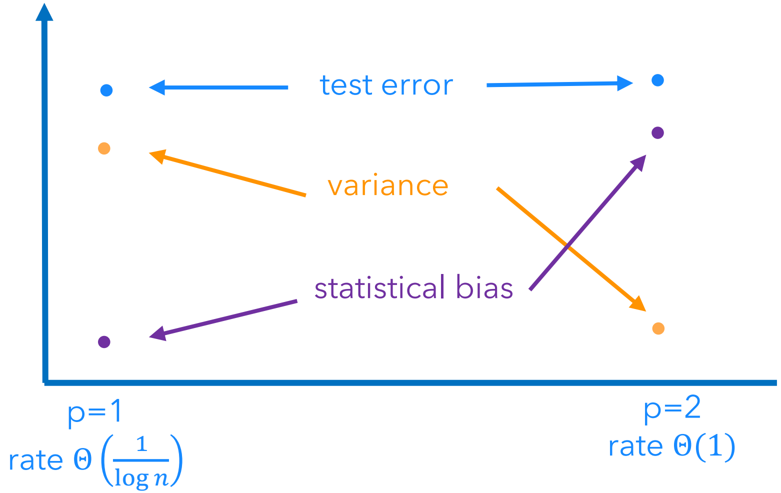


strong inductive bias towards sparsity

no inductive bias towards sparsity

A bias-variance trade-off for $p \in [1,2]$

Min-ℓ_p-norm interpolation $\hat{w}_p = \mathrm{argmin}_w \, ||w||_p \; s.t. \, y = Xw$

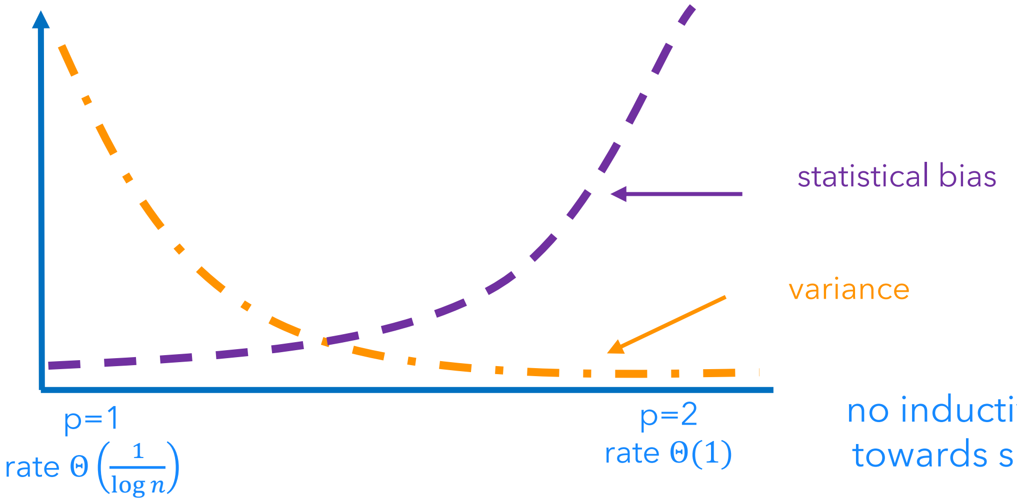


strong inductive bias towards sparsity

no inductive bias towards sparsity

A bias-variance trade-off for $p \in [1,2]$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$



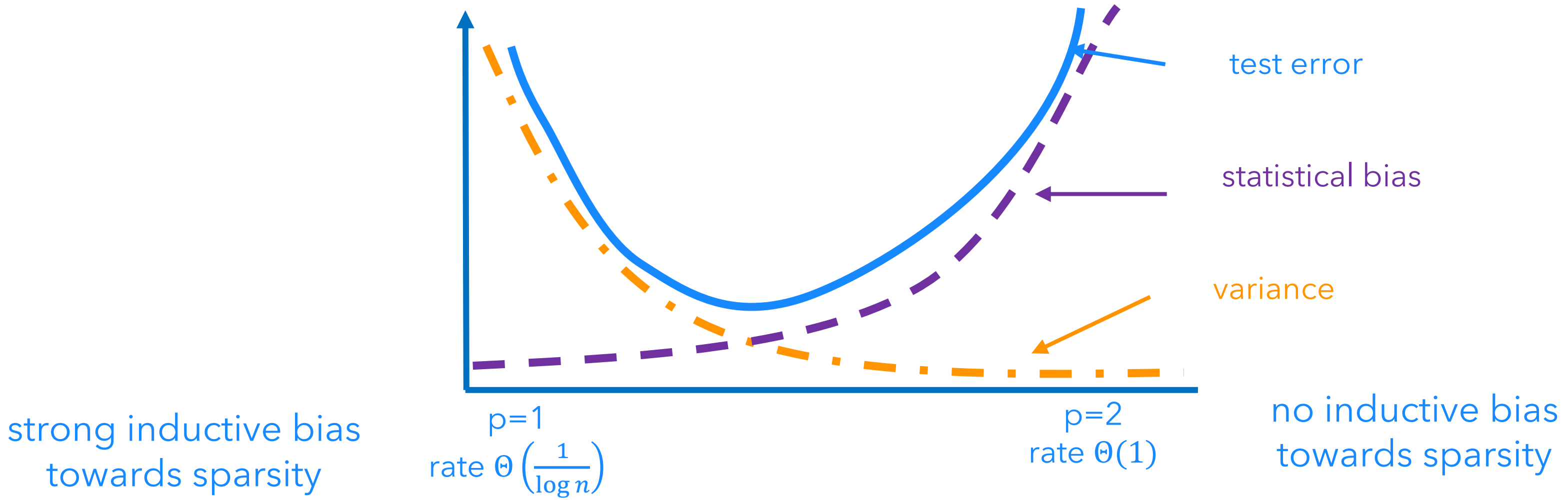
strong inductive bias towards sparsity

no inductive bias towards sparsity

Trade-off between bias and variance for interpolators via strength of inductive bias!

A bias-variance trade-off for $p \in [1,2]$

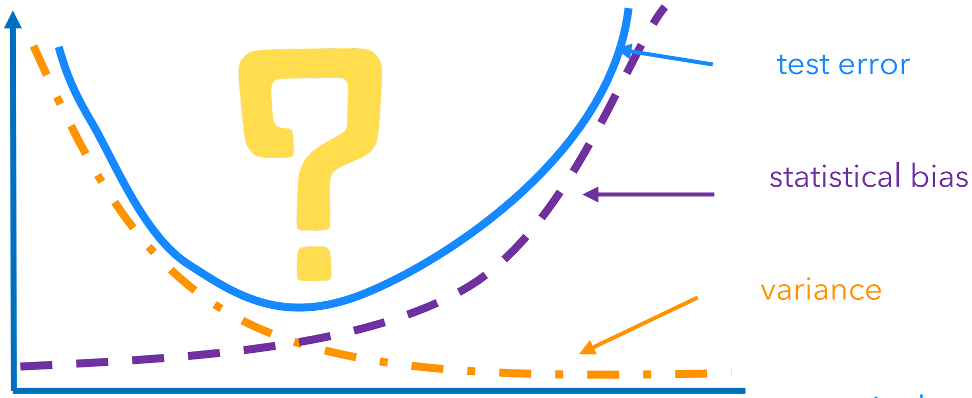
Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$



Trade-off between bias and variance for interpolators via strength of inductive bias!

A bias-variance trade-off for $p \in [1,2]$

Min-ℓ_p-norm interpolation $\hat{w}_p = \text{argmin}_w \, ||w||_p \;\; s.t. \, y = Xw$



strong inductive bias towards sparsity

p=1
rate $\Theta\left(\frac{1}{\log n}\right)$

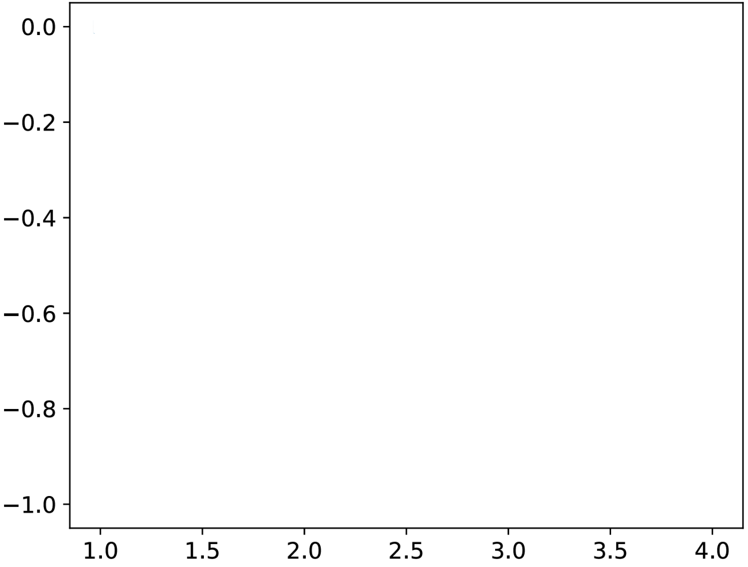
Which rates?

p=2
rate $\Theta(1)$

no inductive bias towards sparsity

Trade-off between bias and variance for interpolators via strength of inductive bias!

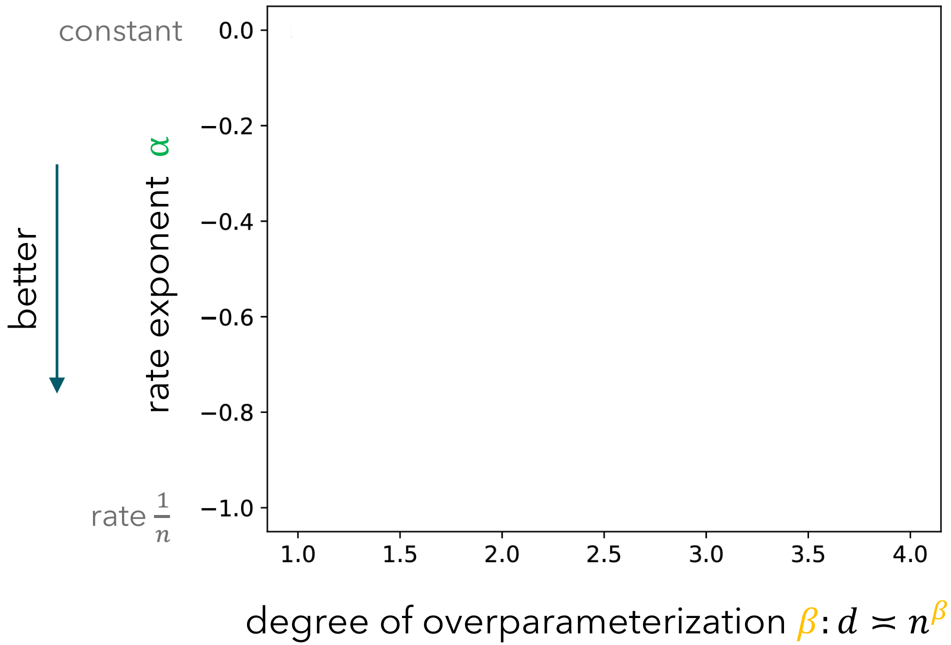
Tight bounds for $p \in [1, 2]$



degree of overparameterization β: $d \asymp n^{\beta}$

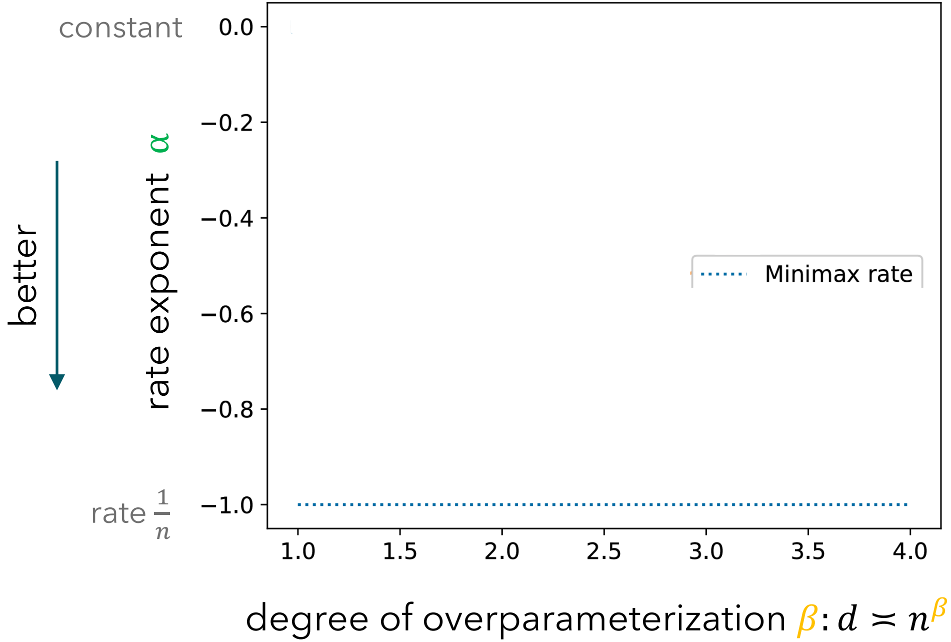
We plot α where $\left\|\widehat{w}_p - w^\star\right\|^2 = \widetilde{\Theta}(n^{\alpha})$ w.h.p.

Tight bounds for $p \in [1, 2]$



We plot α where $\left\|\widehat{w}_p - w^\star\right\|^2 = \widetilde{\Theta}(n^{\alpha})$ w.h.p.

Tight bounds for $p \in [1, 2]$

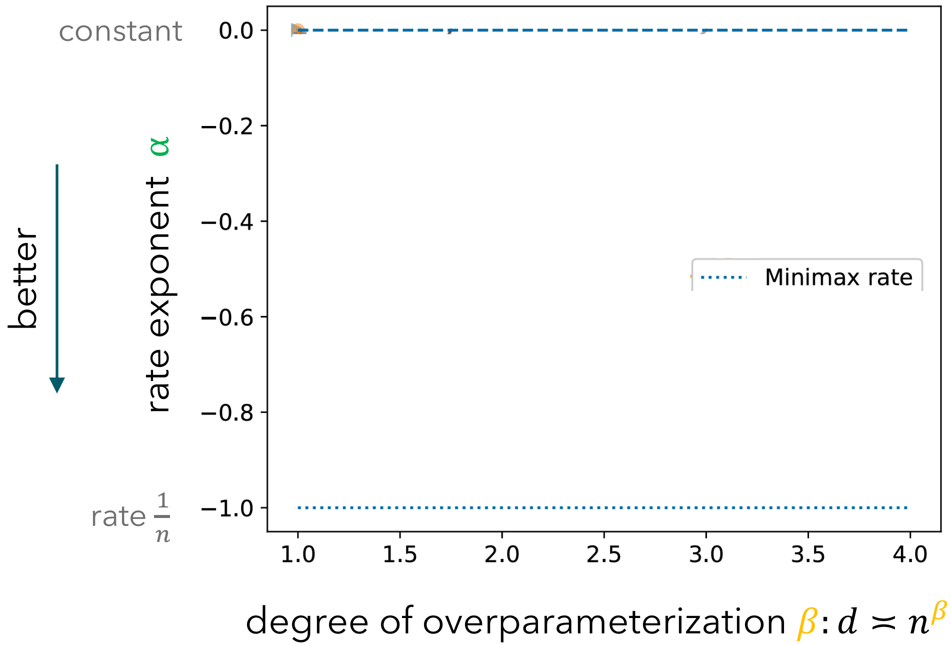


- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO)

  $\left\|\widehat{w}_\lambda - w^\star\right\|^2 = \widetilde{\Theta}(n^{-1}) \rightarrow \alpha = -1$

We plot α where $||\widehat{w}_p - w^\star||^2 = \widetilde{\Theta}(n^\alpha)$ w.h.p.

Tight bounds for $p \in [1, 2]$



- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO)

  $||\widehat{w}_\lambda - w^\star||^2 = \widetilde{\Theta}(n^{-1}) \ \rightarrow \ \alpha = -1$

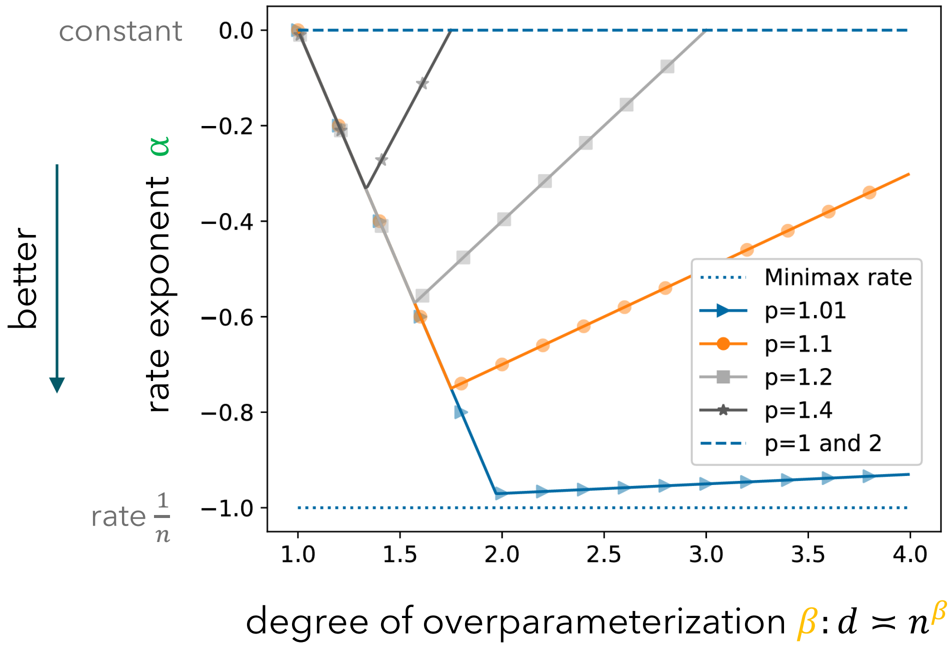
- Interpolators with $p = 1, 2$:

  $$||\widehat{w}_p - w^\star||^2 = \widetilde{\Theta}(1) \rightarrow \ \alpha = 0$$

For $p \geq 1$: [Wang, Donhauser, Yang '22]

We plot α where $||\widehat{w}_p - w^\star||^2 = \widetilde{\Theta}(n^\alpha)$ w.h.p.

Tight bounds for $p \in [1, 2]$



- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO)

  $||\widehat{w}_\lambda - w^\star||^2 = \widetilde{\Theta}(n^{-1}) \rightarrow \alpha = -1$

- Interpolators with $p = 1, 2$:

  $$||\widehat{w}_p - w^\star||^2 = \widetilde{\Theta}(1) \rightarrow \alpha = 0$$

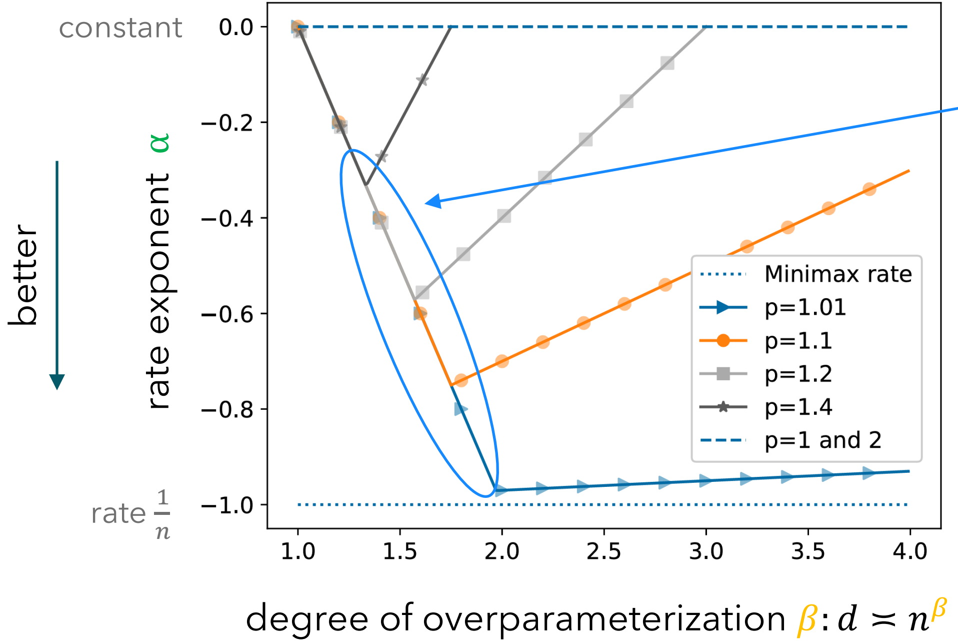
- Interpolators for $p \in (1,2)$:

  $$||\widehat{w}_p - w^\star||^2 = \widetilde{\Theta}(n^\alpha) \text{ with } \alpha < 0$$

For $p \in [1,2)$: [Wang, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

We plot α where $\left\|\widehat{w}_p - w^\star\right\|^2 = \widetilde{\Theta}(n^{\alpha})$ w.h.p.

Tight bounds for $p \in [1, 2]$

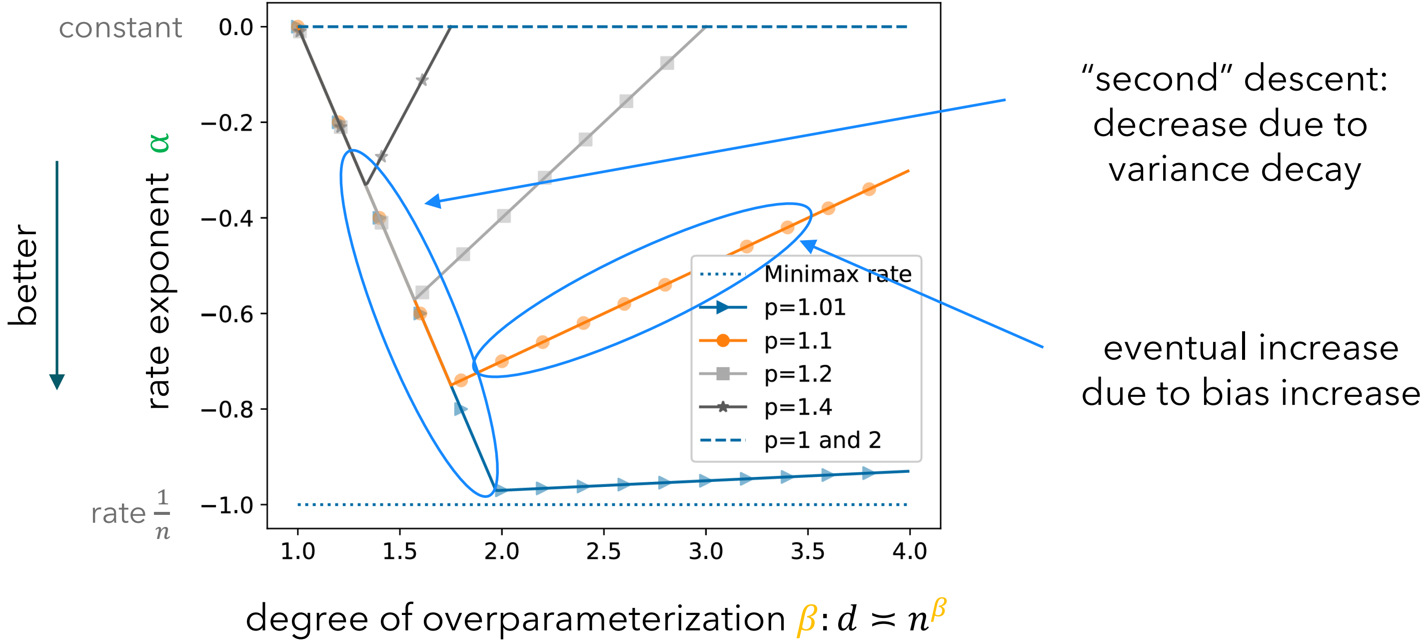


"second" descent:
decrease due to
variance decay

For $p \in [1,2)$: [Wang, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

We plot α where $\left\|\widehat{w}_p - w^\star\right\|^2 = \widetilde{\Theta}(n^\alpha)$ w.h.p.

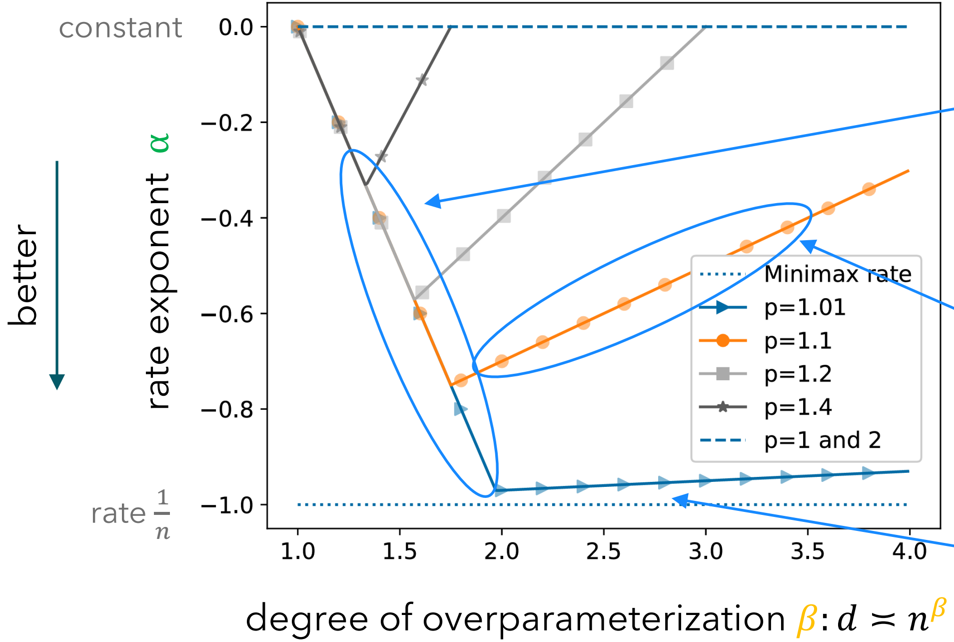
Tight bounds for $p \in [1, 2]$



For $p \in [1,2)$: [Wang, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

We plot α where $\left\|\widehat{w}_p - w^\star\right\|^2 = \widetilde{\Theta}(n^{\alpha})$ w.h.p.

Tight bounds for $p \in [1, 2]$



"second" descent: decrease due to variance decay

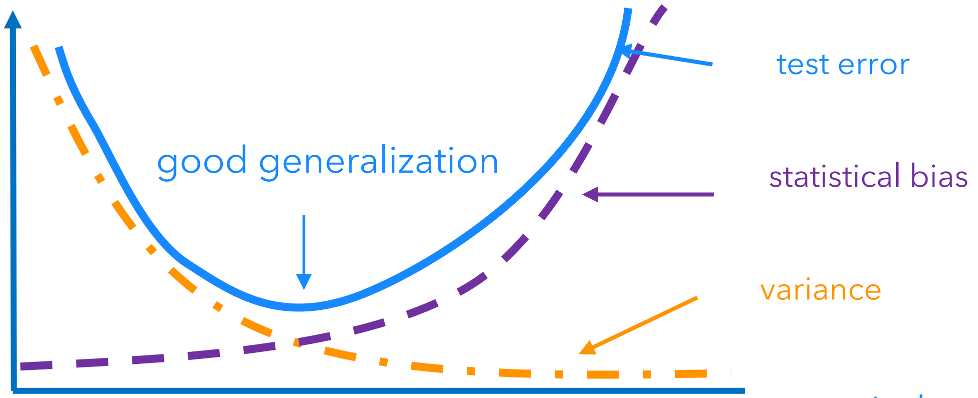
eventual increase due to bias increase

good generalization & ≈ harmless interpolation

For $p \in [1,2)$: [Wang, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

A new bias-variance trade-off for interpolators

Min-ℓ_p-norm interpolation $\hat{w}_p = \operatorname{argmin}_w ||w||_p \; s.t.\, y = Xw$



strong inductive bias towards sparsity

p=1
rate $\Theta\left(\frac{1}{\log n}\right)$

1<p<2
rate $\Theta(n^{\alpha})$
$-1 < \alpha < 0$

p=2
rate $\Theta(1)$

no inductive bias towards sparsity

Take-away: medium strength of inductive bias is best when interpolating noise

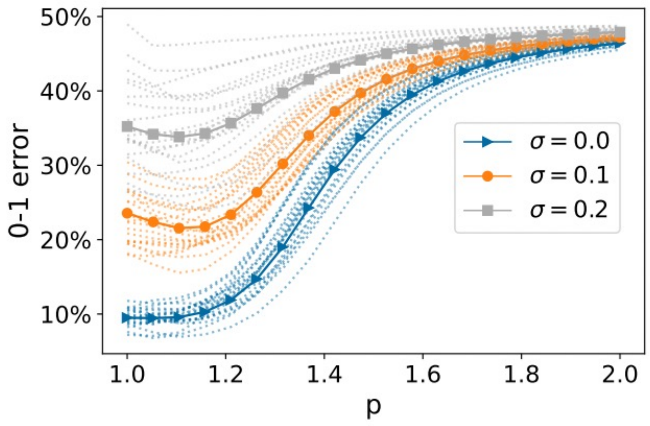
How transferable is this "new" intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi '15] with localized convergence [Koehler, Zhou, Sutherland, Srebro '21] carries over to lin. classification

[Stojanovic, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

How transferable is this “new” intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi ’15] with localized convergence [Koehler, Zhou, Sutherland, Srebro ’21] carries over to lin. classification

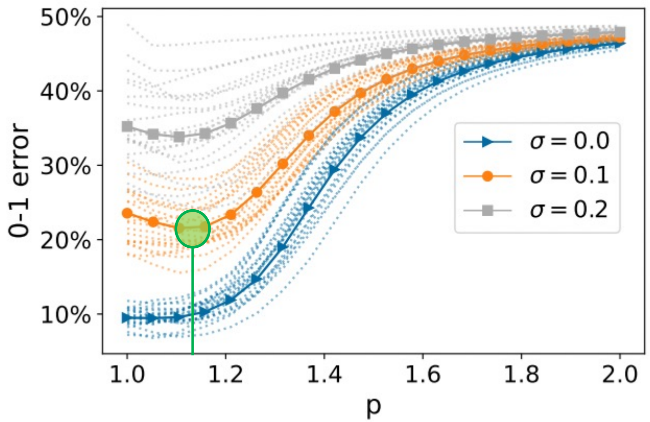


Synthetic experiment:
Isotropic Gaussians with $d \sim 5000, n \sim 100$

[Stojanovic, Donhauser, Yang ’22], [Donhauser, Ruggeri, Stojanovic, Yang ’22]

How transferable is this “new” intuition?

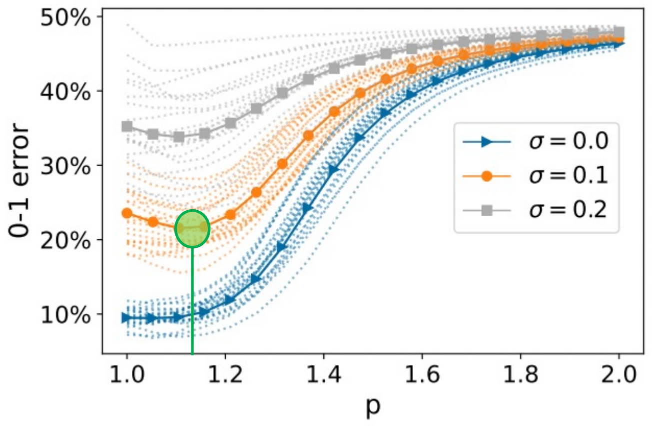
- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi ’15] with localized convergence [Koehler, Zhou, Sutherland, Srebro ’21] carries over to lin. classification



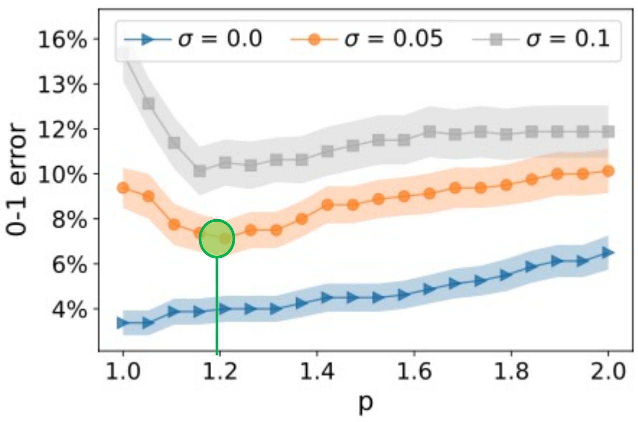
Synthetic experiment:
Isotropic Gaussians with $d \sim 5000, n \sim 100$

How transferable is this "new" intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi '15] with localized convergence [Koehler, Zhou, Sutherland, Srebro '21] carries over to lin. classification



Synthetic experiment:
Isotropic Gaussians with $d \sim 5000, n \sim 100$



Real-world experiment:
Leukemia dataset with $d \sim 7000, n \sim 70$

[Stojanovic, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

How transferable is this "new" intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi '15] with localized convergence [Koehler, Zhou, Sutherland, Srebro '21] carries over to lin. classification [Donhauser, Ruggeri, Stojanovic, Yang '22]

open: theory is still incomplete and restricted to Gaussians!

How transferable is this "new" intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi '15] with localized convergence [Koehler, Zhou, Sutherland, Srebro '21] carries over to lin. classification [Donhauser, Ruggeri, Stojanovic, Yang '22]

open: theory is still incomplete and restricted to Gaussians!

- Intuition carries over to high-dimensional kernel learning with convolutional kernels where bias and variance vary with inductive bias [Aerni, Milanta, Donhauser, Yang '23]

How transferable is this "new" intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi '15] with localized convergence [Koehler, Zhou, Sutherland, Srebro '21] carries over to lin. classification [Donhauser, Ruggeri, Stojanovic, Yang '22]

  open: theory is still incomplete and restricted to Gaussians!

- Intuition carries over to high-dimensional kernel learning with convolutional kernels where bias and variance vary with inductive bias [Aerni, Milanta, Donhauser, Yang '23]

- Preliminary experiments for neural networks also suggest this behavior for rotational invariance and filter size…

Nonlinear structure: Rotational invariance for WideResNet

- Satellite images (EuroSAT) to be classified in terms of type of land usage

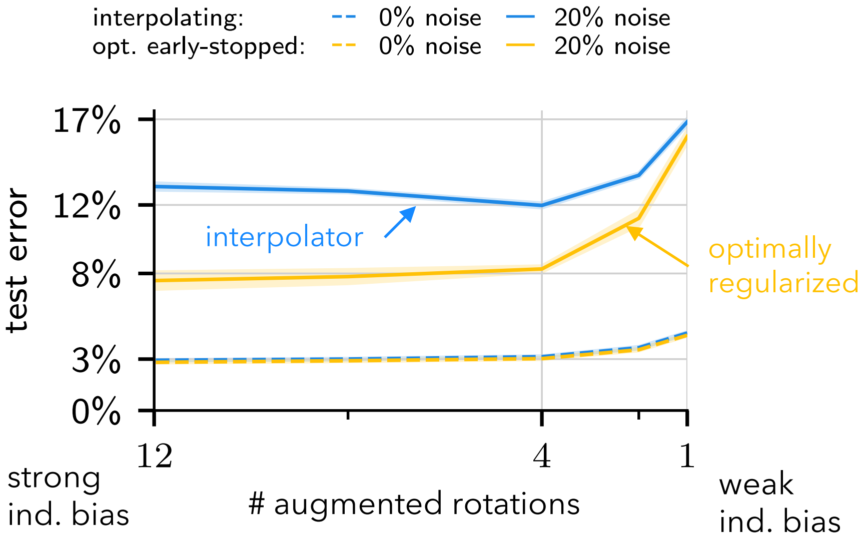
- strength of rotational invariance via "amount of" data augmentation

[Aerni, Milanta, Donhauser, Yang '23]

Nonlinear structure: Rotational invariance for WideResNet

- Satellite images (EuroSAT) to be classified in terms of type of land usage

- strength of rotational invariance via "amount of" data augmentation

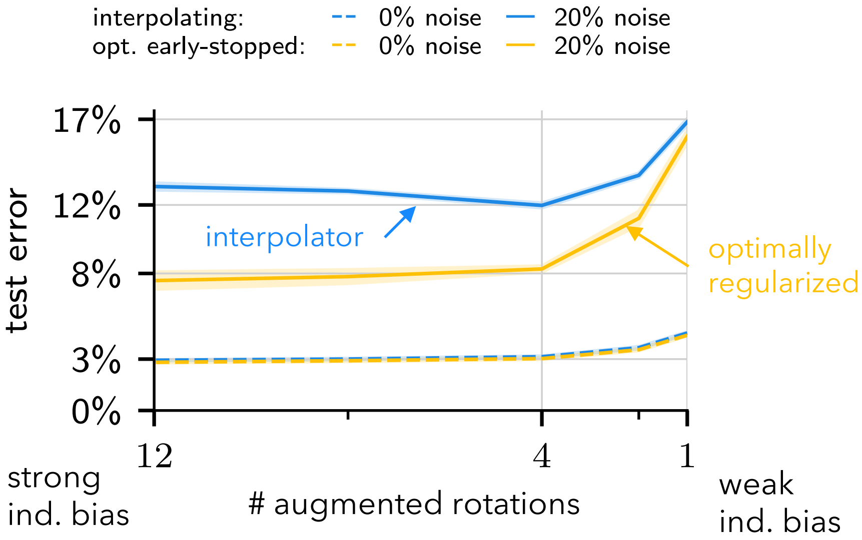


[Aerni, Milanta, Donhauser, Yang '23]

Nonlinear structure: Rotational invariance for WideResNet

- Satellite images (EuroSAT) to be classified in terms of type of land usage

- strength of rotational invariance via "amount of" data augmentation



Confirmed: medium strength of inductive bias is best when interpolating noise

[Aerni, Milanta, Donhauser, Yang '23]

Open: How transferable is this "new" intuition?

- Proof technique using Convex Gaussian Minmax Theorem [Thrampoulidis, Oymak, Hassibi '15] with localized convergence* [Koehler, Zhou, Sutherland, Srebro '21] carries over to classification [Donhauser, Ruggeri, Stojanovic, Yang '22]

  open: theory is still incomplete and restricted to Gaussians!

- Intuition carries over to high-dimensional kernel learning with convolutional kernels where bias and variance vary with inductive bias [Aerni, Milanta, Donhauser, Yang '23]

- Preliminary experiments for neural networks also suggest this behavior for rotational invariance and filter size

  open: comprehensive experimental NN study!

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

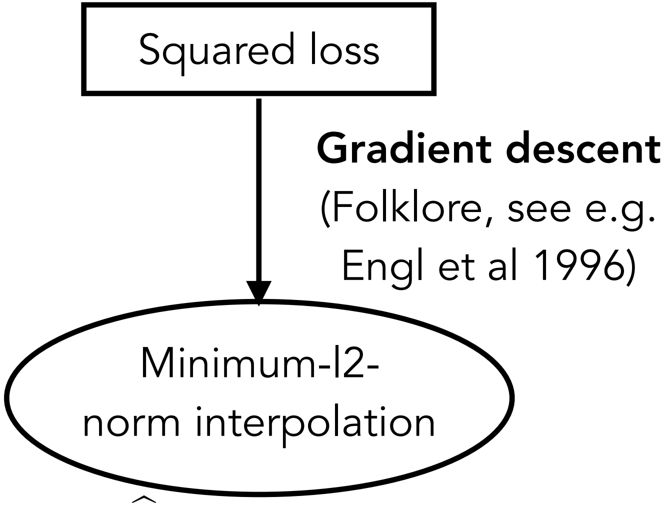
Classification-vs-regression: A tale of two loss functions

	0-1 loss	Squared loss
Logistic loss		
Squared loss		Regression

Classification-vs-regression: A tale of two loss functions

	0-1 loss	Squared loss
Logistic loss	Classification, most popular	
Squared loss	Classification, less popular	Regression

Differences in training loss functions



$$\widehat{\theta}_2 = \arg\min \|\theta\|_2$$

$$\text{subject to}$$

$$X_i^\top \theta = Y_i, i \in [n].$$

- Closed-form
- Linked to MLE under additive noise

Differences in training loss functions

Logistic loss

Gradient descent
(Soudry et al,
Ji & Telgarsky, 2018)

Hard-margin SVM

$$\widehat{\theta}_{\mathsf{SVM}} = \arg\min \|\theta\|_2 \text{ subject to } Y_i \cdot X_i^\top \theta \geq 1, i \in [n].$$

Squared loss

Gradient descent
(Folklore, see e.g.
Engl et al 1996)

Minimum-l2-norm interpolation

$$\widehat{\theta}_2 = \arg\min \|\theta\|_2 \text{ subject to } X_i^\top \theta = Y_i, i \in [n].$$

- Closed-form
- Linked to MLE under additive noise

Differences in training loss functions

Logistic loss → **Gradient descent** (Soudry et al, Ji & Telgarsky, 2018) → **Hard-margin SVM**

$$\widehat{\theta}_{\text{SVM}} = \arg\min \|\theta\|_2$$
$$\text{subject to}$$
$$Y_i \cdot X_i^\top \theta \geq 1, i \in [n].$$

- Not closed-form
- Linked to MLE under logistic noise

Squared loss → **Gradient descent** (Folklore, see e.g. Engl et al 1996) → **Minimum-l2-norm interpolation**

$$\widehat{\theta}_2 = \arg\min \|\theta\|_2$$
$$\text{subject to}$$
$$X_i^\top \theta = Y_i, i \in [n].$$

- Closed-form
- Linked to MLE under additive noise

Differences in test loss functions

Regression: Test MSE

$$\mathcal{E}_{\mathsf{MSE}} = \mathbb{E}\left[(X^\top(\widehat{\theta} - \theta^*))^2\right]$$

Classification: Test 0-1 error

$$\mathcal{E}_{0-1} = \mathbb{E}\left[\mathbb{I}[\mathrm{sgn}(X^\top\widehat{\theta}) \neq \mathrm{sgn}(X^\top\theta^*)]\right]$$

Differences in test loss functions

Regression: Test MSE

$$\mathcal{E}_{\mathsf{MSE}} = \mathbb{E}\left[(X^\top(\widehat{\theta} - \theta^*))^2\right]$$

Classification: Test 0-1 error

$$\mathcal{E}_{0-1} = \mathbb{E}\left[\mathbb{I}[\mathrm{sgn}(X^\top\widehat{\theta}) \neq \mathrm{sgn}(X^\top\theta^*)]\right]$$

Two core challenges when analyzing classification:

1. Hard-margin SVM does not have a closed-form solution, unlike minimum-l2-norm interpolation
2. 0-1 error metric challenging to sharply analyze as compared to MSE

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
    - For fixed interpolator, certain problem instances/distributions are more benign
    - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

One analysis path for l2, step 1: showing that **SVM = interpolation**

Fourier features
on 1-dimensional data,
isotropic covariance

n = 32,
d = 1000

One analysis path for l2, step 1: showing that **SVM = interpolation**

Fourier features
on 1-dimensional data,
isotropic covariance

n = 32,
d = 1000

Result (Hsu, Muthukumar and Xu 2021)**: hard margin SVM = minimum-l2-norm interpolation on binary labels** in spiked covariance ensemble if $d \gg n \log n$ and $R \ll \frac{d}{n}$

Conditions for general anisotropic covariances also provided in terms of "effective ranks" in Hsu et al (2021)

One analysis path for l2, step 1: showing that **SVM = interpolation**

Fourier features
on 1-dimensional data,
isotropic covariance

n = 32,
d = 1000

Result (Hsu, Muthukumar and Xu 2021)**: hard margin SVM = minimum-l2-norm interpolation on binary labels** in spiked covariance ensemble if $d \gg n \log n$ and $R \ll \frac{d}{n}$

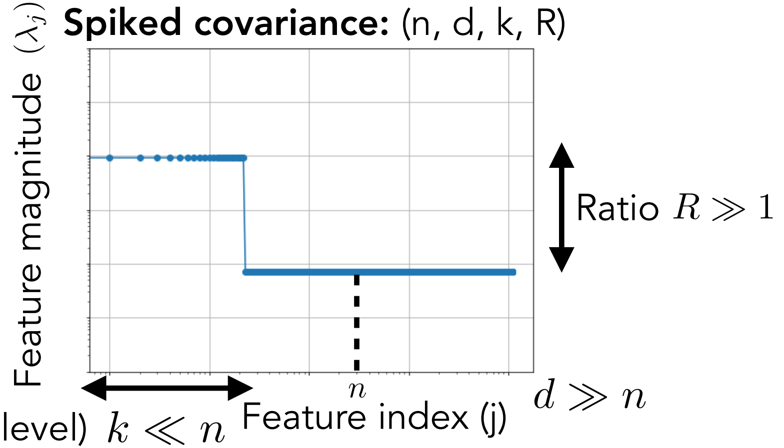
Implication: SVM has a closed-form expression, can be more easily analyzed!

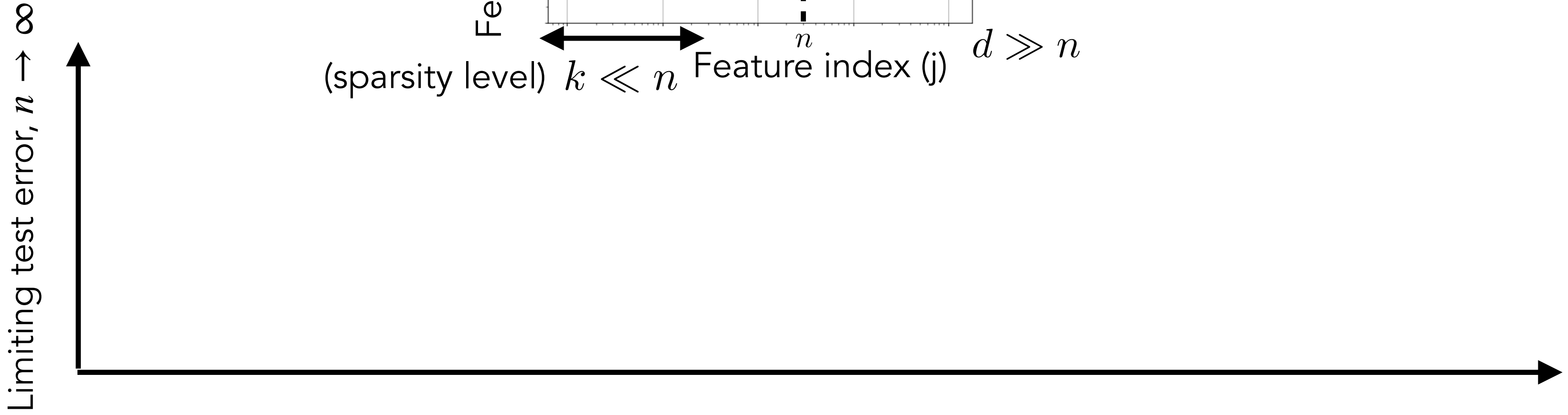
Conditions for general anisotropic covariances also provided in terms of "effective ranks" in Hsu et al (2021)

One analysis path for l2, step 2: analyzing 0-1 error of **interpolator**

Spiked covariance: (n, d, k, R)

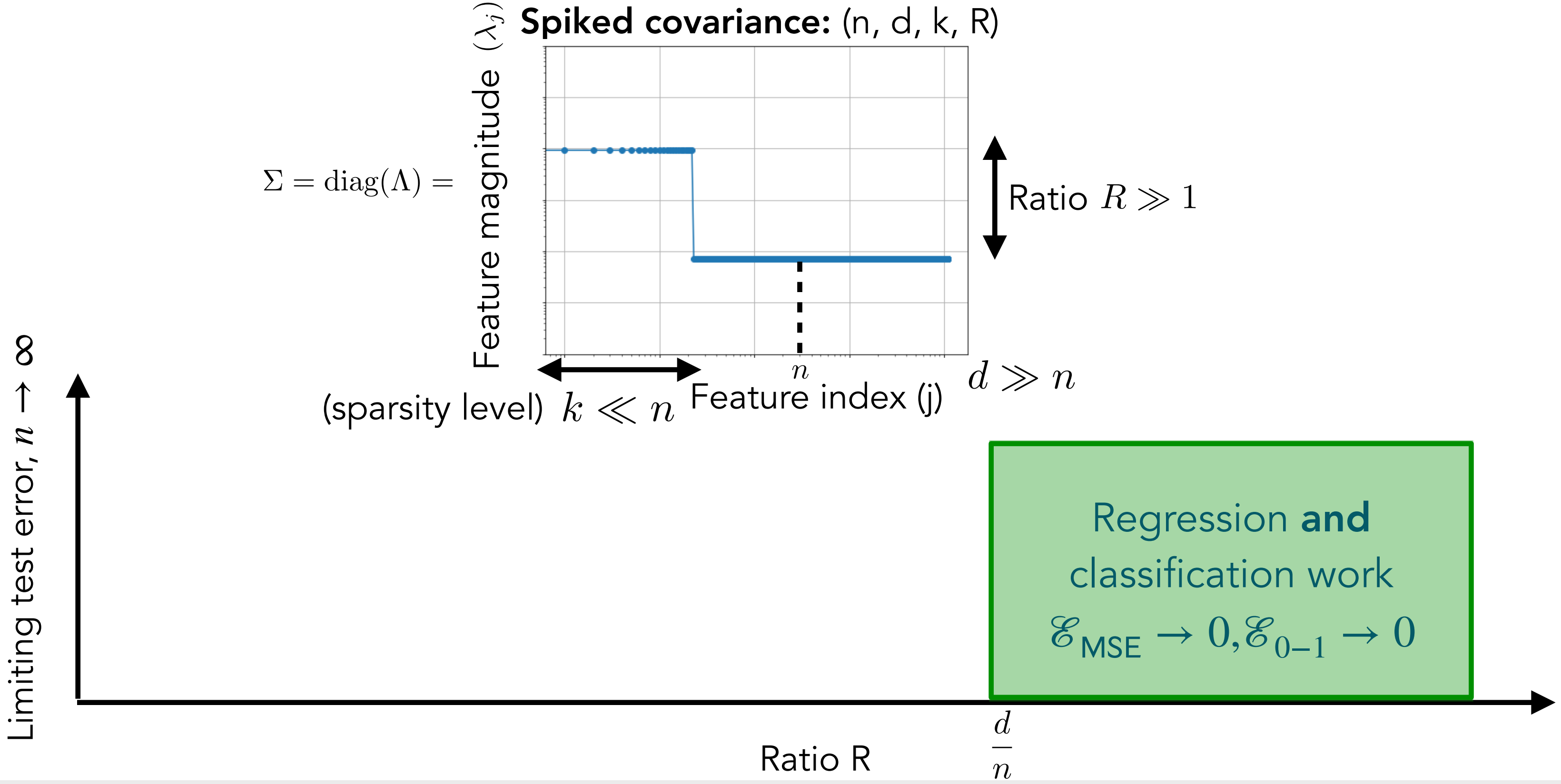
$\Sigma = \text{diag}(\Lambda) =$





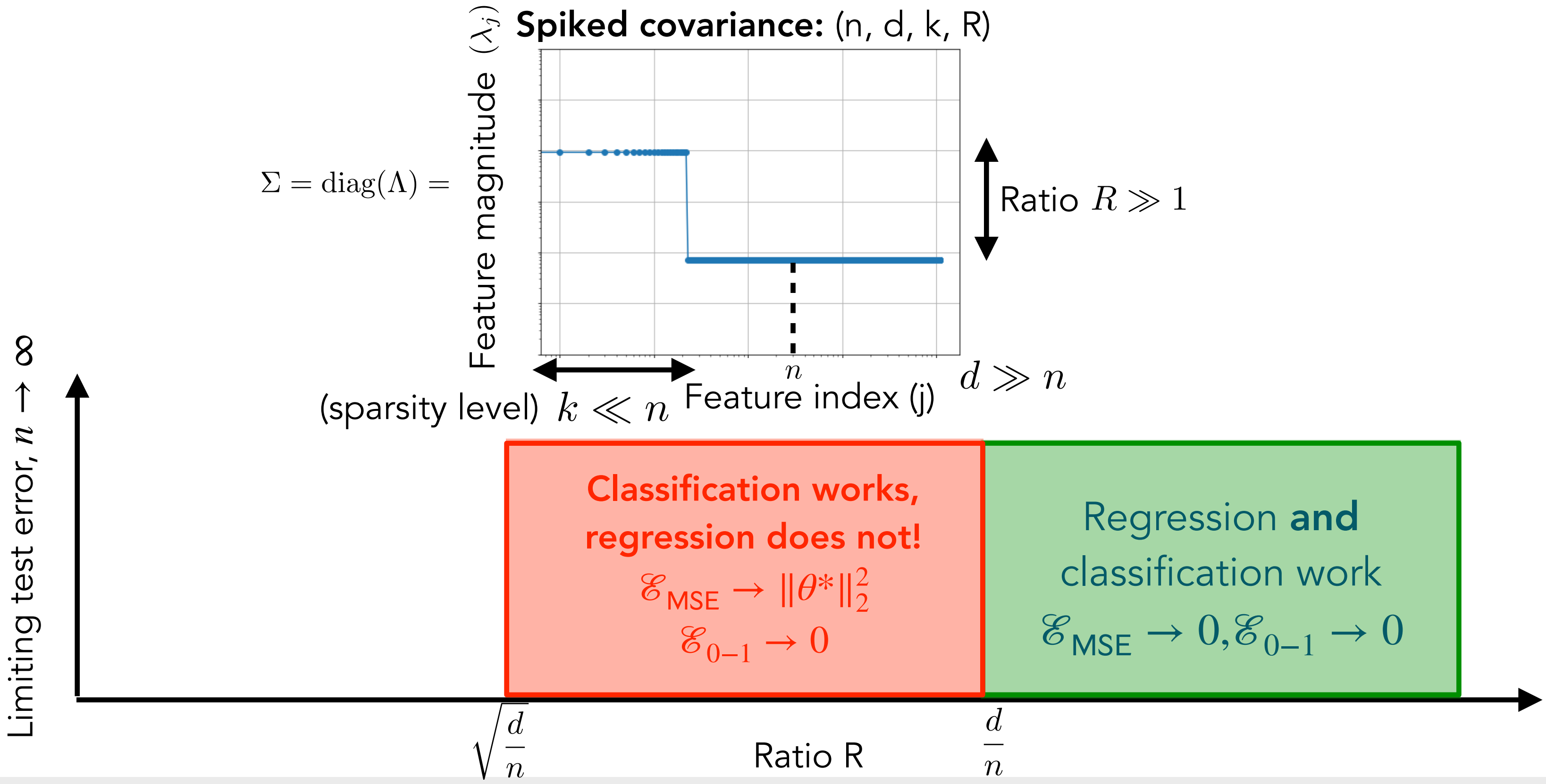
[Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR'21]

One analysis path for l2, step 2: analyzing 0-1 error of **interpolator**



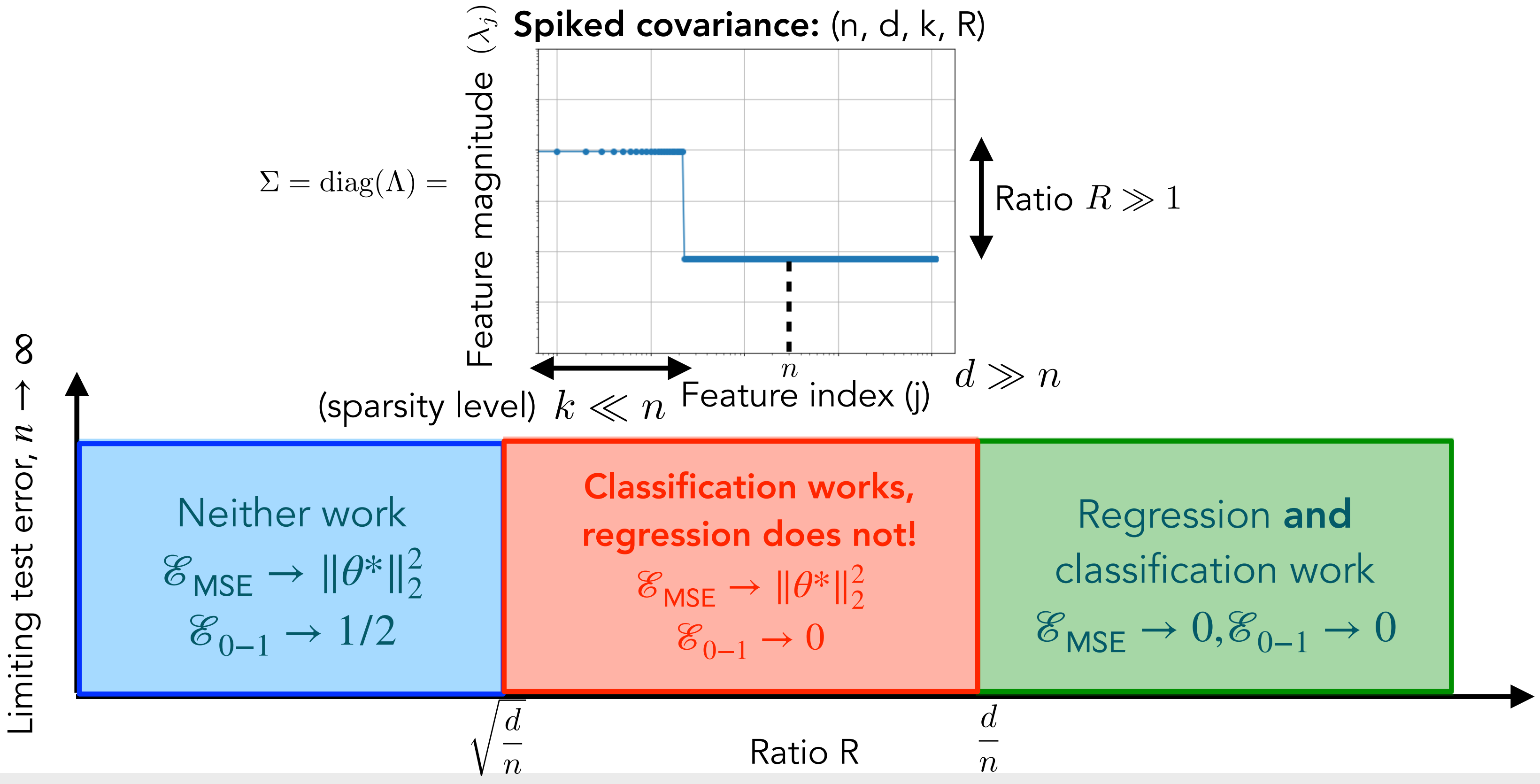
[Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR'21]

One analysis path for l2, step 2: analyzing 0-1 error of **interpolator**



[Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR’21]

One analysis path for l2, step 2: analyzing 0-1 error of **interpolator**



[Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR'21]

Takeaways for classification with l2-minimizing solutions

- Different **training loss functions** could yield **similar or even identical solutions**

Takeaways for classification with l2-minimizing solutions

- Different **training loss functions** could yield **similar or even identical solutions**
- Classification 0-1 test loss is **much more benign than regression MSE**; so l2-inductive bias could work better for classification tasks

Plan today…

Part I: For linear regression, we discuss how

- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

Part II: For classification, we discuss the

- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data

Benign overfitting in neural networks

- Most theoretical works on benign overfitting focus on linear/kernel setting.
- We’ll discuss recent works in neural networks and open questions.

Benign overfitting in neural networks

- Most theoretical works on benign overfitting focus on linear/kernel setting.
- We'll discuss recent works in neural networks and open questions.
- Notably: all results on benign overfitting in neural nets require ambient dimension $d \gg n$
- Very unsatisfying: neural nets can be overparameterized in $d \ll n$ regime, when is overfitting benign in this setting?

Which estimators do we care about?

Model	Algorithm	Setting	Estimator
Linear	Gradient descent	Classification	ℓ_2 max-margin
Linear	Gradient descent	Regression	ℓ_2 min-norm interpolator
Linear	Adaboost	Classification	ℓ_1 max-margin
Linear	Basis pursuit	Regression	ℓ_1 min-norm interpolator
Neural nets	Gradient descent	Classification	?
Neural nets	Gradient descent	Regression	?

Which estimators do we care about?

Model	Algorithm	Setting	Estimator
Linear	Gradient descent	Classification	ℓ_2 max-margin
Linear	Gradient descent	Regression	ℓ_2 min-norm interpolator
Linear	Adaboost	Classification	ℓ_1 max-margin
Linear	Basis pursuit	Regression	ℓ_1 min-norm interpolator
Neural nets	Gradient descent	Classification	?
Neural nets	Gradient descent	Regression	?

- Next: implicit bias of GD in neural net classification.
- After: "trajectory analysis", directly analyzing properties of neural nets trained by GD

Telgarsky'13, Soudry-Hoffer-Nacson-Gunasekar-Srebro'18, Ji-Telgarsky'18, …

Implicit bias in neural networks

- Which interpolators do we care about for neural nets?
- We'll focus on classification tasks, training by GD/GF on logistic loss.
    - Very little known about implicit bias of GD for neural nets in regression setting.

Implicit bias in neural networks

- Which interpolators do we care about for neural nets?
- We'll focus on classification tasks, training by GD/GF on logistic loss.
    - Very little known about implicit bias of GD for neural nets in regression setting.

Theorem

For large class of neural nets, if GD/GF $\theta(t)$ reaches a small enough loss, then $\theta(t)$ converges in direction to a first-order stationary point (KKT point) of the ℓ^2-max margin problem,

$$\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i f(x_i;\theta) \geq 1,\ \forall i \in [n]. \tag{1}$$

Implicit bias in neural networks

- Which interpolators do we care about for neural nets?
- We'll focus on classification tasks, training by GD/GF on logistic loss.
  - Very little known about implicit bias of GD for neural nets in regression setting.

Theorem

For large class of neural nets, if GD/GF $\theta(t)$ reaches a small enough loss, then $\theta(t)$ converges in direction to a first-order stationary point (KKT point) of the ℓ^2-max margin problem,

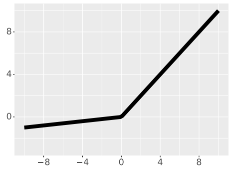
$$\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i f(x_i;\theta) \geq 1, \forall i \in [n]. \tag{1}$$

- KKT point does not imply even local optimality in general.
- In general, very little is known about KKT points of (1).

Implicit bias in neural networks

- A setting where we understand KKT points of max-margin: two-layer leaky ReLU nets with nearly-orthogonal data. ($\phi(q) = \max(\gamma q, q)$)

Implicit bias in neural networks



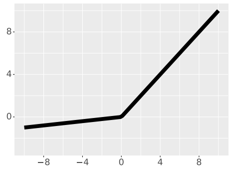
- A setting where we understand KKT points of max-margin: two-layer leaky ReLU nets with nearly-orthogonal data. ($\phi(q) = \max(\gamma q, q)$)

$$f(x;\theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad a_j \in \{\pm 1/\sqrt{m}\},$$

$$\boxed{\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|}.$$

- Satisfied in many settings w.h.p. when $d \gg n^2$ and $(x_i, y_i) \overset{\text{i.i.d.}}{\sim}$ P (e.g., $x \sim \mathsf{N}(0, I_d)$)

Implicit bias in neural networks



- A setting where we understand KKT points of max-margin: two-layer leaky ReLU nets with nearly-orthogonal data. ($\phi(q) = \max(\gamma q, q)$)

$$f(x;\theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x\rangle), \quad a_j \in \{\pm 1/\sqrt{m}\},$$

$$\boxed{\|x_i\|^2 \gg n \max_{k\neq j} |\langle x_j, x_k\rangle|}.$$

- Satisfied in many settings w.h.p. when $d \gg n^2$ and $(x_i, y_i) \overset{\text{i.i.d.}}{\sim}$ P (e.g., $x \sim \mathsf{N}(0, I_d)$)

Theorem

Suppose data is nearly orthogonal. If θ satisfies KKT conditions for ℓ^2-max-margin, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \mathrm{sgn}\big(f(x;\theta)\big) = \mathrm{sgn}\big(\langle \textstyle\sum_{i=1}^{n} s_i y_i x_i, x\rangle\big),$$

where $s_i > 0$ satisfy $\max_{i,j} s_i/s_j = O(1)$.

Frei-Vardi-Bartlett-Srebro-Hu'23

Implicit bias in neural networks

Theorem

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If θ satisfies KKT conditions for ℓ^2-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \mathrm{sgn}\big(f(x;\theta)\big) = \mathrm{sgn}\big(\langle \textstyle\sum_{i=1}^n s_i y_i x_i, x \rangle\big),$$

where $s_i > 0$ satisfy $\max_{i,j} s_i/s_j = O(1)$.

Implicit bias in neural networks

Theorem

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If θ satisfies KKT conditions for ℓ^2-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \operatorname{sgn}\big(f(x;\theta)\big) = \operatorname{sgn}\big(\langle \textstyle\sum_{i=1}^n s_i y_i x_i, x \rangle\big),$$

where $s_i > 0$ satisfy $\max_{i,j} {}^{s_i}/_{s_j} = O(1)$.

- Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.

Implicit bias in neural networks

Theorem

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If θ satisfies KKT conditions for ℓ^2-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \text{sgn}\big(f(x;\theta)\big) = \text{sgn}\big(\langle \textstyle\sum_{i=1}^n s_i y_i x_i, x \rangle\big),$$

where $s_i > 0$ satisfy $\max_{i,j} s_i/s_j = O(1)$.

- Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.
- Decision boundary is very simple, $\approx$ uniform average of data.

Implicit bias in neural networks

Theorem

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If θ satisfies KKT conditions for ℓ^2-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \operatorname{sgn}\big(f(x;\theta)\big) = \operatorname{sgn}\big(\langle \textstyle\sum_{i=1}^n s_i y_i x_i, x \rangle\big),$$

where $s_i > 0$ satisfy $\max_{i,j} {}^{s_i}/_{s_j} = O(1)$.

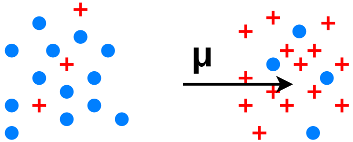
- Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.
- Decision boundary is very simple, $\approx$ uniform average of data.
- Linear model can capture behavior of nonlinear net, trained beyond NTK.

Benign overfitting of neural nets in mixture model

- KKT points for 2-layer leaky nets $\approx \sum_{i=1}^{n} y_i x_i$, when training data is nearly-orthogonal $\left(\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|\right)$.

Benign overfitting of neural nets in mixture model

- KKT points for 2-layer leaky nets $\approx \sum_{i=1}^n y_i x_i$, when training data is nearly-orthogonal $\left(\|x_i\|^2 \gg n \max_{k\neq j} |\langle x_j, x_k\rangle|\right)$.



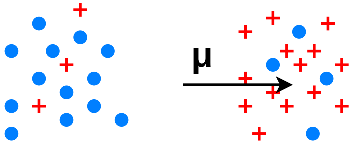
- Near-orthogonality typically holds in low-SNR, $d \gg n$ settings, e.g. mixture model:

$$\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \text{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

- Holds if $\|\mu\| = O(d^{1/2})$ and $d \gg n^2$.

Benign overfitting of neural nets in mixture model

- KKT points for 2-layer leaky nets $\approx \sum_{i=1}^{n} y_i x_i$, when training data is nearly-orthogonal $\left(\|x_i\|^2 \gg n \max_{k\neq j} |\langle x_j, x_k\rangle|\right)$.



- Near-orthogonality typically holds in low-SNR, $d \gg n$ settings, e.g. mixture model:

$$\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \text{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

- Holds if $\|\mu\| = O(d^{1/2})$ and $d \gg n^2$.
- Following results will only hold in this low-SNR, high-dimensional regime
  - We'll see consistency is still possible in this setting

Benign overfitting of neural nets in mixture model

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem (informal)

Suppose labels flipped w.p. $p < 1/2$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point θ of 2-layer leaky ReLU net ℓ_2-max-margin problem satisfies

$$\boxed{\forall k \in [n], \quad y_k = \mathsf{sgn}(f(x_k;\theta))}, \quad \text{and} \quad \boxed{p \le \mathbb{P}(y \ne \mathsf{sgn}(f(x;\theta))) \le p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right)}.$$

Benign overfitting of neural nets in mixture model

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem (informal)

Suppose labels flipped w.p. $p < 1/2$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point θ of 2-layer leaky ReLU net ℓ_2-max-margin problem satisfies

$$\boxed{\forall k \in [n], \quad y_k = \mathrm{sgn}(f(x_k;\theta))}, \quad \text{and} \quad \boxed{p \le \mathbb{P}(y \ne \mathrm{sgn}(f(x;\theta)) \le p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right)}.$$

- No dependence on number of neurons in network.

Benign overfitting of neural nets in mixture model

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem (informal)

Suppose labels flipped w.p. $p < 1/2$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point θ of 2-layer leaky ReLU net ℓ_2-max-margin problem satisfies

$$\forall k \in [n], \quad y_k = \mathsf{sgn}(f(x_k;\theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \mathsf{sgn}(f(x;\theta)) \leq p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right).$$

- No dependence on number of neurons in network.
- Overfitting: perfectly fit training data, even though $\approx pn$ labels are flipped

Benign overfitting of neural nets in mixture model

$$\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \text{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem (informal)

Suppose labels flipped w.p. $p < 1/2$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point θ of 2-layer leaky ReLU net ℓ_2-max-margin problem satisfies

$$\boxed{\forall k \in [n], \quad y_k = \text{sgn}(f(x_k;\theta))}, \quad \text{and} \quad \boxed{p \le \mathbb{P}(y \ne \text{sgn}(f(x;\theta)) \le p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right)}.$$

- No dependence on number of neurons in network.
- Overfitting: perfectly fit training data, even though $\approx pn$ labels are flipped
- Benign overfitting: if $n\|\mu\|^4 \gg d$, test error $\approx$ noise rate.

Benign overfitting of neural nets in mixture model

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem (informal)

Suppose labels flipped w.p. $p < 1/2$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point θ of 2-layer leaky ReLU net ℓ_2-max-margin problem satisfies

$$\forall k \in [n], \quad y_k = \mathsf{sgn}(f(x_k;\theta)), \quad \text{and} \quad p \le \mathbb{P}(y \ne \mathsf{sgn}(f(x;\theta)) \le p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right).$$

- No dependence on number of neurons in network.
- Overfitting: perfectly fit training data, even though $\approx pn$ labels are flipped
- Benign overfitting: if $n\|\mu\|^4 \gg d$, test error $\approx$ noise rate.
- Low-SNR requires $\|\mu\| = O(d^{1/2})$, so results hold for $\|\mu\| = \Theta(d^{\varepsilon})$ for $\varepsilon \in (1/4, 1/2)$

Benign overfitting of neural nets in mixture model

$$\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \text{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem (informal)

Suppose labels flipped w.p. $p < 1/2$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point θ of 2-layer leaky ReLU net ℓ_2-max-margin problem satisfies

$$\forall k \in [n], \quad y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \le \mathbb{P}(y \neq \text{sgn}(f(x;\theta)) \le p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right).$$

- No dependence on number of neurons in network.
- Overfitting: perfectly fit training data, even though $\approx pn$ labels are flipped
- Benign overfitting: if $n\|\mu\|^4 \gg d$, test error $\approx$ noise rate.
- Low-SNR requires $\|\mu\| = O(d^{1/2})$, so results hold for $\|\mu\| = \Theta(d^{\varepsilon})$ for $\varepsilon \in (1/4, 1/2)$
- $\exp(-\Omega(n\|\mu\|^4/d))$ is minimax-optimal!

Benign overfitting of neural nets in mixture model

Recall $\text{sgn}(f(x;\theta)) = \text{sgn}(\langle \sum_{i=1}^{n} y_i x_i, x \rangle)$. What does this estimator look like? Since $x_i = \tilde{y}_i \mu + z_i$,

Benign overfitting of neural nets in mixture model

Recall $\text{sgn}(f(x;\theta)) = \text{sgn}(\langle \sum_{i=1}^n y_i x_i, x\rangle)$. What does this estimator look like? Since $x_i = \tilde{y}_i \mu + z_i$,

$$\begin{aligned}
\sum_{i=1}^n y_i x_i &= \sum_{i\in \text{clean}} \tilde{y}_i(\tilde{y}_i \mu + z_i) + \sum_{i \in \text{noisy}} -\tilde{y}_i(\tilde{y}_i\mu + z_i) \\
&= \left(|\text{clean}| - |\text{noisy}|\right)\mu + \sum_{i=1}^n \tilde{y}_i z_i \\
&\approx \underbrace{(1-2p)n\cdot\mu}_{\text{signal}} + \underbrace{\sum_{i=1}^n \tilde{y}_i z_i}_{\text{overfitting component}}
\end{aligned}$$

Benign overfitting of neural nets in mixture model

Recall $\text{sgn}(f(x;\theta)) = \text{sgn}(\langle \sum_{i=1}^n y_i x_i, x\rangle)$. What does this estimator look like? Since $x_i = \tilde{y}_i \mu + z_i$,

$$\begin{aligned}\sum_{i=1}^n y_i x_i &= \sum_{i\in\text{clean}} \tilde{y}_i(\tilde{y}_i\mu + z_i) + \sum_{i\in\text{noisy}} -\tilde{y}_i(\tilde{y}_i\mu + z_i) \\ &= (|\text{clean}| - |\text{noisy}|)\,\mu + \sum_{i=1}^n \tilde{y}_i z_i \\ &\approx \underbrace{(1-2p)n\cdot\mu}_{\text{signal}} + \underbrace{\sum_{i=1}^n \tilde{y}_i z_i}_{\text{overfitting component}}\end{aligned}$$

Overfitting component helps interpolation, signal helps generalization:

Training data: classify (x_i, y_i) correctly		Test data: classify $(x, \tilde{y})$ correctly	
$\langle y_i x_i, \sum_{i=1}^n \tilde{y}_i z_i\rangle$	is large, positive,	$\langle \tilde{y}x, \sum_{i=1}^n \tilde{y}_i z_i\rangle$	is small, random $\pm$,
$\langle y_i x_i, n\mu\rangle$	is small, noisy labels make $\pm$.	$\langle \tilde{y}x, n\mu\rangle$	is (optimally) large, positive.

- Signal and overfitting component balanced to allow both interpolation + generalization

Other approaches for benign overfitting in neural nets

- Analysis of implicit bias (KKT conditions, minimum norm interpolation, …)

  Frei-Vardi-Bartlett-Srebro'23; Kornowski-Yehudai-Shamir'23; Kou-Chen-Gu'23; …

  - Kornowski-Yehudai-Shamir'23 look at local and global minima of margin-maximization problems (rather than just KKT points)
  - Only applies to ∞-time limit of training

Other approaches for benign overfitting in neural nets

- Analysis of implicit bias (KKT conditions, minimum norm interpolation, …)
  Frei-Vardi-Bartlett-Srebro'23; Kornowski-Yehudai-Shamir'23; Kou-Chen-Gu'23; …
  - Kornowski-Yehudai-Shamir'23 look at local and global minima of margin-maximization problems (rather than just KKT points)
  - Only applies to ∞-time limit of training
- "Trajectory analysis": directly track the weights of neural net trained by GD/GF from random initialization on noisy data, show that it achieves small train and test error Frei-Chatterji-Bartlett'22; Xu-Gu'23; Kou-Chen-Chen-Gu ICML'23; Xu-Wang-Frei-Vardi-Hu'23; Meng-Zou-Cao'23; …
  - Characterizes finite time performance
  - More narrow, less clear "why" benign overfitting happens

Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

$$f(x;\theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'\big(y_i f(x_i;\theta^{(t)})\big) \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

$$f(x;\theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'\big(y_i f(x_i;\theta^{(t)})\big) \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Tasks:
  - Analyze weights $\theta^{(t)}$ and empirical risk $\hat{L}(\theta^{(t)})$ (training example margins $y_i f(x_i;\theta^{(t)})$)

Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

$$f(x;\theta) = \sum_{j=1}^m a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^n -\ell'\big(y_i f(x_i;\theta^{(t)})\big) \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Tasks:
  - Analyze weights $\theta^{(t)}$ and empirical risk $\hat{L}(\theta^{(t)})$ (training example margins $y_i f(x_i;\theta^{(t)})$)
  - Track test error $\mathbb{P}(y \neq \mathrm{sgn}(f(x;\theta^{(t)}))$ (test example margin $y f(x;\theta^{(t)})$)

Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

$$f(x;\theta) = \sum_{j=1}^m a_j \phi(\langle \theta_j, x\rangle), \quad \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^n -\ell'\big(y_i f(x_i;\theta^{(t)})\big)\cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Tasks:
  - Analyze weights $\theta^{(t)}$ and empirical risk $\hat{L}(\theta^{(t)})$ (training example margins $y_i f(x_i;\theta^{(t)})$)
  - Track test error $\mathbb{P}(y \neq \mathrm{sgn}(f(x;\theta^{(t)}))$ (test example margin $y f(x;\theta^{(t)})$)
  - These two must be very different for benign overfitting to occur

Benign overfitting from trajectory analysis

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem

Suppose labels flipped w.p. $p = O(1)$, low SNR and $d \gg n^2$. Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init $\theta^{(0)}$, learning rate), for all $t \geq 1$,

$$\boxed{\hat{L}(\theta^{(t)}) \leq O\left({}^{1}\!/_{t}\right)}, \quad \text{and} \quad \boxed{p \leq \mathbb{P}(y \neq \mathsf{sgn}(f(x;\theta^{(t)})) \leq p + \exp\left(-\Omega(n\|\mu\|^4/d)\right)}.$$

Benign overfitting from trajectory analysis

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem

Suppose labels flipped w.p. $p = O(1)$, low SNR and $d \gg n^2$. Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init $\theta^{(0)}$, learning rate), for all $t \geq 1$,

$$\boxed{\hat{L}(\theta^{(t)}) \leq O\left({}^{1}\!/_{t}\right)}, \quad \text{and} \quad \boxed{p \leq \mathbb{P}(y \neq \mathsf{sgn}(f(x;\theta^{(t)})) \leq p + \exp\left(-\Omega(n\|\mu\|^4/d)\right)}.$$

- No dependence on number of neurons in network.

Benign overfitting from trajectory analysis

$$\tilde{y} \sim \mathsf{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathsf{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem

Suppose labels flipped w.p. $p = O(1)$, low SNR and $d \gg n^2$. Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init $\theta^{(0)}$, learning rate), for all $t \geq 1$,

$$\boxed{\hat{L}(\theta^{(t)}) \leq O\left(1/t\right)}, \quad \text{and} \quad \boxed{p \leq \mathbb{P}(y \neq \mathsf{sgn}(f(x;\theta^{(t)})) \leq p + \exp\left(-\Omega(n\|\mu\|^4/d)\right)}.$$

- No dependence on number of neurons in network.
- Benign overfitting if t is large and $n\|\mu\|^4 \gg d$.

Benign overfitting from trajectory analysis

$$\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \text{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p.$$

Theorem

Suppose labels flipped w.p. $p = O(1)$, low SNR and $d \gg n^2$. Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init $\theta^{(0)}$, learning rate), for all $t \geq 1$,

$$\boxed{\hat{L}(\theta^{(t)}) \leq O\left(1/t\right)}, \quad \text{and} \quad \boxed{p \leq \mathbb{P}(y \neq \text{sgn}(f(x;\theta^{(t)})) \leq p + \exp\left(-\Omega(n\|\mu\|^4/d)\right)}.$$

- No dependence on number of neurons in network.
- Benign overfitting if t is large and $n\|\mu\|^4 \gg d$.
- Same generalization bound as KKT analysis, but now holds throughout GD trajectory.
  - Only tolerates $p = O(1)$, rather than $p < 1/2$ from KKT analysis.

Benign overfitting from trajectory analysis

$$f(x;\theta) = \textstyle\sum_{j=1}^{m} a_j \phi(\langle \theta_j, x\rangle), \quad \hat{L}(\theta) = \frac{1}{n}\textstyle\sum_{i=1}^{n} \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^{n} \underbrace{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}_{\geq 0} \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

Benign overfitting from trajectory analysis

$$f(x;\theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^{n} \underbrace{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}_{\geq 0} \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Difficulty arises: “clean label” examples (in principle) are easier, larger margin $y_i f(x_i;\theta^{(t)})$, while “noisy label” examples harder, smaller margin

Benign overfitting from trajectory analysis

$$f(x;\theta) = \sum_{j=1}^m a_j \phi(\langle \theta_j, x\rangle), \quad \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^n \underbrace{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}_{\geq 0} \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Difficulty arises: "clean label" examples (in principle) are easier, larger margin $y_i f(x_i;\theta^{(t)})$, while "noisy label" examples harder, smaller margin
- Since $-\ell'$ is decreasing, implies noisy labels could have outsized influence on training dynamics $\longrightarrow$ hard for overfitting to be 'benign'

Benign overfitting from trajectory analysis

$$f(x;\theta) = \sum_{j=1}^m a_j \phi(\langle \theta_j, x\rangle), \quad \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^n \underbrace{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}_{\geq 0} \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Difficulty arises: "clean label" examples (in principle) are easier, larger margin $y_i f(x_i;\theta^{(t)})$, while "noisy label" examples harder, smaller margin
- Since $-\ell'$ is decreasing, implies noisy labels could have outsized influence on training dynamics $\longrightarrow$ hard for overfitting to be 'benign'
- Key technical lemma shown in most trajectory analyses: loss ratio bound,

$$\sup_{t\geq 0} \max_{i,j} \frac{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}{-\ell'\big(y_j f(x_j;\theta^{(t)})\big)} = O(1).$$

Benign overfitting from trajectory analysis

$$f(x;\theta) = \sum_{j=1}^m a_j \phi(\langle \theta_j, x\rangle), \quad \hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i;\theta)\big),$$

$$\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n}\sum_{i=1}^n \underbrace{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}_{\geq 0} \cdot y_i \nabla f(x_i;\theta^{(t)}).$$

- Difficulty arises: "clean label" examples (in principle) are easier, larger margin $y_i f(x_i;\theta^{(t)})$, while "noisy label" examples harder, smaller margin
- Since $-\ell'$ is decreasing, implies noisy labels could have outsized influence on training dynamics $\longrightarrow$ hard for overfitting to be 'benign'
- Key technical lemma shown in most trajectory analyses: loss ratio bound,

$$\sup_{t\geq 0} \max_{i,j} \frac{-\ell'\big(y_i f(x_i;\theta^{(t)})\big)}{-\ell'\big(y_j f(x_j;\theta^{(t)})\big)} = O(1).$$

- Known proofs all rely on nearly-orthogonal data ($d \gg n$) to show this

Chatterji-Long'21; Frei-Chatterji-Bartlett'22

"Blessing of Dimensionality"

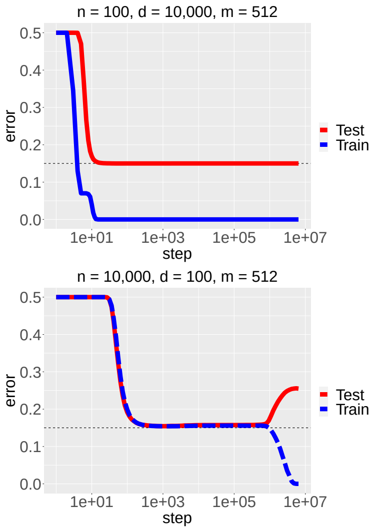
- $d/n \to \infty$ necessary for benign overfitting in linear models, but unknown if necessary for neural networks.

“Blessing of Dimensionality”

- $d/n \to \infty$ necessary for benign overfitting in linear models, but unknown if necessary for neural networks.
- Consider again the Gaussian mixture model, with $p = 0.15$ labels flipped (train and test), $m = 512$ neurons, vary d/n.

“Blessing of Dimensionality”

- $d/n \to \infty$ necessary for benign overfitting in linear models, but unknown if necessary for neural networks.
- Consider again the Gaussian mixture model, with $p = 0.15$ labels flipped (train and test), $m = 512$ neurons, vary d/n.
- Learning dynamics different in $n > d$ setting; overfitting less ‘benign’
  $\longrightarrow$ “Blessing of dimensionality”? See also: [Kornowski-Yehudai-Shamir’23]



Benign, tempered, and catastrophic overfitting

- There is a spectrum of generalization behavior when overfitting.

Benign, tempered, and catastrophic overfitting

- There is a spectrum of generalization behavior when overfitting.
- Let R_n be test error for interpolator (train error = 0) using n samples, R^* best possible test error.

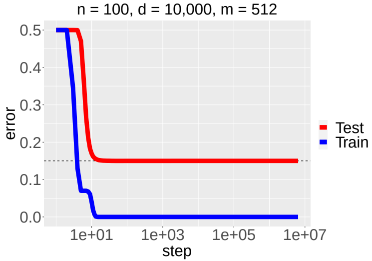
	Regression	Binary Classification
Benign	$\lim_{n\to\infty} R_n = R^*$	$\lim_{n\to\infty} R_n = R^*$
Tempered	$\lim_{n\to\infty} R_n \in (R^*, \infty)$	$\lim_{n\to\infty} R_n \in (R^*, 1/2)$
Catastrophic	$\lim_{n\to\infty} R_n = \infty$	$\lim_{n\to\infty} R_n = 1/2$

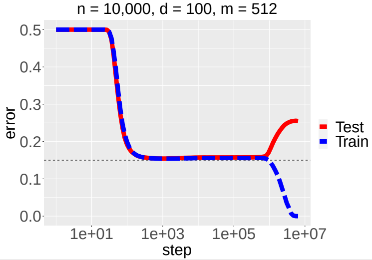
Benign, tempered, and catastrophic overfitting

- There is a spectrum of generalization behavior when overfitting.
- Let R_n be test error for interpolator (train error = 0) using n samples, R^* best possible test error.

	Regression	Binary Classification
Benign	$\lim_{n\to\infty} R_n = R^*$	$\lim_{n\to\infty} R_n = R^*$
Tempered	$\lim_{n\to\infty} R_n \in (R^*, \infty)$	$\lim_{n\to\infty} R_n \in (R^*, 1/2)$
Catastrophic	$\lim_{n\to\infty} R_n = \infty$	$\lim_{n\to\infty} R_n = 1/2$

- Neural net trained on high-dimensional mixture model: (provably) benign; low-dimensional: tempered?





Open questions

- Is benign overfitting in neural nets possible in low dimensions ($n \gg d$)?
    - Overparameterization through wider nets could help, but does it? When? Why?

Open questions

- Is benign overfitting in neural nets possible in low dimensions ($n \gg d$)?
  - Overparameterization through wider nets could help, but does it? When? Why?
- Which neural net interpolators do we care about in regression?

Open questions

- Is benign overfitting in neural nets possible in low dimensions ($n \gg d$)?
  - Overparameterization through wider nets could help, but does it? When? Why?
- Which neural net interpolators do we care about in regression?
- Necessary and sufficient conditions for benign overfitting in linear classification?
  - Fairly complete picture of min-ℓ^2 linear regression, but mostly sufficiency guarantees in classification.
  - Dream: data-dependent, signal-dependent, tight guarantees.

Thanks!