Fast rates for noisy interpolation require rethinking the effects of inductive bias

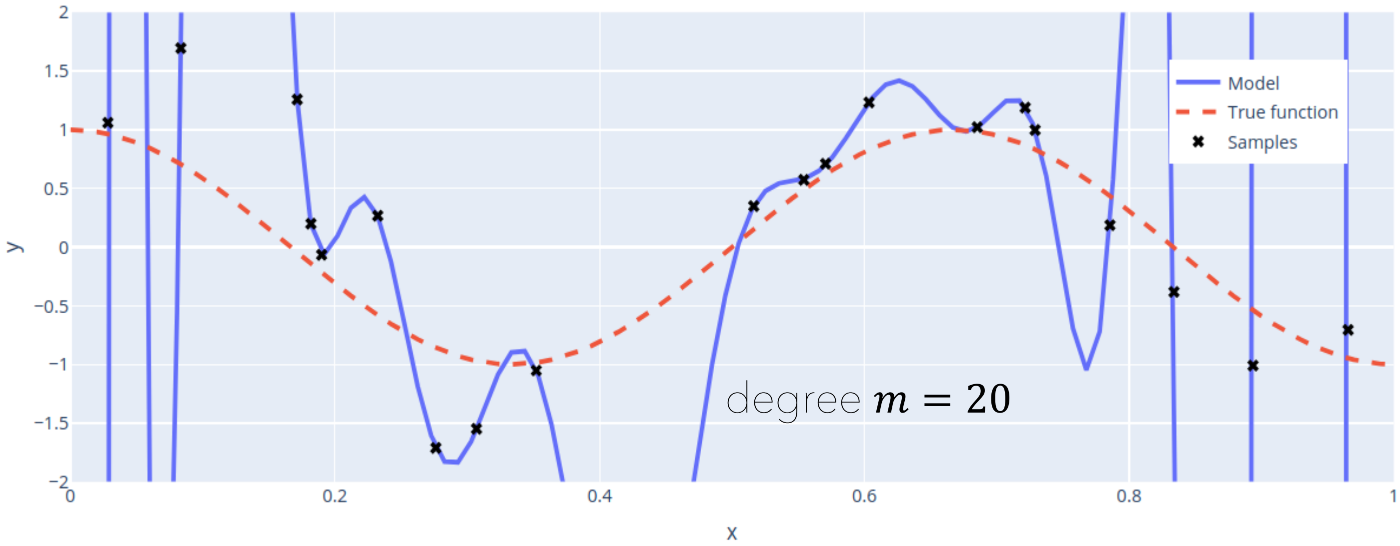
October 25th 2022, Mathematics of Machine Learning, BCAM Bilbao

Fanny Yang, **K. Donhauser**

joint with G. Wang, S. Stojanovic, Marco Milanta, N. Ruggeri, Michael Aerni

Statistical Machine Learning group, CS department, ETH Zurich

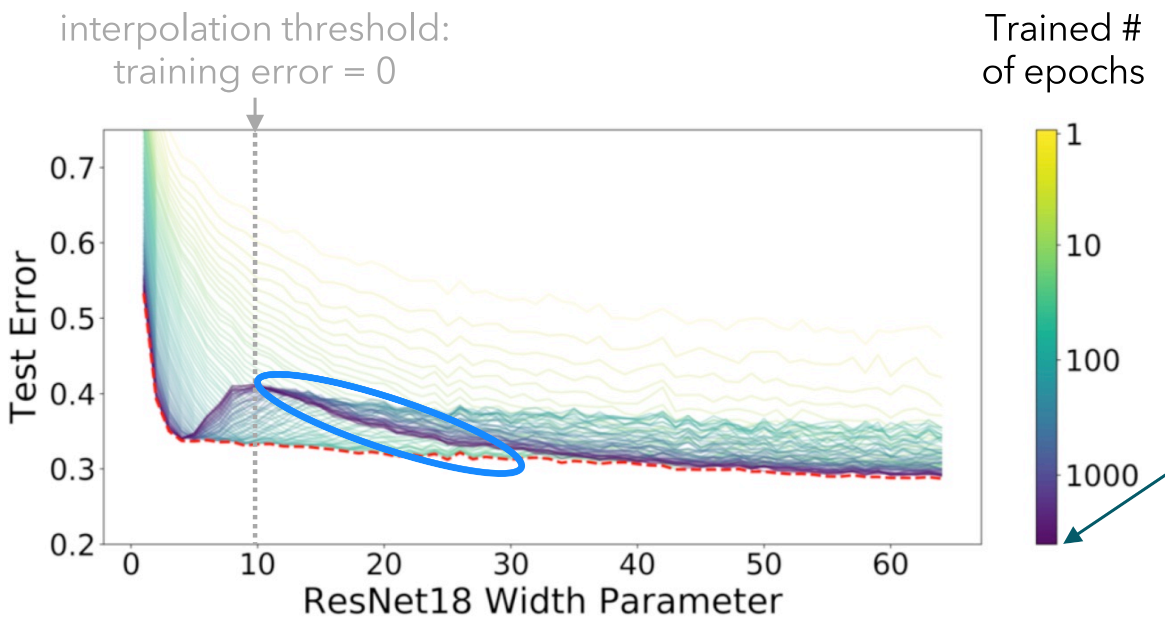
Classical wisdom: Avoid fitting noise



Traditionally: want to avoid fitting noise perfectly for better (optimal) generalization.

Double descent on neural networks

Classification using neural networks and first-order methods on CIFAR-10 with 15% label noise



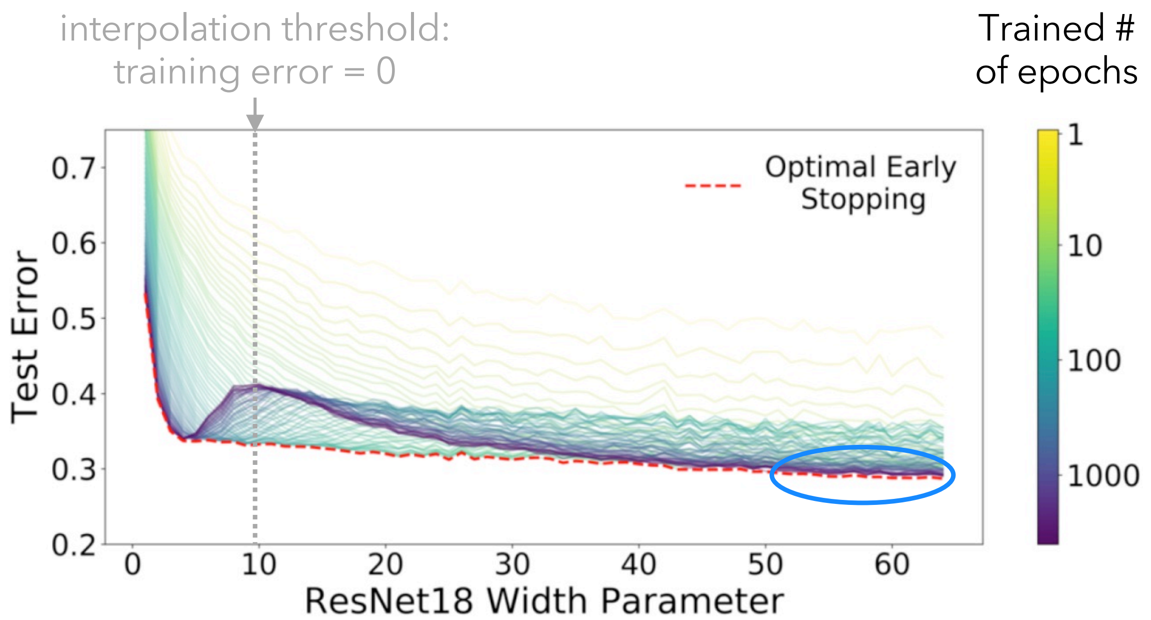
follow blue curve:
model at convergence

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

Source: [NKBYBS '20]

Harmless interpolation on neural networks

Classification using neural networks and first-order methods on CIFAR-10 with 15% label noise



compare
blue (at convergence)
with red curve
(best stopping time)

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

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

Story of this talk…

Question today: What are "mechanisms" so that interpolators $\hat{f}$ with $\hat{f}(x_i) = y_i$ exhibit

1 second descent 2 harmless interpolation 3 good generalization, focusing on 2 3

Our observation: One key mechanism is the "simplicity of the structure" of the interpolator

Further, the strength of the "simplicity/inductive bias" has counterintuitive effect on interpolators compared to classical wisdom on regularized estimators!

We don't: propose to use interpolators in practice → optimally regularized can't be beat

Examples for strong inductive biases

- Strong inductive bias ≜ strong bias towards simple structure of "optimal" model ≜ less flexibility
- Examples for strong structural biases we discuss today:

Linear interpolators:

sparsity $||w||_0 \ll d$

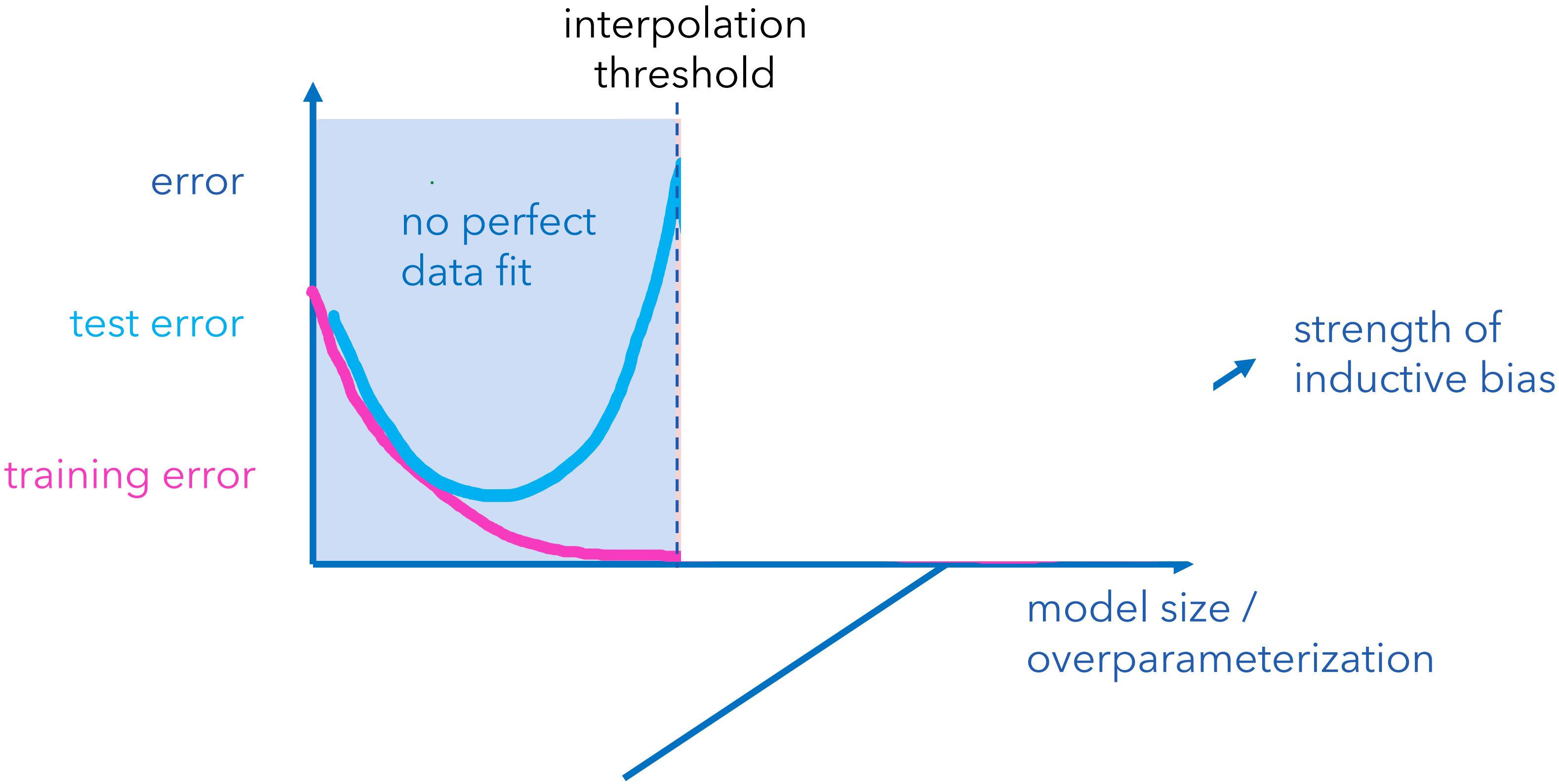
Kernel interpolators:

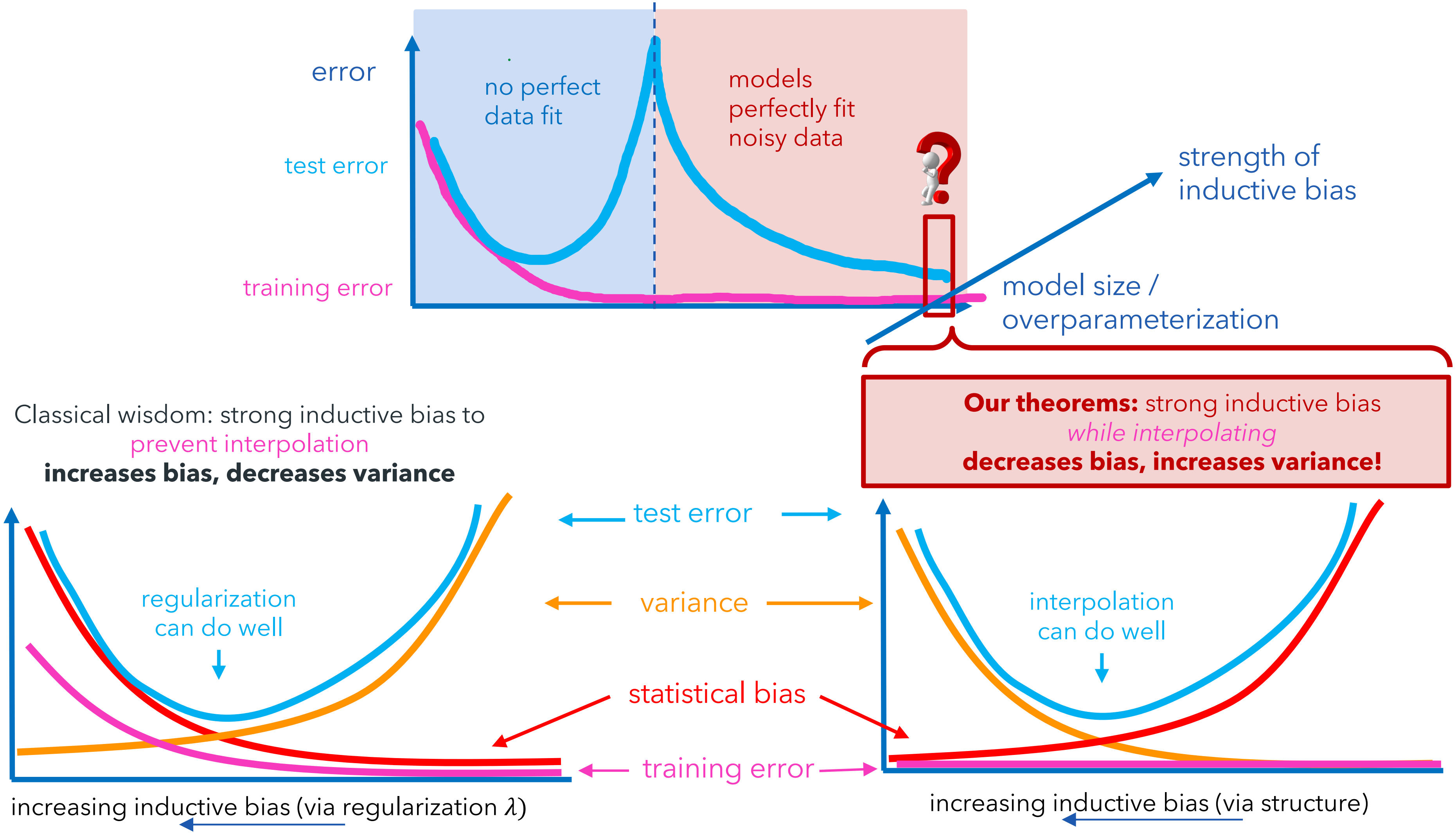
Neural networks:

filter size for convolutional models

rotational invariance

The role of the inductive bias for interpolators




error
no perfect
data fit
models
perfectly fit
noisy data
test error
training error
strength of
inductive bias
model size /
overparameterization
Classical wisdom: strong inductive bias to
prevent interpolation
increases bias, decreases variance
Our theorems: strong inductive bias
while interpolating
decreases bias, increases variance!
test error
variance
statistical bias
training error
regularization
can do well
interpolation
can do well
increasing inductive bias (via regularization λ)
increasing inductive bias (via structure)

Examples for strong inductive biases

Strong inductive bias ≜ strong bias towards simple structure of "optimal" model ≜ less flexibility

Examples for strong structural biases we discuss today:

Part I

Part II: Latest and on-going work

Linear interpolators:

sparsity $||w||_0 \ll d$

Kernel interpolators:

Neural networks:

filter size for convolutional models

rotational invariance

Tight bounds for the risk

Controlled experiments

Linear regression setting (for this talk)

- **Function space:** linear models $f(x) = \langle w, x\rangle$ with $x, w \in \mathbb{R}^d$
- **Data model for $\boldsymbol{n}$ samples**: $y_i = \langle w^\star, x_i\rangle + \xi_i$ with $x_i \sim N(0, I)$ and noise $\xi_i \sim N(0, \sigma^2)$

  with sparse $w^\star = (1,0,\ldots,0)$ with unknown location (for simplicity of presentation)
- **Degree of overparameterization (high-dimensional regime)**: $d \asymp n^\beta$, $\beta > 1$
- **Linear estimators we compare:** **for $\boldsymbol{p} \in [\mathbf{1}, \mathbf{2}]$**

  implicit bias of 1st order methods

  - **Minimum-ℓ_p-norm interpolators**: $\hat{w} = \mathrm{argmin}_w \|w\|_p$ s.t. $y = Xw$
  - **compared against classical regularized estimators**: $\hat{w}_\lambda = \mathrm{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_p^p$
- **Performance measure**: prediction error $\mathbb{E}_{x\sim N(0,I)} (\langle x, \hat{w} - w^\star\rangle)^2 = \|\hat{w} - w^\star\|^2$

(Similar bounds also hold for max-ℓ_p-margin classification $\hat{w} = argmin_w \|w\|_p$ s.t. $y_i\langle x_i, w\rangle \geq 1\ \forall i$)

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

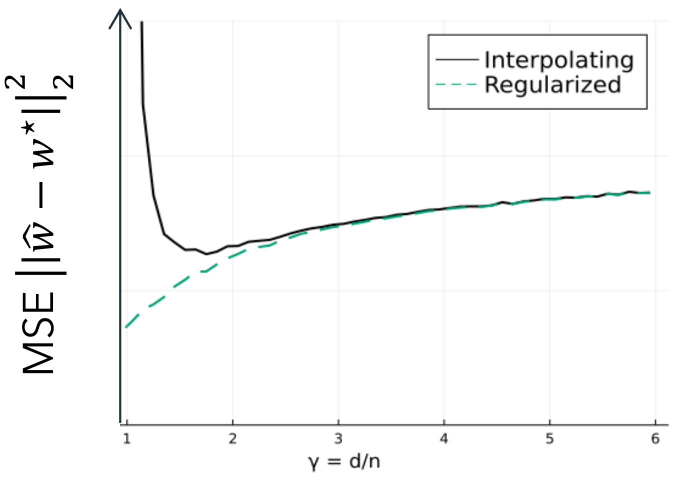
Min-ℓ_p-norm interpolation $\hat{w} = \mathrm{argmin}_w \; ||w||_p \;\; s.t.\, y = Xw$

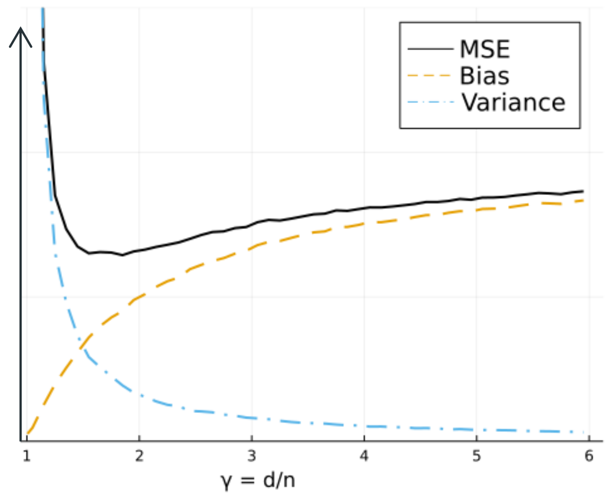
Goal today: populate with high-dimensional tight non-asymptotic rates

Weak inductive bias: $p = 2$ (inconsistent)

Interpolators $\hat{w} = \mathrm{argmin}_w \,||w||_2$ s.t. $y = Xw$ vs. Regularized estimator: $\hat{w}_\lambda = ||y - Xw||_2^2 + \lambda ||w||_2^2$

Linear model $y_i = \langle w^\star, x_i \rangle + \xi_i$ with i.i.d. $x_i \sim N(0, \mathrm{I})$, some $\xi_i \sim N(0, \sigma^2)$



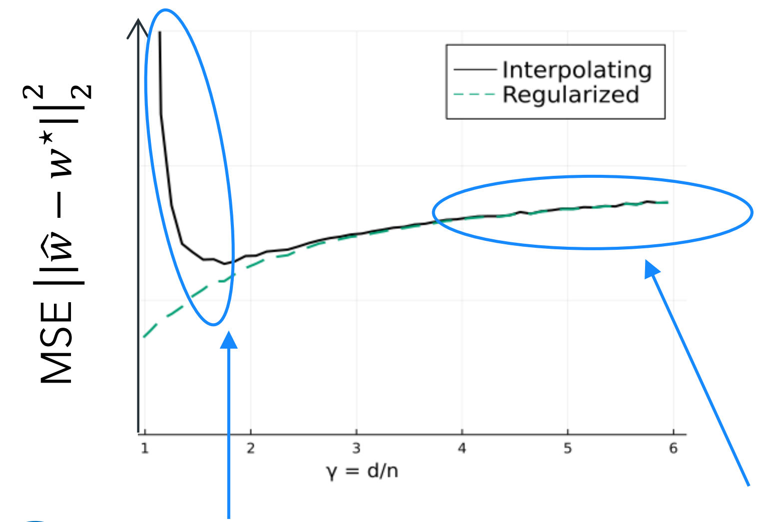


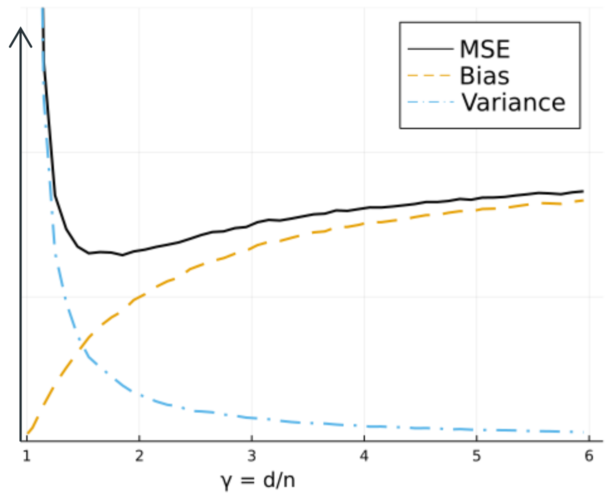
Increasing d/n (≈ "overparameterization) is "implicitly regularizing" as variance decreases

Weak inductive bias: $p = 2$ (inconsistent)

Interpolators $\hat{w} = \mathrm{argmin}_w \, ||w||_2$ s.t. $y = Xw$ vs. Regularized estimator: $\hat{w}_\lambda = ||y - Xw||_2^2 + \lambda ||w||_2^2$

Linear model $y_i = \langle w^\star, x_i \rangle + \xi_i$ with i.i.d. $x_i \sim N(0, \mathrm{I})$, some $\xi_i \sim N(0, \sigma^2)$



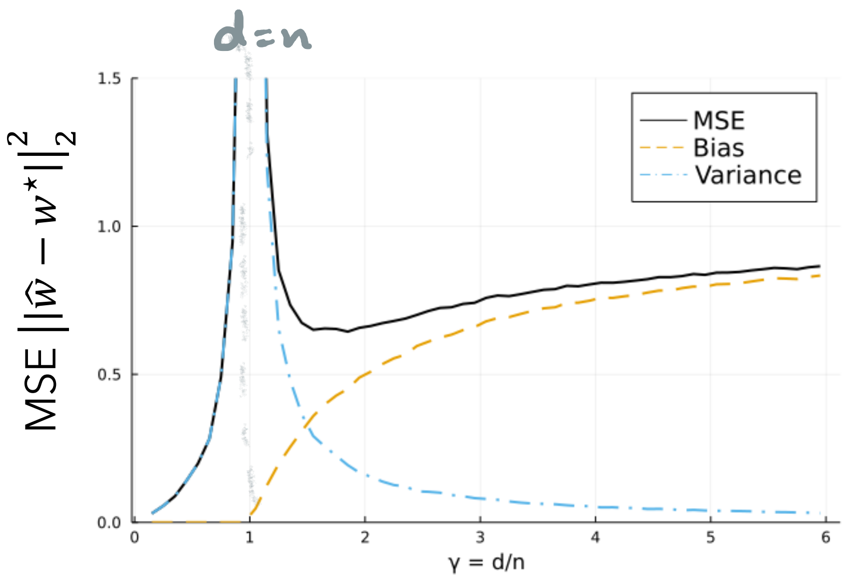


1 "second" descent

Weak inductive bias: $p = 2$ (second descent)

Interpolators $\hat{w} = \operatorname{argmin}_w \|w\|_2$ s.t. $y = Xw$ for $y_i = \langle w^\star, x_i\rangle + \xi_i$ *with* $w^\star = 0$ some $\xi_i \sim N(0, \sigma^2)$

Hence $\hat{w} = \operatorname{argmin}_w \|w\|_2$ s.t. $\xi = Xw$



As $\frac{d}{n}$ increases (assume fixed n and increase $d \to d+1$):

Variance decreases: if $w^\star = 0$,

given min-norm solution $\hat{w}_d$ for d, for $d+1$ we know

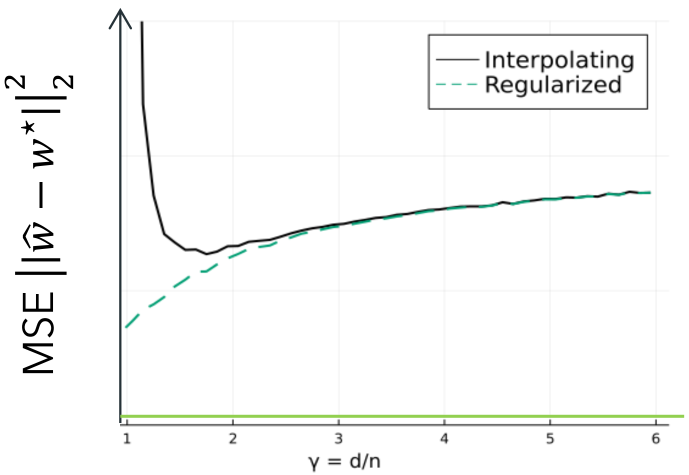
$(\hat{w}_d, 0)$ is also interpolating $\rightarrow \|\hat{w}_{d+1}\|_2 \leq \|\hat{w}_d\|_2$

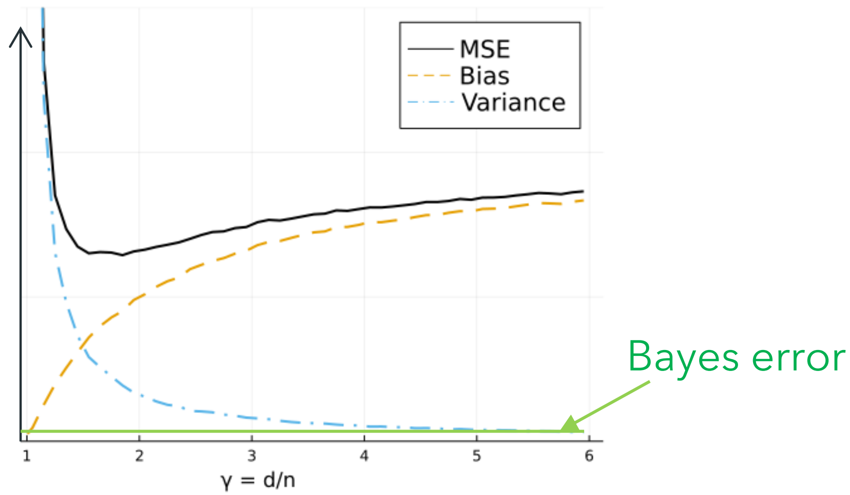
Bias increases because "harder to find signal" in $d+1$

Weak inductive bias: $p = 2$ (inconsistent)

Interpolators $\hat{w} = \mathrm{argmin}_w \, ||w||_2$ s.t. $y = Xw$ vs. Regularized estimator: $\hat{w}_\lambda = ||y - Xw||_2^2 + \lambda ||w||_2^2$

Linear model $y_i = \langle w^\star, x_i \rangle + \xi_i$ with i.i.d. $x_i \sim N(0, \mathrm{I})$, some $\xi_i \sim N(0, \sigma^2)$





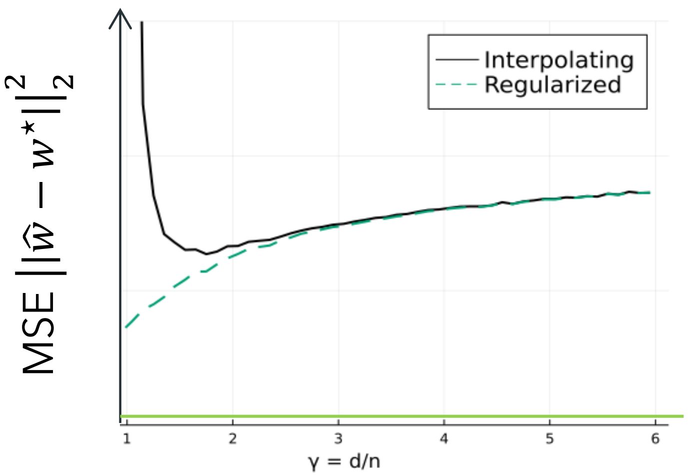
For isotropic Gaussians, $||\hat{w} - w^\star||^2 > c > 0$ for any $\beta > 1$ ($d \asymp n^\beta$) even as $n \to \infty$ due to high bias!

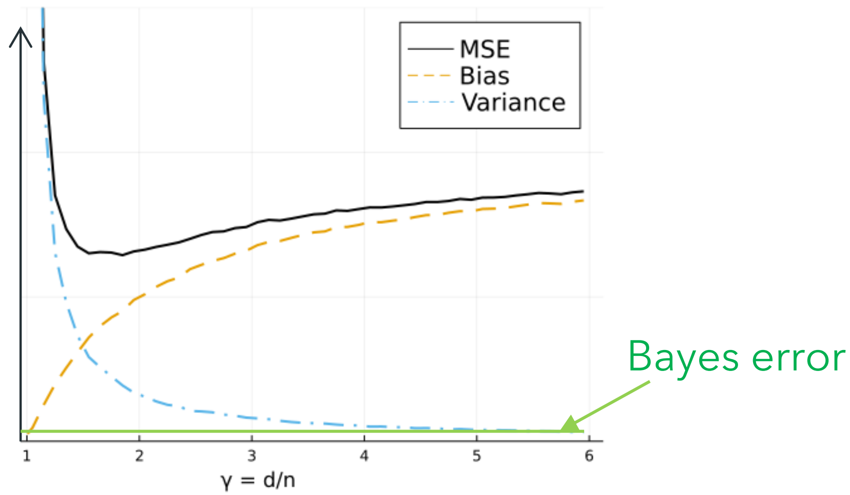
*consistent only for very spiked covariance Σ [HMRT'19, MM'19, BLLT '19, MVSS '20] ⚡ in practice Σ is fixed!

Weak inductive bias: $p = 2$ (inconsistent)

Interpolators $\hat{w} = \text{argmin}_w \, ||w||_2 \text{ s.t. } y = Xw$ vs. Regularized estimator: $\hat{w}_\lambda = ||y - Xw||_2^2 + \lambda ||w||_2^2$

Linear model $y_i = \langle w^\star, x_i \rangle + \xi_i$ with i.i.d. $x_i \sim N(0, \mathrm{I})$, some $\xi_i \sim N(0, \sigma^2)$



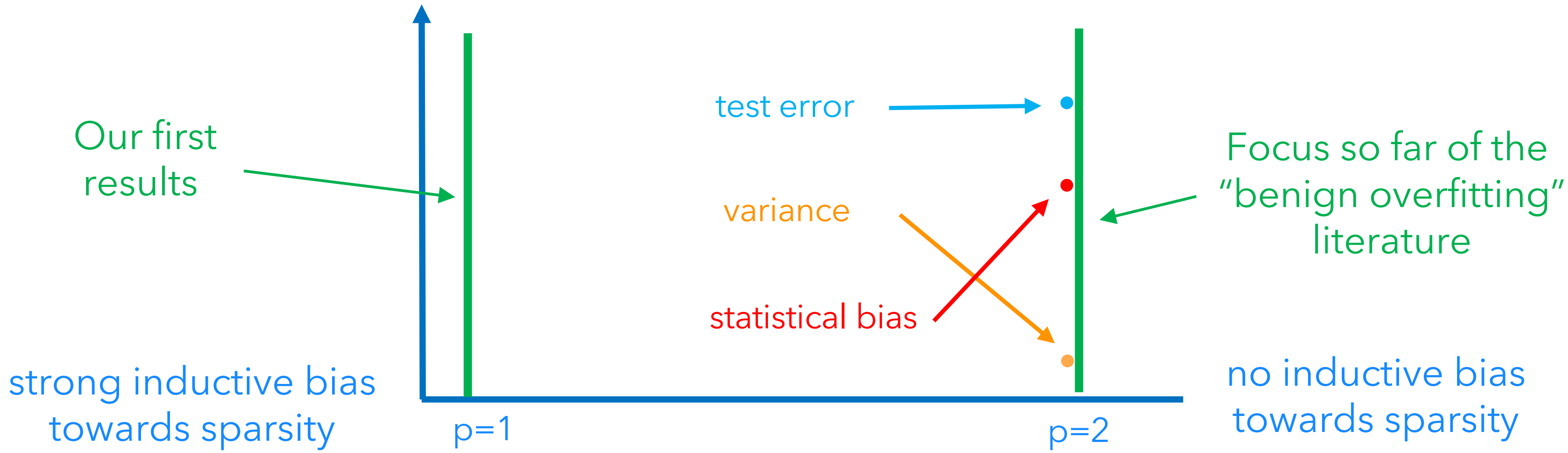


1 second descent ✓ 2 harmless interpolation ✓ 3 good generalization ✗

*consistent only for very spiked covariance Σ [HMRT'19, MM'19, BLLT '19, MVSS '20] ⚡ in practice Σ is fixed!

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

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



Benefits of strong inductive bias (recap)

Remember structural simplicity of ground truth: sparsity $||w^\star||_0 = s \ll d$

Weak (no) inductive bias: encouraging small $||w||_2$ norm

Matching strong inductive bias : small $||w||_0/||w||_1$ norm encouraging sparsity structure

Noiseless
$y = Xw^\star$

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

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

when observations are noisy

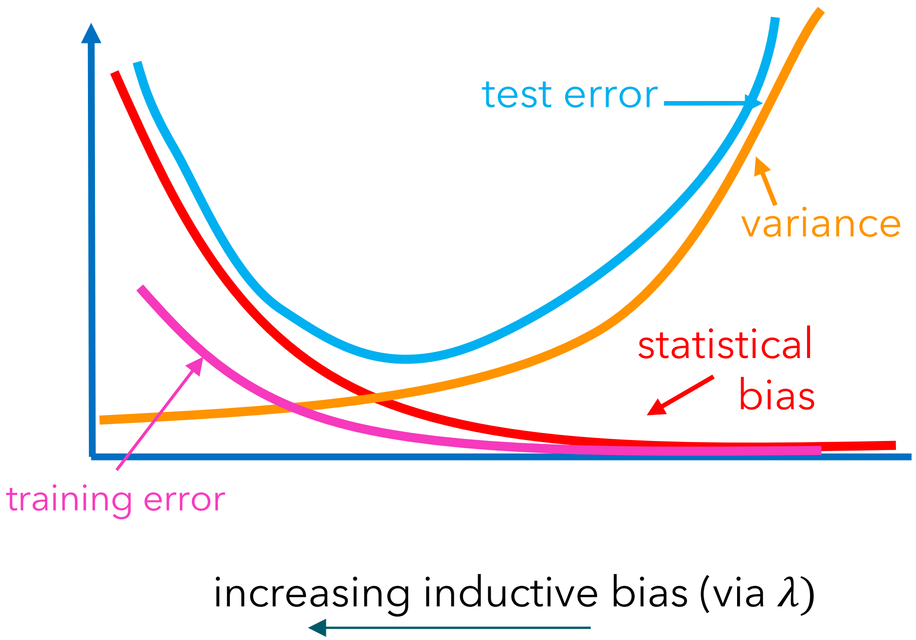
Noisy
$y = Xw^\star + \xi$

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

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

Interpolators are forced to fit noise!

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



- Classical theorems for ℓ_1-penalized:
  - good rates by trading off via λ
    fitting noise (variance) vs
    fit of noiseless function (bias)
- But interpolators have to fit noise perfectly

  → cannot attenuate noise-fitting using λ

Open problem: How much does min-ℓ_1-norm interpolation suffer from noise fitting?

Strong inductive bias: $p = 1$ (consistent but slow)

Previous work for the i.i.d. noise case:

$\Omega\left(\sigma^2/\log\left(\frac{d}{n}\right)\right)$ lower bounds [MVSS '19]

$O(\sigma^2)$ upper bounds [KZSS '21, CLG '20]

(who studied adversarial, vanishing noise)

Theorem [WDY' 21](simplified) – Tight bounds for min-ℓ_1-norm interpolators

There exists a universal constant $c > 0$, s.t. whenever $d \asymp n^\beta$ with $\beta > 1$, $n \geq c$ w.h.p.

$$||\hat{w} - w^\star||^2 = \frac{\sigma^2}{\log(d/n)} + O\left(\frac{\sigma^2}{\log^{3/2}(d/n)}\right)$$

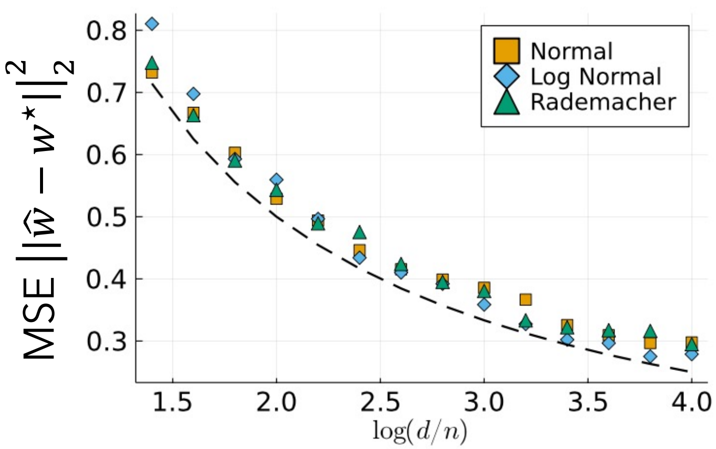
The proof is based on localized uniform convergence and CGMT [KZSS '21]
- who however don't show tight bounds and hence consistency

Strong inductive bias: $p = 1$ (consistent but slow)

Theorem [WDY**' 21](simplified) – Tight bounds for min-ℓ_1-norm interpolators**

There exists a universal constant $c > 0$, s.t. whenever $d \asymp n^{\beta}$ with $\beta > 1$, $n \geq c$ w.h.p.

$$||\hat{w} - w^{\star}||^2 = \frac{\sigma^2}{\log(d/n)} + O\left(\frac{\sigma^2}{\log^{3/2}(d/n)}\right)$$



- This is a lower & upper bound for Gaussian X
- Experimentally, the bound is also tight beyond Gaussian X, but hard to show!

Note: The same bound holds for classification

Strong inductive bias: $p = 1$ (consistent but slow)

Theorem [WDY' 21](simplified) – Tight bounds for min-ℓ_1-norm interpolators

There exists a universal constant $c > 0$, s.t. whenever $d \asymp n^\beta$ with $\beta > 1$, $n \geq c$ w.h.p.

$$||\hat{w} - w^\star||^2 = \frac{\sigma^2}{(\beta-1)\log n} + O\left(\frac{\sigma^2}{((\beta-1)\log n)^{3/2}}\right) \quad \text{(plugging in } d, n \text{ relation)}$$

1 second descent

2 harmless interpolation

3 good generalization

Yes! Variance decreases,

similar intuition as for $p = 2$

No! Variance too large!

Interpolator $\Omega\left(\frac{1}{\log n}\right)$

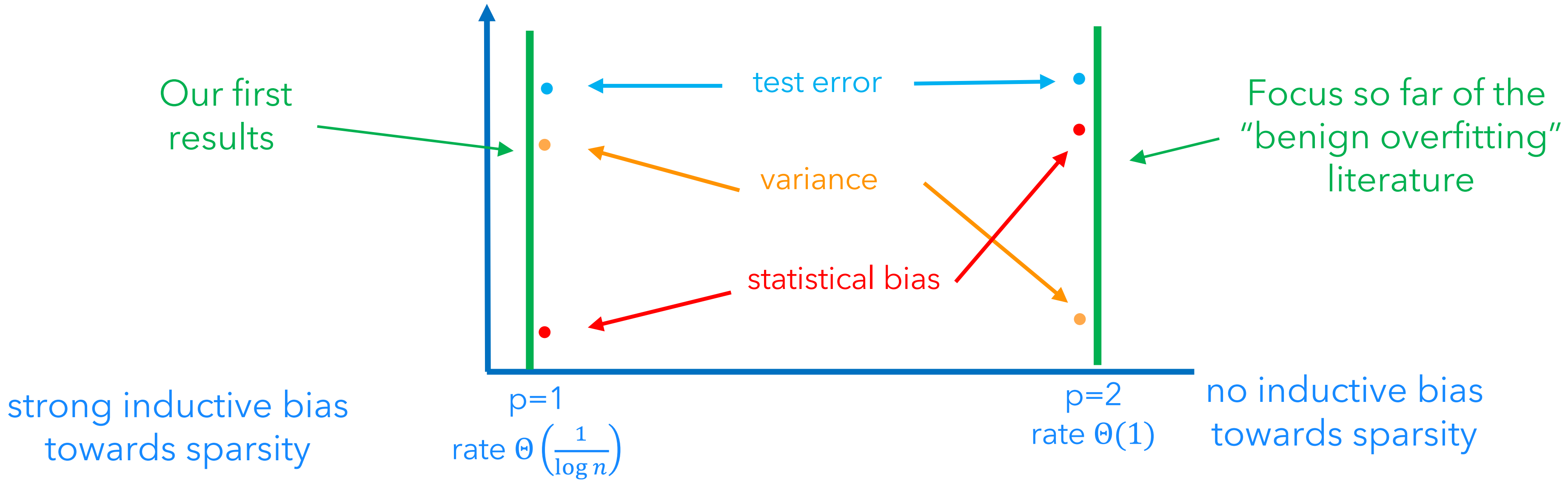
vs. regularized $O\left(\frac{s \log n}{n}\right)$

Consistent but

still slow rate!

So far: Interpolators still poor for $p = 1$

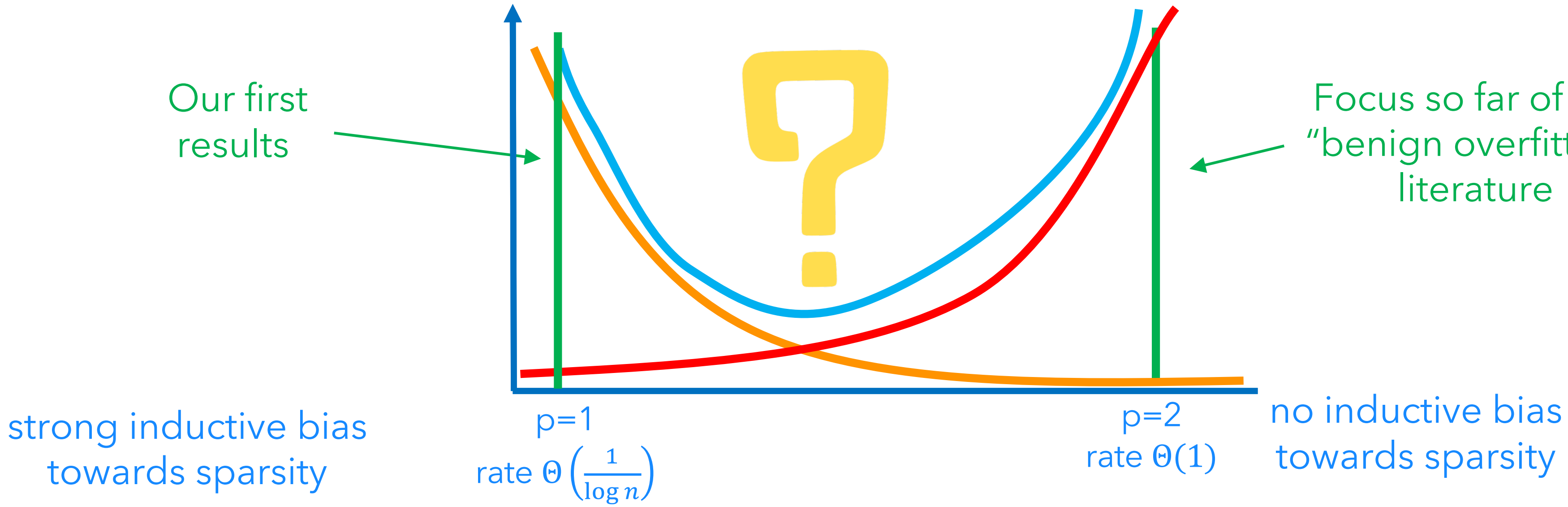
Min-ℓ_p-norm interpolation $\hat{w} = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$



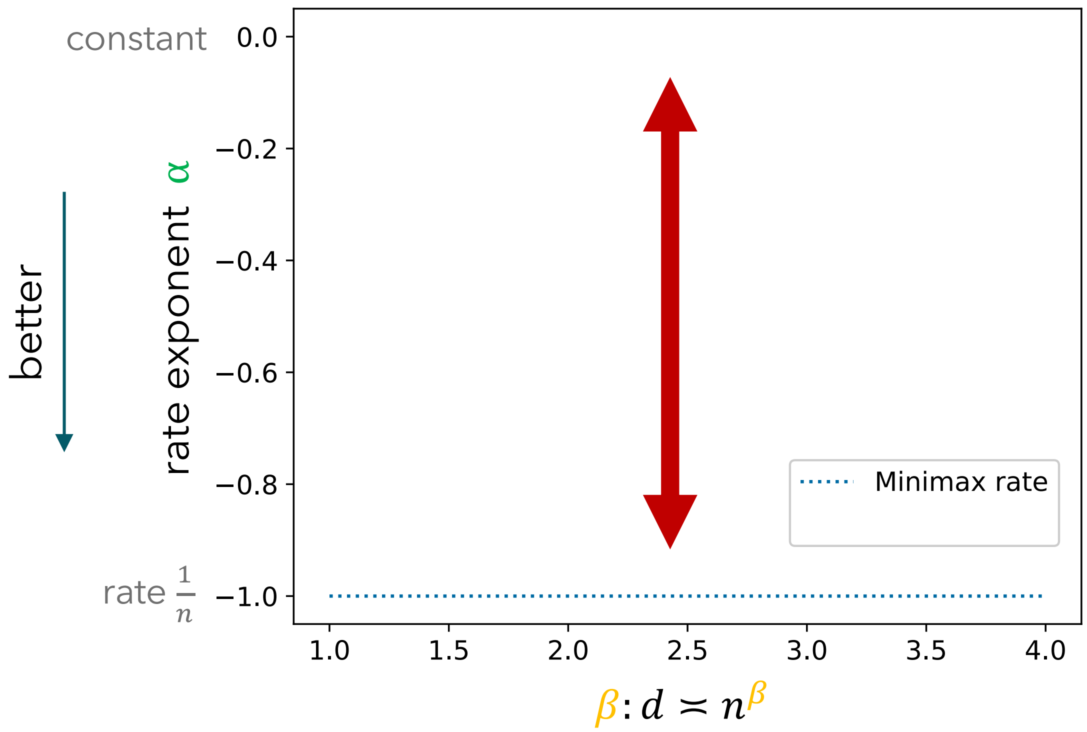
… but now due to high variance!

So far: Interpolators are poor for $p = 1, 2$

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



So far: Interpolators are poor for $p = 1, 2$



- Evaluate MSE $||\widehat{w} - w^\star||^2 \sim \widetilde{\Theta}(n^\alpha)$ with rate exponent α
- 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 $p = 1, 2$: $||\widehat{w} - w^\star||^2 = \widetilde{\Theta}(1) \rightarrow \alpha = 0$

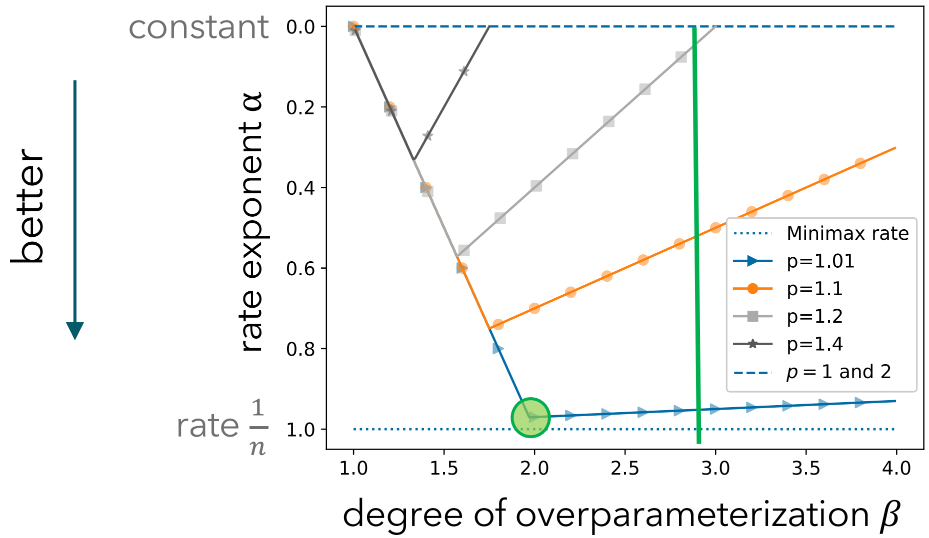
How close can we get to $\alpha = -1$ with ℓ_p-norm interpolators with $p \in (1,2)$?

Medium inductive bias: Fast rates with $p \in (1,2)$

Theorem [DRSY' 22] (informal) – Upper & lower bounds for min-ℓ_p-norm interpolators

For $d \asymp n^\beta$ with $1 < \beta \leq \frac{p/2}{p-1}$, and min-$\ell_p$-norm interpolators with $1 < p < 2$ and n large enough,

we obtain with high probability, error rates of order $\widetilde{\Theta}(n^{-\alpha})$ with α as in graph below



- order-matching upper & lower bound
- for fixed β, some $p > 1$ close to 1 gets best rate
- for $\beta \approx 2$, rates close to $\widetilde{\Theta}\left(\frac{1}{n}\right)$

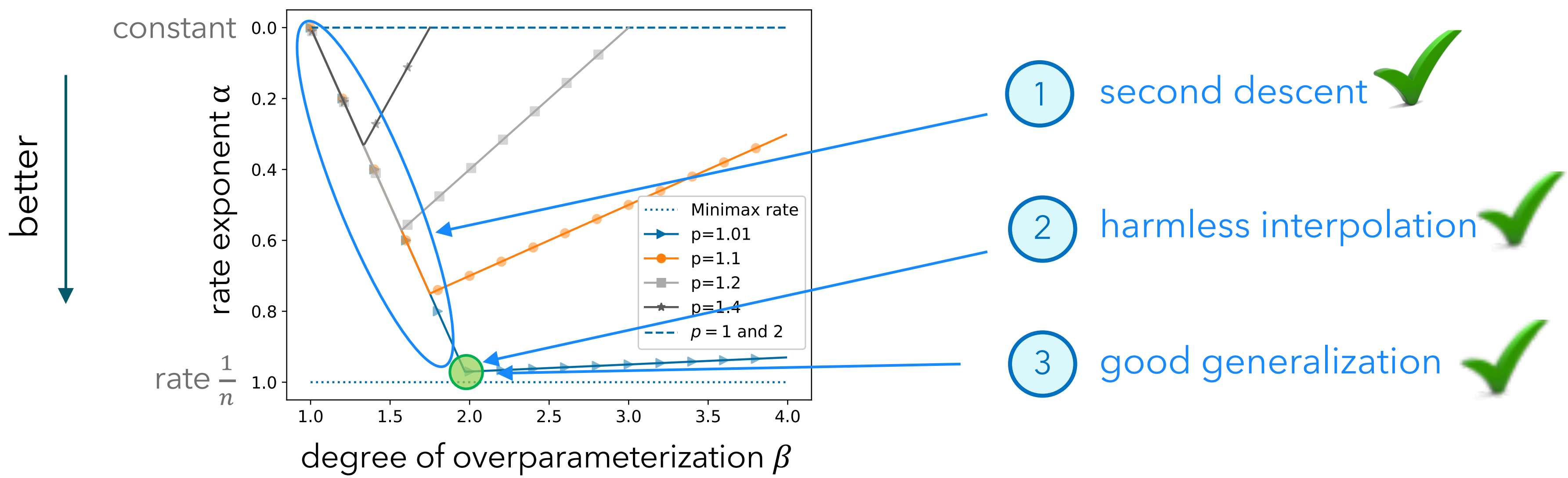
Note: technique applies to classification (see paper) and allows extension to $\Sigma \neq I$ and s-sparse $w^\star$

Medium inductive bias: Fast rates with $p \in (1,2)$

Theorem [DRSY' 22] (informal) – Upper & lower bounds for min-ℓ_p-norm interpolators

For $d \asymp n^\beta$ with $1 < \beta \leq \frac{p/2}{p-1}$, and min-$\ell_p$-norm interpolators with $1 < p < 2$ and n large enough,

we obtain with high probability, error rates of order $\tilde{O}(n^{-\alpha})$ with α as in graph below

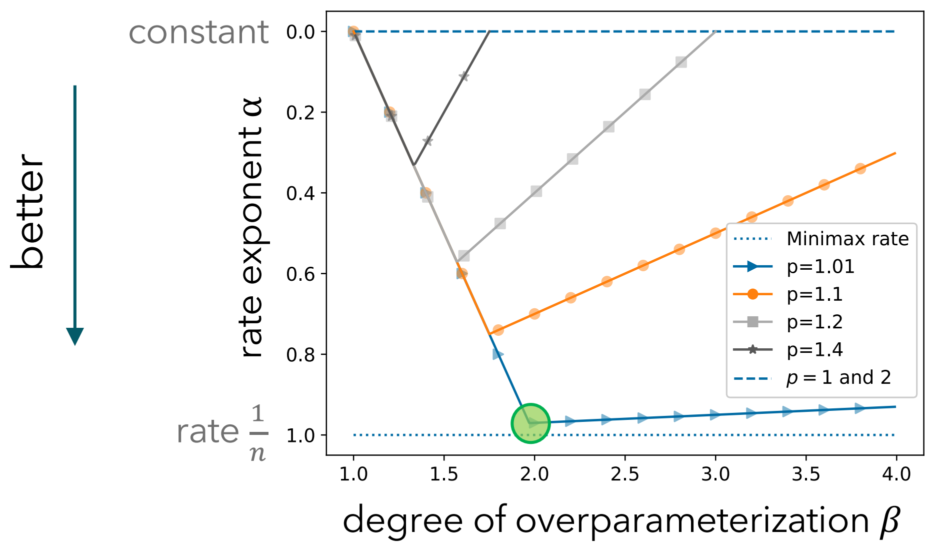


Fast rates with $p \in (1,2)$ - caveat…

Theorem [DRSY' 22] (informal) – Upper & lower bounds for min-ℓ_p-norm interpolators

For $d \asymp n^\beta$ with $1 < \beta \leq \frac{p/2}{p-1}$, and min-$\ell_p$-norm interpolators with $1 < p < 2$ and n large enough,

we obtain with high probability, error rates of order $\tilde{O}(n^{-\alpha})$ with α as in graph below



Caveat:

- "Large enough" actually requires

$$\frac{1}{\log \log d} \lesssim p - 1 \to \text{very large d}$$

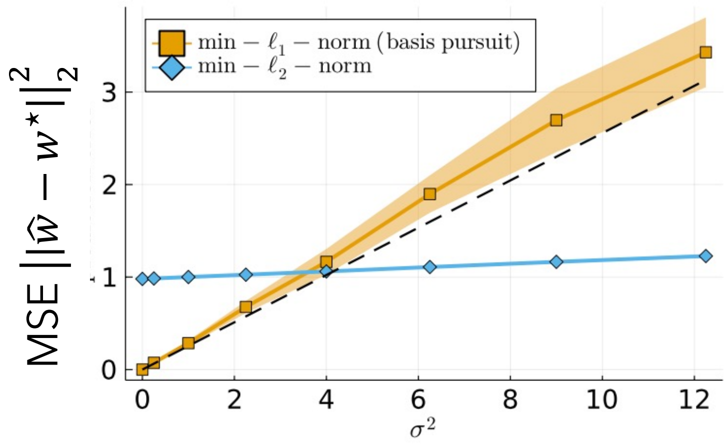
- Only holds for Gaussians

➡ cannot obtain best p for given β

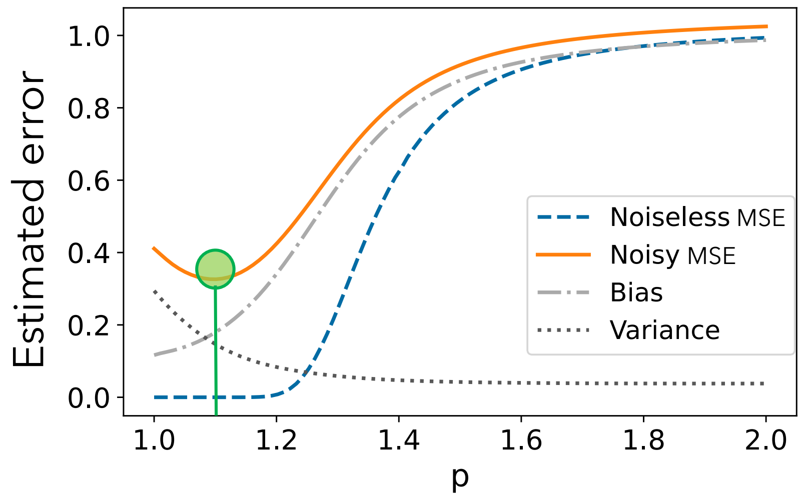
Experimental results for $p \in [1,2]$ (synthetic)

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

→ increasing d *vs.* n does not regularize enough even though it has relatively small bias.



for $d = 20000, n = 400$

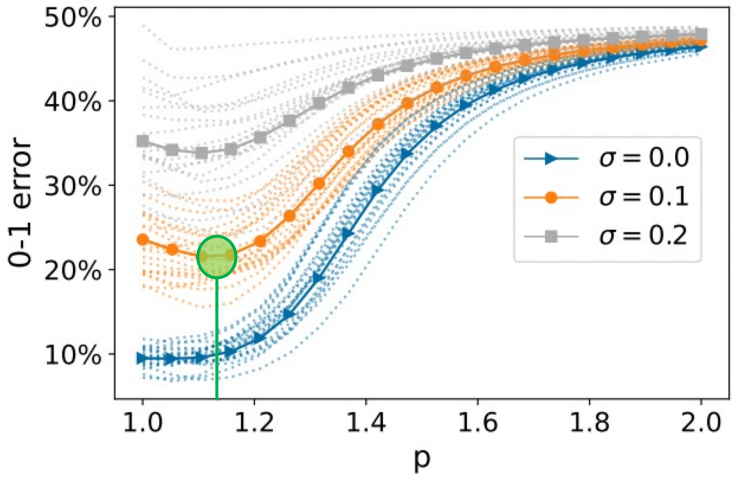


for $d = 5000, n = 100$

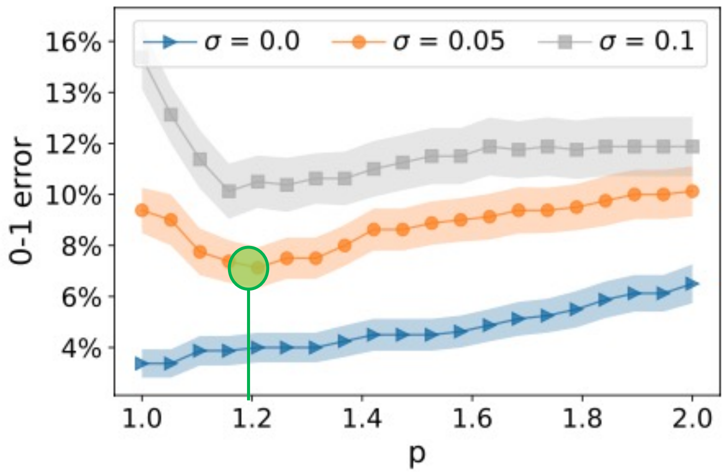
Trade-off between bias and variance for interpolators via strength of inductive bias!

Experimental results for classification (real-world)

Experimental results: hard-ℓ_p-margin SVM for σ: proportion of random label flips



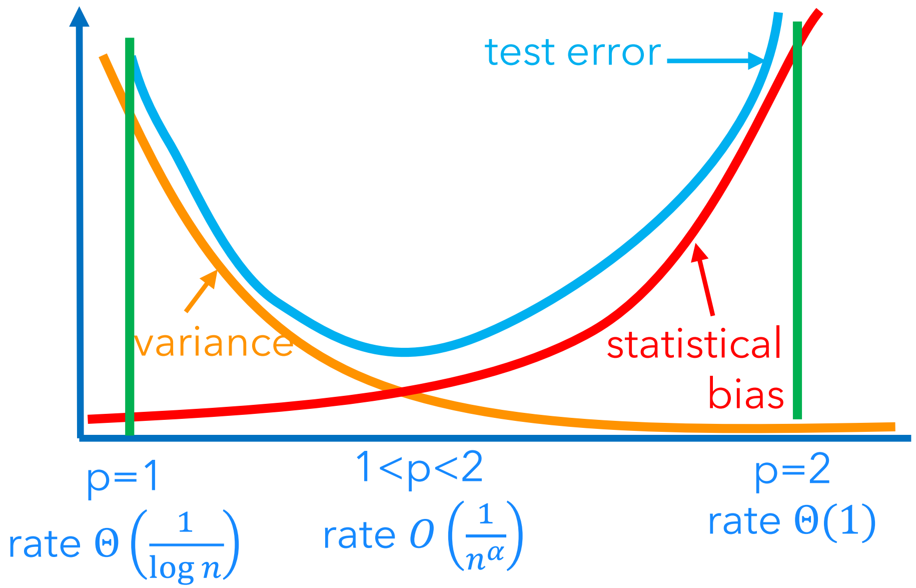
Synthetic experiment:
Isotropic Gaussians with $d \sim 5000, n \sim 100$



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

Full picture for $p \in [1, 2]$

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



- $p = 1$ best for noiseless interpolation
  but $p = 1 + \epsilon$ best for noisy interpolation!
- New bias-variance trade-off that
  shows for medium inductive bias:

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

From linear to non-linear

Bulk of talk

Part II: not yet published

Linear interpolators:

sparsity $||\widehat{w}||_0 \ll d$

Kernel interpolators:

Neural networks:

filter size for convolutional models

rotational invariance

Tight bounds for the risk

Controlled experiments

1 second descent

2 harmless interpolation

3 good generalization

Example IIa: Filter size of convolutional kernels

- Convolutional kernel with filter size q:

some regular κ e.g. exponential

  - consider patches $\left\{x_k^{(q)}\right\}_{k=1}^{d}$ of size q of vector $x \in R^d$
  - and average of nonlinear function over these patches $\mathcal{K}(x,z) = \frac{1}{d}\sum_{i=1}^{d} \kappa\left(\frac{\left\langle x_k^{(q)}, z_k^{(q)}\right\rangle}{q}\right)$
- $x \sim \mathcal{U}(\{-1,1\}^d)$ and $y = f^\star(x) + \sigma\epsilon$ with Gaussian $\epsilon \sim N(0,1)$ and $f^\star(x) = x_1 x_2$

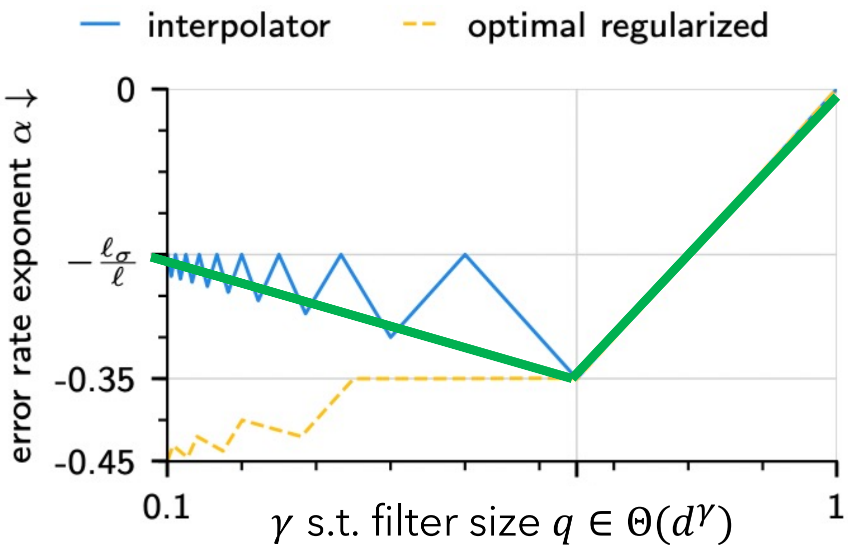
optimal model depends only on small patch → small filter size strongest inductive bias

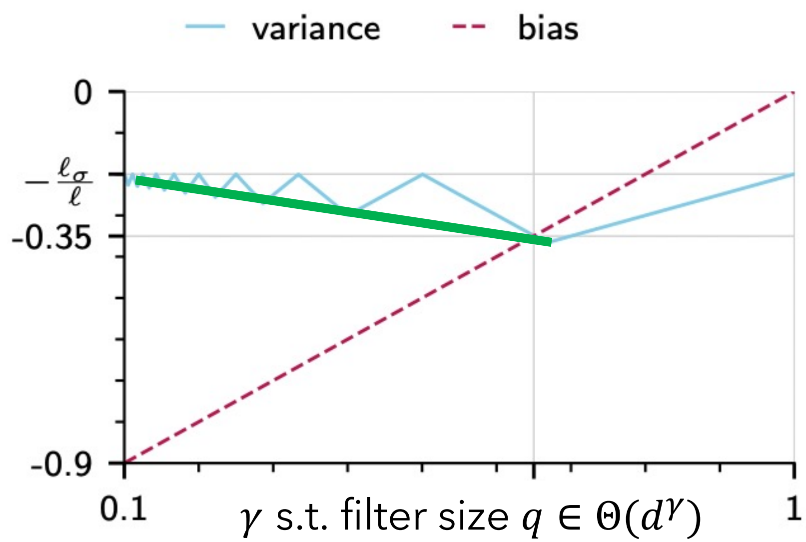
- High-dimensional kernel learning: $n \in \Theta\left(d^{\ell}\right), \sigma^2 \in \Theta(d^{-\ell_\sigma})$ and $q \in \Theta(d^{\gamma})$ with $\ell, \ell_\sigma, \gamma \geq 0$
- Interpolator: $\min \left|\left|f\right|\right|_H \; s.t. \; \forall i\colon \; f(x_i) = y_i$ vs. ridge regularized: $\min \left|\left|y - f(x_1^n)\right|\right|_2^2 + \lambda \left|\left|f\right|\right|_H^2$

Example IIa: Filter size of convolutional kernels

Illustration of our tight bounds (order Θ) for $n \in \Theta(d^{\ell}), \sigma^2 \in \Theta(d^{-\ell_\sigma})$, $q \in \Theta(d^{\gamma})$

where smaller γ / smaller filter size → stronger inductive/structural bias

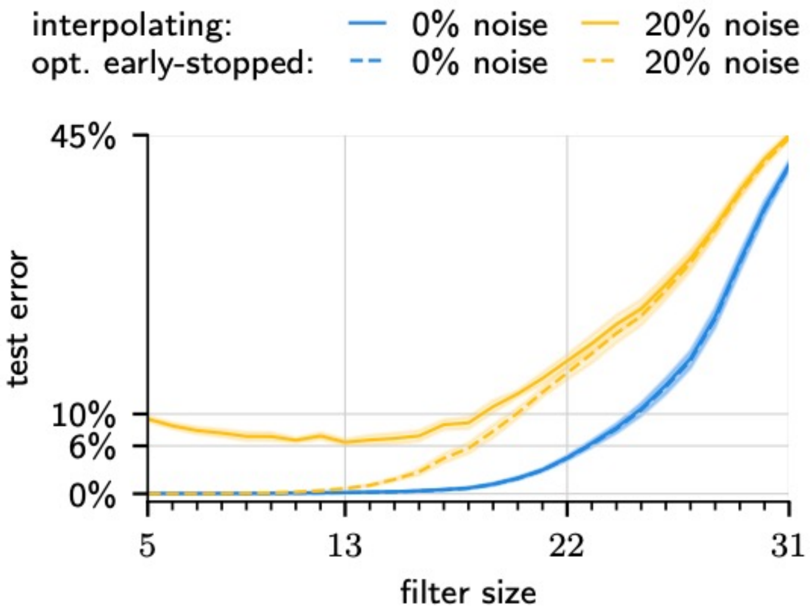




increasing inductive bias ←

Harmless interpolation only for weak inductive bias!

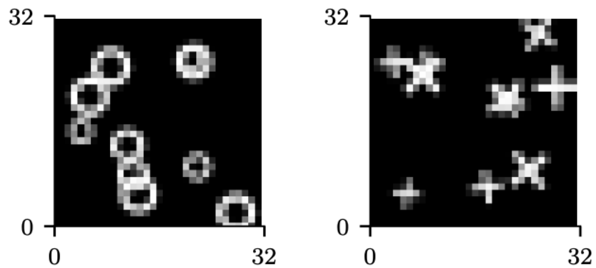
Example III: Filter size for convolutional NNs



- Synthetic image dataset

  allowing controlled experiments

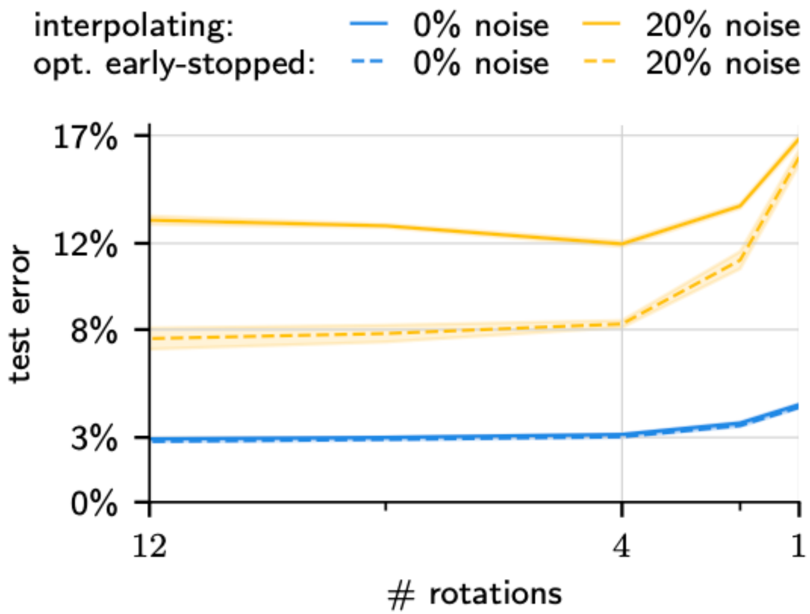
  where ground truth has small filter size



- simple NN with one convolutional layer

strongest inductive bias (smallest filter size) best for noiseless case, slightly weaker best for noisy
Harmless interpolation only for weak inductive bias!

Example III: Rotational invariance for WideResNet

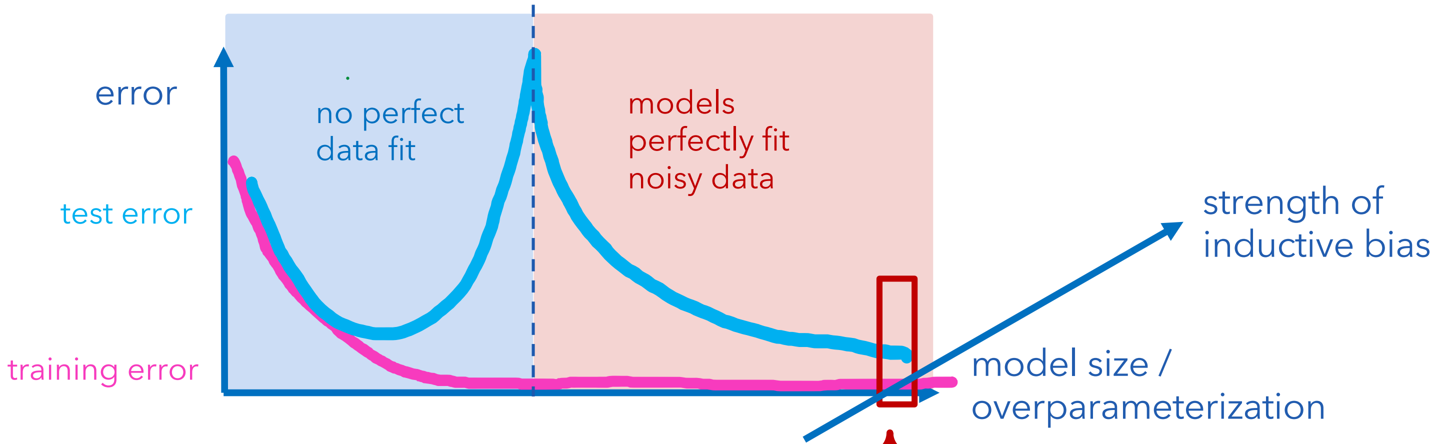


- Satellite images (EuroSAT) to be classified in terms of type of land usage

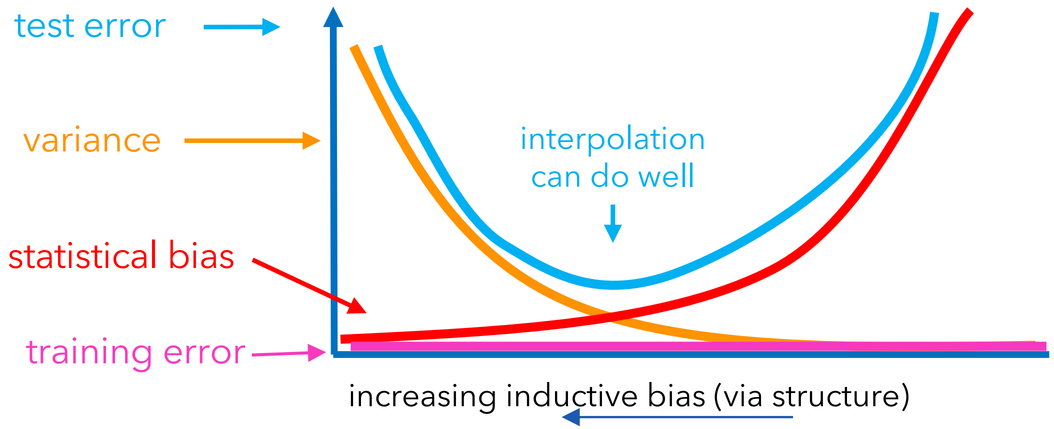
- strength of rotational invariance via "amount of" data augmentation

strongest inductive bias (largest #rotations) best for noiseless case, slightly weaker best for noisy

Take-aways…



Our theorems: increasing inductive bias
while interpolating
decreases bias, increases variance!



Interpolator can generalize well when

- known (noiseless case):
  there is strong inductive bias
  towards simple structure
  matching optimal model.

- new (noisy case):
  there is some but not too much
  inductive bias

Papers discussed in the talk

SML group: sml.inf.ethz.ch

- Wang*, Donhauser*, Yang *"Tight bounds for minimum l1-norm interpolation of noisy data"*, AISTATS '22
- Stojanovic, Donhauser, Yang *"Tight bounds for maximum ℓ1-margin classifiers"*, on arxiv soon
- Donhauser, Ruggeri, Stojanovic, Yang *"Fast rates for noisy interpolation require rethinking the effects of inductive bias"*, ICML '22
- Aerni*, Milanta*, Donhauser, Yang *"Strong inductive biases provably prevent harmless interpolation"*, on arxiv soon..