Strong inductive biases provably prevent harmless interpolation

January 6rd 2023, SlowDNN Workshop, Abu Dhabi

Fanny Yang, **K. Donhauser**

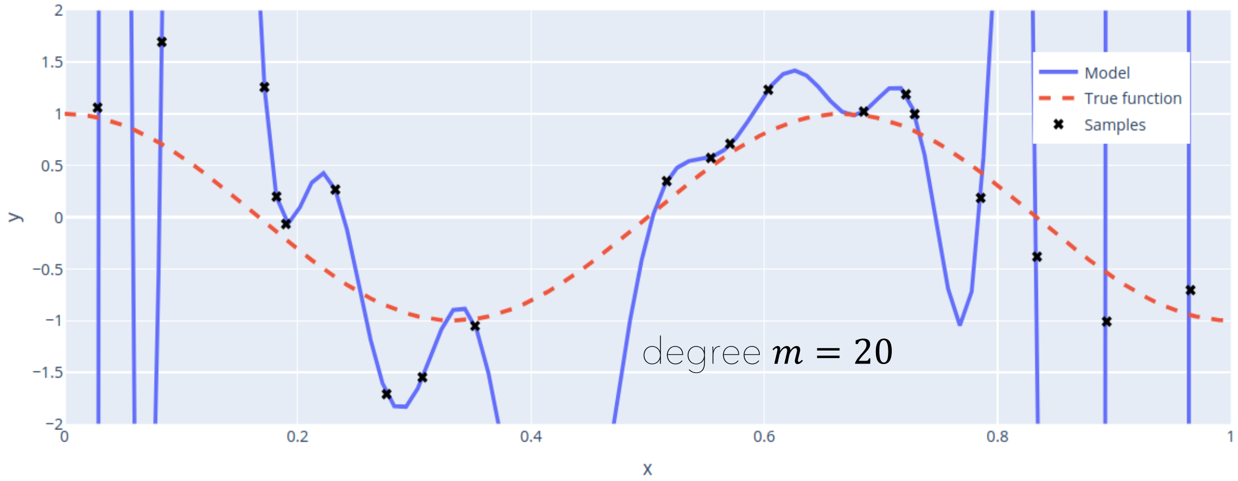
joint with G. Wang, S. Stojanovic, Marco Milanta, N. Ruggeri, Michael Aerni

Statistical Machine Learning group, CS department, ETH Zurich

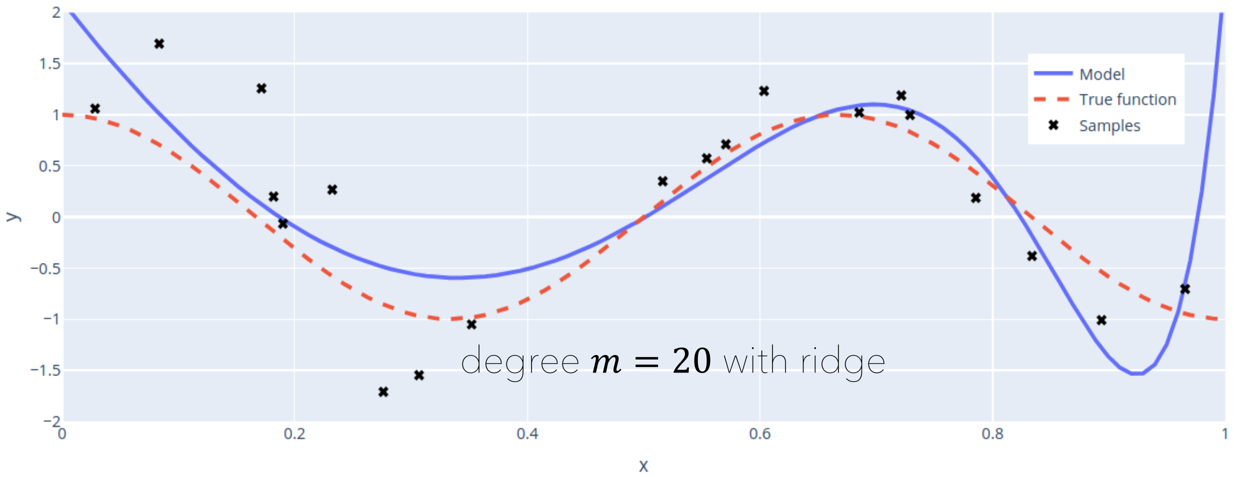
ETH zürich

DINFK

Classical wisdom: Avoid fitting noise



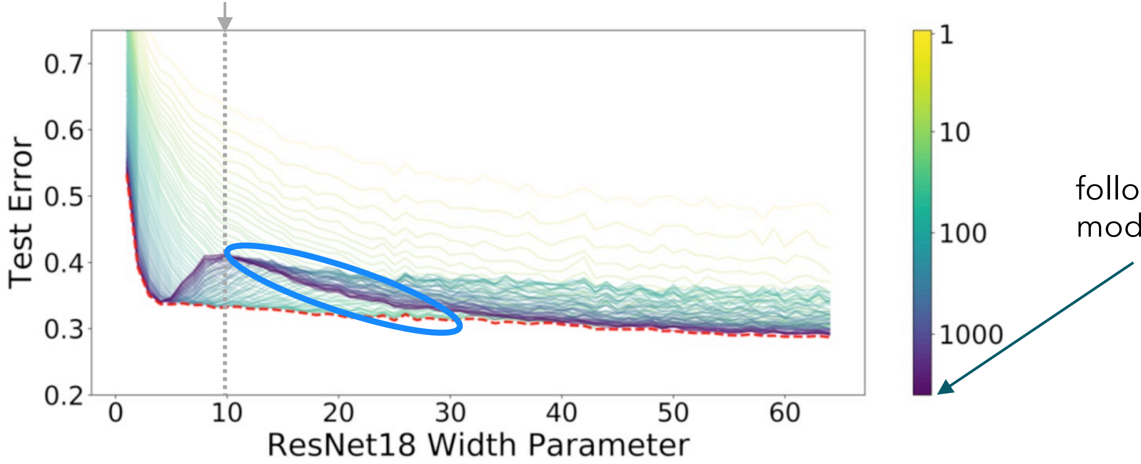
Classical wisdom: Avoid fitting noise



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

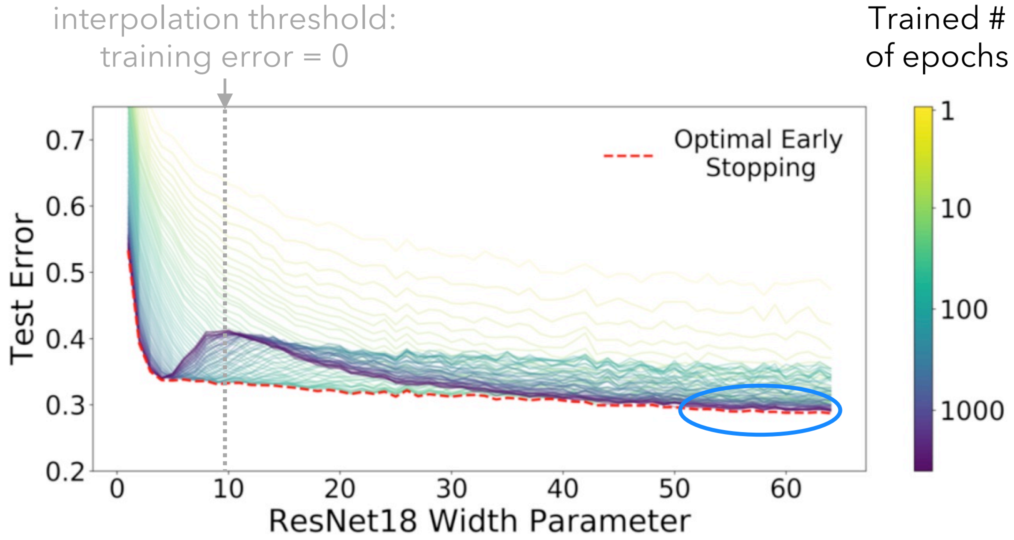


1 After interpolation threshold, we have a second "descent" – overparameterization helps

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: [NKBYBS '20]

Interpolators with certain structural/inductive bias

Question today: What kind of interpolators $\hat{f}$ with $\hat{f}(x_i) = y_i$ exhibit all of

(1) overparameterization helps interpolators (2) harmless interpolation (3) good generalization

well understood in linear case (see Misha's talk, Ohad's)

Focus in this talk

opt. algorithm minimizing loss → has implicit bias towards → certain interpolator (e.g. min norm) → has structural/inductive bias towards → e.g. sparsity, invariances, filter sizes

Good generalization for high-dim. diverse covariates (e.g. isotropic) only possible when interpolator "has clue" where to search (i.e. via structural bias aligned with optimal parameters)

Story of this talk…

Question today: What kind of interpolators $\hat{f}$ with $\hat{f}(x_i) = y_i$ exhibit all of

(1) overparameterization helps interpolators (2) harmless interpolation (3) good generalization

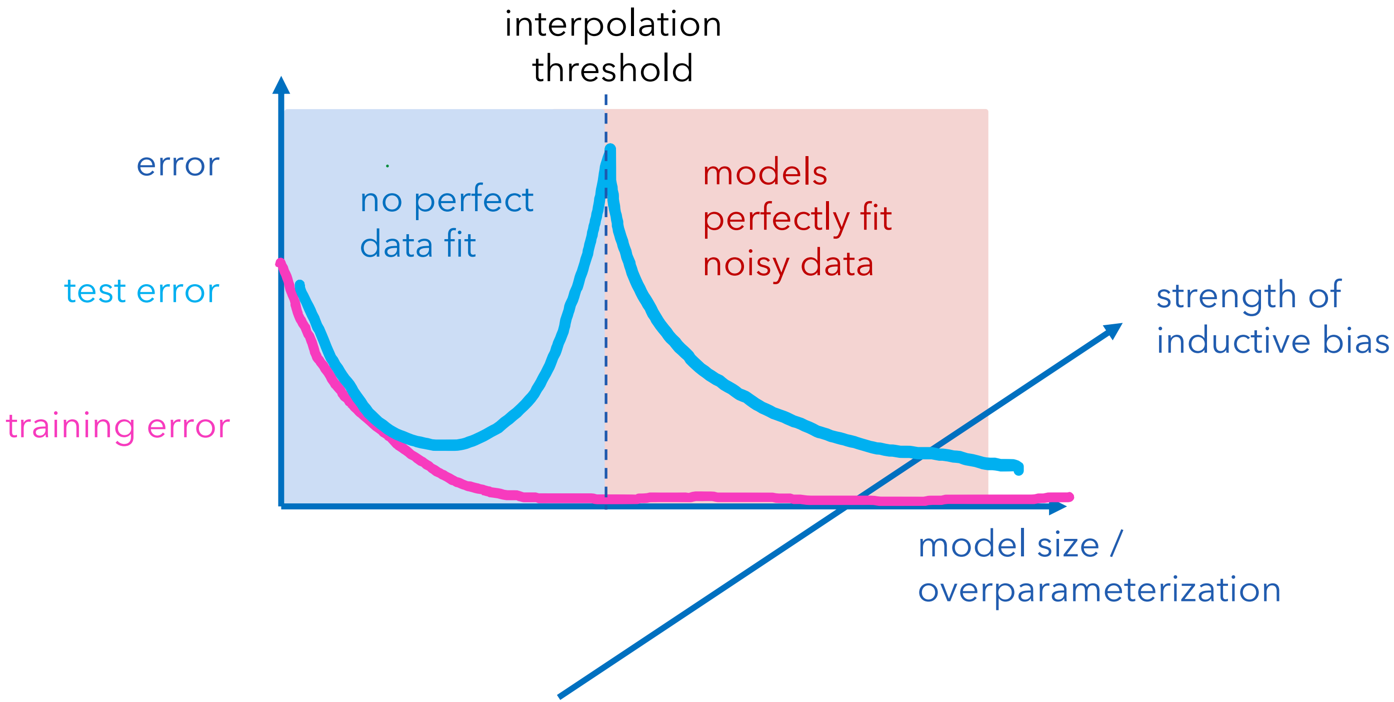
well understood in linear case (see Misha's talk, Ohad's) — (1), (2)

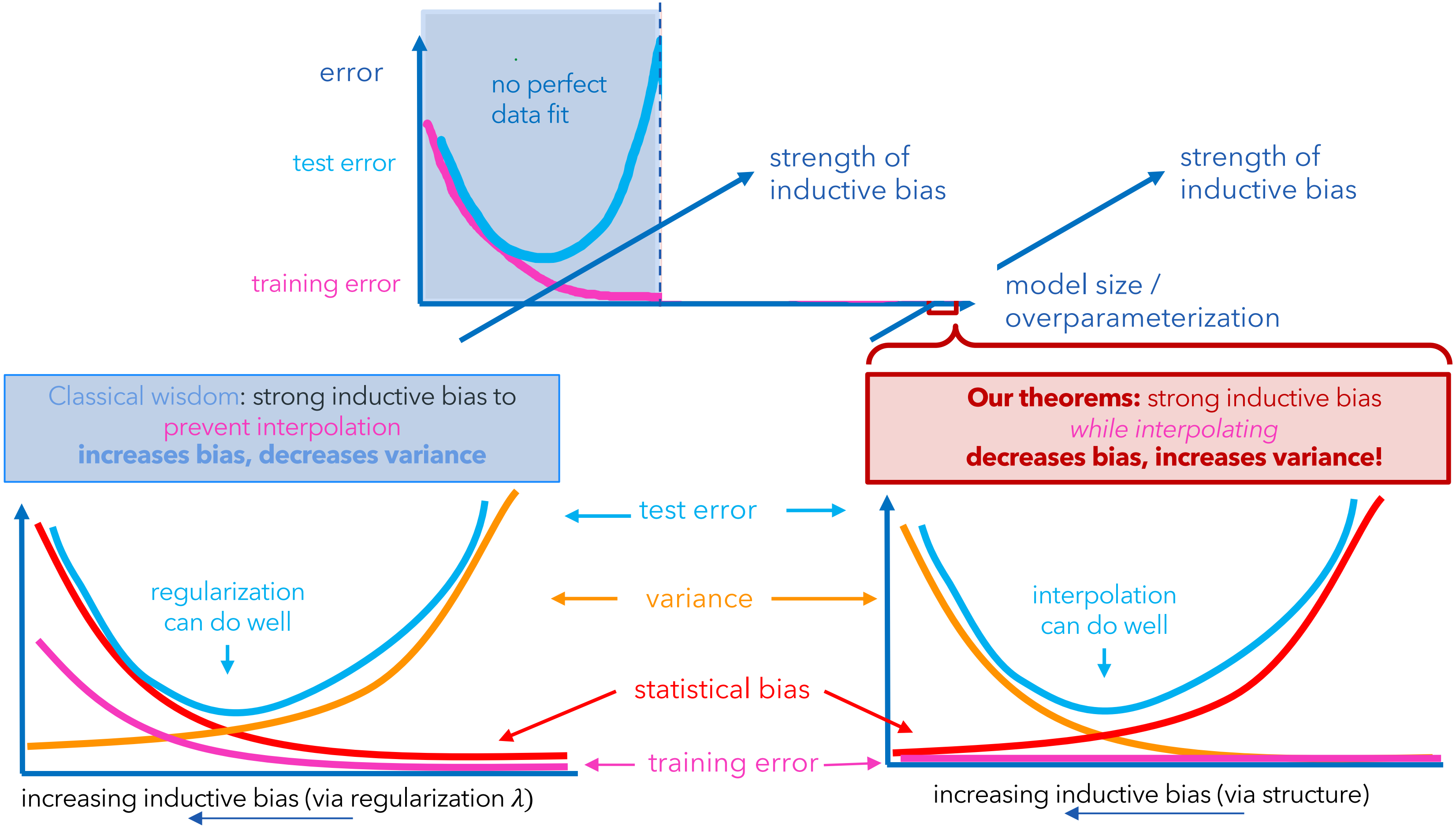
Focus in this talk — (2), (3)

Our take-away: One key mechanism to achieve (2) (3) is the degree of the "simplicity of the structure" of the interpolator that matches with optimal parameters,

i.e. the **strength of the "simplicity/inductive bias"**

The role of the inductive bias for interpolators




error
no perfect
data fit
test error
training error
strength of
inductive bias
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 n samples:** $y_i = \langle w^\star, x_i \rangle + \xi_i$ with $\boldsymbol{x_i} \sim \boldsymbol{N(0, I)}$ and noise $\xi_i \sim N(0, \sigma^2)$

  simple structure: sparse $w^\star = (1, 0, \dots, 0)$ with unknown location (for simplicity of presentation)
- **Degree of overparameterization (high-dimensional regime)**: $d \asymp n^\beta$, $\beta > 1$
- **Linear estimators we compare: for $p \in [1, 2]$**
  - **Minimum-ℓ_p-norm interpolators**: $\hat{w} = \mathrm{argmin}_w \, ||w||_p$ s.t. $y = Xw$ ← implicit bias of 1st order methods
  - **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 $\hat{w} = \text{argmin}_w \; ||w||_p \; s.t.\, y = Xw$

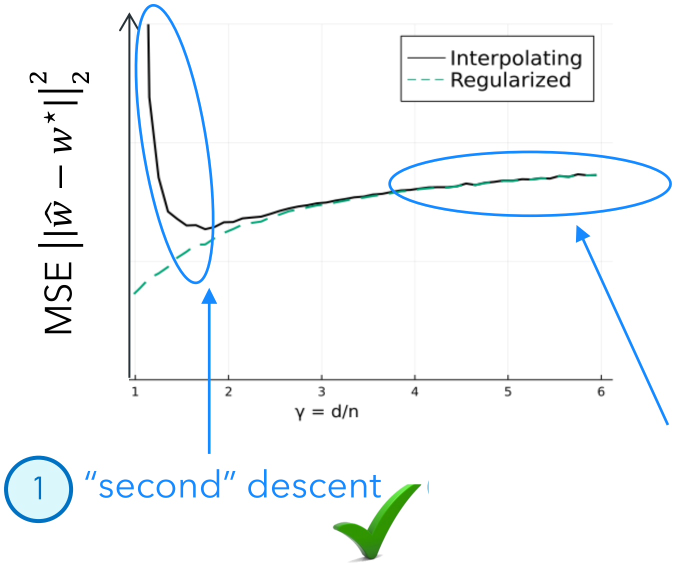
Goal today: populate for $p \leq 2$ with high-dimensional tight **non-asymptotic rates**

Weak inductive bias: $p = 2$ (prior work)

can analyze closed-form-solution!

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

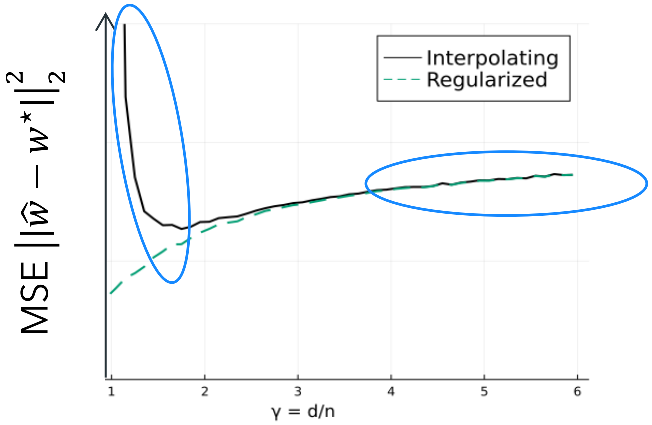
Linear model $y_i = \langle w^\star, x_i \rangle + \xi_i$ with i.i.d. $x_i \sim N(0, \mathrm{I})$, some $\xi_i \sim N(0, \sigma^2)$

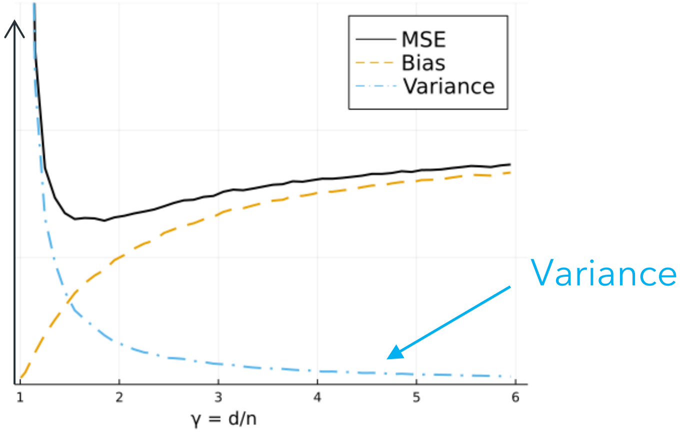


Weak inductive bias: $p = 2$ (prior work)

Interpolators $\hat{w} = \text{argmin}_w \, ||w||_2$ 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 $\xi_i \sim N(0, \sigma^2)$





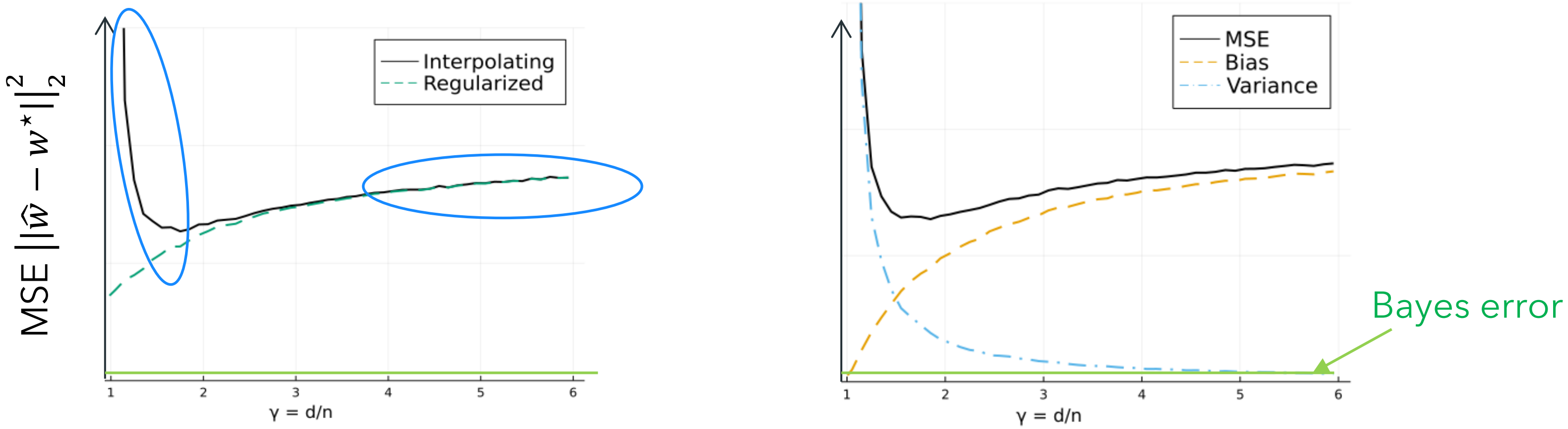
Increasing overparameterization via $\frac{d}{n}$ decreases variance ("implicitly regularizing")

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

Weak inductive bias: $p = 2$ (prior work)

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 $\xi_i \sim N(0, \sigma^2)$



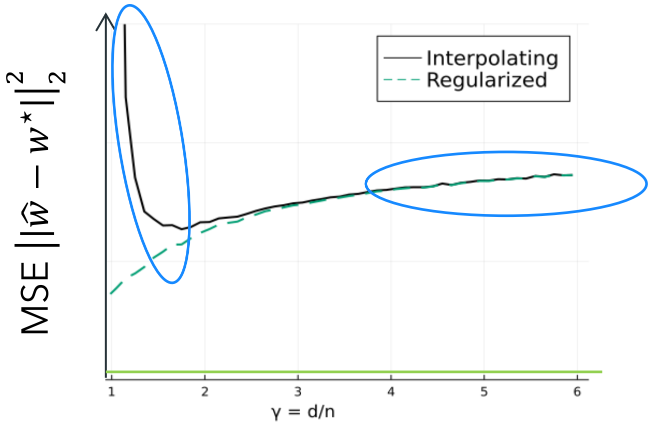
For isotropic Gaussians, $||\hat{w} - w^\star||^2 > c > 0$ for any $\beta > 1$ ($d \asymp n^\beta$) even as $n \to \infty$ due to high bias!

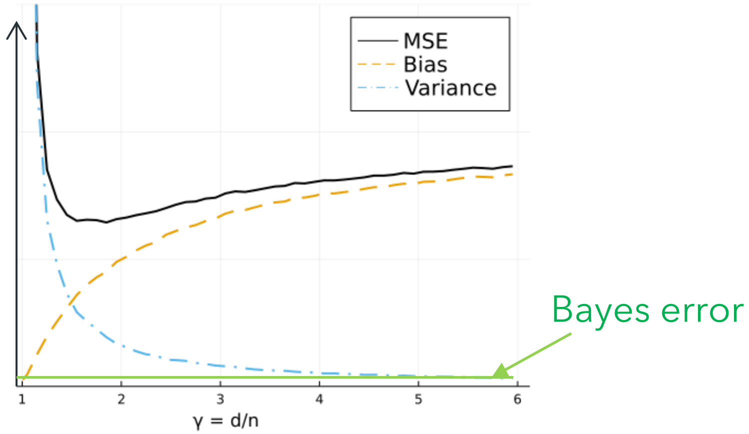
*consistent only for very spiked covariance Σ [HMRT'19, MM'19, BLLT '19, MVSS '20] ⚡ in practice Σ is fixed!

Weak inductive bias: $p = 2$ (prior work)

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 $\xi_i \sim N(0, \sigma^2)$



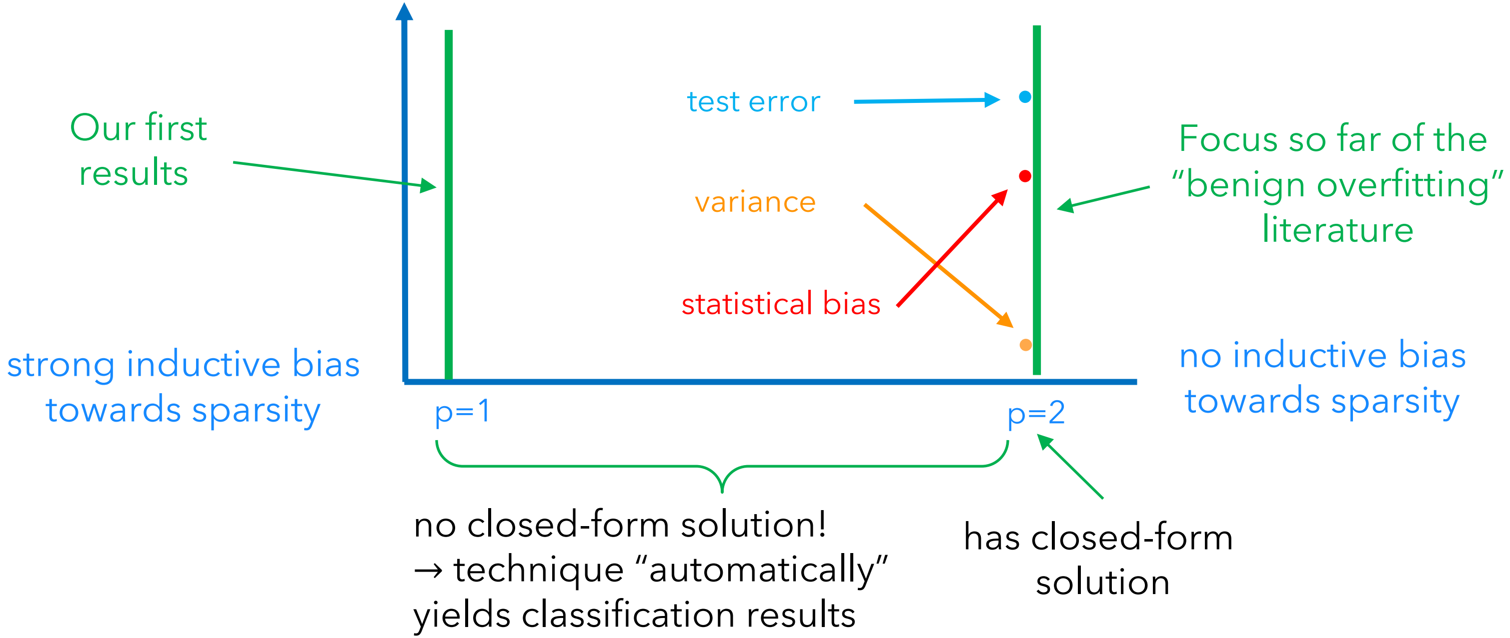


1 second descent ✓ 2 harmless interpolation ✓ 3 good generalization ✗

*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 $p = 1$ (classical)

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

Corresponding 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

Open problem: How much does min-ℓ_1-norm interpolation suffer when forced to fit noise?

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

Previous non-asymptotic bounds 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.

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

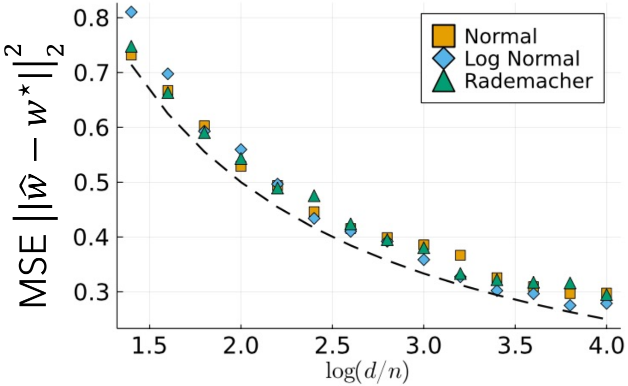
The proof is based on localized uniform convergence and CGMT [KZSS '21]
- who however don't show tight bounds and hence don't prove consistency

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 ⚠

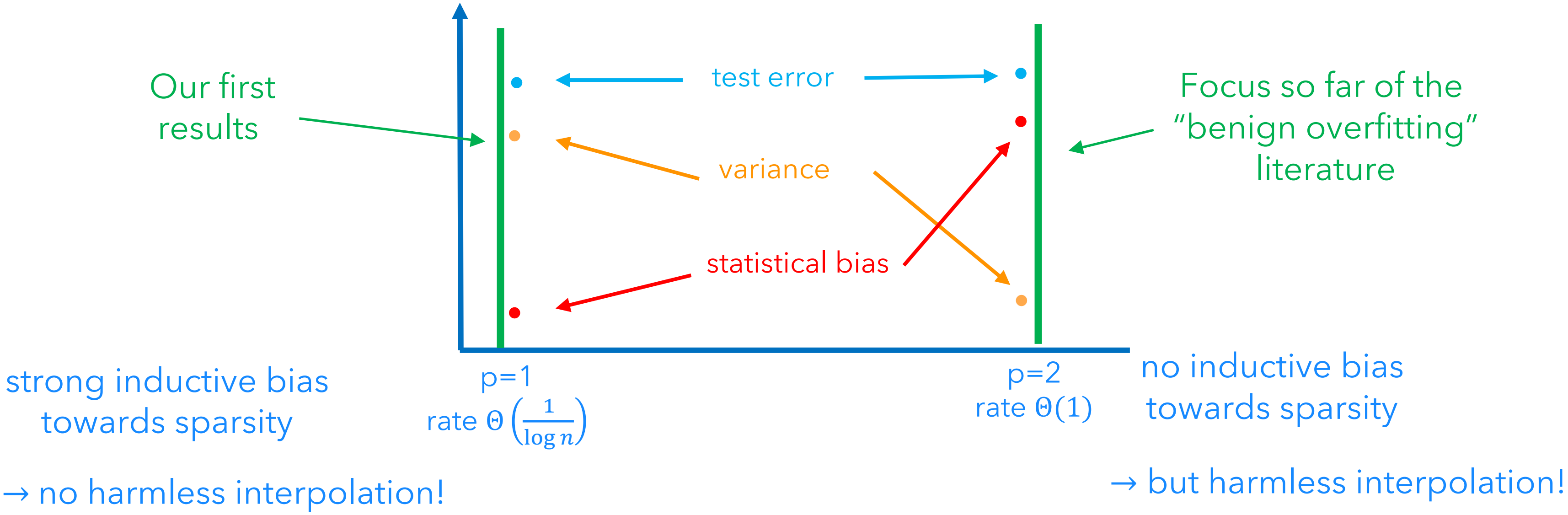
Yes! Variance decreases, similar intuition as for $p = 2$

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 are poor for $p = 1, 2$

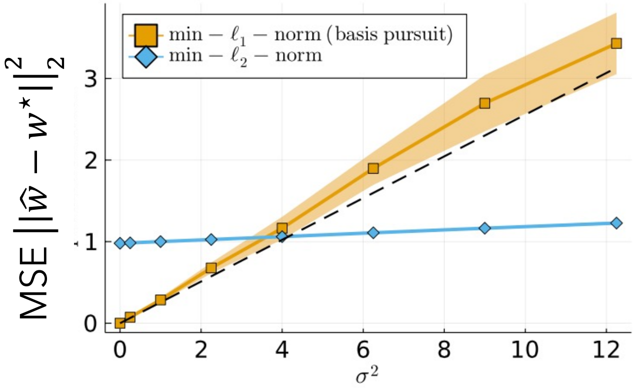
Min-ℓ_p-norm interpolation $\hat{w} = \text{argmin}_w \, ||w||_p \; s.t. \, y = Xw$



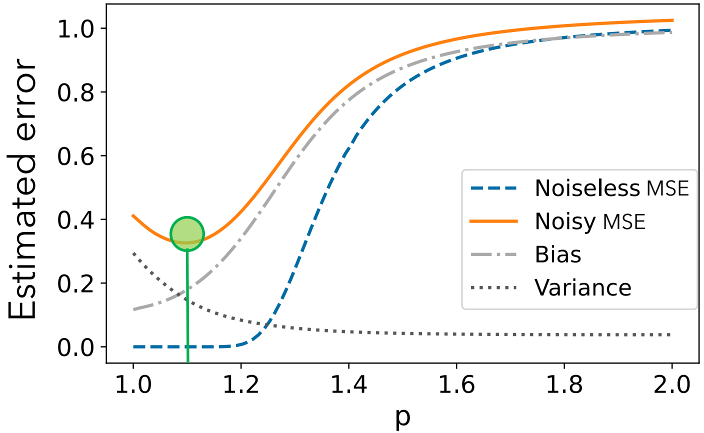
Higher noise sensitivity for $p = 1$ (synthetic)

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 $d = 20000, n = 400$



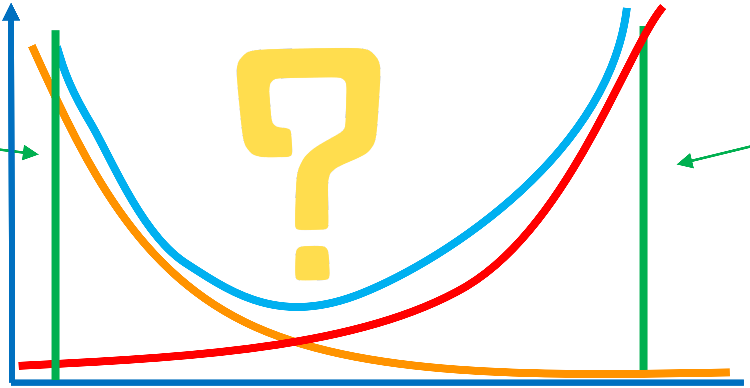
for $d = 5000, n = 100$

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

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

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

Our first results:



Focus so far of the "benign overfitting" literature

strong inductive bias towards sparsity

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

Which rates?

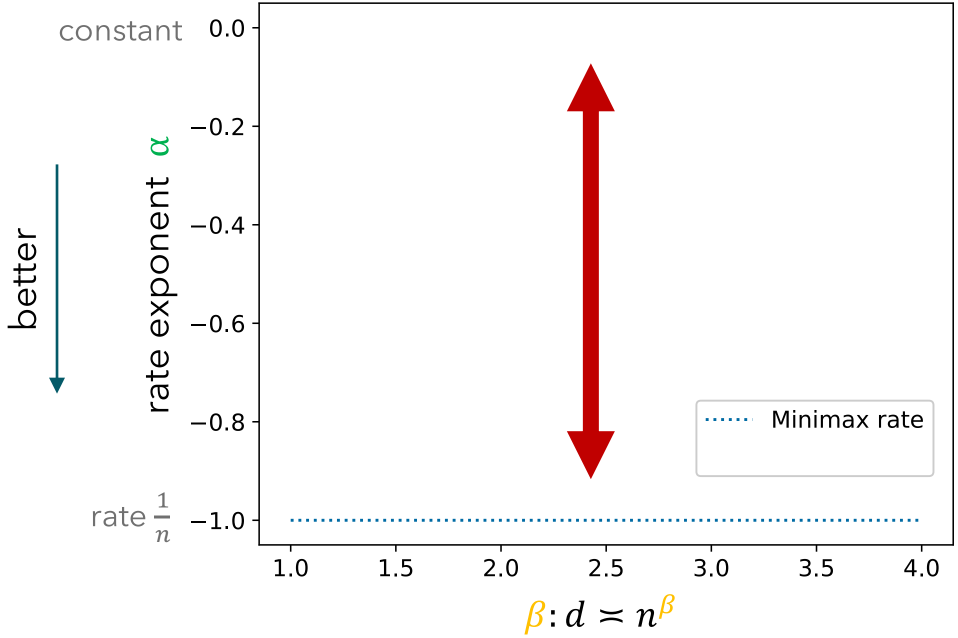
p=2
rate $\Theta(1)$

no inductive bias towards sparsity

→ no harmless interpolation!

→ but harmless interpolation!

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



- Evaluate MSE $||\hat{w} - w^{\star}||^2 \sim \widetilde{\Theta}(n^{\alpha})$ with rate exponent α
- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO)
  $||\hat{w}_{\lambda} - w^{\star}||^2 = \widetilde{\Theta}(n^{-1}) \rightarrow \alpha = -1$
- Interpolators with $p = 1, 2$:
  $||\hat{w} - w^{\star}||^2 = \widetilde{\Theta}(1) \rightarrow \alpha = 0$

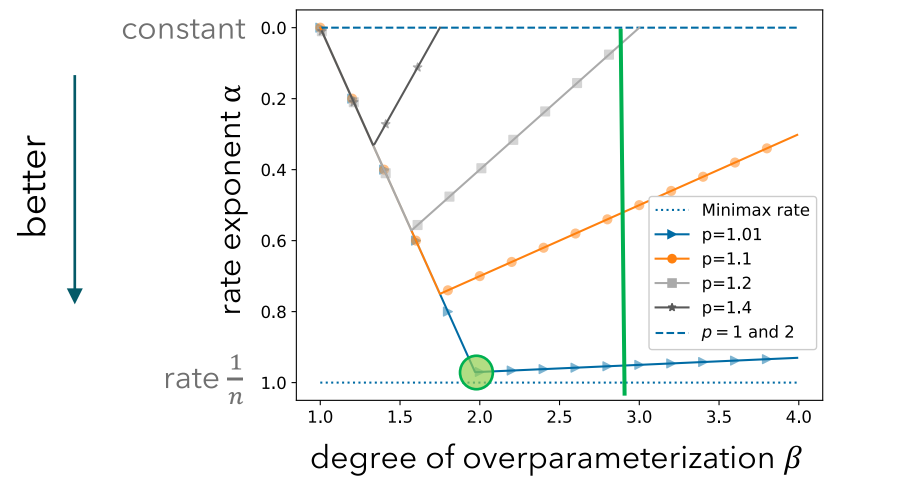
How close can we get to $\alpha = -1$ with ℓ_p-norm interpolators with $p \in (1,2)$?

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 $\tilde{\Theta}(n^{-\alpha})$ with α as in graph below



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

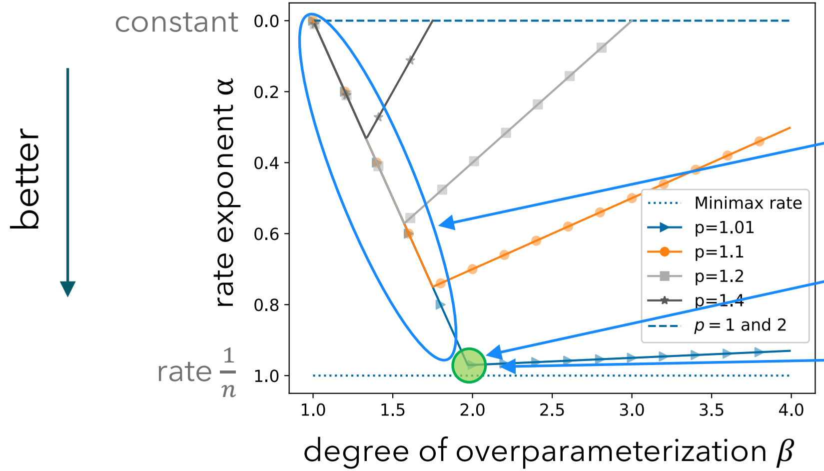
Note: technique applies to classification (see paper) and allows extension to $\Sigma \neq I$ and s-sparse $w^\star$

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



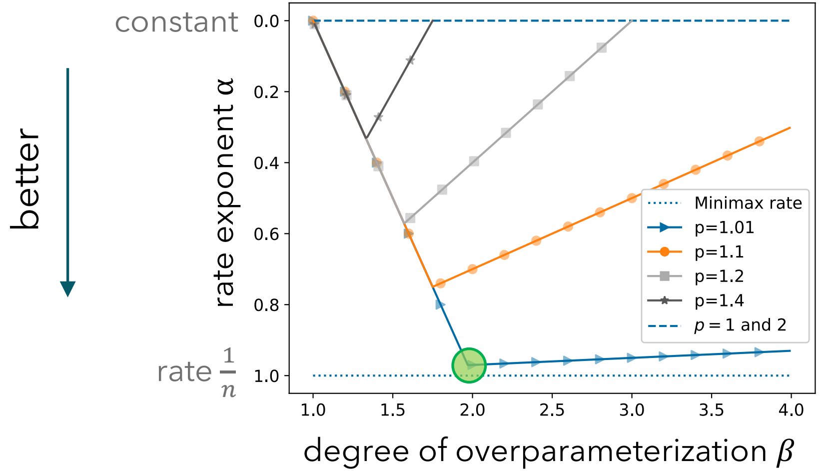
1. second descent ✓
2. harmless interpolation ✓
3. good generalization ✓

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

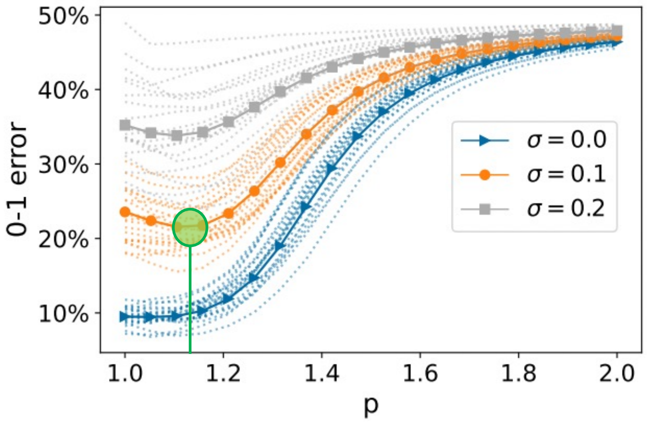
  $$\frac{1}{\log \log d} \lesssim p - 1 \to \text{very large d}$$

- Only holds for Gaussians

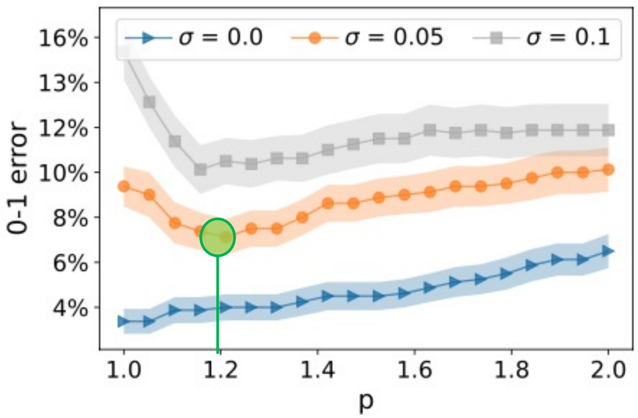
➡ cannot obtain best p for given β

Experimental results for classification (real-world)

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



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

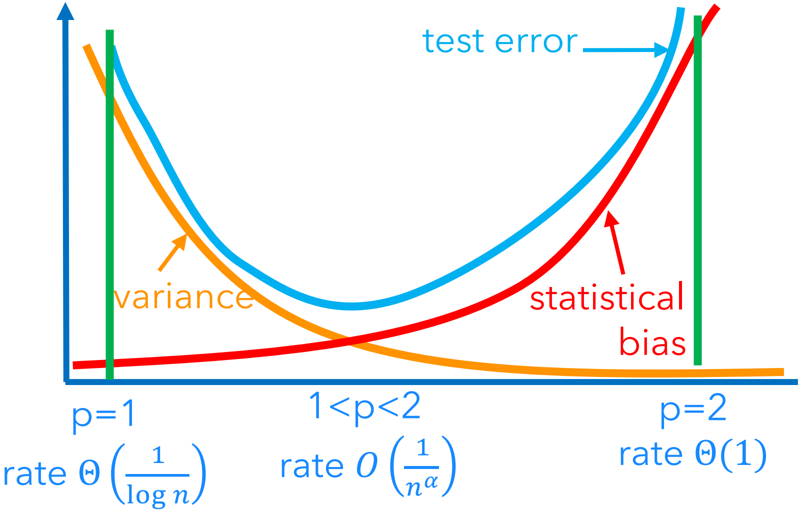


Real-world experiment:
Leukemia dataset with $d \sim 7000, n \sim 70$

Strong ind. bias best to interpolate noiseless data, medium ind. bias best to interpolate **noisy** data!

Conclusions for full picture $p \in [1, 2]$

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



A $p = 1$ (strongest bias) best for **noiseless** interpolation
$p = 1 + \epsilon$ (medium bias) best for **noisy** interpolation!

B Concrete non-asymptotic rates that show for medium-strength inductive bias:

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

Analogous phenomenon for non-linear models?

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

Nonlinear structure: Filter size of convolutional kernels

- Convolutional kernel with filter size q:
  - consider patches $\left\{x_k^{(q)}\right\}_{k=1}^{d}$ of size q of vector $x \in R^d$

    some regular κ e.g. exponential

  - 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 consider $f^\star(x) = x_1 \dots x_{L^\star}$

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

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

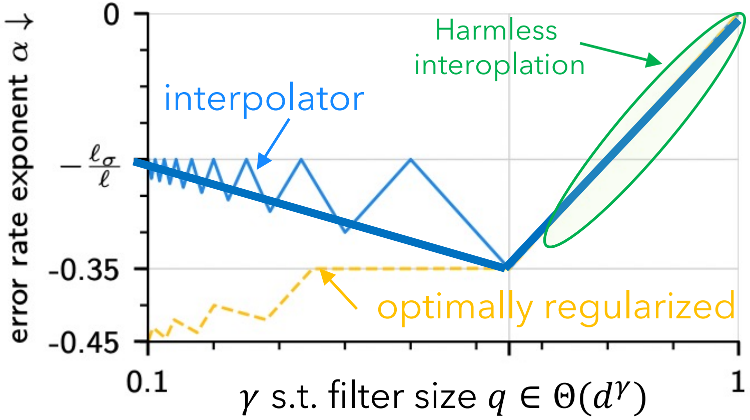
Nonlinear structure: Filter size of convolutional kernels

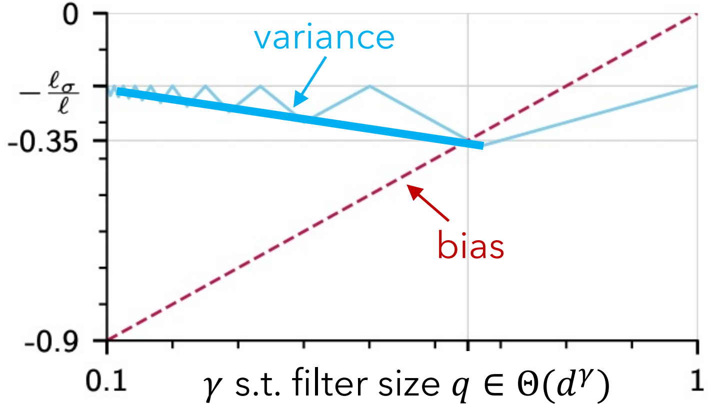
Theorem [AMDY' 22] (informal) – Upper & lower bounds for high-dim kernel learning

For $n \in \Theta(d^{\ell}), \sigma^2 \in \Theta(d^{-\ell_\sigma})$, $q \in \Theta(d^{\gamma})$, $\lambda \in \Theta(d^{\ell_\lambda})$ or $\lambda \to 0$ w.h.p., we obtain tight bounds

$Var(\hat{f}_\lambda) \in \Theta(n^{\frac{-\ell_\sigma - \ell_\lambda}{\ell} - \frac{\gamma}{\ell}\min\{\delta, 1-\delta\}})$ and $Bias^2(\hat{f}_\lambda) \in \Theta(n^{-2}\, n^{-\frac{2}{\ell}(\ell_\lambda + 1 + \gamma(L^\star - 1))})$ with $\delta = \frac{(\ell - \ell_\lambda - 1)}{\gamma} - \left\lfloor \frac{(\ell - \ell_\lambda - 1)}{\gamma} \right\rfloor$

yielding prediction error rates of order $\tilde{O}(n^{-\alpha})$ with α as in graph below for fixed $\ell, \ell_\lambda, \ell_\sigma$

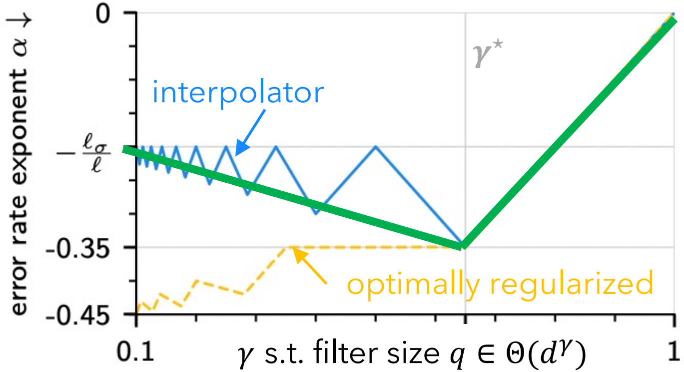


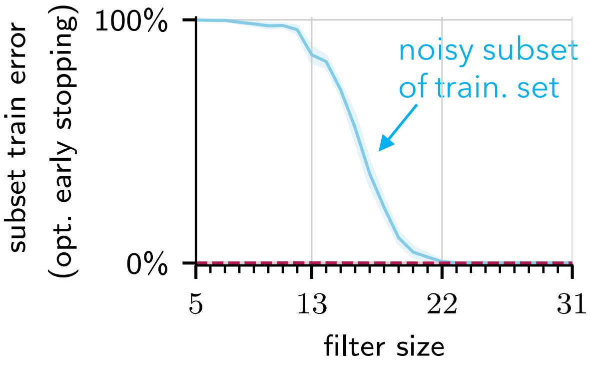


Prior work: [LRZ '19] showed multiple descent as a function of overparameterization

Fitting noise is necessary for weak inductive bias

- $\lambda^\star(\ell, \ell_\sigma, \gamma)$ minimizes population risk for $n \in \Theta(d^\ell), \sigma^2 \in \Theta(d^{-\ell_\sigma})$, $q \in \Theta(d^\gamma)$
- $\gamma^\star$: filter size exponent at which bias = variance (medium inductive bias)





on CNN
& synthetic
image data

Theorem [AMDY' 22] (informal) – Training error for optimally regularized model

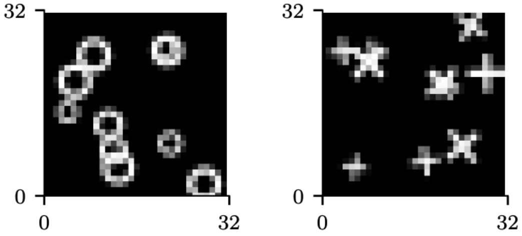
It holds for $\lambda^\star(\ell, \ell_\sigma, \gamma)$ that $E_\epsilon\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{\lambda^\star}(x_i) - y_i\right)^2\right] \to \tau_\gamma \sigma^2$ with $\tau_\gamma = 1$ if $\gamma < \gamma^\star$ and $\tau_\gamma < 1$ if $\gamma \geq \gamma^\star$

strong ind. bias   weak ind. bias

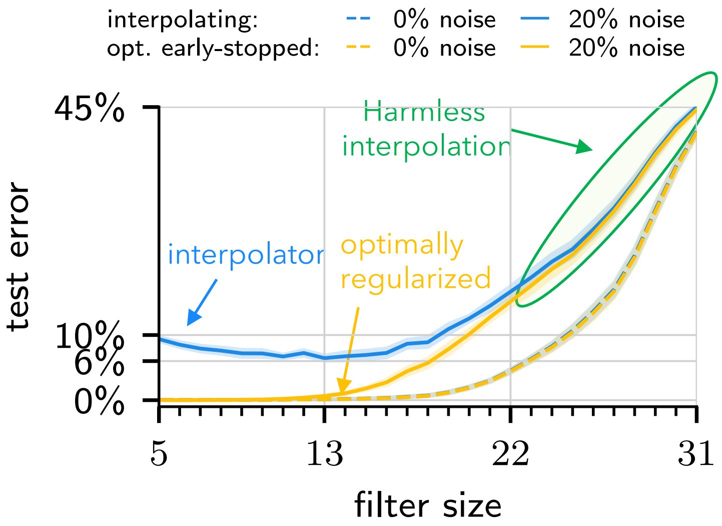
→ any noise fitting harmful for strong inductive bias vs. some noise fitting optimal for weak inductive bias

Nonlinear structure: Filter size of convolutional NN

- Synthetic image dataset
  allowing controlled experiments
  where ground truth has small filter size



- simple NN with one convolutional layer

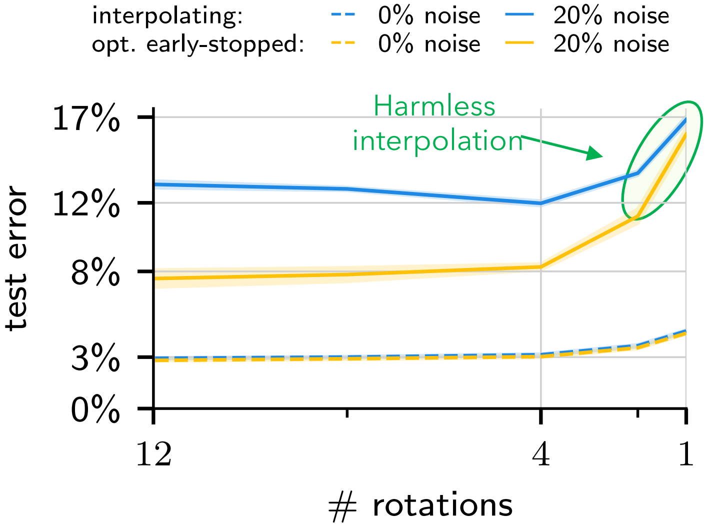


A strongest inductive bias (smallest filter size) best for noiseless case, slightly weaker best for noisy

B harmless interpolation only for weak inductive bias!

Nonlinear structure: Rotational invariance for WideResNet

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

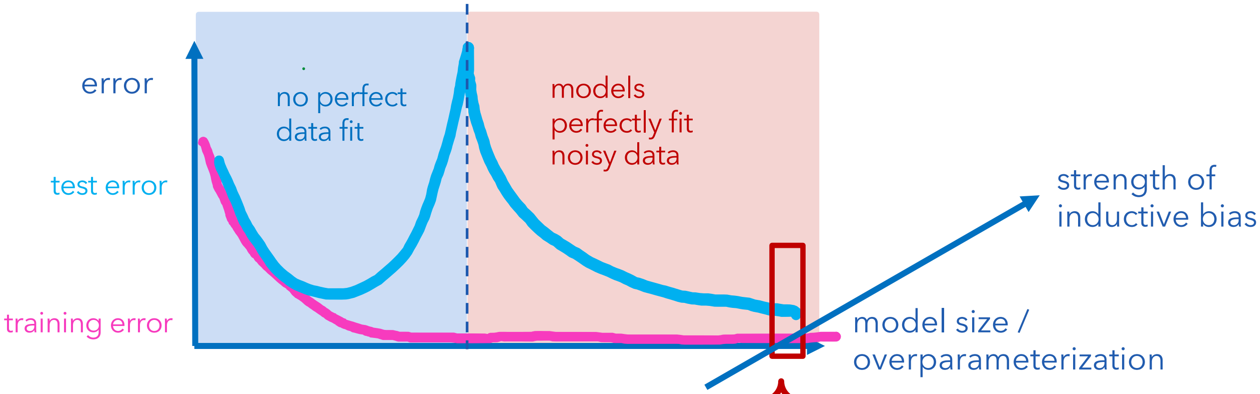


- strength of rotational invariance via "amount of" data augmentation

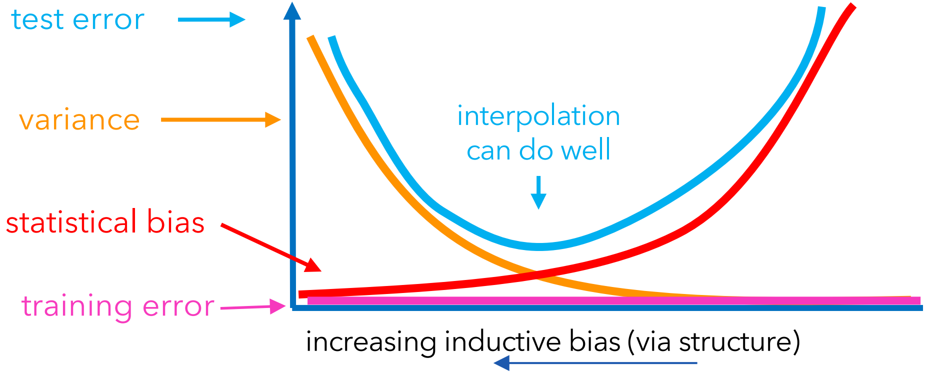
A strongest inductive bias (largest # rotations) best for noiseless case, slightly weaker best for noisy

B harmless interpolation only for weak inductive bias!

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

Open questions

For linear

- Technical: Going beyond Gaussians – seems surprisingly difficult

For non-linear:

- Technical: going beyond toy covariate distributions (or toy kernels)
- Proof for neural networks?
- **Experimental: What are other natural structural biases & datasets for NN one could test our hypothesis on?**

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*", arxiv preprint
- 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*", hopefully ICLR '23…

Clean theorem statement for min-ℓ_p

Theorem 1. *Let the data distribution be as described in Section 2.1 and assume that $\sigma \asymp 1$. Further, let q be such that $\frac{1}{p}+\frac{1}{q}=1$. Then there exist universal constants $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6, \kappa_7 > 0$ such that for any $n \geq \kappa_1$ and any $p \in \left(1+\frac{\kappa_2}{\log\log(d)}, 2\right)$ and $n\log(n)^{\kappa_3} \lesssim d \lesssim n^{q/2}\log(n)^{-\kappa_4 q}$, the estimation error of the min-ℓ_p-norm interpolator 1 is upper and lower bounded by*

$$\frac{\sigma^{4-2p}q^p d^{2p-2}}{n^p} \vee \frac{\sigma^2 n}{d} \lesssim \mathrm{R}_{\mathcal{R}}(\hat{w}) \lesssim \frac{\sigma^{4-2p}q^p d^{2p-2}}{n^p} \vee \frac{\sigma^2 n \exp(\kappa_5 q)}{qd}, \tag{2}$$

with probability at least $1-\kappa_6 d^{-\kappa_7}$ over the draws of the data set.

Theorem 4. *Let the data distribution be as described in Section 3.1 and assume that the noise model $\mathbb{P}_\sigma$ is independent of n, d and p. Let q be such that $\frac{1}{p}+\frac{1}{q}=1$. There exist universal constants $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6, \kappa_7 > 0$ such that for any $n \geq \kappa_1$, any $p \in \left(1+\frac{\kappa_2}{\log\log(d)}, 2\right)$ and any $n\log^{\kappa_3}(n) \lesssim d \lesssim \frac{n^{q/2}}{\log^{\kappa_4 q}(n)}$, the prediction error of the max-ℓ_p-norm interpolator 4 is upper bounded by*

$$\mathrm{R}_{\mathcal{C}}(\hat{w}) \lesssim \frac{\log^{3/2}(d) q^{\frac{3}{2}p} d^{3p-3}}{n^{\frac{3}{2}p}} \vee \frac{n\exp(\kappa_4 q)}{qd} \vee \frac{\log^{\kappa_5}(d)}{n}, \tag{6}$$

with probability at least $1-\kappa_6 d^{-\kappa_7}$ over the draws of the data set.