DINFK

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

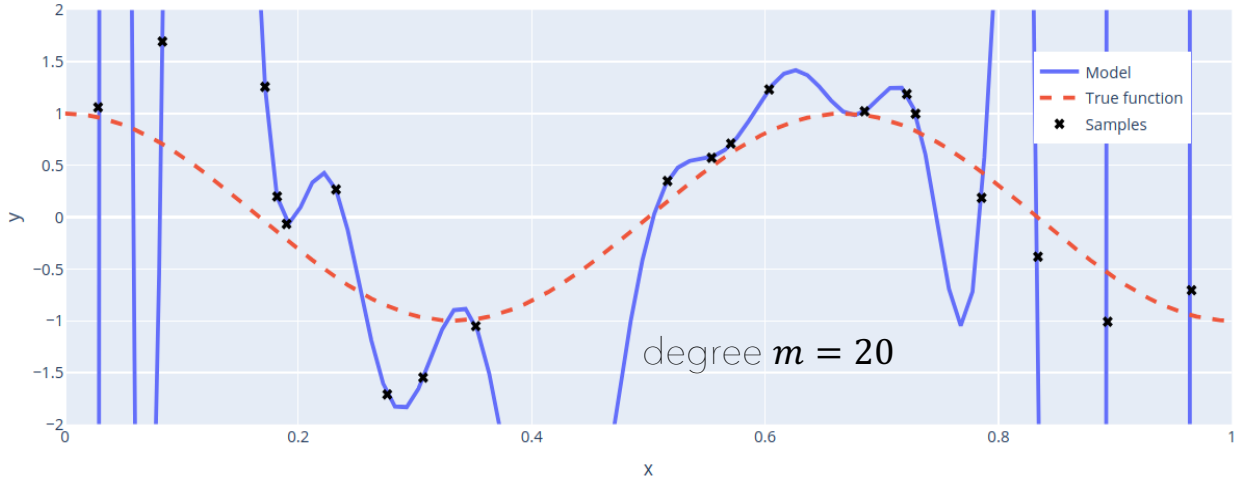
June 9nd 2022, 1W-MINDS seminar

Fanny Yang, joint work with **K. Donhauser**, G. Wang, S. Stojanovic, N. Ruggeri

Statistical Machine Learning group, CS department, ETH Zurich

ETH zürich

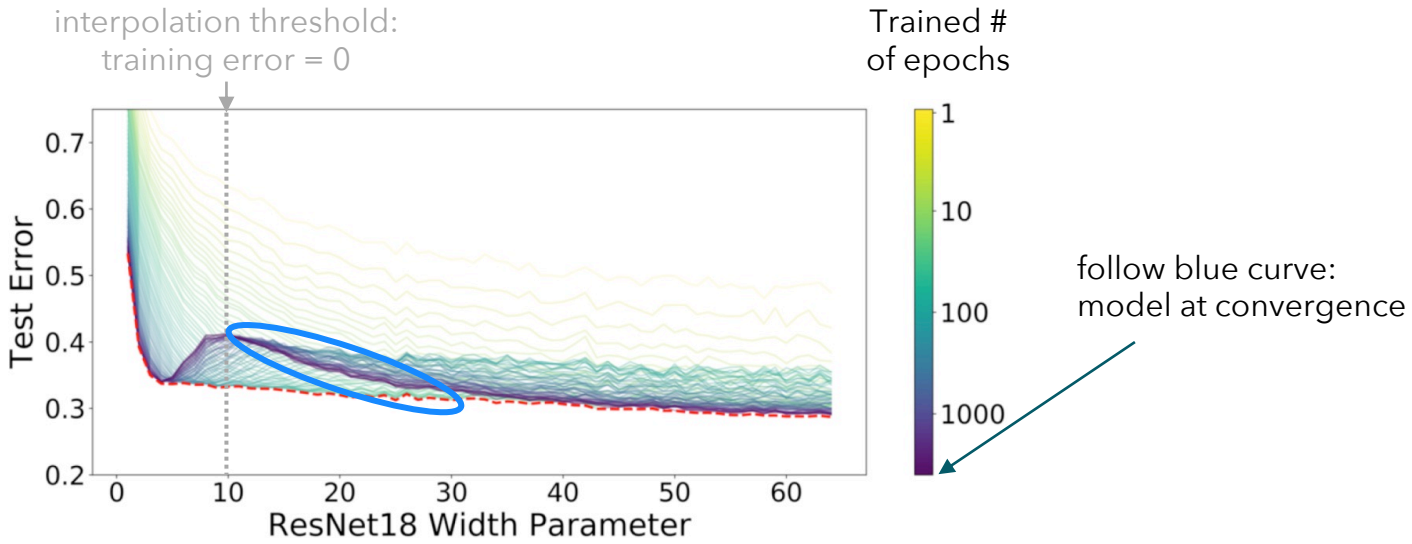
Classical wisdom: Avoid fitting noise



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

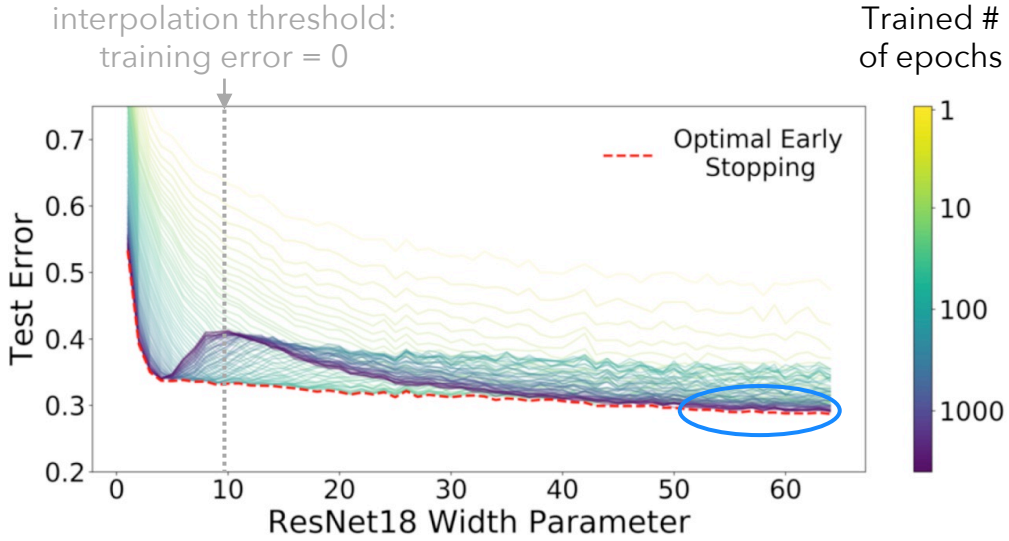


1 After interpolation threshold, we have a second "descent" (double descent)

Source: [Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever '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)

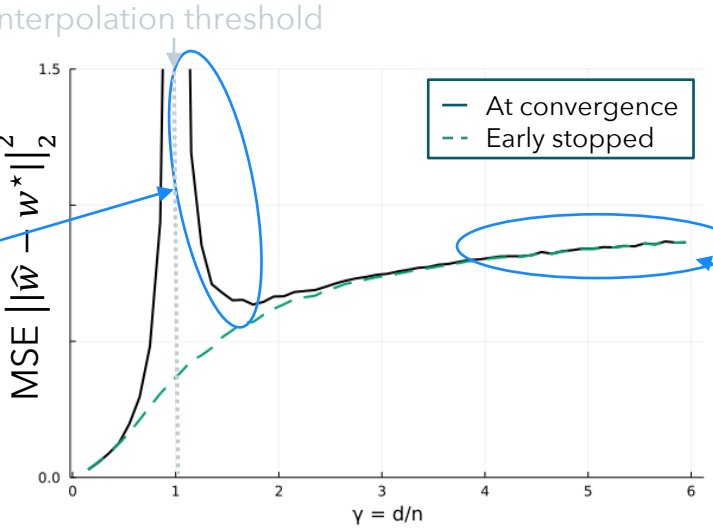
2 For large models, interpolation is not worse than regularization (harmless interpolation)

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

Are these observations unique to neural networks?

Analogous surprising observations can be made for **linear models!**

Here we run gradient descent on $\left||y - Xw|\right|_2^2$ with $w_0 = 0$ for $y = Xw^\star + \xi$ with X, ξ standard Gaussians



d: # features
n: # samples

1 Second Descent after interpolation

2 Harmless interpolation for large d/n

Interpolating models on two ends of analyzability

Neural network interpolators

- are found by using 1st order methods to minimize **non-convex losses**
- feature learning with overparameterization ~ e.g. width of hidden layers

Linear interpolators

- are interpolators that minimize **convex optimization problems**
- fixed d features with overparameterization ~ in terms of how much larger $d \gg n$

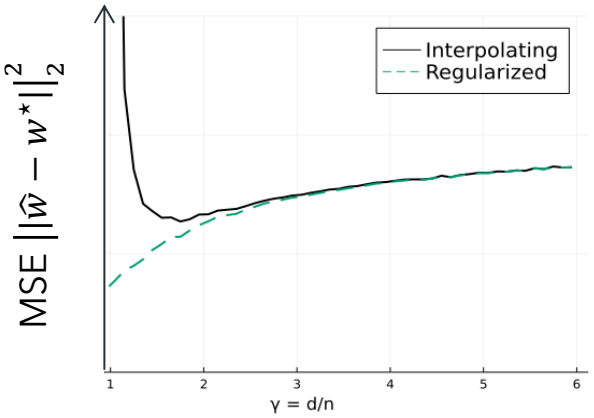
High-level talk outline

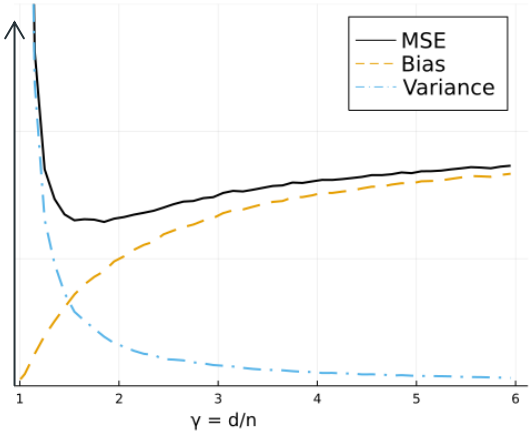
- Observations for interpolating models
  - "Second" descent
  - Harmless interpolation
- Explanation for the phenomena for linear interpolators
  - Previous intuition: Bias-variance trade-off by varying degree overparameterization
  - New complementary intuition: Bias-variance trade-off by varying inductive bias
  - Tight bounds show how moderate inductive bias can yield fast error rates

Previous work for min-ℓ_2-norm interpolators

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

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





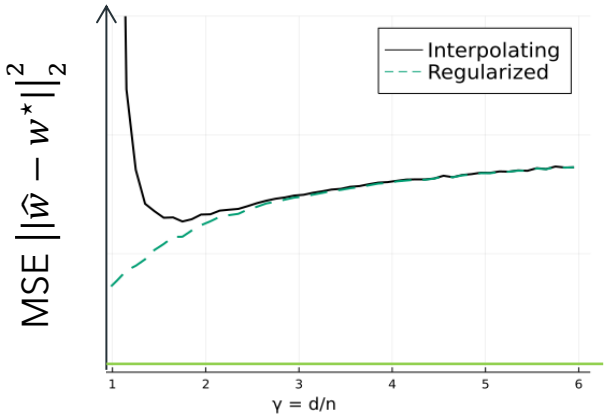
Previous bounds* explain harmless interpolation & second descent: As $\frac{d}{n}$ increases, **variance decreases!**

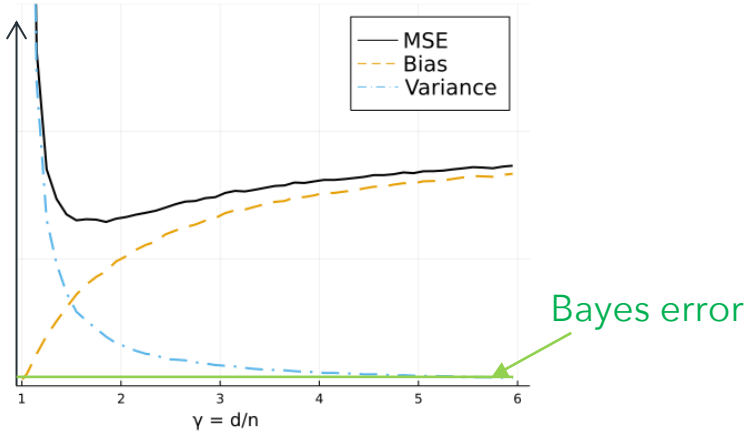
*[HMRT'19, MM'19, BLLT '19, MVSS '20]

Previous work for min-ℓ_2-norm interpolators

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

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





Are we happy? No, as opposed to NN, error overall is actually high for large $\frac{d}{n}$ **as the bias increases!**

What's missing? Structure…

Problem: Estimator has no "clue" where to search (all directions are equally valid)

- Line of work shows that min-ℓ_2-interpolator can generalize well*

  but only for very specialized covariance Σ  in practice Σ is fixed!

Question: What kind of interpolators can learn $w^\star$ well for large $d \gg n$?

Classical intuition for $d \gg n$: good estimation only possible if

- we assume **simple structure of $w^\star$** (such as sparsity) and
- the estimator has a **strong matching inductive bias** encouraging structural simplicity

*[BLLT '19, TB '20, MNSBHS '21, LS '22]

Benefits of strong inductive bias (recap)

Example for structural simplicity: sparsity $||w^\star||_0 = s \ll d$

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

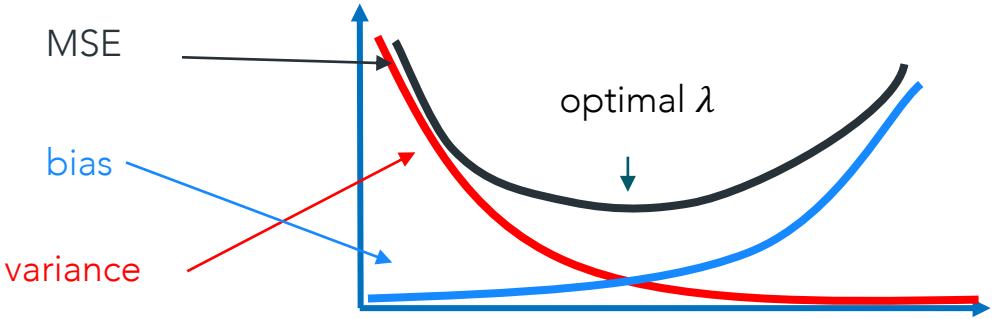
Estimation error
minimax rate $O\left(\frac{s \log d}{n}\right)$
for optimal λ

Old: Bias-variance trade-off via model complexity

Noisy
$y = Xw^{\star} + \epsilon$

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

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



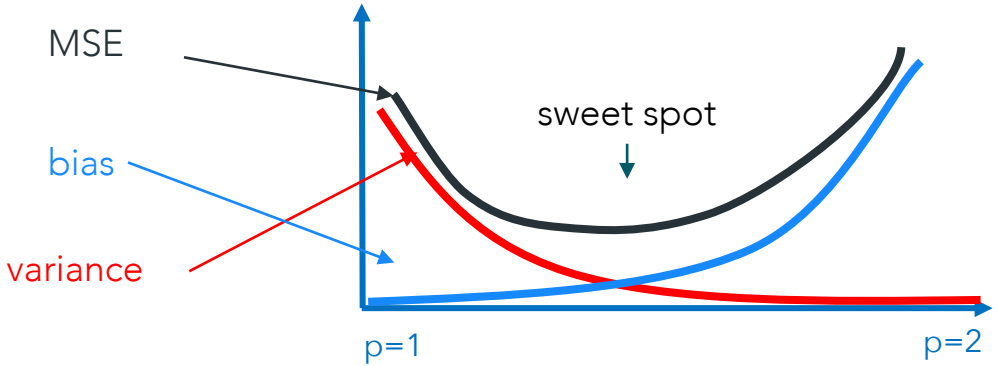
decreasing fit of noise via increasing λ

But interpolators cannot attenuate noise-fitting by choosing an optimal λ!

Our work: Bias-variance trade-off via inductive bias

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

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



decreasing "strength of inductive bias" via increasing p

Interpolators cannot attenuate by increasing p (decreasing structural bias)!

Setting for presentation of our results (simplified)

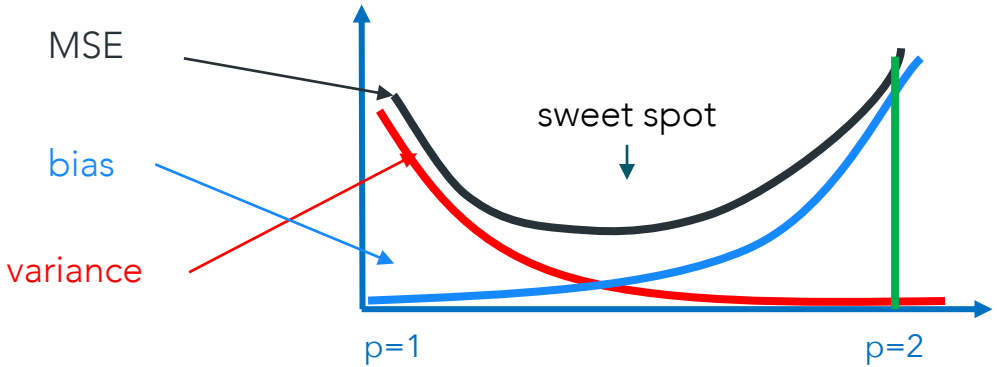
- **Function space:** linear models $f(x) = \langle w, x \rangle$ with $x, w \in \mathbb{R}^d$
- **Data model for n samples:** $y_i = \langle w^\star, x_i \rangle + \xi_i$ with $x_i \sim N(0, I)$ and noise $\xi_i \sim N(0, \sigma^2)$

  with sparse $w^\star = (1, 0, \ldots, 0)$ with unknown location (for simplicity of presentation)
- **Degree of overparameterization**: $d \asymp n^\beta$, $\beta > 1$
- **Minimum-ℓ_p-norm interpolators for $p \in [1, 2]$:** $\hat{w} = \operatorname{argmin}_w ||w||_p$ s.t. $y = Xw$
- **Performance measure**: prediction error $\mathbb{E}_{x \sim N(0,I)} (\langle x, \hat{w} - w^\star \rangle)^2 = ||\hat{w} - w^\star||^2$

Recap for $p = 2$

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

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



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

(Slow) rate for $p = 1$

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.

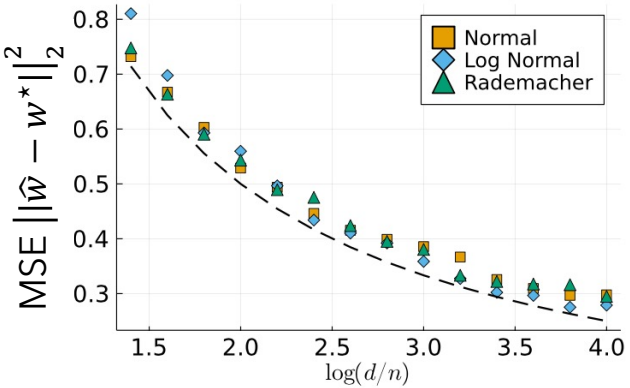
$$\left|\left|\widehat{w} - w^\star\right|\right|^2 = \frac{\sigma^2}{\log\left(d/n\right)} + O\left(\frac{\sigma^2}{\log^{3/2}\left(d/n\right)}\right)$$

(Slow) rate for $p = 1$

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.

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



- 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

*in [DRSY '22]

(Slow) rate for $p = 1$

Theorem [WD$\boldsymbol{Y}$**′ 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.

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

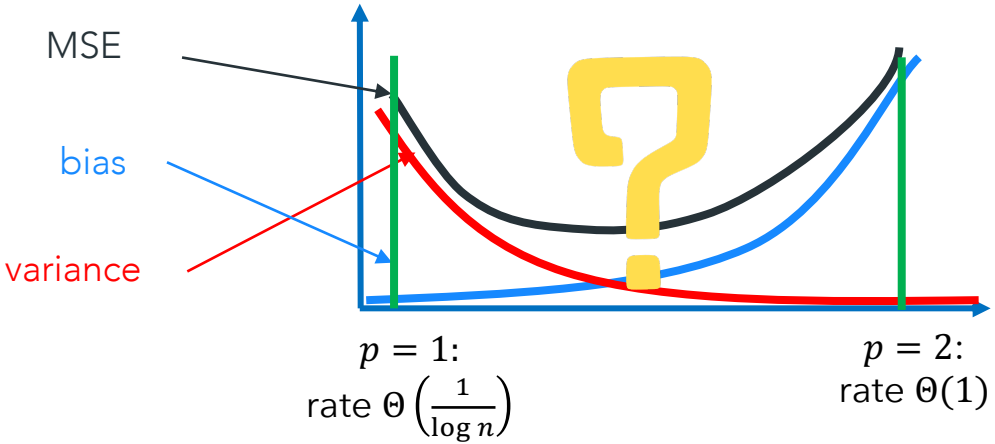
1. Second Descent after interpolation — ✓ Yes! Variance decreases, similar intuition as for $p = 2$

2. Harmless interpolation for large β — ✗ No! Interpolator $\Omega\left(\frac{1}{\log n}\right)$ vs. regularized $O\left(\frac{s \log n}{n}\right)$

Our work: Bias-variance trade-off via inductive bias

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

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



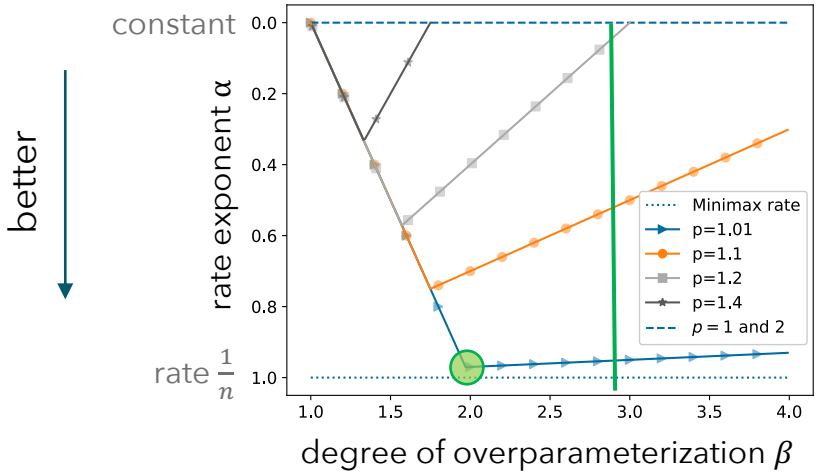
So far the extremes of very strong ($p = 1$) and no ($p = 2$) inductive bias perform badly

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



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

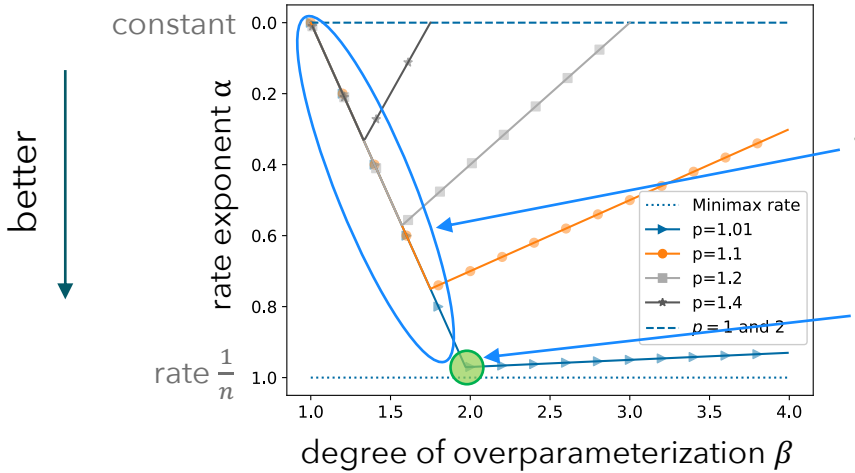
Note: technique applies to classification (see paper) and allows extension to $\Sigma \neq I$ and s-sparse $w^{\star}$

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 after interpolation

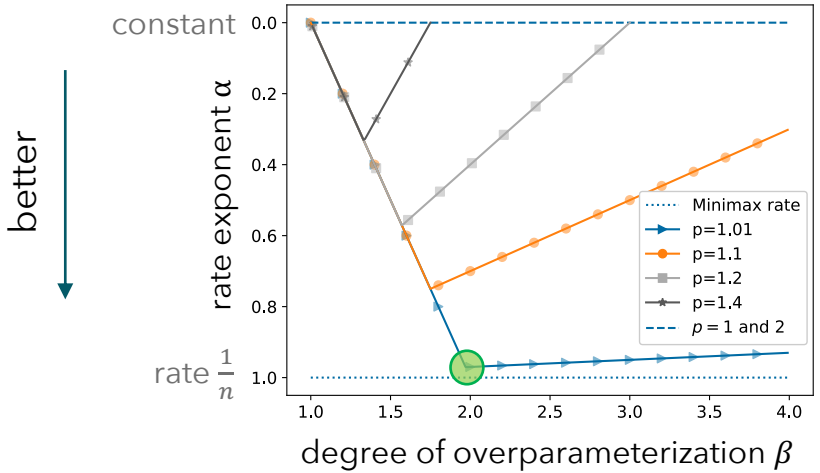
2 Harmless interpolation

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

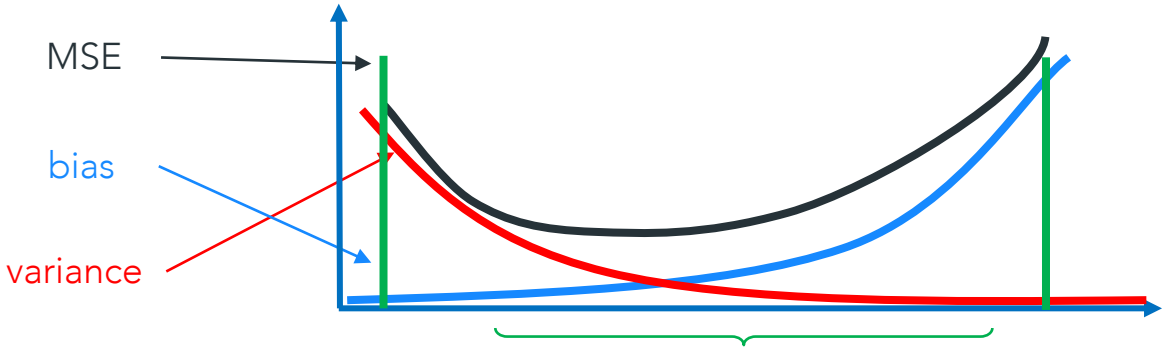
$$\frac{1}{\log\log d} \lesssim p - 1 \rightarrow \text{very large } \mathbf{d}$$

➡ cannot obtain best p for given β

Our work: Bias-variance trade-off via inductive bias

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

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



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

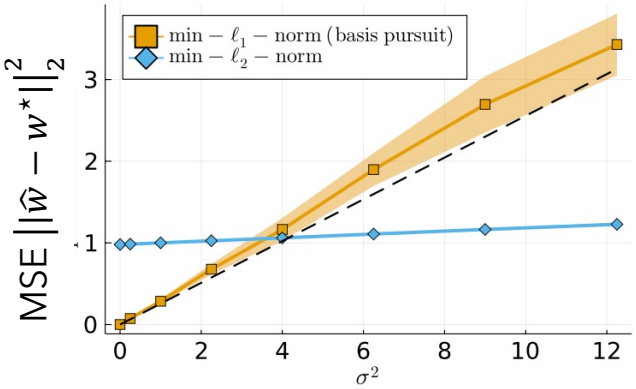
$1 < p < 2$:
$O\left(\frac{1}{n^\alpha}\right)$

$p = 2$:
rate $\Theta(1)$

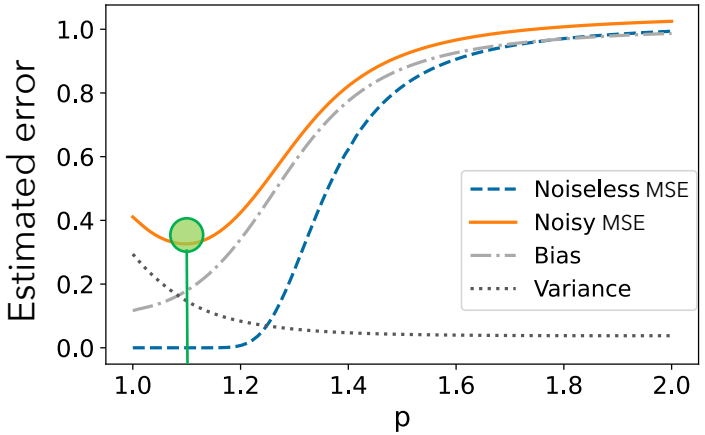
Bias-variance tradeoff for $p \in (1,2)$

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 $d = 20000, n = 400$

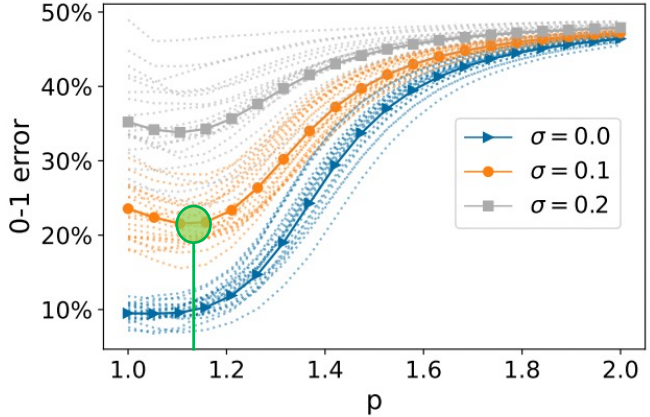


for $d = 5000, n = 100$

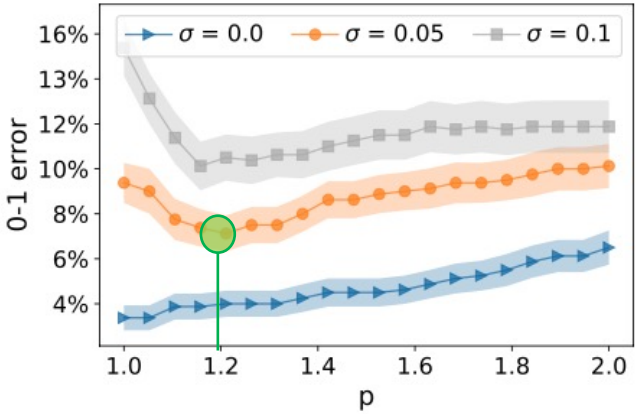
Trade-off between bias and variance for interpolators via strength of inductive bias!

Experimental results for classification

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 $d \sim 7000, n \sim 70$

The tale of two "new" bias-variance trade-offs

Previous intuition for interpolators:

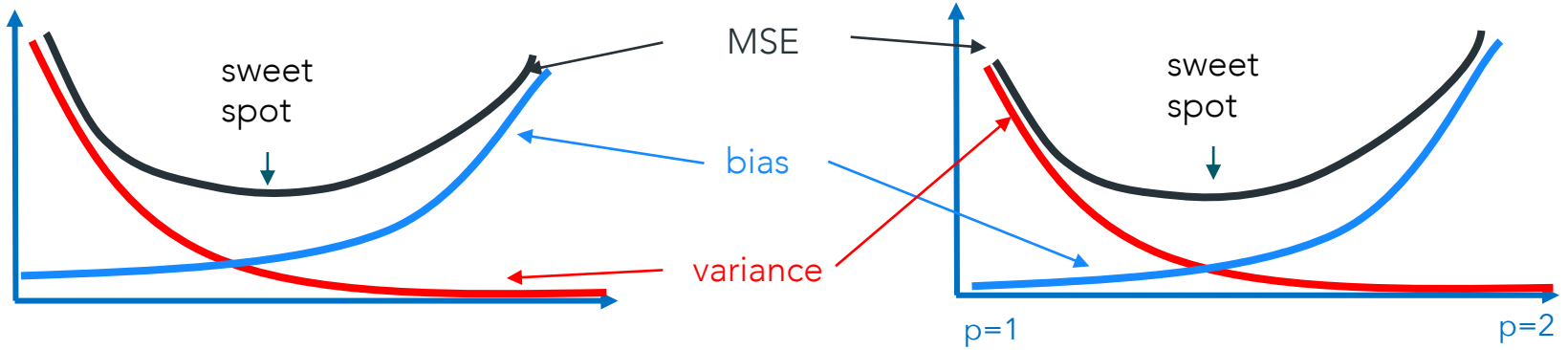
$$\hat{w} = \text{argmin}_w \, ||w||_p \; s.t. \; y = Xw$$

Bias-variance trade-off via overparameterization

Our new intuition for interpolators

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

Bias-variance tradeoff via inductive bias



decreasing effect of noise via increasing d_{eff}/n

decreasing "strength of inductive bias" via increasing p

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
- Donhauser, Ruggeri, Stojanovic, Yang *"Fast rates for noisy interpolation require rethinking the effects of inductive bias"*, ICML '22