Prospects and perils of interpolating models

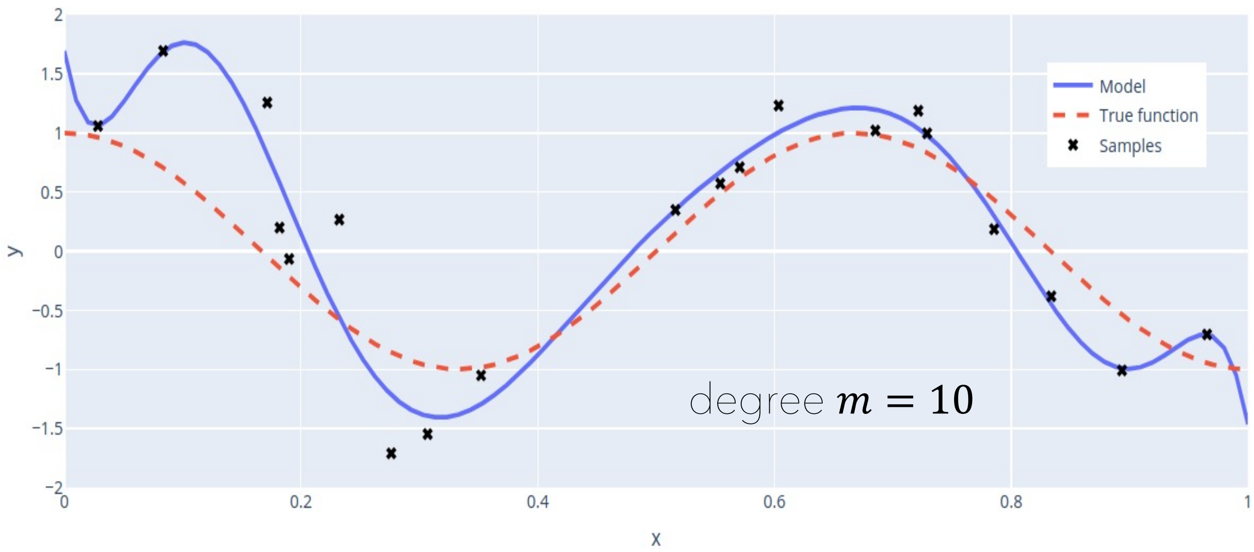
March 9th 2022, MSRI Workshop

Fanny Yang, Assistant Professor at CS department, ETH Zurich

Statistical Machine Learning group

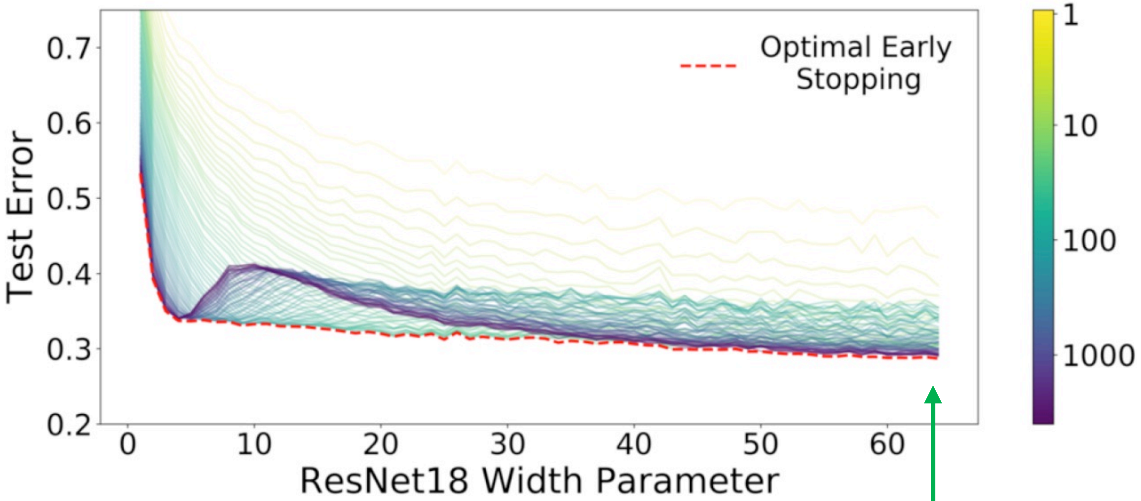
Regularization is good in low dimensions

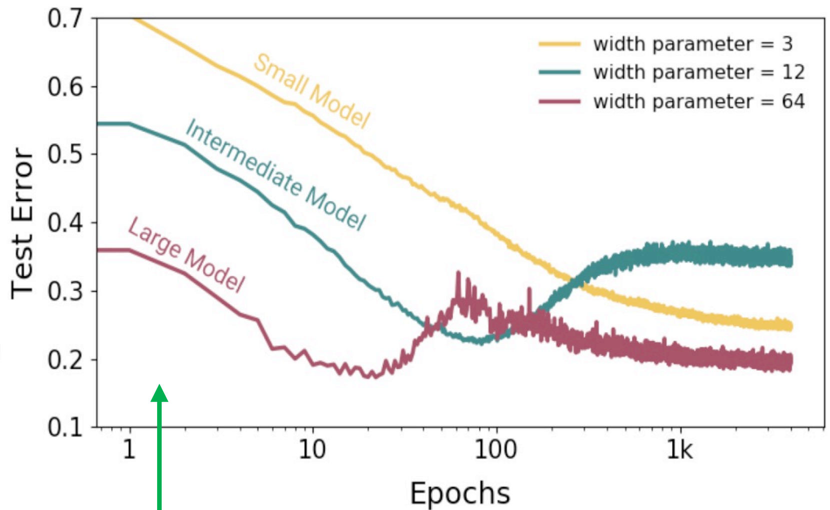
- Traditionally: want to avoid fitting noise perfectly for better (optimal) generalization.
- For example, here is the typical example used in my Intro to ML lecture



Provocation: Interpolation seems fine for deep learning

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

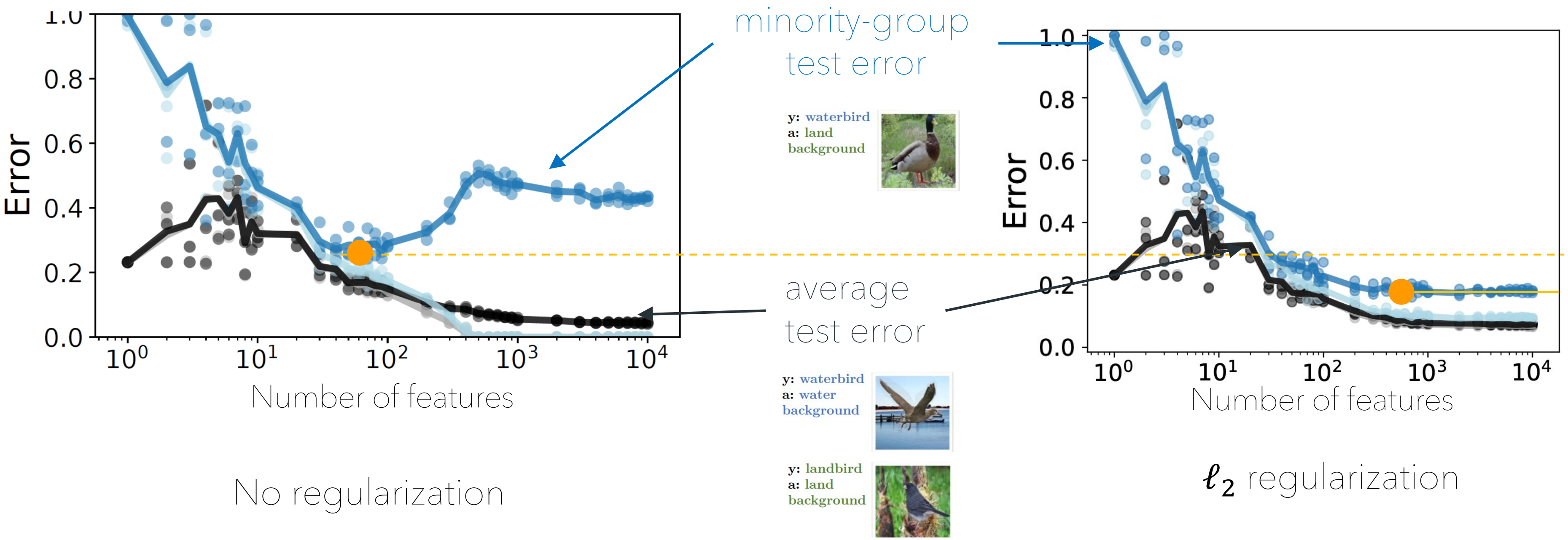




- The more parameters/width, the smaller the test error
- For large models, regularization does not decrease test error

But interpolation hurts worst-group accuracy

Training: First-order method on reweighted loss according to group size



For large models, regularization boosts worst-group accuracy!

This talk: formalizable intuition when interpolation may be a good idea (and when it might not)

Neural networks are hard . . .

Interpolators we discuss today

large models ≜ large $\frac{d}{n}$

- Function space: High-dimensional linear models $f(x) = w^\top x$ with $x, w \in R^d$ and $d \gg n$ samples

- Data model:
  for samples (x_i, y_i)

  Regression:
  $y_i = \langle w^\star, x_i \rangle + \xi_i$ with $x_i \sim N(0, I)$
  and noise $\xi_i \sim N(0, \sigma^2)$

  Classification:
  $y_i = \mathrm{sgn}\langle w^\star, x_i \rangle \xi_i$ with $x_i \sim N(0, I)$
  and noise $\xi_i = -1$ w.p. $\sigma\%$
  random label flips or logistic noise

- Interpolators:
  for $p \in [1,2]$

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

  max-ℓ_p-margin interpolator (hard-ℓ_p SVM)
  $$\hat{w} = \mathrm{argmin}_w \, ||w||_p \; s.t. \; y_i \langle x_i, w \rangle \geq 1 \; \forall i$$

these interpolators arise at convergence of first order methods on the square and logistic loss*

*implicit bias of GD e.g. [Telgarsky '13, Soudry et al. '18, Telgarsky, Ji '19], classification vs. regression e.g. [Muthukumar et al. '21]

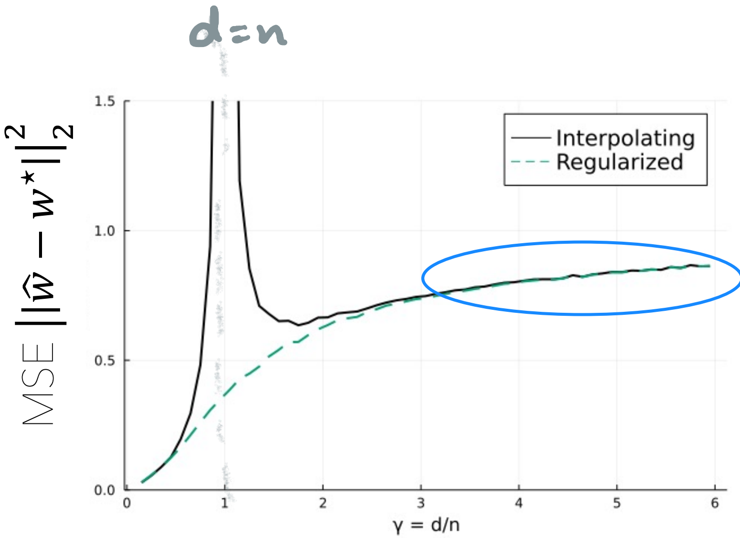
Overview of today on a high-level

- Prospects: How well can we do interpolation in the noisy case
  - previous work: high-dimensionality acts as "implicit regularizer" reducing variance at the cost of bias
  - our results: "moderate" inductive bias → fast rates for estimation error even for noisy interpolation
- Perils: Interpolation might be problematic for robustness
  - previous work: surprising empirical observations in adversarial robustness setting
  - our results: proof for some of these peculiar phenomena even in the linear and noiseless setting

Previous: Some established intuition for min-ℓ_2-norm interpolation

Implicit regularization: Variance decreases as $d/n \uparrow$

Increasing d/n is often said be "implicitly regularizing" because variance decreases (with $\frac{n}{d-n}$)



regularization
$\approx$
interpolation

pure noise fitting

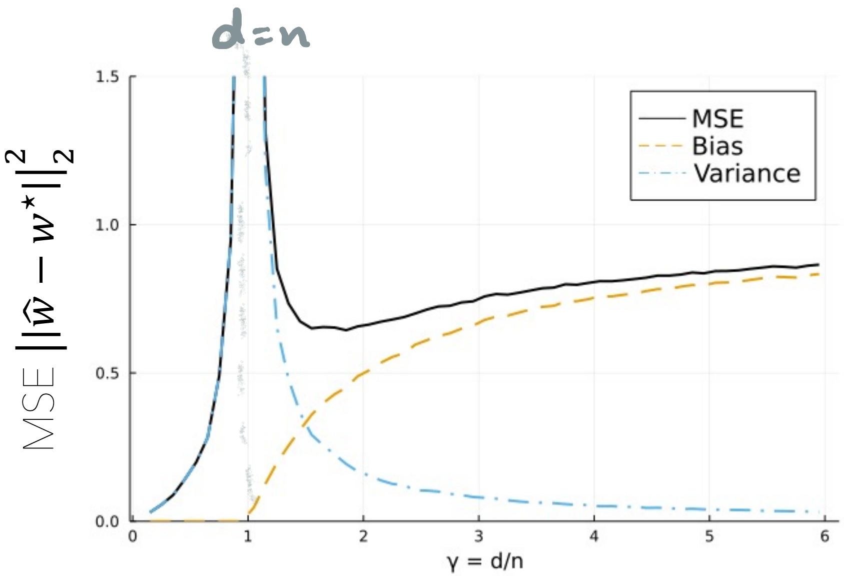
Simple intuition: Assume fixed n and $w^\star = 0$ such that min-norm solution $\hat{w} = \operatorname{argmin}_w ||w||_2 \; s.t. \boxed{Xw = \epsilon}$

→ The min-norm solution $\hat{w}_d$ for d, yields interpolating solution $(\hat{w}_d, 0)$ for $d + 1 \rightarrow ||\hat{w}_{d+1}||_2 \leq ||\hat{w}_d||_2$

Bias increases as $d/n \uparrow \Rightarrow$ "bad" trade-off

- On the other hand, bias has to increase with d/n as you have less information about your data.
- Back-of-the-envelope: in the noiseless case, $\widehat{w}$ is projection of $w^\star$ onto the n-dim span of rows(X)

→ If all directions are equally likely (isotropic $\Sigma = I$), on average it captures $\frac{n}{d}$ of $w^\star$ → $||\widehat{w} - w^\star||_2 \approx 1 - \frac{n}{d}$



→ as $\frac{d}{n}$ grows: Variance ↓, Bias ↑

→ MSE $\approx 1 - \frac{n}{d} + \frac{n}{d-n}$ gives you a "deadlock"

i.e. does not decrease with n

argument is e.g. in [Hastie et al. '18]; *for spiked covariances, prediction error can be consistent, see e.g. [Bartlett et al. 19']

Consistency or rates of prediction error?

- Obviously in high dimensions should assume structure to have any hope even for noiseless!

→ For the rest of the first half assume sparsity $||w^{\star}||_0 = s \ll d$. Well-known literature:

Basis pursuit (noiseless): $\mathbf{argmin}_w \; ||w||_1 \; s.t. \; y = Xw$

→ right inductive bias encouraging sparsity

Lasso (noisy): $\mathbf{argmin}_w \; ||y - Xw||_2^2 + \lambda ||w||_1$

→ right bias using explicit regularization $O\left(\frac{s \log d}{n}\right)$

Open questions:

- are **consistent or fast rates** possible for basis pursuit on noisy data for sparse $w^{\star}$?
- is the strongest inductive bias,. i.e. ℓ_1-norm, the best choice for noisy interpolation?

*So far: only non-vanishing prediction error bounds for isotropic, i.i.d. noise setting for min-ℓ_1-norm**

*[Wojtaszczyk '10, Chinot et al. '21, Koehler et al. '21]

Our results: Consistency and fast rates
for min-ℓ_p-norm/max-ℓ_p-margin interpolation

for $p \in [1,2)$

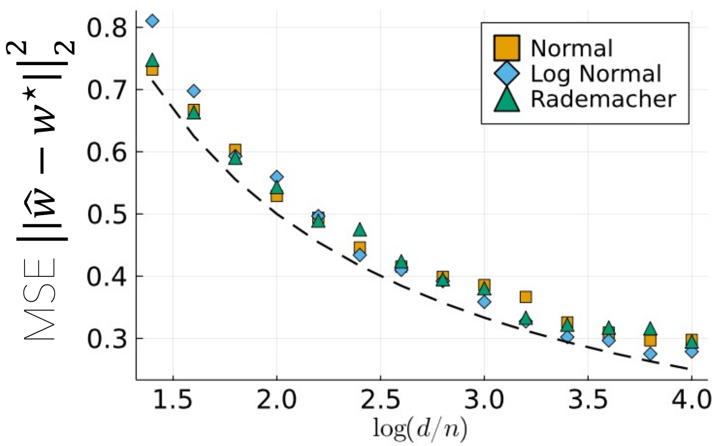
Consistency for noisy basis pursuit

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

For a sparse ground truth $||w^{\star}||_0 \leq \frac{n}{\log\left(\frac{d}{n}\right)}$, isotropic Gaussians, if $n \log n \lesssim d \lesssim e^n$

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

that is, as $n \to \infty$, the error vanishes (asymptotic consistency).



- This is a lower + upper bound for Gaussian X
  (experimentally bound also tight beyond Gaussian X)
- For classification, the directional estimation error

$$\left|\left|\frac{\hat{w}}{||\hat{w}||_2} - \frac{w^{\star}}{||w^{\star}||_2}\right|\right|_2^2 = O\left(\frac{\kappa(\sigma)}{\log d/n}\right)$$ when $w^{\star}$ is $\mathbf{1}$-sparse*

- Make no mistake: this is a slow rate! Lasso: $O\left(\frac{s \log d}{n}\right)$

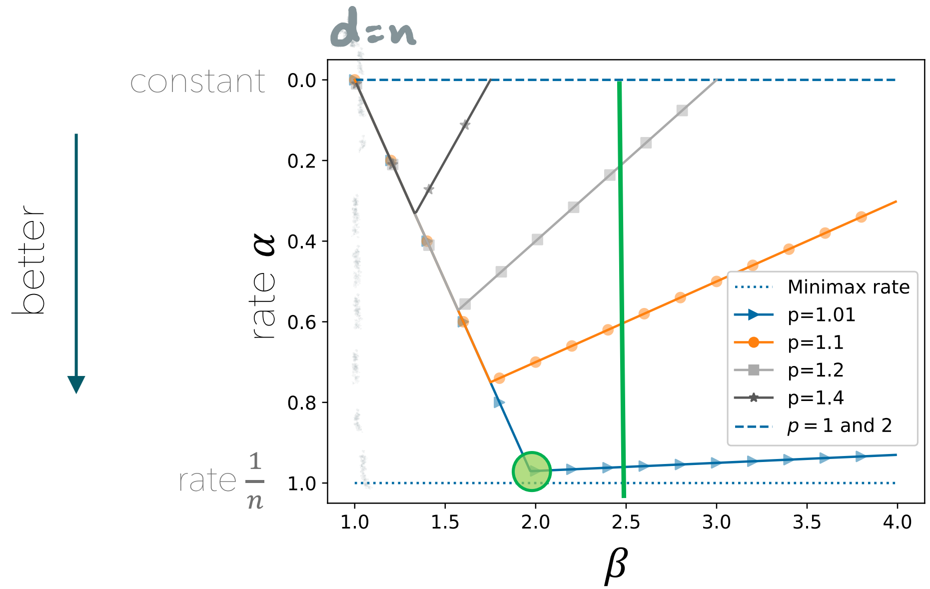
*in [DRSY '22]

Fast rates with modest inductive bias for regression

Theorem [DRSY' 22] – Tight bounds for min-ℓ_p-norm interpolators

For a 1-sparse ground truth $d \asymp n^{\beta}$ and isotropic Gaussians, for d large enough, $1 < p < 2$ and $1 < \beta \leq \frac{p/2}{p-1}$,

we obtain with probability at least $1 - d^{-c}$ prediction error rates $\tilde{O}(n^{-\alpha})$ with α as in graph below

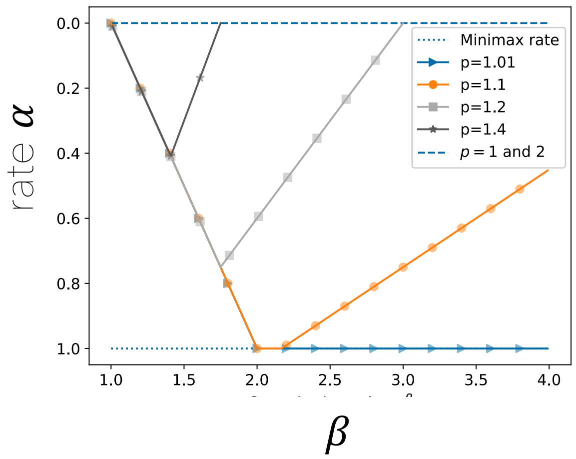


- for $\beta \approx 2$, we get rates close to $\frac{1}{n}$!
- for fixed β, some $p > 1$ close to 1 gets best rate
- Caveat: Large enough actually requires $\frac{1}{\log \log d} \lesssim p - 1 \to$ very large d

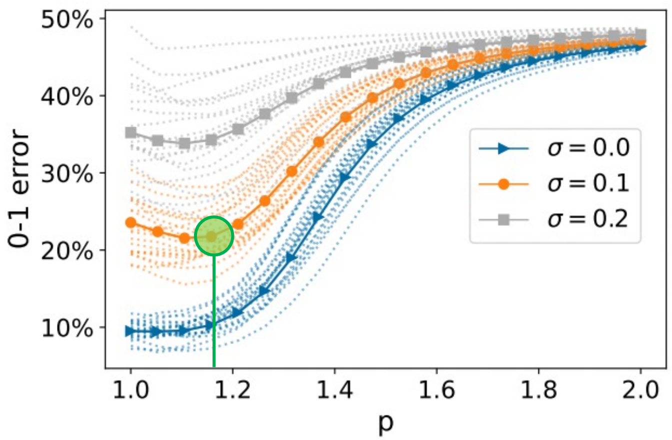
Fast rates with modest inductive bias for classification

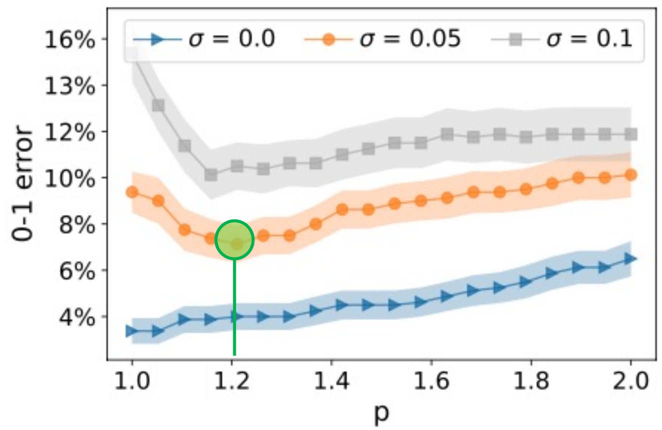
Theorem [DRSY' 22] – Upper bounds for max-ℓ_p-margin interpolators

For a 1-sparse ground truth $d \asymp n^{\beta}$ and $\Sigma = \mathrm{I}$, for d large enough and $1 < \beta \leq \frac{p/2}{p-1}$,
we obtain rates $\tilde{O}(n^{-\alpha})$ w/ probability at least $1 - d^{-c}$ for classification with α as in graph



Theoretical bounds





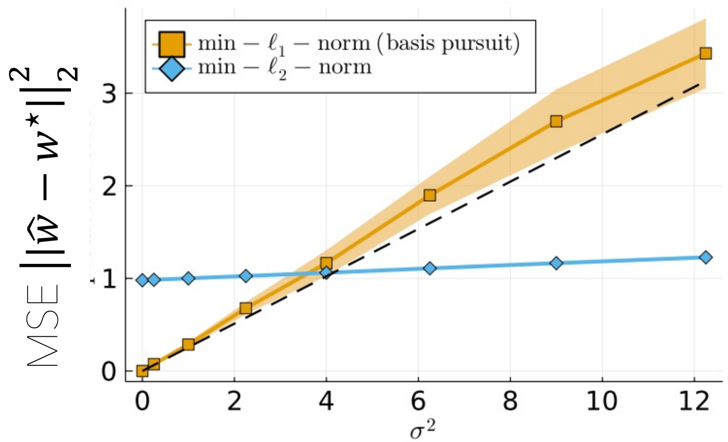
Experimental results: hard-ℓ_p-margin SVM for σ: proportion of label flips

(isotropic Gaussians for $d \sim 5000, n \sim 100$) (real leukemia dataset for $d \sim 7000, n \sim 70$)

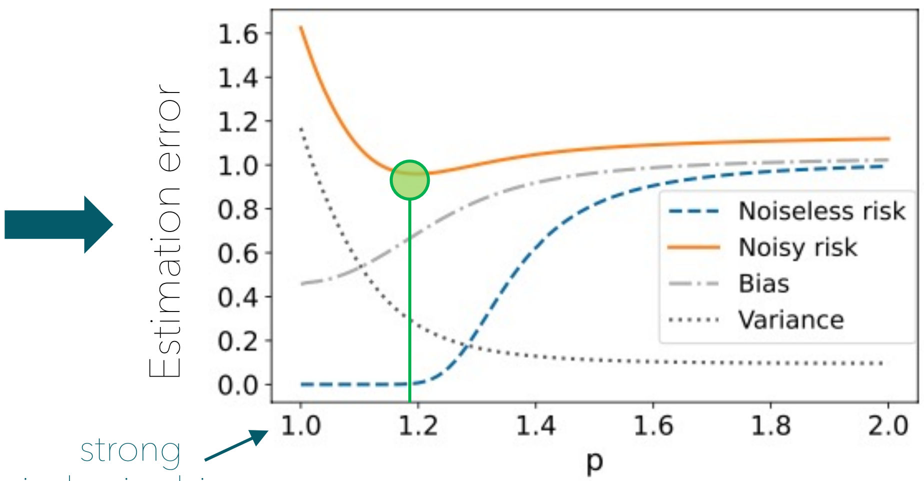
Intuition: a "new" bias-variance tradeoff

What's wrong with min-ℓ_1-interpolation? Variance and sensitivity to noise is too large

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



for $d = 20000, n = 400$



strong inductive bias

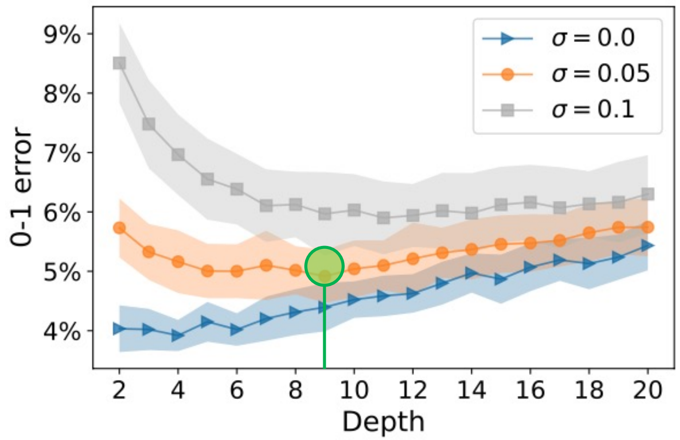
little inductive bias

New trade-off between bias and variance as a function of the strength of inductive bias!

Beyond linear models: Does this intuition transfer?

- Take-away intuition: in the presence of moderate noise, interpolation can do well
  if we use a moderate amount of inductive bias (if ground truth has "simple" structure)
- Back to images and neural networks: does this intuition transfer in any way?
  Question: what is a corresponding "strong" inductive bias? Filter size? depth? width?

Preliminary experiment with CNTK on binarized MNIST using depth:



For noisy (orange/grey) data,
best interpolating estimator has
"medium" inductive bias (depth)

…maybe? … still need much more evidence!

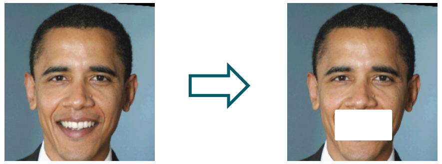
Overview of today on a high-level

- Prospects: How well can we do interpolation in the noisy case
  - previous work: high-dimensionality acts as "implicit regularizer" reducing variance at the cost of bias
  - our results: "moderate" inductive bias → fast rates for estimation error even for noisy interpolation
- Perils: Interpolation might be problematic for robustness
  - previous work: surprising empirical observations in adversarial robustness setting
  - our results: proof for some of these peculiar phenomena even in the linear and noiseless setting

Adversarial robustness primer

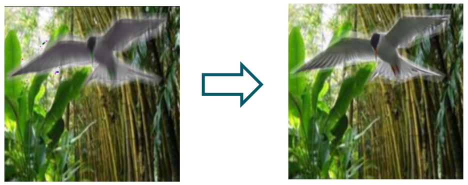
same label/value
for the ground truth

- usually consider consistent perturbations, that is for all $x' \in T(x, \epsilon)$, we have $f^\star(x') = f^\star(x)$



x $\quad$ x'

same person
despite mask



x $\quad$ x'

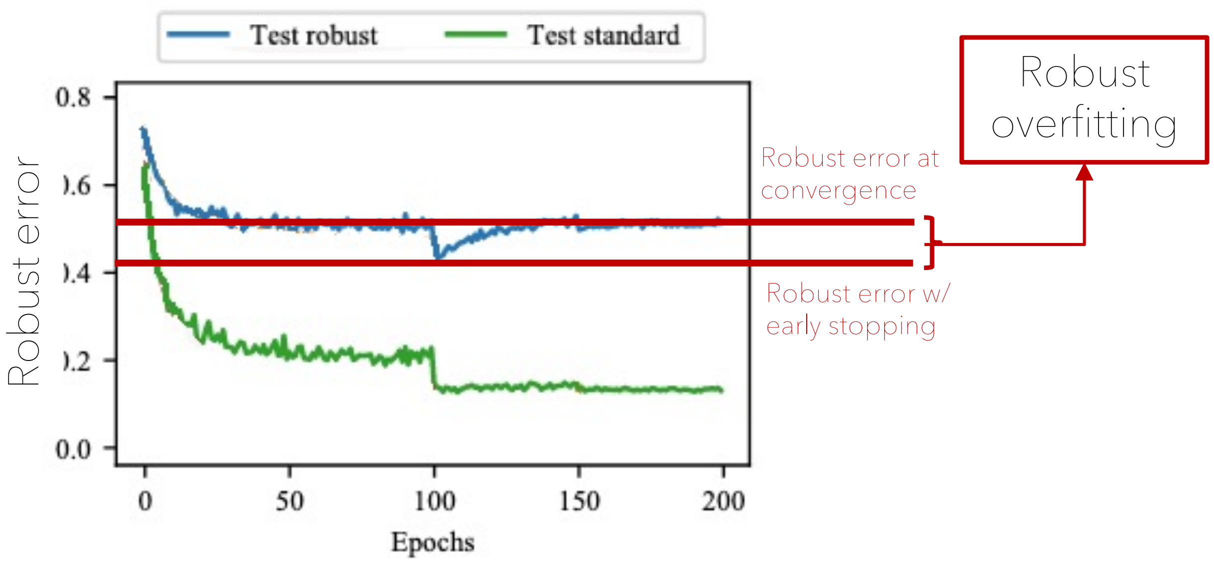
same bird
despite blur

- Goal is to achieve lower robust (test) error $\mathbb{E}_{x,y} \max_{x' \in T(x,\epsilon)} \ell(y, f(x'))$ than standard training
- Adversarial training (AT) minimizes empirical robust risk $\frac{1}{\mathrm{n}} \sum_{i=1}^{n} \max_{x' \in T(x,\epsilon)} L(y, f(x'))$, usually is better
- Interpolating AT: Usually using first-order method on empirical robust risk until convergence

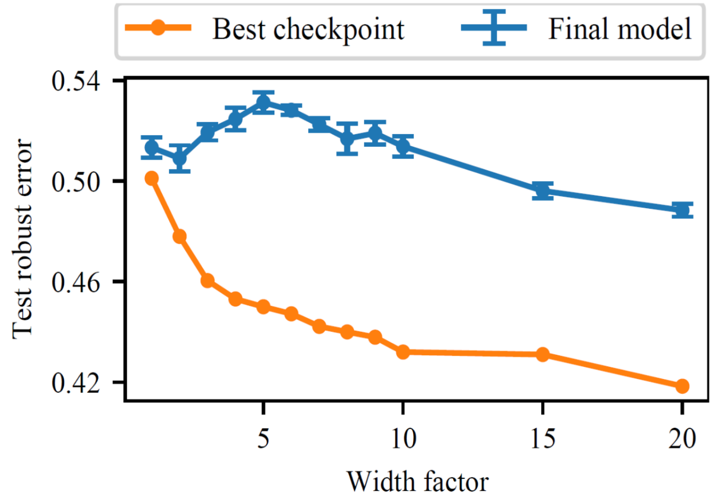
Next: some empirical phenomena that arise with interpolation and adversarial robustness

Interpolating AT yields worse robust risk – than regularized

Regularized adversarial training
(early stopping) yields lower robust risk



"Robust overfitting" persists
for large models!

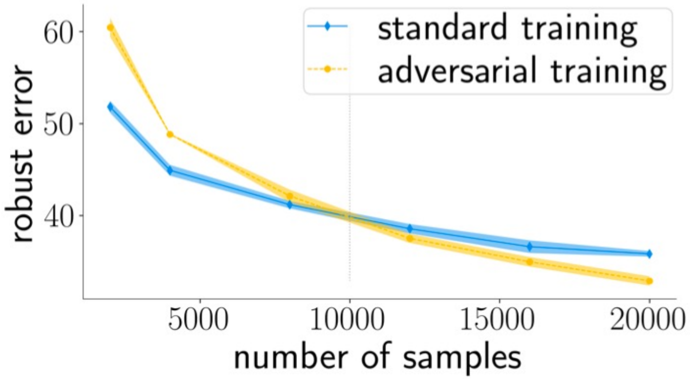


Interpolating AT yields worse robust risk – than standard

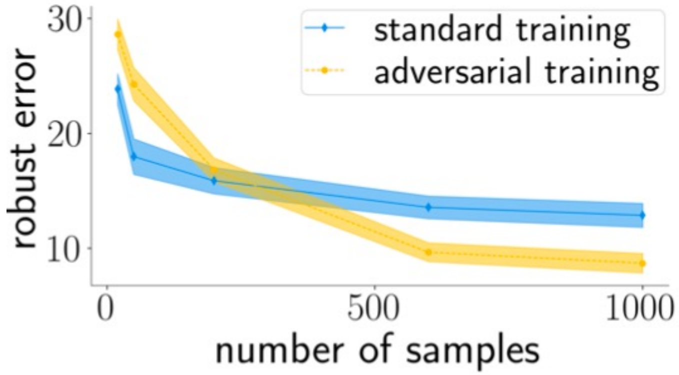
… in the small sample regime for perceptible attacks. Some image examples from [CHY '22]:

Mask

Mask attacks on CIFAR-10



Illumination attacks on Waterbirds



What's happening with robust error when we interpolate?

Many possible reasons for weirdness when training neural networks

Previous work: noise different impact? non-convex optimization? robust estimator complicated?

We find: Lots of weirdness even when noiseless & convex & simple (linear) robust ground truth

...theoretical results for linear models

Adversarial robustness for linear models

- We consider noiseless observations in classification $y = \mathrm{sgn}(\langle w^\star, x\rangle)$ or regression $y = \langle w^\star, x\rangle$
- Different consistent perturbations: $\mathrm{sgn}\langle w^\star, x'\rangle = \mathrm{sgn}\langle w^\star, x\rangle$ *or* $\langle w^\star, x'\rangle = \langle w^\star, x\rangle$ with $x' = x + \delta$

depending on $\boldsymbol{x}$ distribution requires $\boldsymbol{\delta} \perp \boldsymbol{w}^\star$ or just $||\delta||_p \leq \epsilon$

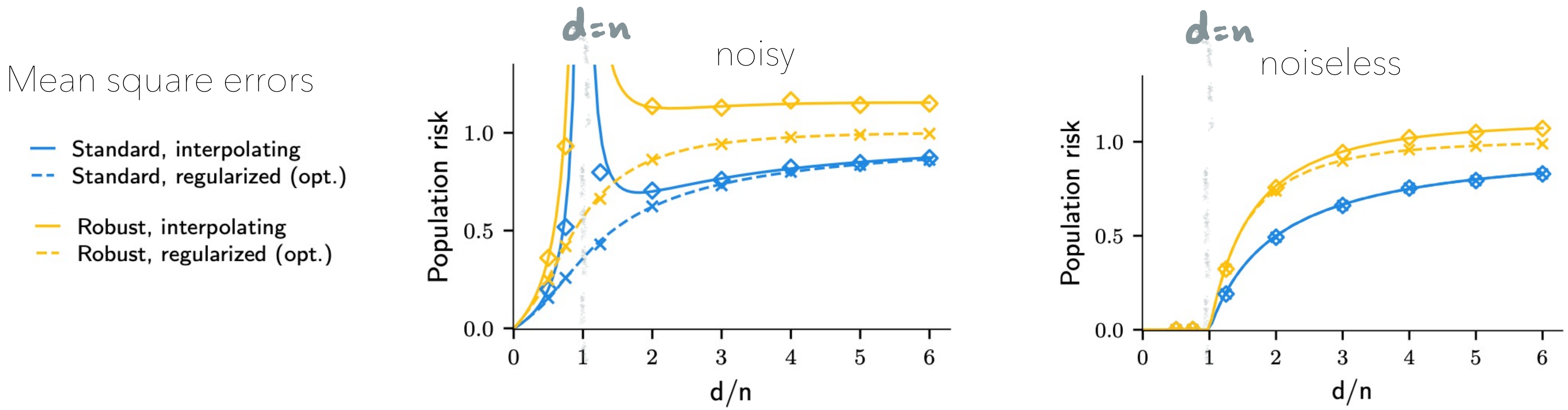
- Interpolating adversarial training (AT): (S)GD on $\frac{1}{\mathrm{n}}\sum_{i=1}^{n} \max_{\delta\in S(\epsilon)} L(y, w^\top(x+\delta))$
- (Ridge)-regularized adversarial training: minimum of $\frac{1}{\mathrm{n}}\sum_{i=1}^{n} \max_{\delta\in S(\epsilon)} L(y, w^\top(x+\delta)) + \lambda||w||_2^2$

Adversarial evaluation benefits from regularization

Robust error: $\mathbb{E}_{x,y} \max_{\delta \in S(\epsilon)} \ell(y, w^\top(x+\delta))$, standard error: $\mathbb{E}_{x,y} \ell(y, w^\top x)$, standard training

Theorem [DTAHY' 22] (informal) – Adversarial accuracy benefits from regularization

Consistent perturbations ($\delta \perp w^\star$) for regression ($\delta \perp w^\star$), $x \sim N(0, I)$: Asymptotically as $\frac{d}{n} \to \gamma$, the min-ℓ_2-norm interpolator has higher robust error than the regularized estimator but the same standard error

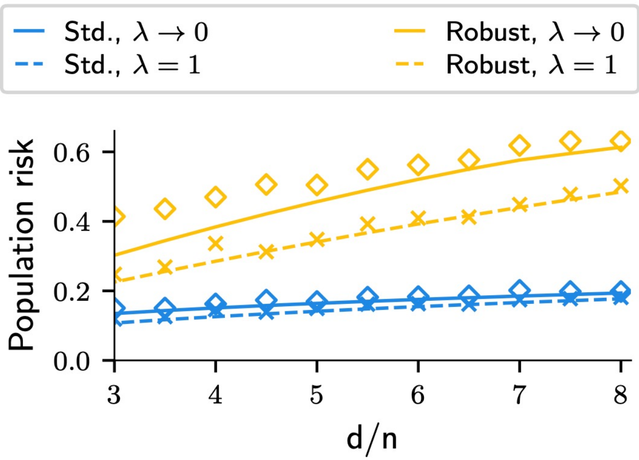


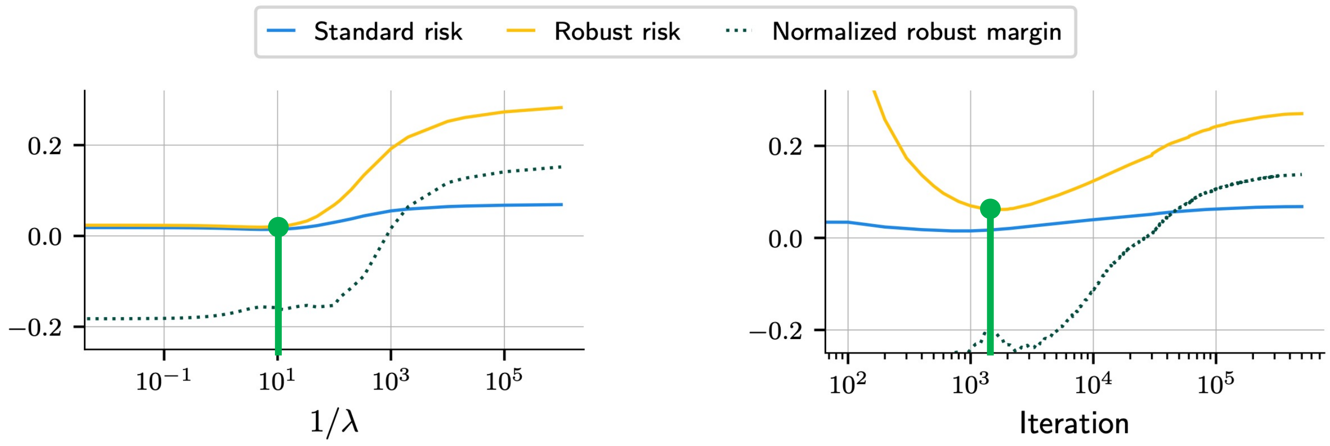
Adversarial training (AT) benefits from regularization

Theorem [DTAHY**' 22] (informal) – Proof for robust overfitting**

Consistent ℓ_∞-perturbations ($\delta \perp w^\star$) for classification w/ sparse ground truth, $x \sim N(0, I)$:

Asymptotically as $\frac{d}{n} \to \gamma$, interpolating AT yields higher robust error than regularized AT.





more regularization

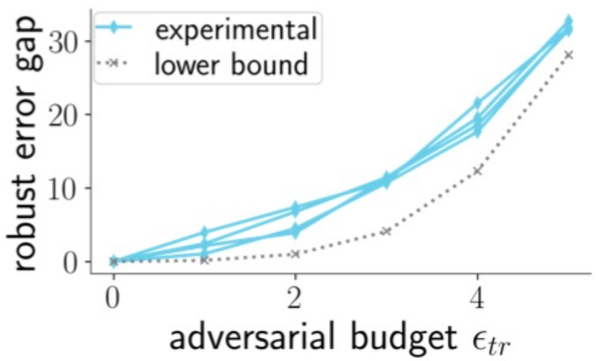
more regularization

Adversarial training worse than standard training

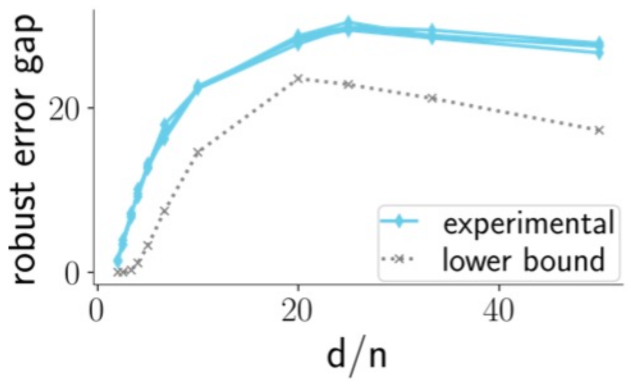
Robust error gap: Robust error (adversarial training) – Robust error (standard training)

Theorem [CHY' 22] (informal) – Non-asymptotic lower bounds for robust error gap

Consistent but directed attacks ($\delta \parallel w^\star$), Gaussian mixture: almost surely, interpolating adversarial training yields higher robust error than the interpolating standard training. More specifically we prove:



Almost surely, robust error gap monotonically increases with attack budget



Lower bound on the gap increases with $\frac{d}{n}$ until adversarial training ≈ random guessing

Take-aways

- interpolation can generalize almost as well as regularized estimators with right amount of inductive bias

  proof for min-ℓ_p-norm interpolation for $p \in [1,2]$ where $p = 1$ is strong, $p = 2$

- for robust evaluation, regularized estimators could generalize better than interpolating estimators

  even in the noiseless and consistent case

  - for standard training (proof for regression)
  - for adversarial training (proof for classification)

- for perceptible, directed attacks, even weirder things can happen for interpolating estimators:
  - adversarial training may be worse than standard training for small samples

Group and references

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", arxiv preprint '22
- Donhauser*, Tifrea*, Aerni, Heckel, Yang "Interpolation can hurt robust generalization even when there is no noise", NeurIPS '21
- Clarysse, Hörmann, Yang "Why adversarial training can hurt robust accuracy", arxiv preprint '22