Surprising phenomena of max-ℓ_p-margin classifiers in high dimensions

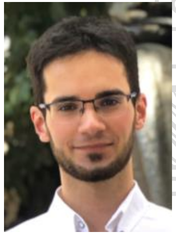
October 17th 2024, IPAM Workshop

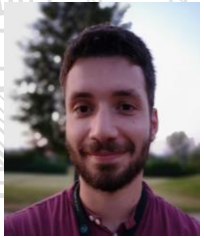
Fanny Yang, **K. Donhauser,** S. Stojanovic, N. Ruggeri

Statistical Machine Learning group, CS department, ETH Zurich

ETH zürich

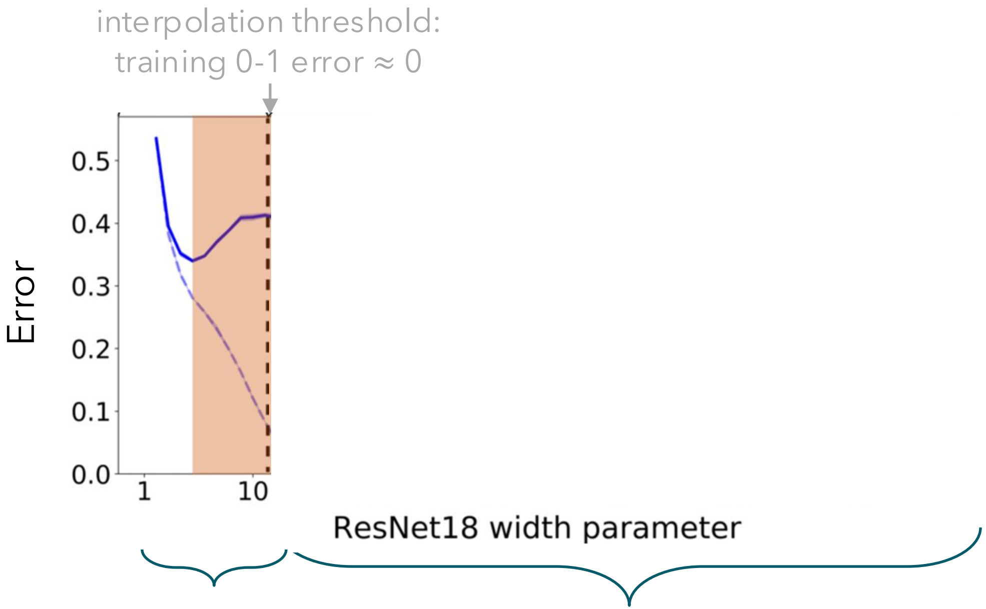
DINFK





Double descent on neural networks

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

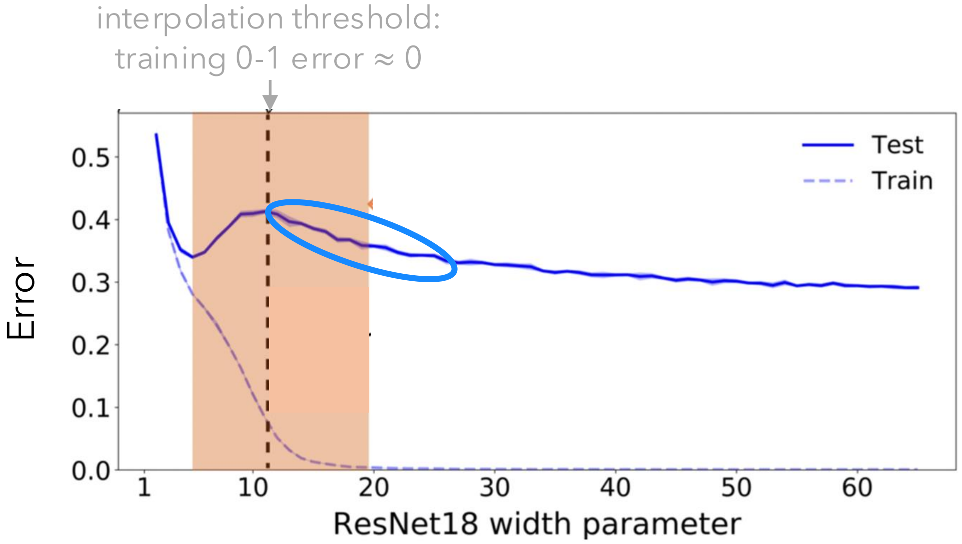


Classical bias-variance regime

Modern overparameterized regime

Interpolation and double descent on neural networks

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

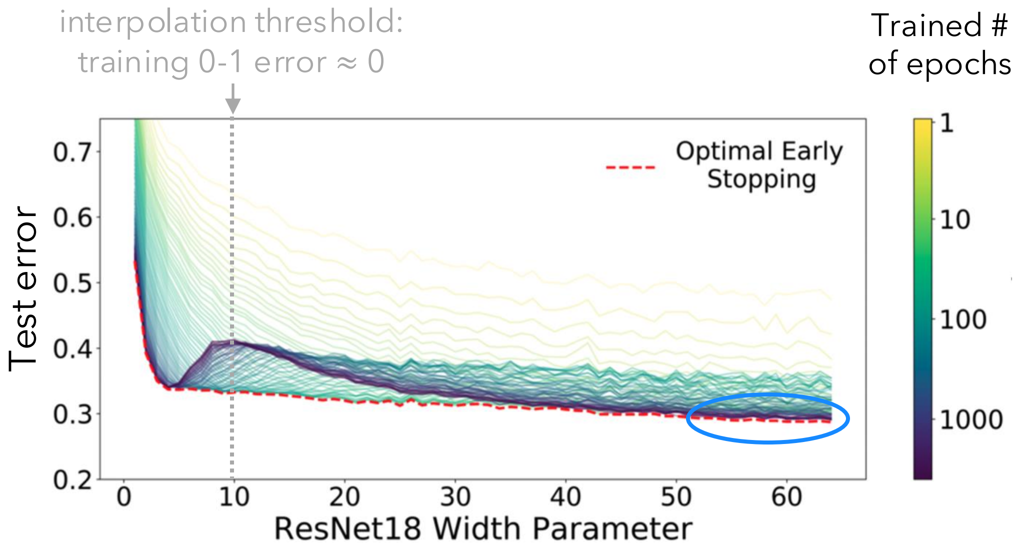


1 After "interpolation" threshold, we have a second "descent" (double descent) for "interpolators"

[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



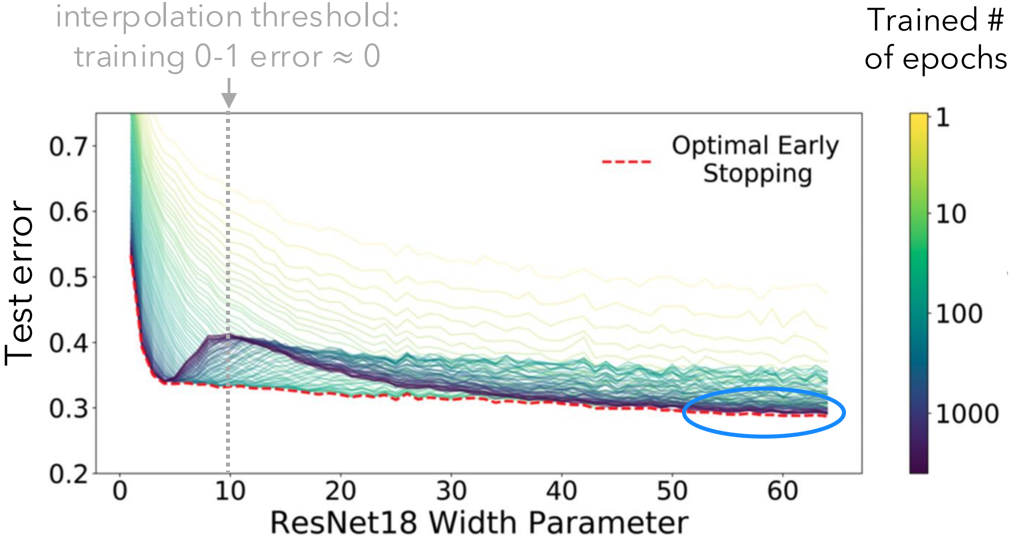
compare blue
(at convergence)
with red curve
(best stopping time)

2 For large models, training until "convergence" is not worse than stopping early

Good accuracy for non-early-stopped classifiers

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

interpolation threshold:
training 0-1 error ≈ 0



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

3 For large models, interpolating models achieve good test accuracy

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

Focus today, philosophically speaking…

Today

Deep Learning experiments

uncovers ⇩

Empirical phenomena (counterintuitive to classical ML theory)

inspires ➡

New statistical theory

uncovers ⇩

theoretical phenomena (too specific, not provable w/ "clean" classical techniques)

inspires ➡

Fundamentally new proof techniques

(beyond uniform convergence)

Goal is ***not to find*** *better interpolators in practice but trying* ***to understand when*** *interpolation is a good idea (out of intellectual curiosity)*

Plan ahead

- Double descent motivates new angle on underdetermined linear models
- Setting: Sparse linear classification
- We study today: Max-ℓ_p-margin linear classifiers
  - Q1: For noisy observations, what are (tight) rates?
  - Q2: For noiseless observations, is it adaptive to sparsity for $p = 1$?

Sparse linear classification

- Goal: Recovery of a sparse unit-norm $w^\star$ from $n \ll d$ measurements with with $\|w^*\|_0 = s \ll n$
- Measurements: via standard **Gaussian matrix** $X \sim N(0, I)$ with labels y that are noisy versions of $Xw^\star$
- Classification (1-bit compressed sensing): $y = \text{sign}(Xw^\star) \odot \xi$ with label noise $\xi_i \in \{-1, +1\}$
  - Noise model $\xi_i = -1 | x_i \sim \mathbb{P}_\sigma\ (.\,; \langle x_i, w^* \rangle)$ can only depend on x_i in the direction of w^*

    Examples: random label flips, logistic regression model
  - Performance measure $\left\| \frac{\hat{w}}{\|\hat{w}\|_2} - w^\star \right\|_2 \approx \pi\, \mathbb{E}_{x \sim N(0,I)} 1[\text{sgn}(\langle w^*, x \rangle) \neq \text{sgn}(\langle \hat{w}, x \rangle)]$

    for small error

Our focus today

Intermezzo: Classical intuition from sparse regression

Find $\hat{w}$ with small error $||\hat{w} - w^\star||_2$ from $y = Xw^\star + \xi$ by inducing sparsity

perfect data fit

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

sacrificing data fit

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

Classical work on sparse classification

Find $\hat{w}$ with small error $\left\|\frac{\hat{w}}{\|\hat{w}\|_2} - w^\star\right\|_2$ from $y = \text{sign}(Xw^\star) \odot \xi$ by inducing sparsity

perfect data fit

nonconvex!

Noiseless
$y = \text{sign}(Xw^\star)$

$$\text{argmin}_w \|w\|_0 \; s.t. \min_i y_i \langle w, x_i \rangle > 0, \|w\|_2 = 1$$

minimax rate $O\left(\frac{s \log d}{n}\right)$

[Jacques-Laska-Boufounos-Baraniuk-13, Matsumoto-Mazumdar-22 for BIHT]

when observations are noisy

Noisy
$y = \text{sign}(Xw^\star) \odot \xi$

$$\text{argmax}_w \; y^\top X w \; s.t. \|w\|_1 \leq \sqrt{s}, \|w\|_2 \leq 1$$

rate of order $O\left(\sqrt{\frac{s \log d}{n}}\right)$

[Plan-Vershynin-13]; [Zhang-Yi-Jin-14]

"sacrificing" data fit
(max-*average*-margin)

Classical work on sparse classification

$$\text{argmin}_w \|w\|_p \quad s.t. \min_i y_i\langle w, x_i\rangle \geq 1 \quad \text{for } p \in [1,2]$$

perfect data fit

convex relaxation!

$$\text{argmin}_w \|w\|_0 \; s.t. \min_i y_i\langle w, x_i\rangle > 0, \|w\|_2 = 1$$

Noiseless
$y = \text{sign}(Xw^\star)$

when observations are noisy

Noisy
$y = \text{sign}(Xw^\star) \odot \xi$

$$\text{argmax}_w \; y^\top X w \; s.t. \|w\|_1 \leq \sqrt{s}, \|w\|_2 \leq 1$$

"sacrificing" data fit

rate of order $O\left(\sqrt{\frac{s \log d}{n}}\right)$

[Plan-Vershynin-13]; [Zhang-Yi-Jin-14]

This talk: For minimizer of a **convex** problem w/ **perfect** fit:

Q1: How about perfect data fit of noisy data?
Q2: How about the performance on noiseless data?

Focus in this work: Maximum ℓ_p-margin classifiers

$$\text{argmin}_w \, ||w||_p \quad s.t. \min_i y_i \langle w, x_i \rangle \geq 1 \quad \text{for } p \in [1,2]$$

- **Natural motivation: Steepest descent** on logistic loss $w^{t+1} = w^t - \eta_t d^t$ with

$$d^t = \text{argmin}_v \langle \nabla L(w^t), v \rangle + \frac{1}{2} ||v||_p^2$$

converges to maximum ℓ_p-margin classifiers *[Telgarsky '13, Gunasekar-Lee-Soudry-Srebro '20]*

- For $p = 1$, can view it as a convex ℓ_1-relaxation of ℓ_0 −objective for perfect fit (optimal for noiseless)

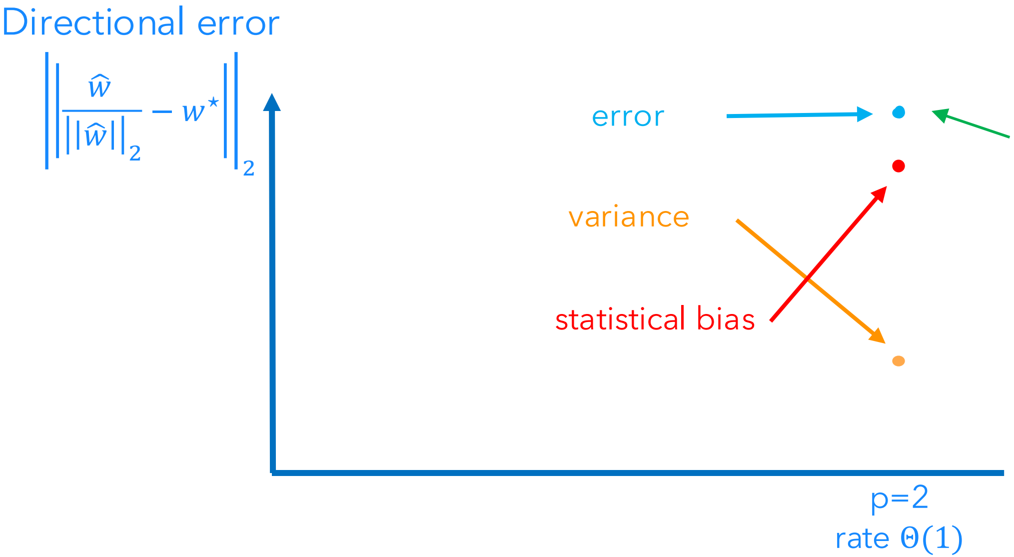
$$\text{argmin}_w \, ||w||_0 \; s.t. \; \min_i y_i \langle w, x_i \rangle > 0, ||w||_2 = 1 \rightarrow \; \text{argmin}_w \, ||w||_1 \;\; s.t. \; \min_i y_i \langle w, x_i \rangle \geq 1$$

Plan ahead

- Double descent motivates new angle on underdetermined linear models
- Setting: Sparse linear classification
- We study today: Max-ℓ_p-margin linear classifiers
  - Q1: For noisy observations, what are (tight) rates?
  - Q2: For noiseless observations, is it adaptive to sparsity for $p = 1$?

Noisy data: Previous work $p = 2$

Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p$ s.t. $\min_i y_i \langle w, x_i \rangle \geq 1$



Focus in previous works:

- consistency only when covariance is spiky
- inconsistent for isotropic Gaussians

[Muthukumar-Narang-Subramanian-Belkin-Hsu-Sahai '21]

$p = 1$ is consistent but still slow

Previous non-asymptotic bounds for the i.i.d. noise case:

$\Theta\left(\sqrt{\sigma^2/\log\left(\frac{d}{n}\right)}\right)$ tight bounds for min-ℓ_1-norm interpolator *for regression* [Wang-Donhauser-Y.'21]

vs.

$O(1)$ upper bounds [Chinot-Loeffler-Kuchelmeister-vandeGeer '22], [Wojtaszczyk '10] (for adversarial, vanishing noise)

Theorem [Stojanovic-Donhauser-Y' 24](simplified) – Tight bounds for max-ℓ_1-margin classifiers

Suppose $\|w^*\|_0 \lesssim \frac{n}{\log\left(\frac{d}{n}\right)^5}$. Assume $c_1 n \leq d \leq \exp(c_2 n^{1/5})$ for some constants c_1, c_2. Then

$$\left\|\frac{\hat{w}}{\|\hat{w}\|_2} - w^\star\right\|_2 = \frac{\kappa_\sigma}{\sqrt{\log(d/n)}} + O\left(\frac{1}{\log^{3/4}(d/n)}\right)$$

where κ_σ only depends on the label noise distribution $\mathbb{P}_\sigma$.

Plugging in $d \asymp n^\beta$ with $\beta > 1$ yields a rate of $\frac{1}{\sqrt{\log n}}$ 😬 other algorithms can achieve lower bound* $O\left(\frac{1}{\sqrt{n}}\right)$!

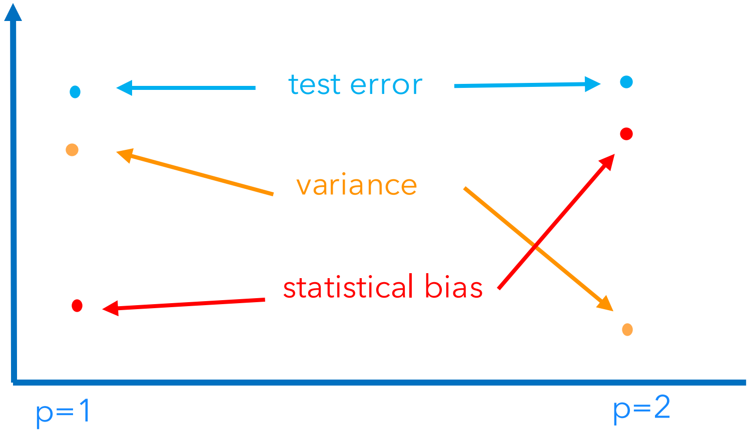
*[Abramovich-Grinshtein '18]

For regime $d \asymp n^\beta$ with $\beta > 1$

So far: Max-ℓ_p-margin classifiers poor for $p = 1, 2$

Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p$ s.t. $\min_i y_i \langle w, x_i \rangle \geq 1$

Directional error $\left\| \frac{\hat{w}}{\|\hat{w}\|_2} - w^\star \right\|_2$



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

rate $\Theta(1)$

bad generalization due to high variance!

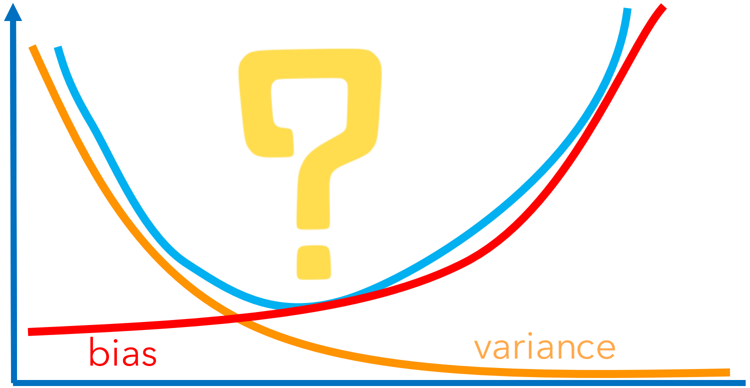
bad generalization due to high bias

For regime $d \asymp n^{\beta}$ with $\beta > 1$

A new bias-variance trade-off for interpolators

Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p \text{ s.t. } \min_i y_i \langle w, x_i \rangle \geq 1$

Directional error $\left\| \frac{\hat{w}}{\|\hat{w}\|_2} - w^\star \right\|_2$



bad generalization due to high variance!

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

Which rates?

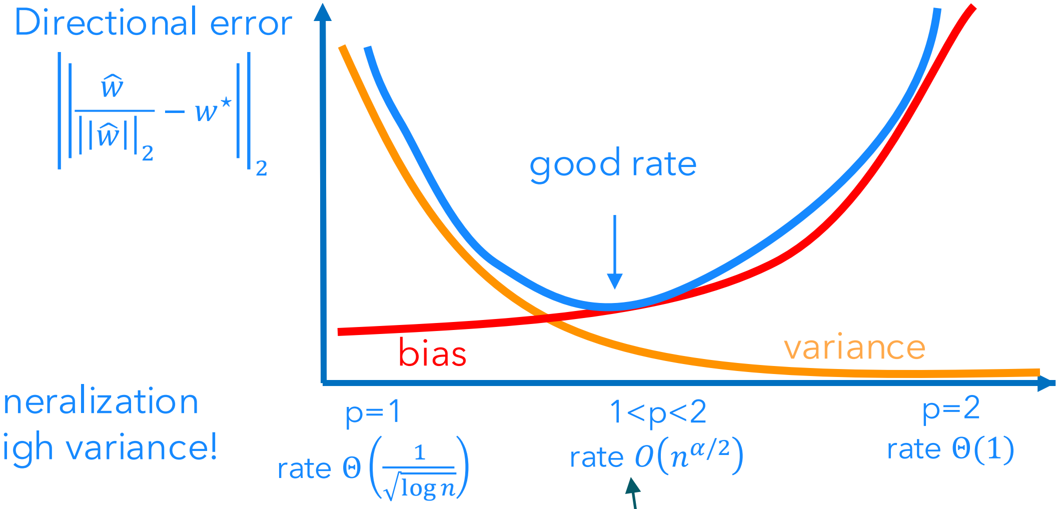
p=2 rate $\Theta(1)$

bad generalization due to high bias

For regime $d \asymp n^\beta$ with $\beta > 1$

A new bias-variance trade-off for interpolators

Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p$ s.t. $\min_i y_i \langle w, x_i \rangle \geq 1$



bad generalization due to high variance!

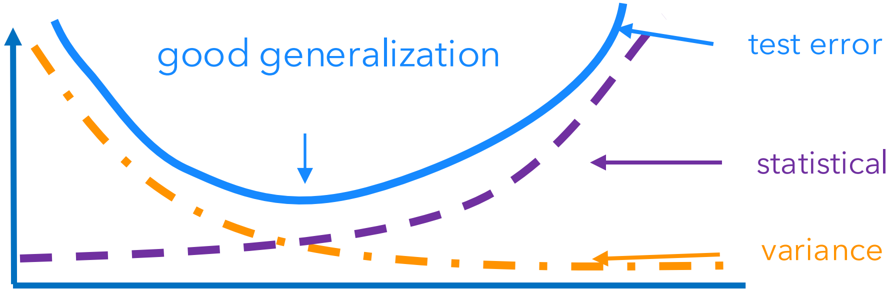
bad generalization due to high bias

for $\beta > 2$ and $1 + \frac{1}{\log\log d} < p < \frac{6\beta-2}{6\beta-3}$, we prove $\alpha = -1$, matching lower bound (up to log factors)

precise theorem in [Donhauser-Ruggeri-Stojanovic-**Y** '22]

A new bias-variance trade-off for interpolators

Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p$ s.t. $\min_i y_i \langle w, x_i \rangle \geq 1$



strong inductive bias towards sparsity

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

1<p<2
[1]rate $\Theta\left(n^{\alpha/2}\right)$
$-1 < \alpha < 0$

p=2
rate $\Theta(1)$

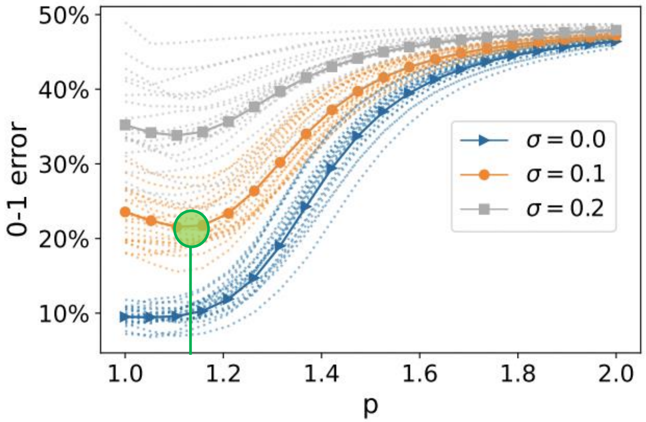
no inductive bias towards sparsity

High-level take-away[2]: whatever strongest inductive bias is best to interpolate noiseless data
medium strength of inductive bias is better when interpolating noise

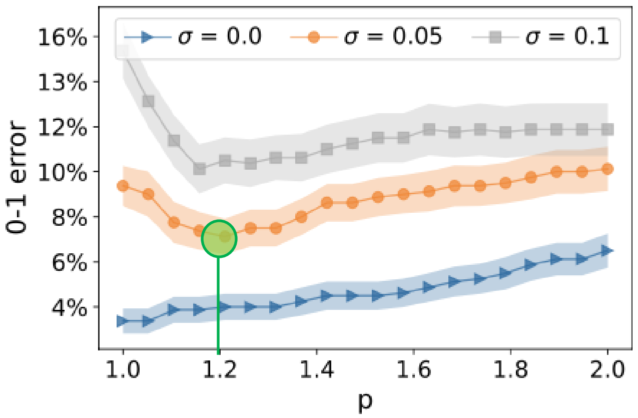
[1]same rates hold for regression; [2]same trend holds for kernel learning, experimentally on NN

Experimental results (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!

Plan ahead

- Double descent motivates new angle on underdetermined linear models
- Setting: Sparse linear classification
- We study today: Max-ℓ_p-margin linear classifiers
  - Q1: For noisy observations, what are (tight) rates?
  - Q2: For noiseless observations, is it adaptive to sparsity for $p = 1$?

Noiseless case: A fundamental question for ℓ_1-relaxations

For sparse regression $y = Xw^\star$

$$\text{argmin}_w \; ||w||_0 \; s.t. \; y = Xw$$

0-error for $n \sim s \log d$

convex relaxation

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

KNOWN:
0-error for $n \sim s \log d$

ℓ_1-relaxations behave like ℓ_0
(adaptive for hard-sparse $w^\star$)

For sparse classification $y = sign(Xw^\star)$

$$\text{argmax}_w \min_i y_i \langle w, x_i \rangle \; s.t. \; ||w||_0 \leq s, ||w||_2 \leq 1$$

error[1] $O\left(\frac{s \log d}{n}\right)$

convex relaxation

$$\underset{w}{\text{argmax}} \min_i y_i \langle w, x_i \rangle \quad s.t. \; ||w||_1 \leq 1$$

UNKNOWN

Open: Do ℓ_1-relaxations behave like ℓ_0 also for classification?

[1][Jacques-Laska-Boufounos-Baraniuk-13, Matsumoto-Mazumdar-22 for BIHT]

Surprisingly: No adaptivity to sparsity

[Chinot-Loeffler-Kuchelmeister-vandeGeer '22],[Wojtaszczyk '10] show an upper bound of order $\tilde{O}\left(\frac{\|w^*\|_1^2}{n}\right)^{1/3}$ **for any $w^\star$** and conjectured faster rate for sparse $w^\star$ should be possible

Theorem [Stojanovic-Donhauser-**Y**' 24] – **Noiseless classification (informal)**

Suppose $\|w^*\|_0 \lesssim n^{\frac{2}{3}} \log(d)^{-5}$. For any $n \geq \kappa_1$, and $\kappa_1 n^{\frac{2}{3}} \leq d \leq \exp(\kappa_3 n^{1/12})$, w.h.p.

$$\left\| \frac{\hat{w}}{\|\hat{w}\|_2} - w^\star \right\|_2 = c \left(\frac{\|w^*\|_1^2}{n \operatorname{polylog}(d/n)} \right)^{1/3} + O\left(\frac{\|w^*\|_1^2}{n \operatorname{polylog}(d/n)} \right)^{1/3}$$

For $y = sign(Xw^\star)$

$$\operatorname*{argmax}_w \min_i y_i \langle w, x_i \rangle \quad s.t. \|w\|_1 \leq 1$$

⇨ error $\tilde{\Theta}\left(\frac{\|w^*\|_1^2}{n}\right)^{1/3}$ even slower than what's possible for noisy data!

Conclusion in the **noiseless** case: A fundamental gap

For sparse regression $y = Xw^\star$

$\mathrm{argmin}_w \, ||w||_0 \; s.t. \; y = Xw$

0-error for $n \sim s \log d$

convex relaxation

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

KNOWN: 0-error for $n \sim s \log d$

ℓ_1-relaxations behave like ℓ_0 (adaptive for hard-sparse $w^\star$)

For sparse classification $y = sign(Xw^\star)$

$\mathrm{argmax}_w \min_i y_i \langle w, x_i \rangle \; s.t. \; ||w||_0 \leq s, ||w||_2 \leq 1$

error[1] $O\left(\frac{s \log d}{n}\right)$

convex relaxation

$\mathrm{argmax}_w \min_i y_i \langle w, x_i \rangle \quad s.t. ||w||_1 \leq 1$

OUR WORK error $\Theta\left(\frac{\|w^*\|_1^2}{n}\right)^{1/3}$

ℓ_1 relaxations worse than ℓ_0 and not adaptive to hard-sparse $w^\star$
(same dependence on n as for non-sparse $\boldsymbol{w}^\star$)

[1][Jacques-Laska-Boufounos-Baraniuk-13, Matsumoto-Mazumdar-22 for BIHT]

What's the intuition behind the "bad" ℓ_1-relaxation?

The ground truth has an order smaller margin than the max-l1-margin solution

- [Chinot-Loeffler-Kuchelmeister-vandeGeer-'22] prove $\max_{||w||_1 \leq 1} \min_i y_i \langle w, x_i \rangle \geq \Omega(n^{-\frac{1}{3}})$
- Take simple ground truth $w^\star = (1,0,0,\ldots,0)$ Then for our specific distribution

  w.h.p. $\min_i y_i \langle w^\star, x_i \rangle \leq O\left(n^{-\frac{1}{2}}\right)$
- Since $n^{-1/2} \ll n^{-1/3}$ the ground truth is not close to maximizing max margin

Our findings suggest many interesting open questions…

How to save the ℓ_1-relaxation for classification?

$$\operatorname{argmax}_w \min_i y_i\langle w, x_i\rangle \; s.t. \; ||w||_0 \le s, ||w||_2 \le 1$$ ⇨ $O\left(\frac{s \log d}{n}\right)$

convex relaxation ⇩

much better than

$O\left(\frac{\sqrt{s}}{n^{1/3}}\right)$ in noiseless case

$$\operatorname*{argmax}_w \min_i y_i\langle w, x_i\rangle \quad s.t. ||w||_1 \le 1$$

swap min-margin for average-margin ⇦

[Plan-Vershynin-13]; [Zhang-Yi-Jin-14]

$O\left(\sqrt{\frac{s \log d}{n}}\right)$ even in noisy case

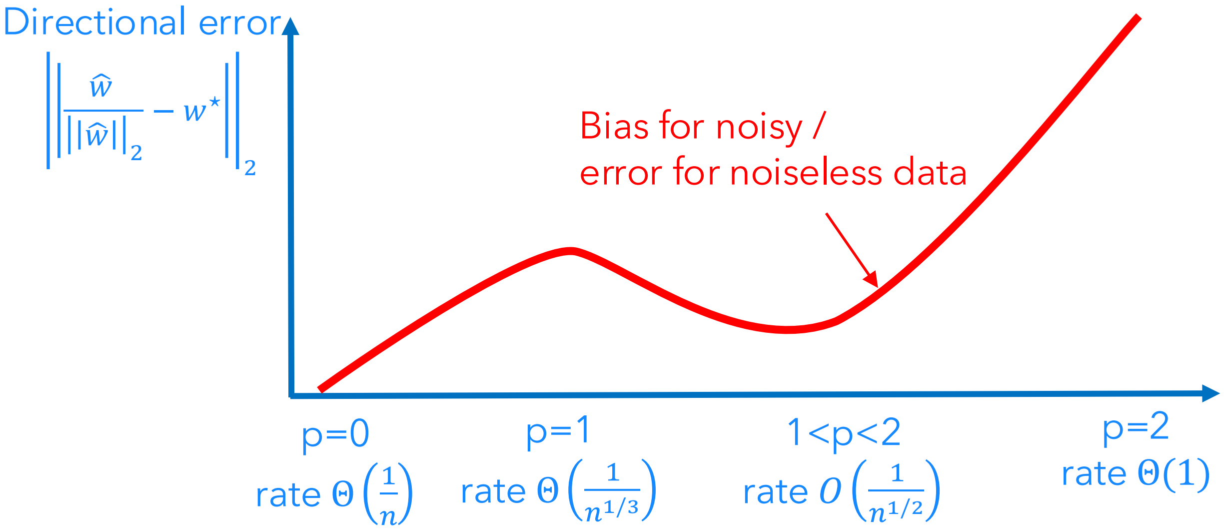
$$\operatorname{argmax}_w y^\top X w \; s.t. ||w||_1 \le \sqrt{s}, ||w||_2 \le 1$$

Open Q1: Does max-average-margin perform better for the noiseless case?

Open Q2: Could the max-average-margin solution be reached via steepest descent?

Understanding the landscape for all p

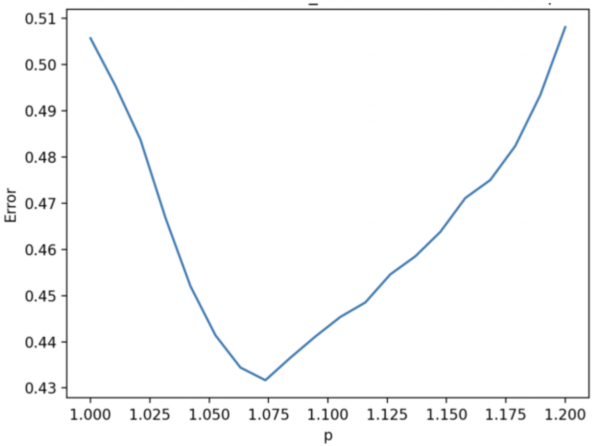
Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p$ s.t. $\min_i y_i \langle w, x_i \rangle \geq 1$



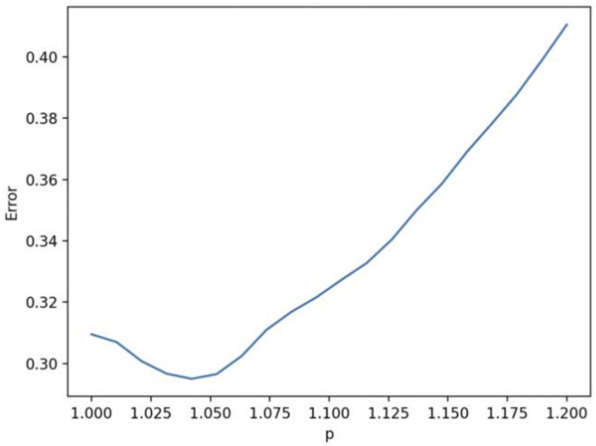
Open Q3: Why does $p \in (0,1)$ do better here?

Understanding the landscape for all p

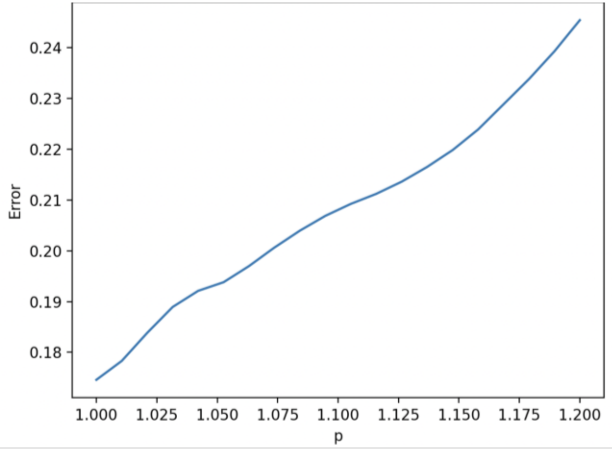
Max-ℓ_p-margin classifier $\hat{w} = \underset{w}{\operatorname{argmin}} \|w\|_p$ s.t. $\min_i y_i \langle w, x_i \rangle \geq 1$



n=50, d=5k



n=100, d=5k



n=400, d=5k

Open Q4: Why is $p > 1$ better than $p = 1$ in the noiseless case?

Papers discussed in the talk

SML group: sml.inf.ethz.ch

Results discussed in the talk:

- Donhauser, Ruggeri, Stojanovic, Yang *“Fast rates for noisy interpolation require rethinking the effects of inductive bias”*, ICML ’22
- Stojanovic, Donhauser, Yang “*Tight bounds for maximum $\ell 1$-margin classifiers*”, ALT ’24

Kernel results and neural network experiments:

- Aerni*, Milanta*, Donhauser, Yang *“Strong inductive biases provably prevent harmless interpolation”*, ICLR ’23