Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent

Yuting Wei

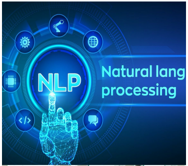
Statistics & Data Science, Wharton
University of Pennsylvania

Harvard Probabilitas Seminar, 2022

Yue Li, CMU Statistics

"Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent," Y. Li, Y. Wei, arxiv.2110.09502, 2021

Successes of deep neural networks



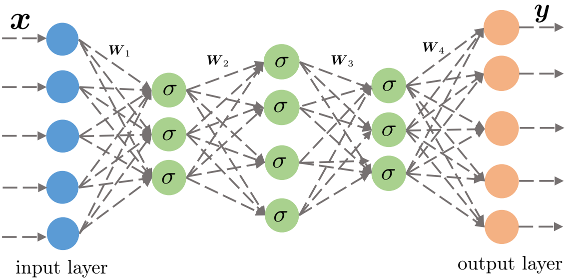


Figure: training deep neural networks (DNN)

$$f(x;\theta) = \sigma(W_L \cdot \sigma(W_{L-1} \cdots \sigma(W_1 \cdot x)))$$

Training deep neural networks (DNN)

- implicit algorithmic benefit by stochastic gradient methods
- training data is of enormous size (in # samples and # dimensions)
- networks are greatly overparametrized (large depth and width)
- networks are trained beyond zero training error

Empirical evidence: Larger models are better

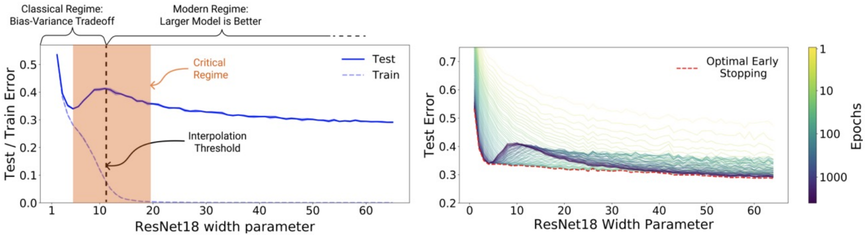


Figure: Nakkiran et al. 2019

See also: Opper (1995, 2001), Neyshabur et al. (2014), Canziani et al. (2016), Advani and Saxe (2017), Spigler et al. (2018), Novak et al. (2018), Geiger et al. (2019), ...

Empirical evidence: Larger models are better

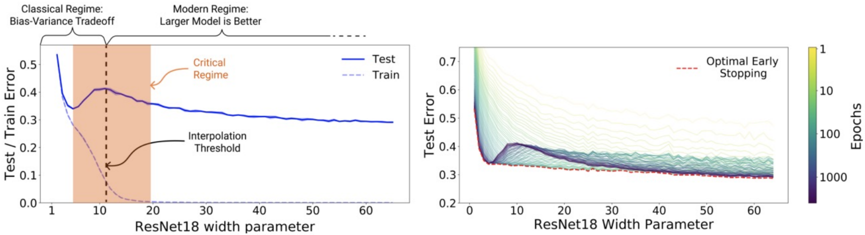
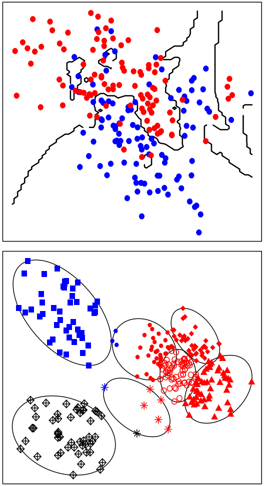


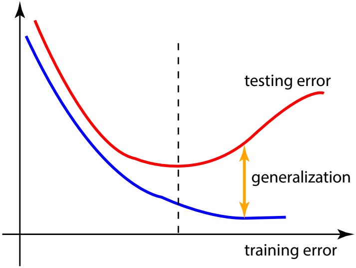
Figure: Nakkiran et al. 2019

See also: Opper (1995, 2001), Neyshabur et al. (2014), Canziani et al. (2016), Advani and Saxe (2017), Spigler et al. (2018), Novak et al. (2018), Geiger et al. (2019), ...

Question: how do these networks manage to generalize?

Classical bias-variance trade-off





"*The elements of statistical learning*" by Hastie, Tibshirani, Friedman

Reconcile bias–variance trade-off

— a curious *double-descent* phenomenon

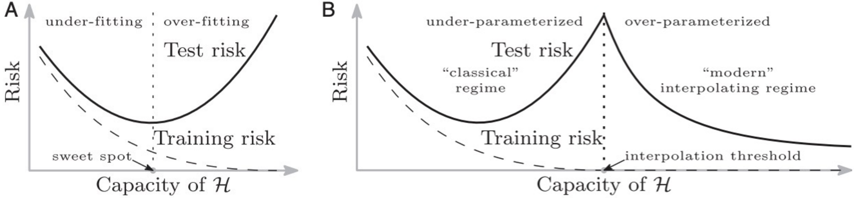


Figure: Belkin, Hsu, Ma, Mandal (2019)

Reconcile bias–variance trade-off

— a curious *double-descent* phenomenon

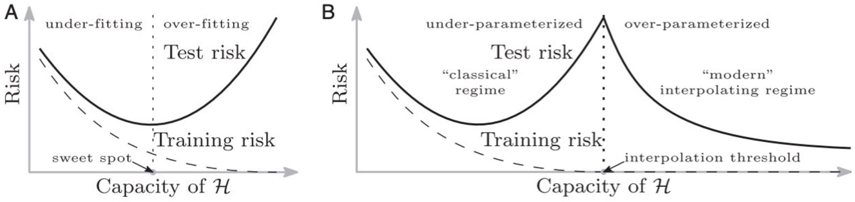


Figure: Belkin, Hsu, Ma, Mandal (2019)

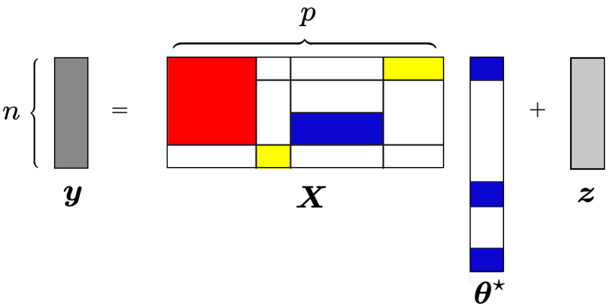
It motivates us to study classical estimators in the modern interpolating regime when interpolation happens!

So far, theoretical understandings are limited...

Limited theoretical understanding

Minimum ℓ_2-norm interpolators

$$\widehat{\boldsymbol{\theta}}^{\ell_2} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2 \right\} \qquad (\lambda \to 0,\ n = \Omega(p))$$



Belkin et al. (2019), Hastie et al. (2019), Mei and Montanari (2019), Muthukumar et al. (2020), Liang and Rakhlin (2020), Belkin et al. (2020), Bartlett et al. (2020, 2021), ...

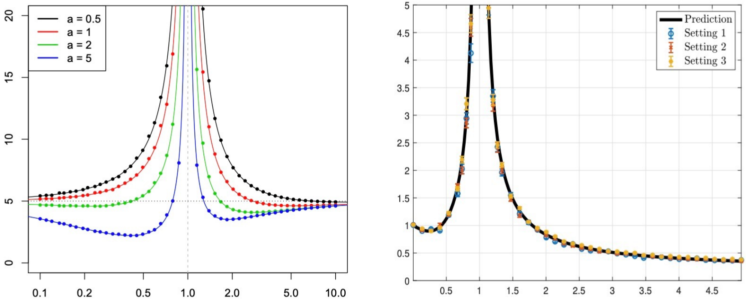


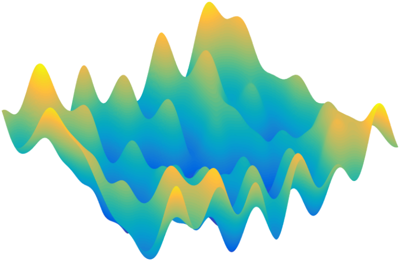
Figure: (left) ridgeless regression for misspecified model Hastie, Montanari, Rosset, Tibshirani (2019), (right) random features regression with ReLU activation Mei and Montanari (2019)

— *resemble the lazy training regime of 2-layer neural nets*

Question: how about other interpolators?

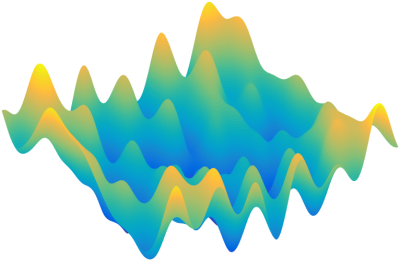
for example: $\widehat{\boldsymbol{\theta}}^{\ell_q} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_q^q \quad (\lambda\to 0, n=\Omega(p))$

Min ℓ_1-norm solutions



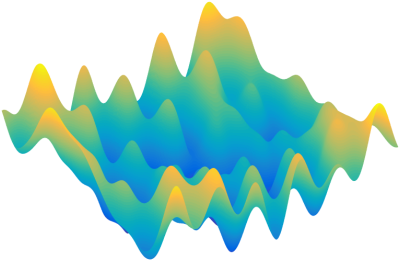
- ℓ_1 penalty encourages sparse solution (for interpretability)

Min ℓ_1-norm solutions



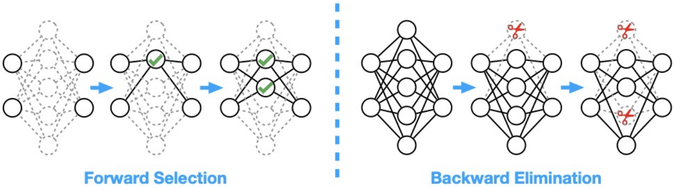
- ℓ_1 penalty encourages sparse solution (for interpretability)
- AdaBoost converges to min ℓ_1-norm solution for linear separable data Rosset, Zhu, Hastie (2004), Zhang and Yu (2005)

Min ℓ_1-norm solutions



- ℓ_1 penalty encourages sparse solution (for interpretability)
- AdaBoost converges to min ℓ_1-norm solution for linear separable data Rosset, Zhu, Hastie (2004), Zhang and Yu (2005)
- Gradient descent on full matrix factorization converges to min nuclear norm solution Gunasekar et al. (2017), Li et al. (2019)

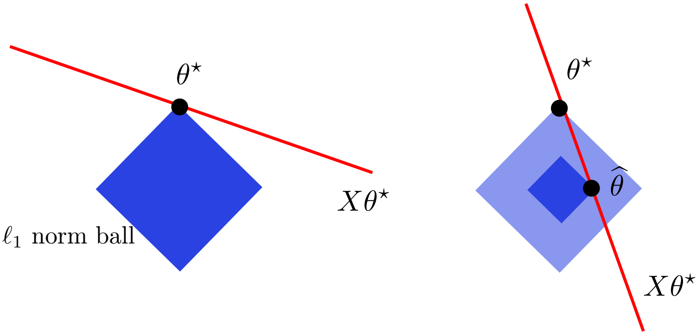
Min ℓ_1-norm solutions



- ℓ_1 penalty encourages sparse solution (for interpretability)
- AdaBoost converges to min ℓ_1-norm solution for linear separable data Rosset, Zhu, Hastie (2004), Zhang and Yu (2005)
- Gradient descent on full matrix factorization converges to min nuclear norm solution Gunasekar et al. (2017), Li et al. (2019)
- empirical successes of dropouts/model-pruning in DL Srivastava, Hinton, et al. (2014), Ye et al. (2020)

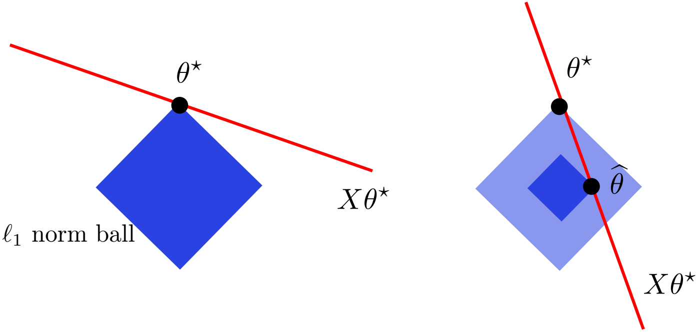
Basis Pursuit for noiseless observations

$$\widehat{\boldsymbol{\theta}}^{\ell_1} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{such that} \quad y_i = \langle \boldsymbol{x}_i,\, \boldsymbol{\theta}\rangle \quad 1 \le i \le n.$$



Basis Pursuit for noiseless observations

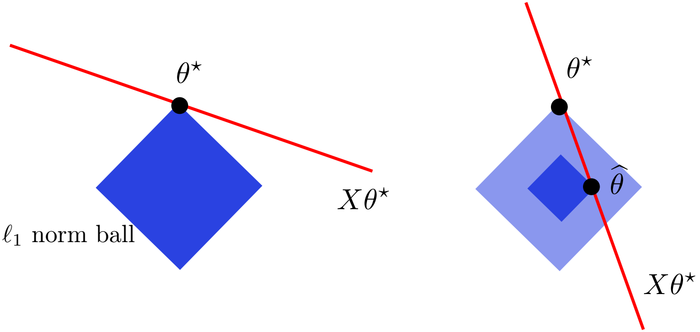
$$\widehat{\boldsymbol{\theta}}^{\ell_1} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{such that} \quad y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle \quad 1 \le i \le n.$$



Chen et al. 2001, Wojtaszczyk, Candes and Tao 2006, Donoho 2006, Donoho et al. 2005, Donoho and Tanner 2009, Amelunxen et al. 2014, Ju et al. 2020, Chinot et al. 2020, Wang et al. 2021, ...

Basis Pursuit for noiseless observations

$$\widehat{\boldsymbol{\theta}}^{\ell_1} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{such that} \quad y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle \quad 1 \leq i \leq n.$$



In the noisy and over-parametrized case ($p > n$), how does *generalization error* of min ℓ_1 solution depend on p/n?

A multi-descent phenomenon

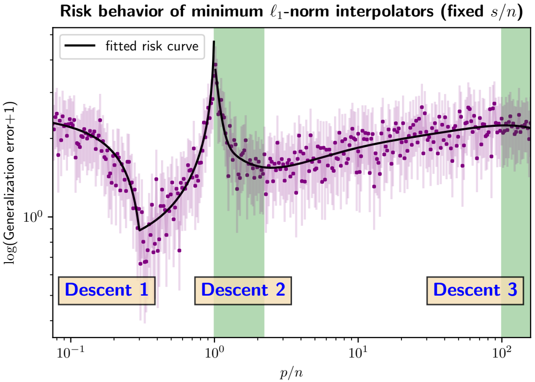


Figure: Multiple descent in sparse linear regression. Let the true signal $\boldsymbol{\theta}^\star$ be an s-sparse vector, where M is the magnitude of non-zero entries. Fix $s/n = 0.3$ and $s/n \cdot M^2 = 10$. Set the sample size as $n = 100$, and choose 500 values of p/n.

A multi-descent phenomenon

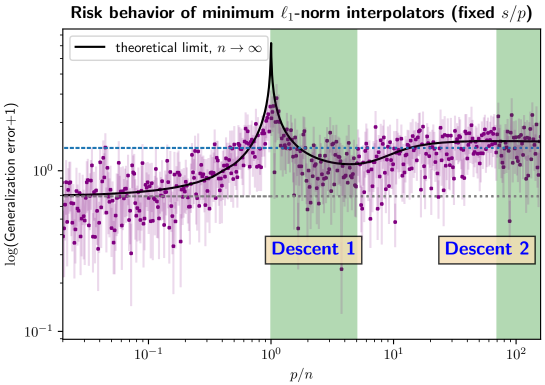


Figure: Multiple descent in sparse linear regression. Let the true signal $\boldsymbol{\theta}^\star$ be an s-sparse vector, where $\sqrt{\delta}M$ is the magnitude of non-zero entries. Fix $s/p = 0.01$ and $s/p \cdot M^2 = 2$. Set the sample size as $n = 100$, and choose 500 values of p/n.

Question:

- How to theoretically characterize $\underbrace{\text{these descents}}_{\text{as a function of } p/n}$?

Challenges:

Question:

- How to theoretically characterize $\underbrace{\text{these descents}}_{\text{as a function of } p/n}$?

Challenges:

- *no* closed-form solutions for min ℓ_1-norm interpolators

Question:

- How to theoretically characterize $\underbrace{\text{these descents}}_{\text{as a function of } p/n}$?

Challenges:

- *no* closed-form solutions for min ℓ_1-norm interpolators
- *no* consistent support recovery in high dimensional regime

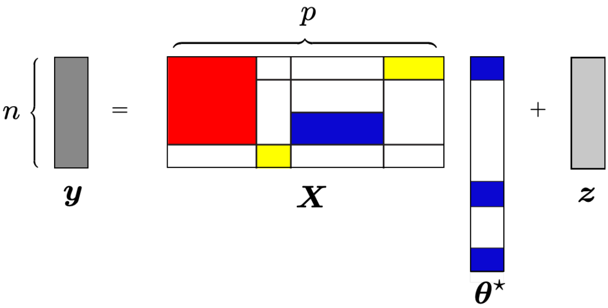
Question:

- How to theoretically characterize $\underbrace{\text{these descents}}_{\text{as a function of } p/n}$?

Challenges:

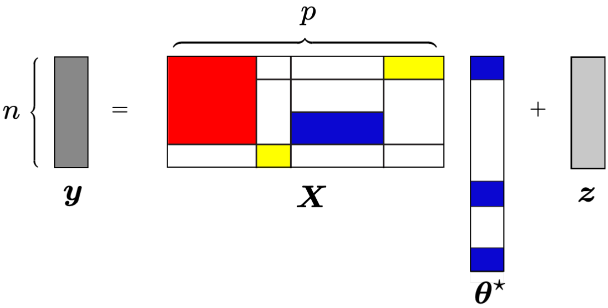
- *no* closed-form solutions for min ℓ_1-norm interpolators
- *no* consistent support recovery in high dimensional regime
- *no* strong convexity in this optimization problem

Model setup and assumptions



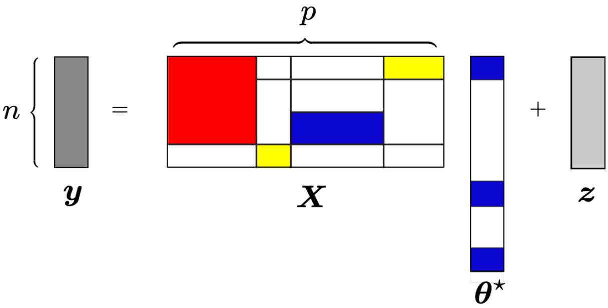
- true signal $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ is s-sparse

Model setup and assumptions



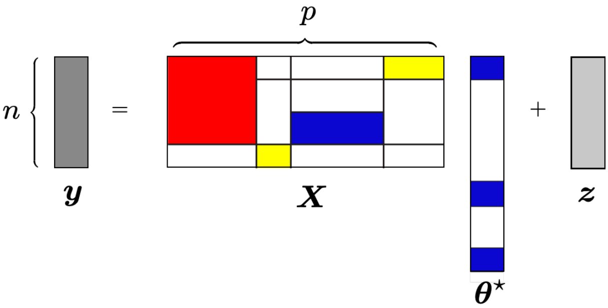
- true signal $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ is s-sparse

Model setup and assumptions



- true signal $\boldsymbol{\theta}^{\star} \in \mathbb{R}^p$ is s-sparse
- proportional regime: $s/p = \epsilon$ (const), $n/p = \delta$ (const)

Model setup and assumptions



- true signal $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ is s-sparse
- proportional regime: $s/p = \epsilon$ (const), $n/p = \delta$ (const)
- Gaussian design and Gaussian noise

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{n}\boldsymbol{I}_p\right), \qquad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Exact asymptotics framework

min ℓ_1-norm interpolator $(n < p)$

$$\widehat{\boldsymbol{\theta}}^{\mathsf{Int}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{subject to} \quad y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle, \quad 1 \le i \le n$$

- generalization error:

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}) := \mathbb{E}\big[(\boldsymbol{x}_{\mathsf{new}}^{\top} \widehat{\boldsymbol{\theta}}^{\mathsf{Int}} - y_{\mathsf{new}})^2\big] = \frac{1}{n}\|\widehat{\boldsymbol{\theta}}^{\mathsf{Int}} - \boldsymbol{\theta}^{\star}\|_2^2 + \sigma^2$$

Exact asymptotics framework

min ℓ_1-norm interpolator $(n < p)$

$$\widehat{\boldsymbol{\theta}}^{\mathsf{Int}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{subject to} \quad y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle, \quad 1 \le i \le n$$

- generalization error:

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}) := \mathbb{E}\big[(\boldsymbol{x}_{\mathsf{new}}^\top \widehat{\boldsymbol{\theta}}^{\mathsf{Int}} - y_{\mathsf{new}})^2\big] = \frac{1}{n}\|\widehat{\boldsymbol{\theta}}^{\mathsf{Int}} - \boldsymbol{\theta}^\star\|_2^2 + \sigma^2$$

- high-dim asymptotics ($\delta = n/p, \; \epsilon = s/p$)

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}, \delta) = \lim_{\substack{n/p = \delta \\ n, p \to \infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}\big) \;=\; ???$$

Exact asymptotics framework

min ℓ_1-norm interpolator $(n < p)$

$$\widehat{\boldsymbol{\theta}}^{\mathsf{Int}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{\theta}\|_1 \qquad \text{subject to} \quad y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle, \;\; 1 \le i \le n$$

- generalization error:

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}) := \mathbb{E}\big[(\boldsymbol{x}_{\mathsf{new}}^\top \widehat{\boldsymbol{\theta}}^{\mathsf{Int}} - y_{\mathsf{new}})^2\big] = \frac{1}{n}\|\widehat{\boldsymbol{\theta}}^{\mathsf{Int}} - \boldsymbol{\theta}^\star\|_2^2 + \sigma^2$$

- high-dim asymptotics ($\delta = n/p, \;\; \epsilon = s/p$)

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}, \delta) = \lim_{\substack{n/p=\delta \\ n,p\to\infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}\big) \;\; = \;\; ???$$

- how does $\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}, \delta)$ vary as a function of δ?

An incomplete literature list on exact asymptotics

$$\mathsf{Risk}(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}, \delta) = \lim_{\substack{n/p=\delta \\ n,p\to\infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}\big) \;=\; ???$$

- compressed sensing and Lasso estimators
  Donoho, Maleki and Montanari (2009), Bayati and Montanari (2011), Stojnic (2013), Oymak et al. (2013), Miolane and Montanari (2018), Bellec and Zhang (2019), Celentano, Montanari and Wei (2020)
- robust regression and ridge regreesion
  Donoho and Montanari (2016), El Karoui (2013, 2018), Thrampoulidis et al. (2018), Dobriban and Wager (2018), Hastie et al. (2019), Mei and Montanari (2019), Patil et al. (2021)
- classification
  Sur, Chen and Candés (2017), Montanari et al. (2019), Liang and Sur (2020), Javanmard and Soltanolkotabi (2020)

Main result: Risk curve for min ℓ_1 solution

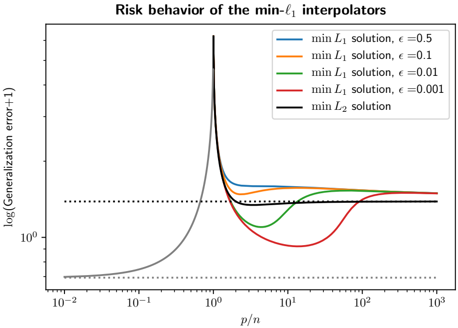
Suppose $\theta_i^\star \overset{\text{i.i.d.}}{\sim} \epsilon \mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0$ (*SNR* $= \frac{1}{\sigma^2}\mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star)^2 = \frac{\epsilon M^2}{\sigma^2}$)

Main result: Risk curve for min ℓ_1 solution

Suppose $\theta_i^\star \overset{\text{i.i.d.}}{\sim} \epsilon \mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0$ (*SNR* $= \frac{1}{\sigma^2}\mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star)^2 = \frac{\epsilon M^2}{\sigma^2}$)

Theorem (Li, W' 21)

- Risk$(\widehat{\boldsymbol{\theta}}^{\text{Int}}; \delta) \to$ Risk$(\mathbf{0})$ *as* p/n *tends to* ∞.

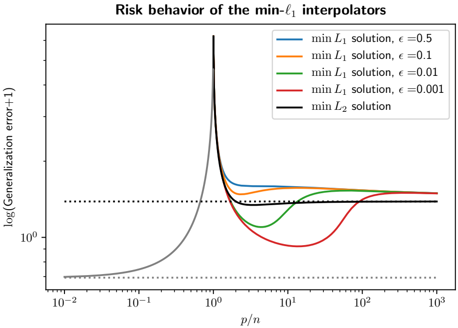


Main result: Risk curve for min ℓ_1 solution

Suppose $\theta_i^\star \overset{\text{i.i.d.}}{\sim} \epsilon\mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0$ (*SNR* $= \frac{1}{\sigma^2}\mathbb{E}(\boldsymbol{x}_i^\top\boldsymbol{\theta}^\star)^2 = \frac{\epsilon M^2}{\sigma^2}$)

Theorem (Li, W' 21)

- Risk$(\widehat{\boldsymbol{\theta}}^{\text{Int}};\delta) \to$ Risk$(\mathbf{0})$ *as* p/n *tends to* ∞.
- *for every given* δ, *there exists* $\tilde{\epsilon}(\delta)$ *st.* Risk$(\widehat{\boldsymbol{\theta}}^{\text{Int}};\delta)$ *decreases with* p/n *at* δ *as long as the sparsity ratio* ϵ *satisfies* $\epsilon \le \tilde{\epsilon}(\delta)$.

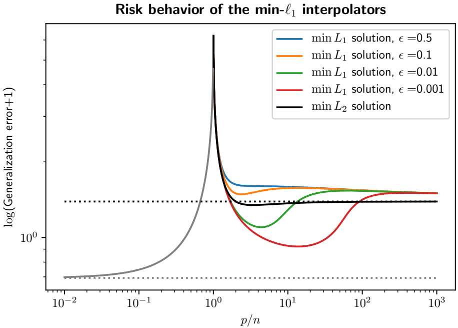


Main result: Risk curve shape (continued)

Suppose $\theta_i^\star \overset{\text{i.i.d.}}{\sim} \epsilon \mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0$ ($SNR = \frac{1}{\sigma^2}\mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star)^2 = \frac{\epsilon M^2}{\sigma^2}$)

Theorem (Li, W '21)

- *there exist two constants* $1 < \eta_1 < \eta_2 < \infty$ *st.* Risk$(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}; \delta)$ *decreases with* p/n *within the range* $p/n \in (1, \eta_1) \cup (\eta_2, \infty)$.

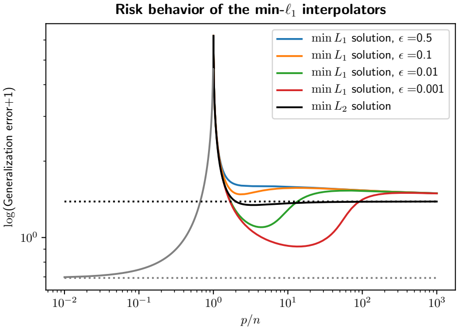


Main result: Risk curve shape (continued)

Suppose $\theta_i^\star \overset{\text{i.i.d.}}{\sim} \epsilon\mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0$ (*SNR* $= \frac{1}{\sigma^2}\mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\theta}^\star)^2 = \frac{\epsilon M^2}{\sigma^2}$)

Theorem (Li, W '21)

- *there exist two constants* $1 < \eta_1 < \eta_2 < \infty$ *st.* Risk$(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}; \delta)$ *decreases with* p/n *within the range* $p/n \in (1, \eta_1) \cup (\eta_2, \infty)$.
- *fix* $\epsilon M^2/\sigma^2$. *There exists* ϵ^* *st. if* $\epsilon < \epsilon^*$, *then there exists region within* (η_1, η_2) *st.* Risk$(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}; \delta)$ *increases with* p/n.



Some heuristic explanations

- Why there is a peak at interpolation ($p = n$)?

Some heuristic explanations

- Why there is a peak at interpolation $(p = n)$?

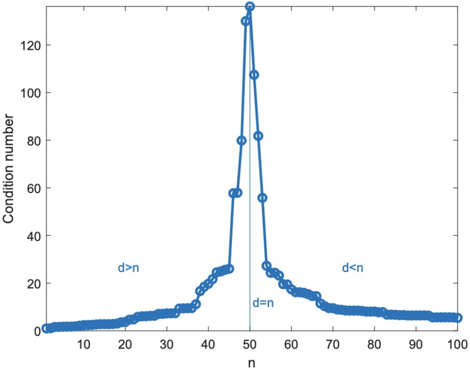


Figure: condition number of $\boldsymbol{X}$

Double descent in condition number, Poggio, Kur and Banburski (2020)

Some heuristic explanations

- Why there is a peak at interpolation $(p = n)$?
- Why there exists a second descent $(p > n)$?

Some heuristic explanations

- Why there is a peak at interpolation $(p = n)$?
- Why there exists a second descent $(p > n)$?

some evidences...

- Bellec, Lecué and Tsybakov (2016) studies optimal-tuned Lasso

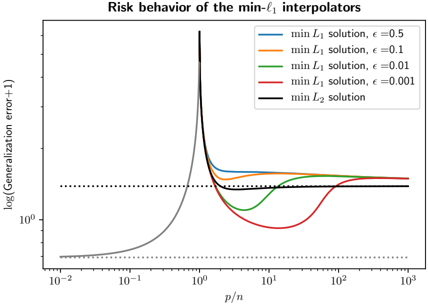
$$\frac{1}{n}\|\widehat{\boldsymbol{\theta}}^{\mathsf{Lasso}} - \boldsymbol{\theta}^\star\|_2^2 \leq c\sigma^2 \cdot \kappa(\boldsymbol{X})^2 \cdot \frac{p}{n} \cdot \epsilon \log(\frac{1}{\epsilon})$$

- Su and Candés, (2015) studies SLOPE estimator for $\epsilon \to 0$

$$\frac{1}{n}\|\widehat{\boldsymbol{\theta}}^{\mathsf{SLOPE}} - \boldsymbol{\theta}^\star\|_2^2 \leq 2\sigma^2 \cdot \frac{p}{n} \cdot \epsilon \log(\frac{1}{\epsilon})$$

Some heuristic explanations

- Why there is a peak at interpolation ($p = n$)?
- Why there exists a second descent ($p > n$)?
- (further increase p/n) wrong support $\rightarrow$ even worse than the zero estimator



interplay between over-parameterized ratio and sparsity

Compare to min ℓ_2-norm interpolators

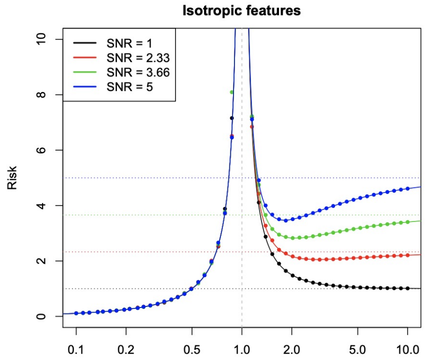


Figure: Hastie et al. (2019)

$$\lim_{\substack{n/p=\delta \\ n,p\to\infty}} \text{Risk}(\widehat{\boldsymbol{\theta}}^{\text{Int},\ell_2}) \overset{\text{a.s.}}{=} \begin{cases} \frac{\delta}{\delta-1}\sigma^2, & \text{if } \delta = n/p > 1 \\ \epsilon M^2(1-\delta) + \frac{1}{1-\delta}\sigma^2, & \text{if } \delta = n/p < 1 \end{cases}$$

Multi-descent in ℓ_2 training

Multi-descent in ℓ_2 training for different reasons Nakkiran et al. (2020), Chen et al. (2020), d'Ascoli et al. (2020), Adlam and Pennington (2020), Li and Meng (2020)

- design matrix has heterogenous structures
- non-linear kernels for kernel regression

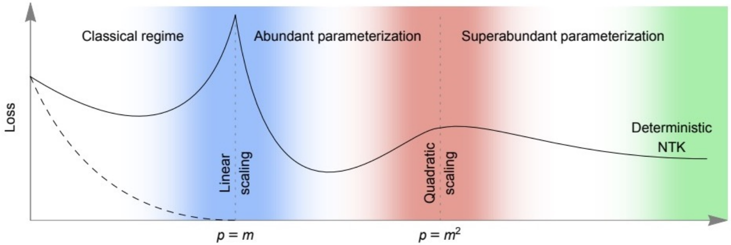


Figure: Adlam and Pennington (2020)

Our analysis framework

Risk characterization

Theorem (Li, Wei '21)

The generalization error of the min ℓ_1-norm interpolator obeys

$$\lim_{\substack{n/p=\delta \\ n,p\to\infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}\big) \overset{\text{a.s.}}{=} \tau^{\star 2}(\delta).$$

— *informally coordinates of* $\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}$ *behave like* $\eta(\Theta + \tau^\star Z; \alpha^\star \tau^\star)$

Here $\eta(x;\zeta) \coloneqq (|x| - \zeta)_+ \,\mathsf{sign}(x)$

$(\tau^\star, \alpha^\star)$ stands for the unique solution to

$$\tau^2 = \frac{1}{\delta}\mathbb{E}\left[\big(\eta(\Theta + \tau Z; \alpha\tau) - \Theta\big)^2\right] + \sigma^2,$$
$$\delta = \mathbb{P}\big(\eta(\Theta + \tau Z; \alpha\tau) = 0\big),$$

where $\Theta \sim P_\Theta$, $Z \sim \mathcal{N}(0,1)$ independent of Θ

Descent analysis

$(\tau^\star, \alpha^\star)$ stands for the unique solution to

$$\tau^2 = \frac{1}{\delta}\mathbb{E}\left[\big(\eta(\Theta + \tau Z; \alpha\tau) - \Theta\big)^2\right] + \sigma^2,$$
$$\delta = \mathbb{P}\big(\eta(\Theta + \tau Z; \alpha\tau) = 0\big),$$

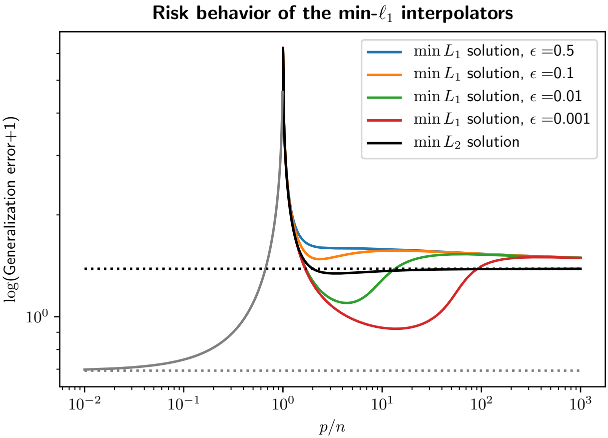
Suppose $\Theta \sim \epsilon\mathcal{P}_{M\sqrt{\delta}} + (1-\epsilon)\mathcal{P}_0$

$$1 = \frac{\nu^2}{M^2}\sigma^2 + \frac{\epsilon}{\delta}\mathbb{E}\left[\big(\eta(\sqrt{\delta}\nu + Z; \alpha) - \sqrt{\delta}\nu\big)^2\right] + \frac{1-\epsilon}{\delta}\mathbb{E}\left[\eta^2(Z;\alpha)\right]$$

$$\delta = \epsilon\mathbb{P}\big(|\nu\sqrt{\delta} + Z| > \alpha\big) + (1-\epsilon)\mathbb{P}(|Z| > \alpha)$$

where $\nu := M/\tau$

Descent analysis



Fix $\mathrm{SNR} = \frac{\epsilon M^2}{\sigma^2}$ and plot $\tau^\star$ as a function of $\frac{p}{n}$

Main tool: Approximate message passing (AMP)

Theorem (Li, W '21)

The generalization error of the min ℓ_1-norm interpolator obeys

$$\lim_{\substack{n/p=\delta \\ n,p\to\infty}} \mathsf{Risk}\big(\widehat{\boldsymbol{\theta}}^{\mathsf{Int}}\big) \overset{\text{a.s.}}{=} \tau^{\star 2}(\delta).$$

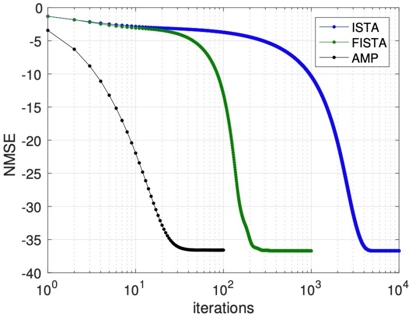
Main tool: Approximate message passing (AMP)

- AMP is an efficient iterative algorithm that has been applied to a broad range of statistical estimation problems Donoho Maleki, Montanari, (2009, 2010a, 2011b), Bayati and Montanari (2011)

Main tool: Approximate message passing (AMP)

- AMP is an efficient iterative algorithm that has been applied to a broad range of statistical estimation problems Donoho Maleki, Montanari, (2009, 2010a, 2011b), Bayati and Montanari (2011)

Example: Solving for Lasso $\left(\mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2+\lambda\|\boldsymbol{\theta}\|_1\right)$



AMP iterates

$$\boldsymbol{\theta}^{t+1} = \eta(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\theta}^t; \zeta_t)$$

$$\boldsymbol{z}^t = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^t + \underbrace{\frac{1}{\delta}\boldsymbol{z}^{t-1}\left\langle \eta'(\boldsymbol{X}^\top \boldsymbol{z}^{t-1} + \boldsymbol{\theta}^{t-1}; \zeta_{t-1})\right\rangle}_{\text{Onsager term}}$$

- $n = 250$, $p = 500$, $\epsilon = 0.1$

Figure credit: Borgerding and Schniter

Applications of AMP

- AMP is successfully applicable in variety of applications
  - imaging Fletcher, Rangan (2014), Vila, Schniter, Meola (2015), Metzler, Mousavi, Baraniuk (2017)
  - communications Schniter (2011), Jeon et al. (2015), Barbier, Krzakala (2017), Rush, Greig, Venkataramanan (2017)

Applications of AMP

- AMP is successfully applicable in variety of applications
  - imaging Fletcher, Rangan (2014), Vila, Schniter, Meola (2015), Metzler, Mousavi, Baraniuk (2017)
  - communications Schniter (2011), Jeon et al. (2015), Barbier, Krzakala (2017), Rush, Greig, Venkataramanan (2017)
- in regression and low rank matrix estimation
  - information-theoretically optimal v.s. computationally feasible estimators Reeves, Pfister (2019), Barbier et al. (2019), Lelarge and Miolane (2019)
  - conjectured to have optimal asymptotic estimation error among all polynomial-time algorithms Celentano and Montanari (2019)

— *tutorial, Feng, Venkataramanan, Rush, Samworth (2021)*

Recipe: AMP for statistical procedures

- dynamics of AMP can be accurately tracked by a simple small-dimensional recursive formula called the *state evolution*

state evolution: $$\tau_{t+1} = F(\tau_t, \alpha^\star \tau_t)$$

$$(\theta_i^{t+1}, \theta_i^\star)_{i=1}^p \overset{d}{\approx} (\eta(\Theta + \tau_t Z; \alpha^\star \tau_t), \Theta)$$

Recipe: AMP for statistical procedures

- dynamics of AMP can be accurately tracked by a simple small-dimensional recursive formula called the *state evolution*

state evolution: $\tau_{t+1} = F(\tau_t, \alpha^\star \tau_t)$

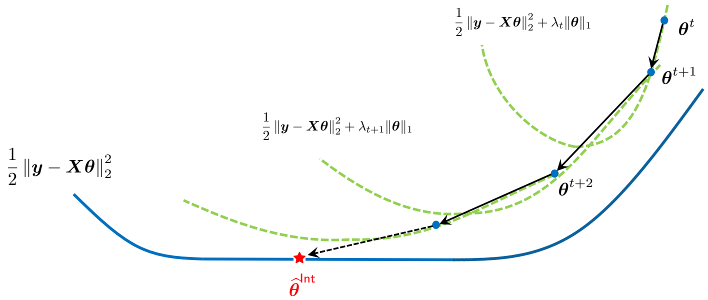
$$(\theta_i^{t+1}, \theta_i^\star)_{i=1}^p \overset{d}{\approx} (\eta(\Theta + \tau_t Z; \alpha^\star \tau_t), \Theta)$$

- construct AMP algorithms that converge to ...

  M-estimators Donoho and Montanari (2013), Lasso Bayati and Montanari (2011), SLOPE estimator Su and Candés. (2015), MLE for generalized linear model Sur, Chen, Candés (2017), lower-rank matrix estimation Montanari, Venkataramanan (2021), ...

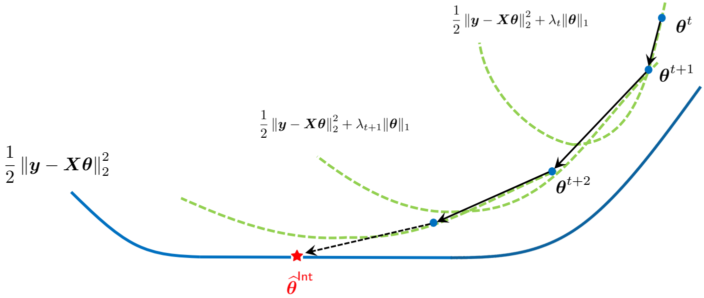
AMP for min ℓ_1 interpolator

Illustration of the AMP updates for the minimum ℓ_1-norm interpolator.



AMP for min ℓ_1 interpolator

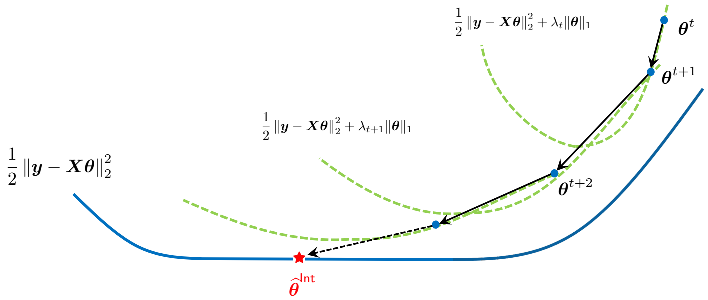
Illustration of the AMP updates for the minimum ℓ_1-norm interpolator.



- structural property when restricted strongly convexity is lacking

AMP for min ℓ_1 interpolator

Illustration of the AMP updates for the minimum ℓ_1-norm interpolator.

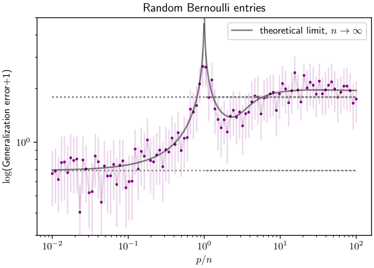


- structural property when restricted strongly convexity is lacking
- proper choice of λ_t sequence

$$\mathsf{dist}(\boldsymbol{\theta}^{t+1},\boldsymbol{\theta}^t)\le\exp(-\lambda_t)\cdot\mathsf{dist}(\boldsymbol{\theta}^t,\boldsymbol{\theta}^{t-1})+c|\lambda_t-\lambda_{t+1}|$$

Several extensions and questions...

Experiments for non-Gaussian designs



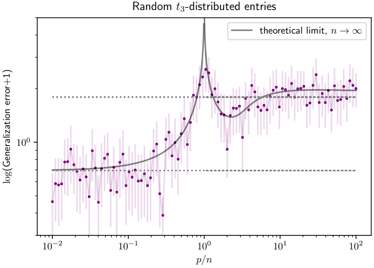


Figure: The entries of the design matrix $\sqrt{n}\boldsymbol{X}$ are i.i.d. sampled from the Bernoulli(0.5) distribution for the left, and from $t(3)/\sqrt{3}$ distribution for the right.

Experiments for non-Gaussian designs

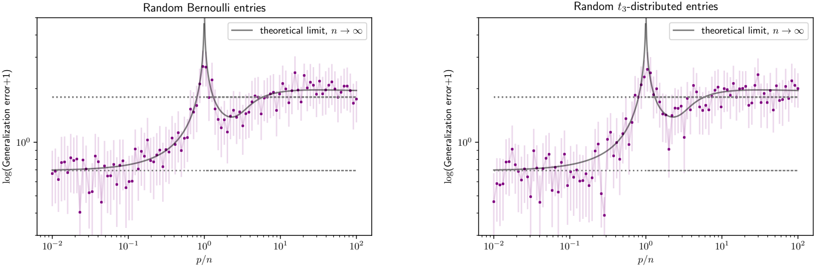


Figure: The entries of the design matrix $\sqrt{n}\boldsymbol{X}$ are i.i.d. sampled from the Bernoulli(0.5) distribution for the left, and from $t(3)/\sqrt{3}$ distribution for the right.

— **universality phenomenon** Bayati et al. (2015), Oymak and Tropp (2018), Montanari and Nguyen (2017), Chen and Lam (2021)

Experiments for non-Gaussian designs

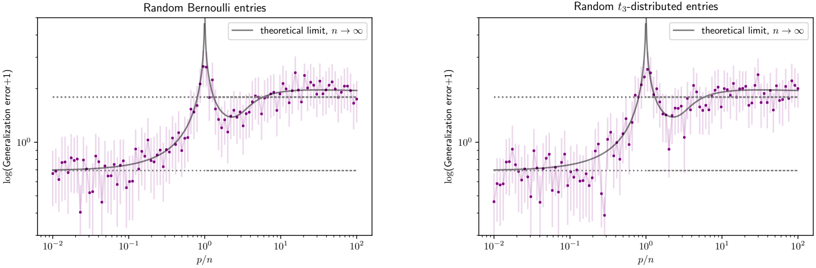
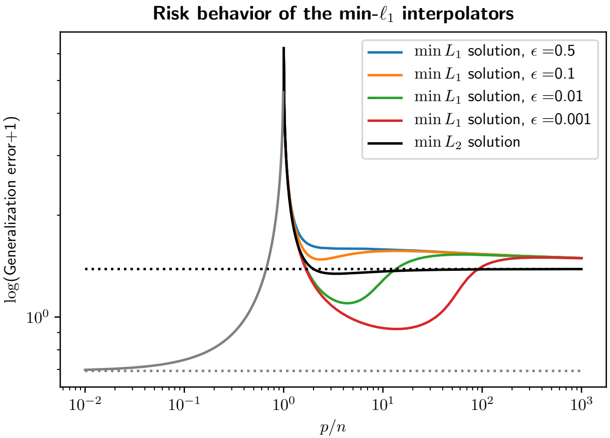


Figure: The entries of the design matrix $\sqrt{n}\boldsymbol{X}$ are i.i.d. sampled from the Bernoulli(0.5) distribution for the left, and from $t(3)/\sqrt{3}$ distribution for the right.

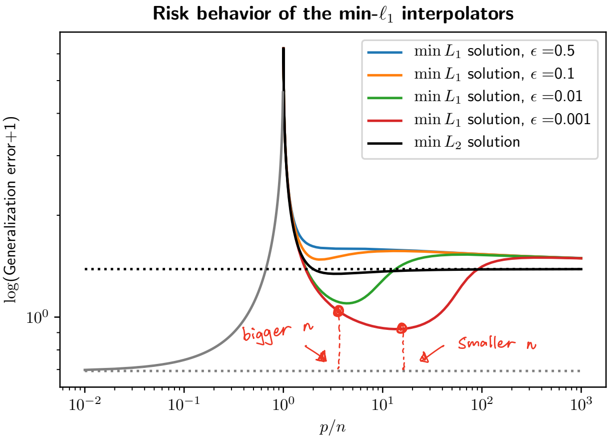
— **universality phenomenon** Bayati et al. (2015), Oymak and Tropp (2018), Montanari and Nguyen (2017), Chen and Lam (2021)

— **beyond i.i.d design** Celentano, Montanari and Wei (2020), Fan (2020)

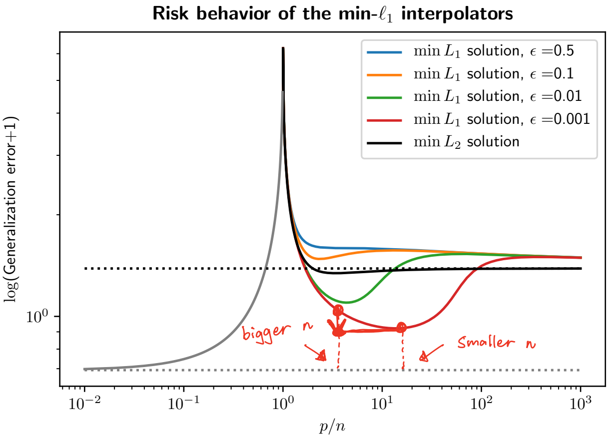
Mitigate multiple descent via cross-validation



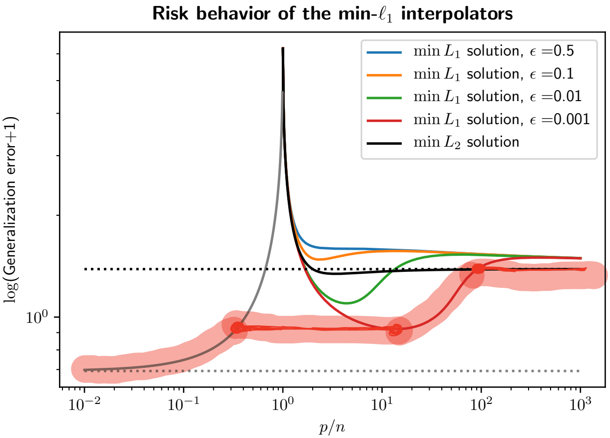
Mitigate multiple descent via cross-validation



Mitigate multiple descent via cross-validation

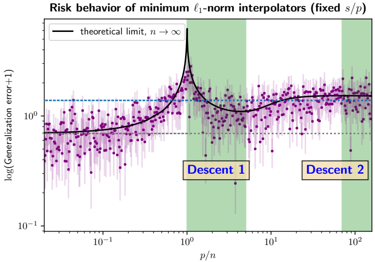


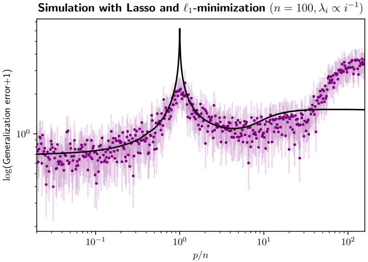
Mitigate multiple descent via cross-validation



Mitigating multiple descents: Model-agnostic risk monotonization in high-dimensional learning — ongoing work with Pratik Patil, Arun Kuchibhotla, Alessandro Rinaldo

Concluding remarks

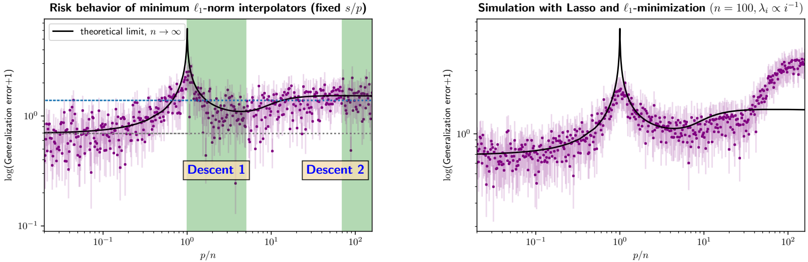




Future directions

- features with general covariance structure $\rightarrow$ even more oscillations in risk curve

Concluding remarks



Future directions

- features with general covariance structure $\to$ even more oscillations in risk curve
- generalize to more complex model $\to$ towards understanding dnn

Thanks for your attention! Questions?

Paper:

"Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent,"
Y. Li, Y. Wei, 2021

"The Lasso with general Gaussian designs with applications to hypothesis testing,"
M. Celentano, A. Montanari, Y. Wei, 2020

Linear sparsity

Example: Genome-wide association studies (GWAS): genetic variants $\rightarrow$ disease

Leading Edge

Perspective

Cell

An Expanded View of Complex Traits: From Polygenic to Omnigenic

Evan A. Boyle,[1,*] Yang I. Li,[1,*] and Jonathan K. Pritchard[1,2,3,*]
[1]Department of Genetics
[2]Department of Biology
[3]Howard Hughes Medical Institute
Stanford University, Stanford, CA 94305, USA

matin regions of immune cells (Maurano et al.; 2012; Farh et al., 2015; Kundaje et al., 2015).

These observations are generally interpreted in a paradigm in which complex disease is driven by an accumulation of weak effects on the key genes and regulatory pathways that drive disease risk (Furlong, 2013; Chakravarti and Turner, 2016). This model has motivated many studies that aim to dissect the functional impacts of individual disease-associated variants

Challenge: True signals might NOT be ultra-sparse

$\longrightarrow$ important features may scale proportionally to the feature dimension

Connections to two-layer network training

$$\mathcal{F}_{2NN}^{N} = \left\{ f(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{W}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w_i}, \, \boldsymbol{x} \rangle) \mid a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^p, \forall i \leq N \right\}$$

Connections to two-layer network training

$$\mathcal{F}_{2NN}^{N} = \left\{ f(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{W}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w_i},\, \boldsymbol{x} \rangle) \mid a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^p, \forall i \leq N \right\}$$

Lazy regime: model estimation stays close to initialization

$$\begin{aligned}
&\frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0 + \epsilon \boldsymbol{a}, \boldsymbol{W}_0 + \epsilon \boldsymbol{W}) \\
&\quad \approx \frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) + \langle \boldsymbol{a},\, \nabla_{\boldsymbol{a}} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) \rangle + \langle \boldsymbol{W},\, \nabla_{\boldsymbol{W}} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) \rangle \\
&\quad \approx \frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) + \underbrace{\sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_{0,i},\, x_i \rangle)}_{\text{random feature model}} + \underbrace{\sum_{i=1}^{N} a_{0,i} \langle w_i,\, x \rangle \sigma(\langle \boldsymbol{w}_{0,i},\, x_i \rangle)}_{\text{neural tangent kernel model}}
\end{aligned}$$

Connections to two-layer network training

$$\mathcal{F}_{2NN}^{N} = \left\{ f(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{W}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w_i}, \, \boldsymbol{x} \rangle) \mid a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^p, \forall i \leq N \right\}$$

Lazy regime: model estimation stays close to initialization

$$\begin{aligned} &\frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0 + \epsilon \boldsymbol{a}, \boldsymbol{W}_0 + \epsilon \boldsymbol{W}) \\ &\quad \approx \frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) + \langle \boldsymbol{a}, \, \nabla_{\boldsymbol{a}} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) \rangle + \langle \boldsymbol{W}, \, \nabla_{\boldsymbol{W}} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) \rangle \\ &\quad \approx \frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) + \underbrace{\sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_{0,i}, \, x_i \rangle)}_{\text{random feature model}} + \underbrace{\sum_{i=1}^{N} a_{0,i} \langle w_i, \, x \rangle \sigma(\langle \boldsymbol{w}_{0,i}, \, x_i \rangle)}_{\text{neural tangent kernel model}} \end{aligned}$$

— *transform into kernel ridge regression with random kernels!*

Connections to two-layer network training

$$\mathcal{F}_{2NN}^{N} = \left\{ f(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{W}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w_i},\, \boldsymbol{x} \rangle) \mid a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^p, \forall i \leq N \right\}$$

Lazy regime: model estimation stays close to initialization

$$\begin{aligned}
&\frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0 + \epsilon \boldsymbol{a}, \boldsymbol{W}_0 + \epsilon \boldsymbol{W}) \\
&\quad \approx \frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) + \langle \boldsymbol{a},\, \nabla_{\boldsymbol{a}} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) \rangle + \langle \boldsymbol{W},\, \nabla_{\boldsymbol{W}} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) \rangle \\
&\quad \approx \frac{1}{\epsilon} f(\boldsymbol{x}, \boldsymbol{a}_0, \boldsymbol{W}_0) + \underbrace{\sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_{0,i},\, x_i \rangle)}_{\text{random feature model}} + \underbrace{\sum_{i=1}^{N} a_{0,i} \langle w_i,\, x \rangle \sigma(\langle \boldsymbol{w}_{0,i},\, x_i \rangle)}_{\text{neural tangent kernel model}}
\end{aligned}$$

— *transform into kernel ridge regression with random kernels!*

Jacob, Gabriel, Hongler (2018), Chizat, Bach (2019), Du et al. (2018), Arora, et al. (2019), Ghorbani, Mei, Misiakiewicz, Montanari (2019), Montanari, Zhong (2020), Allen-Zhu, Li and Liang (2019)...