Minimum-Norm Interpolation in Statistical Learning: new phenomena in high dimensions

Tengyuan Liang

CHICAGO BOOTH
The University of Chicago Booth School of Business

Regression: with Sasha Rakhlin (MIT), Xiyu Zhai (MIT)
Classification: with Pragya Sur (Harvard)

OUTLINE

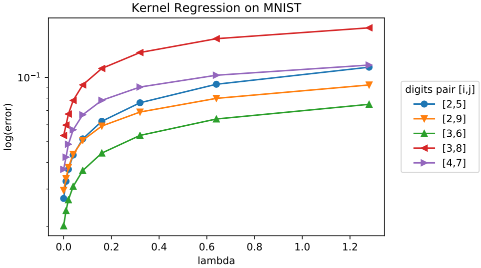
- Motivation: min-norm interpolants for over-parametrized models
- Regression: multiple descent of risk for kernels/neural networks
- Classification: precise asymptotics of boosting algorithms

OVERPARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.

Zhang, Bengio, Hardt, Recht, and Vinyals (2016)

Belkin et al. (2018a,b); Liang and Rakhlin (2018); Bartlett et al. (2019); Hastie et al. (2019)



$\lambda = 0$: the interpolants on training data.

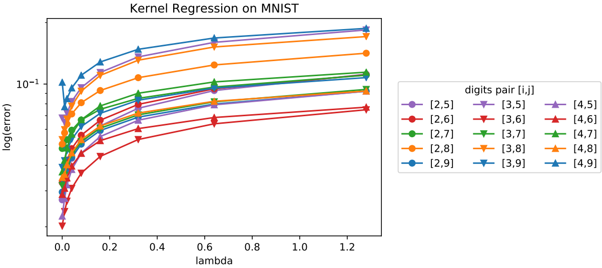
MNIST data from LeCun et al. (2010)

OVERPARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.

Zhang, Bengio, Hardt, Recht, and Vinyals (2016)

Belkin et al. (2018a,b); Liang and Rakhlin (2018); Bartlett et al. (2019); Hastie et al. (2019)

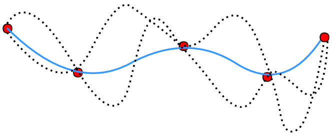


$\lambda = 0$: the interpolants on training data.

MNIST data from LeCun et al. (2010)

OVERPARAMETRIZED REGIME OF STAT/ML

In fact, many models behave the same on training data.



Practical methods or algorithms favor certain functions!

Principle: among the models that **interpolate**,
algorithms favor certain form of **minimalism**.

OVERPARAMETRIZED REGIME OF STAT/ML

Principle: among the models that **interpolate**,
algorithms favor certain form of **minimalism**.

- overparametrized linear model and matrix factorization
- kernel regression
- support vector machines, Perceptron
- boosting, AdaBoost
- two-layer ReLU networks, deep neural networks

OVERPARAMETRIZED REGIME OF STAT/ML

Principle: among the models that **interpolate**,
algorithms favor certain form of **minimalism**.

- overparametrized linear model and matrix factorization
- kernel regression
- support vector machines, Perceptron
- boosting, AdaBoost
- two-layer ReLU networks, deep neural networks

minimalism typically measured in form of **certain norm**
motivates the study of **min-norm interpolants**

MIN-NORM INTERPOLANTS

minimalism typically measured in form of **certain norm**
motivates the study of **min-norm interpolants**

Regression

$$\widehat{f} = \arg\min_{f} \|f\|_{\text{norm}}, \quad \text{s.t. } y_i = f(x_i) \ \forall i \in [n].$$

Classification

$$\widehat{f} = \arg\min_{f} \|f\|_{\text{norm}}, \quad \text{s.t. } y_i \cdot f(x_i) \geq 1 \ \forall i \in [n].$$

Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels

with Sasha Rakhlin (MIT), Xiyu Zhai (MIT)

Regression

$$\widehat{f} = \operatorname*{arg\,min}_{f} \|f\|_{\text{norm}}, \quad \text{s.t.} \;\; y_i = f(x_i) \; \forall i \in [n].$$

SHAPE OF RISK CURVE

Classic: U-shape curve

Recent: double descent curve

Belkin, Hsu, Ma, and Mandal (2018a); Hastie, Montanari, Rosset, and Tibshirani (2019)

Question: shape of the **risk curve** w.r.t. **"over-parametrization"**?

SHAPE OF RISK CURVE

Classic: U-shape curve

Recent: double descent curve

Belkin, Hsu, Ma, and Mandal (2018a); Hastie, Montanari, Rosset, and Tibshirani (2019)

Question: shape of the **risk curve** w.r.t. **"over-parametrization"**?

We model the **intrinsic dim.** $d = n^{\alpha}$ with $\alpha \in (0, 1)$, with feature cov. $\Sigma_d = I_d$.

We consider the **non-linear Kernel Regression** model.

Data Generating Process

DGP.

- $\{x_i\}_{i=1}^n \overset{i.i.d}{\sim} \mu = \mathcal{P}^{\otimes d}$, dist. of each coordinate satisfies weak moment condition.
- target $f_\star(x) := \mathbb{E}[Y|X = x]$, with bounded $\mathrm{Var}[Y|X = x]$.

Kernel.

- $h \in C^\infty(\mathbb{R})$, $h(t) = \sum_{i=0}^\infty \alpha_i t^i$ with $\alpha_i \geq 0$.
- inner product kernel $k(x, z) = h(\langle x, z\rangle / d)$.

Target Function.

- Assume $f_\star(x) = \int k(x, z)\rho_\star(z)\mu(dz)$ with $\|\rho_\star\|_\mu \leq C$.

DATA GENERATING PROCESS

Given n i.i.d. data pairs $(x_i, y_i) \sim \mathcal{P}_{X,Y}$.

Risk curve for minimum RKHS norm $\|\cdot\|_{\mathcal{H}}$ interpolants $\widehat{f}$?

$$\widehat{f} = \arg\min_f \|f\|_{\mathcal{H}}, \quad \text{s.t. } y_i = f(x_i) \; \forall i \in [n].$$

SHAPE OF RISK CURVE

Theorem (L., Rakhlin & Zhai, '19).

For any integer $\iota \geq 1$, consider $d = n^{\alpha}$ where $\alpha \in (\frac{1}{\iota+1}, \frac{1}{\iota})$.

SHAPE OF RISK CURVE

Theorem (L., Rakhlin & Zhai, '19).

For any integer $\iota \geq 1$, consider $d = n^{\alpha}$ where $\alpha \in (\frac{1}{\iota+1}, \frac{1}{\iota})$.

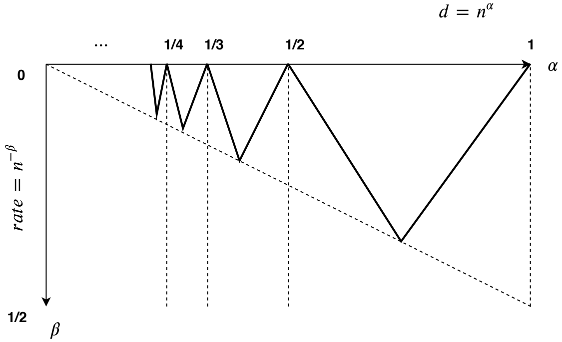
With probability at least $1 - \delta - e^{-n/d^{\iota}}$ on the design $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$\mathbb{E}\left[\| \widehat{f} - f_* \|_{\mu}^2 | \mathbf{X} \right] \leq C \cdot \left(\frac{d^{\iota}}{n} + \frac{n}{d^{\iota+1}} \right) \asymp n^{-\beta},$$

$$\beta := \min\{(\iota+1)\alpha - 1, 1 - \iota\alpha\}.$$

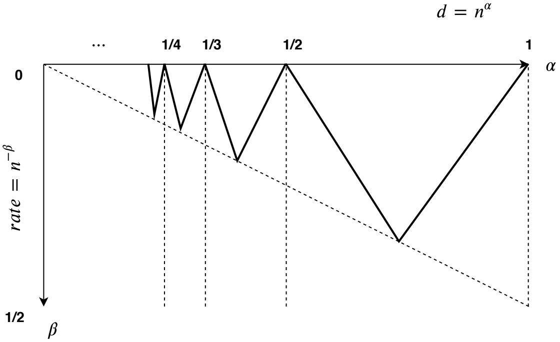
Here the constant $C(\delta, \iota, h, \mathcal{P})$ does not depend on d, n.

MULTIPLE DESCENT



multiple-descent behavior of the rates as the scaling $d = n^{\alpha}$ changes.

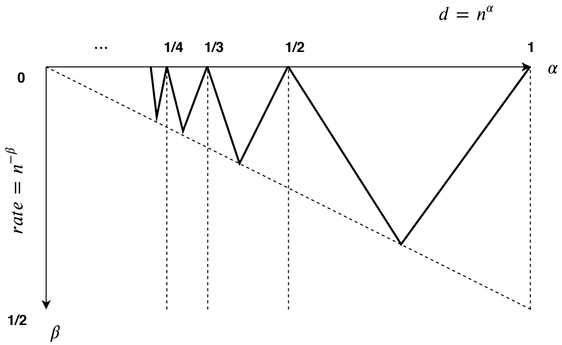
MULTIPLE DESCENT



multiple-descent behavior of the rates as the scaling $d = n^{\alpha}$ changes.

- **valley**: "valley" on the rate curve at $d = n^{\frac{1}{\iota+1/2}}$, $\iota \in \mathbb{N}$

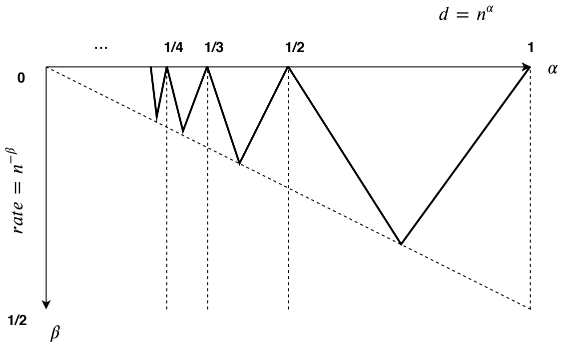
MULTIPLE DESCENT



multiple-descent behavior of the rates as the scaling $d = n^{\alpha}$ changes.

- **valley**: "valley" on the rate curve at $d = n^{\frac{1}{\iota+1/2}}$, $\iota \in \mathbb{N}$
- **over-parametrization**: towards over-parametrized regime, the good rate at the bottom of the valley is better

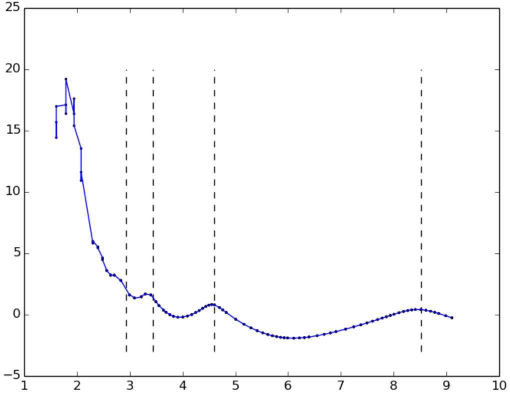
MULTIPLE DESCENT



multiple-descent behavior of the rates as the scaling $d = n^{\alpha}$ changes.

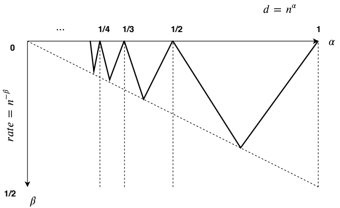
- **valley**: "valley" on the rate curve at $d = n^{\frac{1}{\iota+1/2}}$, $\iota \in \mathbb{N}$
- **over-parametrization**: towards over-parametrized regime, the good rate at the bottom of the valley is better
- **empirical**: preliminary empirical evidence of multiple descent

EMPIRICAL EVIDENCE

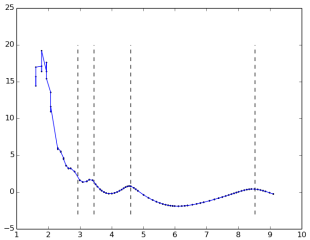


empirical evidence of **multiple-descent behavior** as the scaling $d = n^{\alpha}$ changes.

MULTIPLE DESCENT



theory



empirical

APPLICATION TO WIDE NEURAL NETWORKS

Neural Tangent Kernel (NTK)

Jacot, Gabriel, and Hongler (2018); Du, Zhai, Poczos, and Singh (2018)......

$$k_{\mathrm{NTK}}(x, x') = U\Big(\frac{\langle x, x' \rangle}{\|x\|\|x'\|}\Big), \text{ with } U(t) = \frac{1}{4\pi}\Big(3t(\pi - \arccos(t)) + \sqrt{1 - t^2}\Big)$$

Compositional Kernel of Deep Neural Network (DNN)

Daniely et al. (2016); Poole et al. (2016); Liang and Tran-Bach (2020)

$$k_{\mathrm{DNN}}(x, x') = \sum_{i=0}^{\infty} \alpha_i \cdot \Big(\frac{\langle x, x' \rangle}{\|x\|\|x'\|}\Big)^i$$

APPLICATION TO WIDE NEURAL NETWORKS

Neural Tangent Kernel (NTK)

Jacot, Gabriel, and Hongler (2018); Du, Zhai, Poczos, and Singh (2018)......

$$k_{\mathrm{NTK}}(x, x') = U\Big(\frac{\langle x, x' \rangle}{\|x\| \|x'\|}\Big), \text{ with } U(t) = \frac{1}{4\pi}\Big(3t(\pi - \arccos(t)) + \sqrt{1 - t^2}\Big)$$

Compositional Kernel of Deep Neural Network (DNN)

Daniely et al. (2016); Poole et al. (2016); Liang and Tran-Bach (2020)

$$k_{\mathrm{DNN}}(x, x') = \sum_{i=0}^{\infty} \alpha_i \cdot \Big(\frac{\langle x, x' \rangle}{\|x\| \|x'\|}\Big)^i$$

Corollary (L., Rakhlin & Zhai, '19)**.**

Multiple descent phenomena hold for kernels including NTK, and compositional kernel of DNN.

Precise High-Dimensional Asymptotic Theory for Boosting and Min-ℓ_1-Norm Interpolated Classifiers

with Pragya Sur (Harvard)

Classification

$$\widehat{f} = \arg\min_{f} \|f\|_{\text{norm}}, \quad \text{s.t. } y_i \cdot f(x_i) \geq 1 \ \forall i \in [n].$$

PROBLEM FORMULATION

Given n-i.i.d. data pairs $\{(x_i, y_i)\}_{1 \le i \le n}$, with $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$

$$y_i \in \{\pm 1\} \text{ binary labels}, \quad x_i \in \mathbb{R}^p \text{ feature vector (weak learners)}$$

Consider when data is linearly separable

$$\mathbb{P}\left(\exists \theta \in \mathbb{R}^p,\ y_i x_i^\top \theta > 0 \text{ for } 1 \le i \le n\right) \to 1 \ .$$

Natural to consider overparametrized regime

$$p/n \to \psi \in (0, \infty) \ .$$

BOOSTING / ADABOOST

"... mystery of AdaBoost as the most important unsolved problem in Machine Learning"

Wald Lecture, Breiman (2004)

"An important open problem is to derive more careful and precise bounds which can be used for this purpose. Besides paying closer attention to constant factors, such an analysis might also involve the measurement of more sophisticated statistics."

Schapire, Freund, Bartlett, and Lee (1998)

ℓ_1 GEOMETRY, MARGIN, AND INTERPOLATION

min-ℓ_1-norm interpolation equiv. max-ℓ_1-margin

$$\max_{\|\theta\|_1 \le 1} \min_{1 \le i \le n} y_i x_i^\top \theta =: \kappa_{\ell_1}(X, y) \ .$$

Prior understanding:

$$\text{generalization error} < \frac{1}{\sqrt{n}\kappa} \cdot (\text{log factors, constants})$$

Schapire, Freund, Bartlett, and Lee (1998)

$$\text{optimization steps} < \frac{1}{\kappa^2} \cdot (\text{log factors, constants})$$

Rosset, Zhu, and Hastie (2004); Zhang and Yu (2005); Telgarsky (2013)

ℓ_1 GEOMETRY, MARGIN, AND INTERPOLATION

Prior understanding:

$$\text{generalization error} < \frac{1}{\sqrt{n}\kappa} \cdot (\text{log factors, constants})$$

Schapire, Freund, Bartlett, and Lee (1998)

$$\text{optimization steps} < \frac{1}{\kappa^2} \cdot (\text{log factors, constants})$$

Rosset, Zhu, and Hastie (2004); Zhang and Yu (2005); Telgarsky (2013)

However, many questions remain:

Statistical

- how large is the ℓ_1-margin $\kappa_{\ell_1}(X, y)$?
- angle between the interpolated clasifier $\hat{\theta}$ and the truth $\theta_\star$?
- precise generalization error of Boosting? relation to Bayes Error?

Computational

- effect of increasing overparametrization $\psi = p/n$ on optimization?
- proportion of weak-learners activated by Boosting with zero initialization?

DATA GENERATING PROCESS

DGP. $x_i \sim \mathcal{N}(0, \Lambda)$ i.i.d. with diagonal cov. $\Lambda \in \mathbb{R}^{p\times p}$, and y_i are generated with non-decreasing $f : \mathbb{R} \to [0,1]$,

$$\mathbb{P}(y_i = +1|x_i) = 1 - \mathbb{P}(y_i = -1|x_i) = f(x_i^\top \theta_\star) \ ,$$

with some $\theta_\star \in \mathbb{R}^p$.

Consider high-dim asymptotic regime with overparametrized ratio

$$p/n \to \psi \in (0,\infty), \quad n,p \to \infty.$$

signal strength : $\|\Lambda^{1/2}\theta_\star\| \to \rho \in (0,\infty)$, coordinate : $\bar{w}_j = \sqrt{p}\dfrac{\lambda_j^{1/2}\theta_{\star,j}}{\rho}, 1 \le j \le p.$

Assume

$$\frac{1}{p}\sum_{j=1}^{p} \delta_{(\lambda_j,\bar{w}_j)} \overset{\text{Wasserstein-2}}{\Longrightarrow} \mu, \ \text{a dist. on } \mathbb{R}_{>0} \times \mathbb{R}$$

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Theorem (L. & Sur, '20).

For $\psi \geq \psi^\star$ (separability threshold), sharp asymptotic characterization holds:

$$\text{Margin:} \quad \lim_{\substack{n,p\to\infty \\ p/n\to\psi}} p^{1/2} \cdot \kappa_{\ell_1}(X, y) = \kappa_\star(\psi, \mu) \ , \ a.s.$$

$$\text{Generalization error:} \quad \lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{\ell_1} < 0\right) = \mathrm{Err}_\star(\psi, \mu) \ , \ a.s.$$

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Theorem (L. & Sur, '20).

For $\psi \geq \psi^\star$ (separability threshold), sharp asymptotic characterization holds:

$$\text{Margin:} \quad \lim_{\substack{n,p\to\infty \\ p/n\to\psi}} p^{1/2} \cdot \kappa_{\ell_1}(X, y) = \kappa_\star(\psi, \mu) \ , \ a.s.$$

$$\text{Generalization error:} \quad \lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{\ell_1} < 0\right) = \mathrm{Err}_\star(\psi, \mu) \ , \ a.s.$$

precise asymptotics can also be established on

$$\text{Angle:} \quad \frac{\langle \hat{\theta}_{\ell_1}, \theta_\star \rangle_\Lambda}{\|\hat{\theta}_{\ell_1}\|_\Lambda \|\theta_\star\|_\Lambda}, \qquad \text{Loss:} \quad \sum_{j\in[p]} \ell(\hat{\theta}_{\ell_1,j}, \theta_{\star,j})$$

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Theorem (L. & Sur, '20).

For $\psi \geq \psi^\star$ (separability threshold), sharp asymptotic characterization holds:

$$\text{Margin:} \quad \lim_{\substack{n,p\to\infty \\ p/n\to\psi}} p^{1/2} \cdot \kappa_{\ell_1}(X, y) = \kappa_\star(\psi, \mu) \;, \; a.s.$$

$$\text{Generalization error:} \quad \lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{\ell_1} < 0\right) = \mathrm{Err}_\star(\psi, \mu) \;, \; a.s.$$

precise asymptotics can also be established on

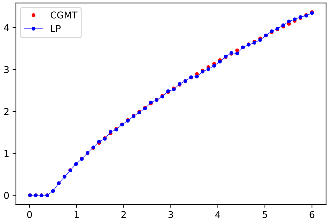
$$\text{Angle:} \quad \frac{\langle \hat{\theta}_{\ell_1}, \theta_\star \rangle_\Lambda}{\|\hat{\theta}_{\ell_1}\|_\Lambda \|\theta_\star\|_\Lambda}, \qquad \text{Loss:} \quad \sum_{j\in[p]} \ell(\hat{\theta}_{\ell_1,j}, \theta_{\star,j})$$

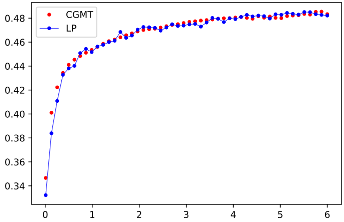
Gaussian comparison: Gordon (1988); Thrampoulidis et al. (2014, 2015, 2018)

ℓ_2-margin: Gardner (1988); Shcherbina and Tirozzi (2003); Deng et al. (2019); Montanari et al. (2019)

THEORY VS. EMPIRICAL

x-axis, varying ψ overparametrization ratio





Margin: $p^{1/2} \cdot \kappa_{\ell_1}(X, y) \to \kappa_\star(\psi, \mu)$

Generalization: $\mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{\ell_1} < 0\right) \to \mathrm{Err}_\star(\psi, \mu)$

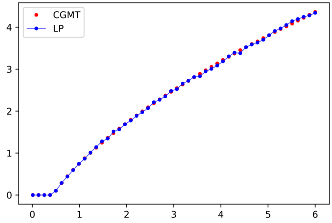
Blue: empirical (numerical solution via linear programming)

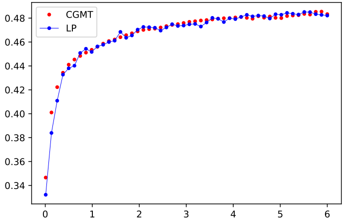
vs.

Red: theoretical (fixed point via non-linear equation system)

THEORY VS. EMPIRICAL

x-axis, varying ψ overparametrization ratio





Margin: $p^{1/2} \cdot \kappa_{\ell_1}(X, y) \to \kappa_\star(\psi, \mu)$

Generalization: $\mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{\ell_1} < 0\right) \to \mathrm{Err}_\star(\psi, \mu)$

Blue: empirical (numerical solution via linear programming)
vs.
Red: theoretical (fixed point via non-linear equation system)

Strikingly Accurate Asymptotics for Breiman's Max Min-Margin!

$$\max_{\|\theta\|_1 \le 1} \min_{1 \le i \le n} \; y_i x_i^\top \theta$$

NON-LINEAR EQUATION SYSTEM: FIXED POINT

[L. & Sur, '20]: $\kappa_\star(\psi, \mu)$ enjoys the analytic characterization via fixed point $c_1(\psi, \kappa), c_2(\psi, \kappa), s(\psi, \kappa)$

define $F_\kappa(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$

$$F_\kappa(c_1, c_2) := \left(\mathbb{E}\left[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2\right]\right)^{\frac{1}{2}} \quad \text{where} \begin{cases} Z_2 \perp (Y, Z_1) \\ Z_i \sim \mathcal{N}(0,1),\ i = 1,2 \\ \mathbb{P}(Y = +1|Z_1) = 1 - \mathbb{P}(Y = -1|Z_1) = f(\rho \cdot Z_1) \end{cases} .$$

NON-LINEAR EQUATION SYSTEM: FIXED POINT

[L. & Sur, '20]: $\kappa_\star(\psi,\mu)$ enjoys the analytic characterization via fixed point $c_1(\psi,\kappa), c_2(\psi,\kappa), s(\psi,\kappa)$

Fixed point equations for $c_1, c_2, s \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ given $\psi > 0$, where the expectation is over $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$

$$c_1 = -\mathop{\mathbb{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left(\frac{\Lambda^{-1/2}W\cdot \mathrm{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)$$

$$c_1^2 + c_2^2 = \mathop{\mathbb{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left(\frac{\Lambda^{-1/2}\,\mathrm{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)^2 .$$

$$1 = \mathop{\mathbb{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left|\frac{\Lambda^{-1}\,\mathrm{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right|$$

$$\text{with}\quad \mathrm{prox}_\lambda(t) = \arg\min_s\left\{\lambda|s| + \frac{1}{2}(s-t)^2\right\} = \mathrm{sgn}(t)\,(|t|-\lambda)_+$$

$$T(\psi,\kappa) := \psi^{-1/2}\,[F_\kappa(c_1,c_2) - c_1\partial_1 F_\kappa(c_1,c_2) - c_2\partial_2 F_\kappa(c_1,c_2)] - s$$

with $c_1(\psi,\kappa), c_2(\psi,\kappa), s(\psi,\kappa)$.

$$\kappa_\star(\psi,\mu) := \inf\{\kappa \ge 0 \,:\, T(\psi,\kappa) \ge 0\}$$

Generalization error, Bayes error, and angle

With $c_i^\star := c_i(\psi, \kappa_\star(\psi, \mu)),\ i = 1, 2.$

$$\mathrm{Err}_\star(\psi, \mu) = \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right)$$

$$\mathrm{BayesErr}(\psi, \mu) = \mathbb{P}\left(Y Z_1 < 0\right)$$

Generalization error, Bayes error, and angle

With $c_i^\star := c_i(\psi, \kappa_\star(\psi, \mu)),\ i = 1, 2.$

$$\mathrm{Err}_\star(\psi, \mu) = \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right)$$

$$\mathrm{BayesErr}(\psi, \mu) = \mathbb{P}\left(Y Z_1 < 0\right)$$

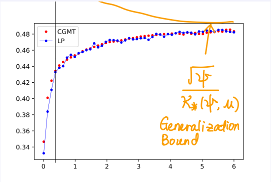
$$\frac{\langle \hat{\theta}_{\ell_1}, \theta_\star \rangle_\Lambda}{\|\hat{\theta}_{\ell_1}\|_\Lambda \|\theta_\star\|_\Lambda} \to \frac{c_1^\star}{\sqrt{(c_1^\star)^2 + (c_2^\star)^2}}$$

Mannor et al. (2002); Jiang (2004); Bartlett and Traskin (2007); Bartlett et al. (2004)

Resolves an open question posed in Breiman '99.

Statistical and Algorithmic implications

significantly improves over prior generalization bounds



overparametrization → faster optimization

overparametrization → sparser solution

SUMMARY

Research agenda: statistical and computational theory for min-norm interpolants

(naive usage of Rademacher complexity, or VC-dim struggles to explain)

SUMMARY

Research agenda: statistical and computational theory for min-norm interpolants

(naive usage of Rademacher complexity, or VC-dim struggles to explain)

- Regression: [L. & Rakhlin '18, AOS], [L., Rakhlin & Zhai '19, COLT]
- Classification: [L. & Sur '20]
- Kernels vs. Neural Networks: [L. & Dou '19, JASA], [L. & Tran-Bach '20]

Thank you!

- **Liang, T.** & Sur, P. (2020). — A Precise High-Dimensional Asymptotic Theory for Boosting and Min-L1-Norm Interpolated Classifiers.

  arXiv:2002.01586

- **Liang, T.**, Tran-Bach, H. (2020). — Mehler's Formula, Branching Process, and Compositional Kernels of Deep Neural Networks.

  arXiv:2004.04767

- **Liang, T.**, Rakhlin, A. & Zhai, X. (2019). — On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels.

  Conference on Learning Theory (COLT), 2020

- **Liang, T.** & Rakhlin, A. (2018). — Just Interpolate: Kernel "Ridgeless" Regression Can Generalize.

  The Annals of Statistics, 2020

- Dou, X. & **Liang, T.** (2019). — Training Neural Networks as Learning Data-adaptive Kernels: Provable Representation and Approximation Benefits.

  Journal of the American Statistical Association, 2020