Theory for Minimum Norm Interpolation: Regression and Classification in High Dimensions

Tengyuan Liang

The University of Chicago Booth School of Business

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

OUTLINE

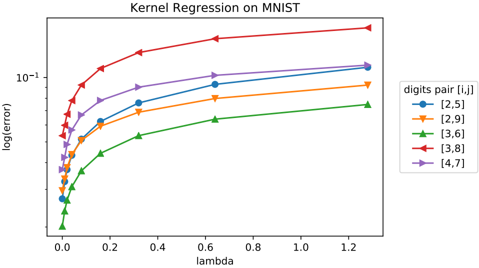
- Motivation: min-norm interpolants
- Regression: multiple descent of risk
- Classification: boosting on separable data

OUTLINE

- Motivation: min-norm interpolants
- Regression: multiple descent of risk
  - application to wide neural networks
  - restricted lower isometry of kernels
  - small-ball property
- Classification: boosting on separable data
  - precise high-dim asymptotics
  - convex Gaussian min-max theorem
  - algorithmic implications on boosting

OVER-PARAMETRIZED REGIME OF STAT / ML

Model class complex enough to interpolate the training data.

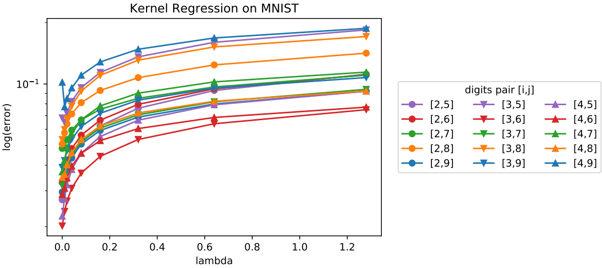


$\lambda = 0$: the interpolants on training data.

MNIST data from LeCun et al. (2010)

OVER-PARAMETRIZED REGIME OF STAT / ML

Model class complex enough to interpolate the training data.



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

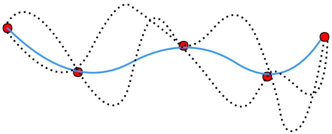
MNIST data from LeCun et al. (2010)

OVER-PARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.

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

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**.

OVER-PARAMETRIZED REGIME OF STAT / ML

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

- over-parametrized linear model and matrix factorization
- kernel machines
- support vector machines
- boosting, AdaBoost
- two-layer ReLU networks

OVER-PARAMETRIZED REGIME OF STAT / ML

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

- over-parametrized linear model and matrix factorization
- kernel machines
- support vector machines
- boosting, AdaBoost
- two-layer ReLU 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

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

Classification

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

REGRESSION

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

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

SHAPE OF RISK CURVE

Classic: U-shape curve

Recent: double descent curve

Belkin, Hsu, Ma, and Mandal (2018); 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 (2018); 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.

SHAPE OF RISK CURVE

We consider the **intrinsic dim.** $d = n^\alpha$ with $\alpha \in (0,1)$.

A **non-linear Kernel Regression** model.

DGP.

- $\{x_i\}_{i=1}^n \overset{i.i.d}{\sim} \mu = \mathcal{P}^{\otimes d}$. distribution of each coordinate $\mathbf{x} \sim \mathcal{P}$ satisfies weak moment $\forall t > 0, \mathbb{P}(|\mathbf{x}| > t) \leq C(1+t)^{-\nu}$.
- 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 $\|\rho_\star\|_\mu \leq C$.

SHAPE OF RISK CURVE

We consider the **intrinsic dim.** $d = n^{\alpha}$ with $\alpha \in (0,1)$.

A **non-linear Kernel Regression** model.

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 $\alpha \in (\frac{1}{\iota+1}, \frac{1}{\iota})$.

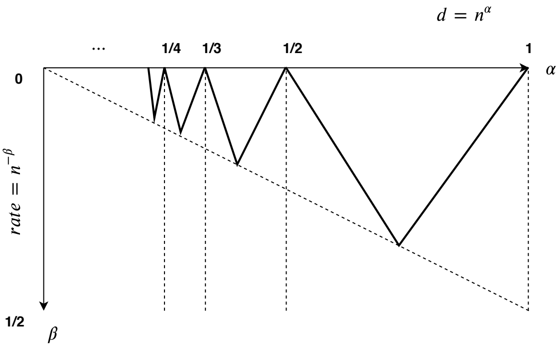
With probability at least $1 - \delta - e^{-n/d^{\iota}}$ on the design $\mathbf{X} \in \mathbb{R}^{n \times d}$,

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

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

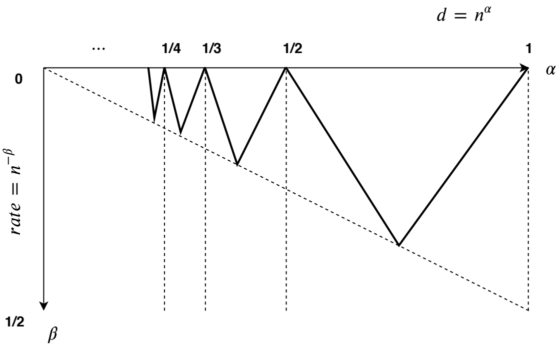
Here the constant $C(\delta, \iota, h, \mathcal{P})$ does not depend on d, n.

MULTIPLE DESCENT



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

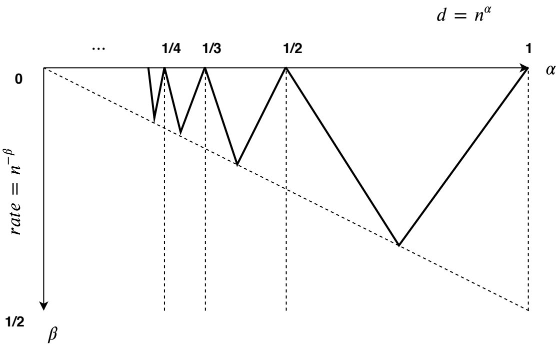
MULTIPLE DESCENT



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

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

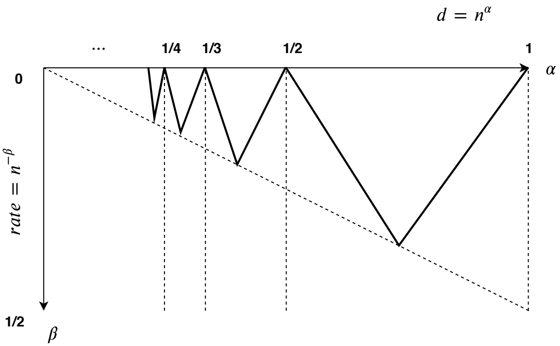
MULTIPLE DESCENT



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

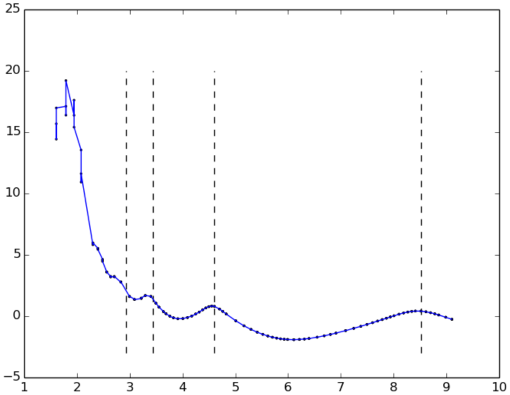
MULTIPLE DESCENT



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

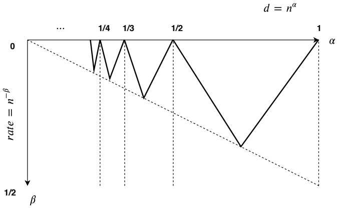
- **valley**: "valley" on the rate curve at $d = n^{\frac{1}{\iota+1/2}}$, $\iota \in \mathbb{N}$
- **over-parametrization**: towards over-parametrized regime, the good rate at the bottom of the valley is better
- **empirical**: preliminary empirical evidence of multiple descent

EMPIRICAL EVIDENCE

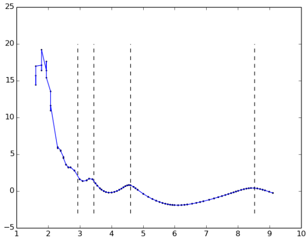


empirical evidence of **multiple-descent behavior** as the scaling $d = n^{\alpha}$ changes.

MULTIPLE DESCENT

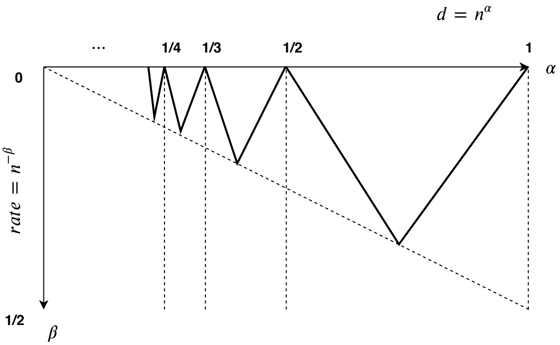


theory



empirical

MULTIPLE DESCENT



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

- $\alpha = 1$: Liang and Rakhlin (2018)
- $\alpha = 0$: Rakhlin and Zhai (2018)
- $\alpha = 1$ double descent: Belkin, Hsu, Ma, and Mandal (2018); Hastie, Montanari, Rosset, and Tibshirani (2019); Bartlett, Long, Lugosi, and Tsigler (2019)
- general α, stair-case, random fourier feature: Ghorbani, Mei, Misiakiewicz, and Montanari (2019)

APPLICATION TO WIDE NEURAL NETWORKS

Neural Tangent Kernel (NTK)

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

$$k_{\mathrm{NTK}}(x, x') = \frac{1}{4\pi} U\left(\frac{\langle x, x'\rangle}{\|x\|\|x'\|}\right)$$

$$U(t) = 3t(\pi - \arccos(t)) + \sqrt{1 - t^2}$$

APPLICATION TO WIDE NEURAL NETWORKS

Neural Tangent Kernel (NTK)

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

$$k_{\text{NTK}}(x, x') = \frac{1}{4\pi} U\left(\frac{\langle x, x'\rangle}{\|x\|\|x'\|}\right)$$

$$U(t) = 3t(\pi - \arccos(t)) + \sqrt{1 - t^2}$$

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

Our results can be generalized to the following type of kernels

$$k(x, x') = \sum_{i=0}^{\infty} \alpha_i \cdot \left(\frac{\langle x, x'\rangle}{\|x\|\|x'\|}\right)^i$$

that include NTK.

Consider integer ι that satisfies $d^{\iota} \log d \lesssim n \lesssim d^{\iota+1} / \log d$, then

$$\text{Risk} \lesssim \frac{d^{\iota}}{n} + \frac{n \log d}{d^{\iota+1}}$$

Ideas behind the proof

Proof Idea: on a **filtration of spaces** indexed by polynomial basis, establish **restricted lower isometry** of the empirical kernel.

filtrated empirical kernel

$$n\mathbf{K}_{ij}^{[\leq \iota]} := \sum_{\substack{r_1,\cdots,r_d \geq 0 \\ r_1+\cdots+r_d \leq \iota}} c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d} p_{r_1\cdots r_d}(x_i) p_{r_1\cdots r_d}(x_j) / d^{r_1+\cdots+r_d}$$

$$n\mathbf{K}^{[\leq \iota]} = \underbrace{\Phi}_{n \times \binom{\iota+d}{\iota}} \cdot \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota} \times n}$$

filtrated sample covariance operator

$$\Theta^{[\leq \iota]} := \frac{1}{n} \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota} \times n} \cdot \underbrace{\Phi}_{n \times \binom{\iota+d}{\iota}}$$

IDEAS BEHIND THE PROOF

Proof Idea: on a **filtration of spaces** indexed by polynomial basis, establish **restricted lower isometry** of the empirical kernel.

filtrated empirical kernel

$$n\mathbf{K}_{ij}^{[\leq \iota]} := \sum_{\substack{r_1,\cdots,r_d \geq 0 \\ r_1+\cdots+r_d \leq \iota}} c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d} p_{r_1\cdots r_d}(x_i) p_{r_1\cdots r_d}(x_j) / d^{r_1+\cdots+r_d}$$

$$n\mathbf{K}^{[\leq \iota]} = \underbrace{\Phi}_{n \times \binom{\iota+d}{\iota}} \cdot \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota} \times n}$$

filtrated sample covariance operator

$$\Theta^{[\leq \iota]} := \frac{1}{n} \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota} \times n} \cdot \underbrace{\Phi}_{n \times \binom{\iota+d}{\iota}}$$

Restricted Lower Isometry of Kernel:
all non-zero eigenvalues of $\mathbf{K}^{[\leq \iota]}$ is lower bounded by $d^{-\iota}$

$$\lambda_{\min}\left(\Theta^{[\leq \iota]}\right) \gtrsim d^{-\iota}$$

IDEAS BEHIND THE PROOF

small-ball approach rather than standard concentration

lower bound $\lambda_{\min}\left(\frac{1}{n}\Psi^\top\Psi\right)$ equiv. $\forall u, \|u\| = 1$, lower bound $\|\Psi u\|^2$

utilize non-negativity

$$\|\Psi u\|^2 = \frac{1}{n}\sum_{i=1}^{n}\langle\Psi(x_i), u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(x_i), u\rangle^2] \cdot \frac{1}{n}\sum_{i=1}^{n} I_{\langle\Psi(x_i),u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(X),u\rangle^2]}$$

small-ball property, $\exists$ constants c_1, c_2

$$\mathbb{P}\left(\langle\Psi(x_i), u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(X), u\rangle^2]\right) \geq c_2$$

Koltchinskii and Mendelson (2015); Mendelson (2014)

which will imply w.p. at least $1 - \exp(-c \cdot n)$

$$\frac{1}{n}\sum_{i=1}^{n} I_{\langle\Psi(x_i),u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(X),u\rangle^2]} \geq c_2/2$$

Non-trivial: verify **small-ball** property for polynomials (weakly dependent) via Paley-Zygmund

CLASSIFICATION

Precise High-Dimensional Asymptotic Theory for Boosting and Min-L_1-Norm Interpolated Classifiers

with Pragya Sur (Harvard)

MIN-L_1-NORM INTERPOLATED CLASSIFIER

Regression so far, what about Classification?

Given n-i.i.d. data pairs $\{x_i, y_i\}_{i=1}^n$ with $y_i \in \{\pm 1\}$ being the labels and $x_i \in \mathbb{R}^p$ being feature vectors.

We consider minimum L_1-norm interpolated classifier:

$$\hat{\theta} = \min_{\theta} \|\theta\|_1, \quad \text{s.t. } y_i x_i^\top \theta \geq 1.$$

when data is separable.

MIN-L_1-NORM INTERPOLATED CLASSIFIER

Regression so far, what about Classification?

Given n-i.i.d. data pairs $\{x_i, y_i\}_{i=1}^n$ with $y_i \in \{\pm 1\}$ being the labels and $x_i \in \mathbb{R}^p$ being feature vectors.

We consider minimum L_1-norm interpolated classifier:

$$\hat{\theta} = \min_{\theta} \|\theta\|_1, \quad \text{s.t. } y_i x_i^\top \theta \geq 1.$$

when data is separable.

min-L_1-norm interpolated classifier agrees with the max-L_1-margin direction

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

Why L_1 margin?

Algorithmic: on separable data, Boosting algorithm $\hat{\theta}^{t,\eta}_{\text{boost}}$ with infinitesimal step-size η agrees with the *min-L_1-norm* direction asymptotically

$$\lim_{\eta \to 0} \lim_{t \to \infty} \hat{\theta}^{t,\eta}_{\text{boost}} / \| \hat{\theta}^{t,\eta}_{\text{boost}} \|_1 = \hat{\theta} \ .$$

Freund and Schapire (1995); Zhang and Yu (2005)

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

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

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

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

Consider high-dim asymptotic regime with over-parametrized ratio

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Statistical.

- how large is the empirical L_1-margin?
- angle between the $\hat{\theta}$ (min-L_1-norm interpolated classifier) and the truth $\theta_\star$?
- generalization properties of Boosting?

Computational.

- iterations of the Boosting (precisely as a function of over-parametrization p/n) are required for an ϵ-approx. to the max-L_1-margin?
- proportion of features activated by Boosting (with zero initialization) when the training error vanishes?

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Theorem (L. & Sur, '20)**.**

Under mild conditions, for $\psi \geq \psi^\star(0)$, the following sharp asymptotic characterization

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

Generalization error

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

Thrampoulidis et al. (2014, 2015, 2018); Gordon (1988)
Montanari et al. (2019); Deng et al. (2019); Shcherbina and Tirozzi (2003); Gardner (1988)

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

$\kappa_\star(\psi, \mu)$ enjoys the analytic characterization: [L. & Sur, '20]

define $F_\kappa : \mathbb{R} \times \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

$\kappa_\star(\psi, \mu)$ enjoys the analytic characterization: [L. & Sur, '20]

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 $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$

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

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

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

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

$\kappa_\star(\psi, \mu)$ enjoys the analytic characterization: [L. & Sur, '20]

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 $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$

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

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

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

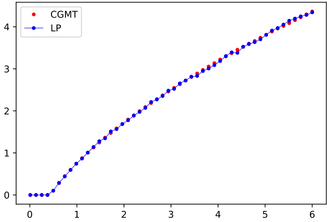
$$\text{with} \quad \mathrm{prox}_\lambda(t) = \arg\min_s \left\{\lambda|s| + \frac{1}{2}(s-t)^2\right\} = \mathrm{sgn}(t)\,(|t| - \lambda)_+$$

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

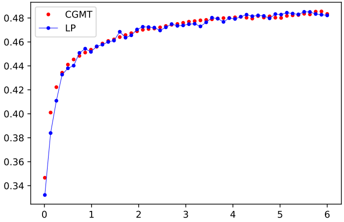
with $c_1(\psi,\kappa), c_2(\psi,\kappa), s(\psi,\kappa)$.

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

THEORY VS. EMPIRICAL



Max-L_1-Margin.



Generalization Error for Min-L_1-Interpolated Classifier.

TECHNICAL REMARKS

Our results builds upon Convex Gaussian Minimax Theorem Thrampoulidis et al. (2014, 2015, 2018); Gordon (1988) and the work on the L_2-margin by Montanari et al. (2019)

L_1 case introduce some technical issues to overcome

- we prove a stronger uniform deviation result that suits the L_1 case, self-normalization property.
- different fixed point equation systems.
- (normalized) max L_1 margin much larger than max L_2 margin.

ALGORITHMIC: BOOSTING

Theorem (L. & Sur, '20).

With proper (non-vanishing) learning rate, the sequence $\{\hat{\theta}^t\}_{t=0}^{\infty}$ obtained by the Boosting algorithm satisfy:

for any $0 < \epsilon < 1$, when the number if iterations

$$t \geq T_\epsilon(p) \quad \text{with} \lim_{n,p\to\infty} \frac{T_\epsilon(p)}{p \log^2 n} = \frac{12\epsilon^{-2}}{\kappa_\star^2(\psi,\mu)} \ ,$$

the solution $\hat{\theta}^t / \|\hat{\theta}^t\|_1$ is an $(1-\epsilon)$-approximation to the Min-L_1-Interpolated Classifier

$$p^{1/2} \cdot \min_{i\in[n]} \frac{y_i x_i^\top \hat{\theta}^t}{\|\hat{\theta}^t\|_1} \in \left[(1-\epsilon)\cdot \kappa_\star(\psi,\mu), \kappa_\star(\psi,\mu)\right] \ .$$

ALGORITHMIC: ACTIVATED FEATURES BY BOOSTING

Theorem (L. & Sur, '20).

Let $S_0(p)$ be the number of features selected when Boosting (for the first time at t) obtains zero training error with $\hat{\theta}^0 = 0$ initialization,

$$\frac{1}{n}\sum_{i=1}^{n} I_{y_i x_i^\top \hat{\theta}^t \leq 0} = 0$$

with

$$S_0(p) := \# \left\{ j \in [p] : \hat{\theta}_j^t \neq 0 \right\} \ .$$

We show

$$\limsup_{n,p\to\infty} \frac{S_0(p)}{p \log^2 p} \leq \frac{12}{\kappa_\star^2(\psi, \mu)} \wedge 1$$

PROOF SKETCH

Step 1: $\sqrt{p}$-rescaling of L_1 ball

$$\xi_{\psi,\kappa}^{(n,p)} := \min_{\|\theta\|_1 \leq \sqrt{p}} \max_{\|\lambda\|_2 \leq 1, \lambda \geq 0} \frac{1}{\sqrt{p}} \lambda^T (\kappa \mathbf{1} - (y \odot X)\theta)$$

It is not hard to see that

$$\xi_{\psi,\kappa}^{(n,p)} = 0, \quad \text{if and only if} \quad \kappa \leq p^{1/2} \cdot \kappa_{\ell_1}\left(\{x_i, y_i\}_{i=1}^n\right) \; ,$$

$$\xi_{\psi,\kappa}^{(n,p)} > 0, \quad \text{if and only if} \quad \kappa > p^{1/2} \cdot \kappa_{\ell_1}\left(\{x_i, y_i\}_{i=1}^n\right) \; .$$

PROOF SKETCH

Step 1: $\sqrt{p}$-rescaling of L_1 ball

$$\xi_{\psi,\kappa}^{(n,p)} := \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T (\kappa \mathbf{1} - (y \odot X)\theta)$$

$$\xi_{\psi,\kappa}^{(n,p)} := \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle\right) - \frac{1}{\sqrt{p}} \boxed{\lambda^T Z \Pi_{w^\perp}(\Lambda^{1/2}\theta)}$$

Step 2: reduction via Gordon's comparison (convex Gaussian min-max theorem)

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

$$\begin{aligned}
&\hat{\xi}_{\psi,\kappa}^{(n,p)} \\
&:= \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle - \tilde{z}\|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right) + \frac{1}{\sqrt{p}} \|\lambda\|_2 \langle g, \Pi_{w^\perp}(\Lambda^{1/2}\theta)\rangle \\
&= \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa \left(\langle w, \Lambda^{1/2}\theta\rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta\right\rangle\right]
\end{aligned}$$

TECHNICAL CHALLENGES IN L_1 CASE

Step 3: large n, p limit

The empirical problem (finite-dim optimization)

$$\hat{\xi}^{(n,p)}_{\psi,\kappa} = \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa \left(\langle w, \Lambda^{1/2}\theta \rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2 \right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right]$$

Let's naively take the limit (infinite-dim optimization)

$$\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa} := \min_{\|h\|_{L_1(\mathcal{Q})} \le 1} \left[\psi^{-1/2} F_\kappa \left(\langle w, \Lambda^{1/2} h \rangle_{L_2(\mathcal{Q})}, \|\Pi_{w^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q})} \right) + \left\langle \Pi_{w^\perp}(G), \Lambda^{1/2} h \right\rangle_{L_2(\mathcal{Q})} \right]$$

One needs to show

$$\lim_{p\to\infty, p/n(p)\to\psi} \hat{\xi}^{(n,p)}_{\psi,\kappa} \overset{\text{a.s.}}{=} \tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}$$

TECHNICAL CHALLENGES IN L_1 CASE

Step 3: large n, p limit

The empirical problem (finite-dim optimization)

$$\hat{\xi}_{\psi,\kappa}^{(n,p)} = \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa \left(\langle w, \Lambda^{1/2}\theta \rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2 \right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right]$$

Let's naively take the limit (infinite-dim optimization)

$$\tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)} := \min_{\|h\|_{L_1(\mathcal{Q})} \le 1} \left[\psi^{-1/2} F_\kappa \left(\langle w, \Lambda^{1/2} h \rangle_{L_2(\mathcal{Q})}, \|\Pi_{w^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q})} \right) + \left\langle \Pi_{w^\perp}(G), \Lambda^{1/2} h \right\rangle_{L_2(\mathcal{Q})} \right]$$

One needs to show

$$\lim_{p\to\infty, p/n(p)\to\psi} \hat{\xi}_{\psi,\kappa}^{(n,p)} \overset{\text{a.s.}}{=} \tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)}$$

L_1 vs. L_2 geometry: for the constraint set $\|\theta\|_1 \le \sqrt{p}$, define

$$c_1 = \langle w, \Lambda^{1/2}\theta \rangle, c_2 = \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2$$

c_2 could be $\sqrt{p} \to \infty$.

TECHNICAL CHALLENGES IN L_1 CASE

Step 3: large n, p limit

The empirical problem (finite-dim optimization)

$$\hat{\xi}^{(n,p)}_{\psi,\kappa} = \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa \left(\langle w, \Lambda^{1/2}\theta \rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2 \right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right]$$

Let's naively take the limit (infinite-dim optimization)

$$\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa} := \min_{\|h\|_{L_1(\mathcal{Q})} \le 1} \left[\psi^{-1/2} F_\kappa \left(\langle w, \Lambda^{1/2} h \rangle_{L_2(\mathcal{Q})}, \|\Pi_{w^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q})} \right) + \left\langle \Pi_{w^\perp}(G), \Lambda^{1/2} h \right\rangle_{L_2(\mathcal{Q})} \right]$$

One needs to show

$$\lim_{p\to\infty, p/n(p)\to\psi} \hat{\xi}^{(n,p)}_{\psi,\kappa} \overset{\text{a.s.}}{=} \tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}$$

L_1 vs. L_2 geometry: for the constraint set $\|\theta\|_1 \le \sqrt{p}$, define

$$c_1 = \langle w, \Lambda^{1/2}\theta \rangle, c_2 = \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2$$

c_2 could be $\sqrt{p} \to \infty$.

[L. & Sur '20] shows uniform deviation over unbounded domain for the fixed-point equation (KKT), using a key self-normalization property of $\partial_i F_\kappa(c_1, c_2)$.

For $i = 1, 2$, we have w.p. at least $1 - n^{-2}$,

$$\sup_{|c_1| \le M, \boxed{c_2 > 0}} |\partial_i \hat{F}_\kappa(c_1, c_2) - \partial_i F_\kappa(c_1, c_2)| \le \frac{C \log n}{\sqrt{n}}$$

[Backup] Convex Gaussian Minimax Theorem

Let $C_1 \subset \mathbb{R}^n, C_2 \subset \mathbb{R}^p$ be two compact sets and let $R : C_1 \times C_2 \to \mathbb{R}$ be a continuous function. Let $X = (X_{i,j}) \in \mathbb{R}^{n\times p}, g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_p)$ be independent vectors and matrices with standard Gaussian entries. Define

$$Q_1(X) = \min_{w_1 \in C_1} \max_{w_2 \in C_2} w_1^\top X w_2 + R(w_1, w_2)$$

$$Q_2(g, h) = \min_{w_1 \in C_1} \max_{w_2 \in C_2} \|w_2\| g^\top w_1 + \|w_1\| h^\top w_2 + R(w_1, w_2).$$

Then

1. For all $t \in \mathbb{R}$,

$$\mathbb{P}(Q_1(X) \leq t) \leq 2\mathbb{P}(Q_2(g, h) \leq t).$$

2. Suppose C_1 and C_2 are both convex, and R is convex concave in (w_1, w_2). Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}(Q_1(X) \geq t) \leq 2\mathbb{P}(Q_2(g, h) \geq t).$$

SUMMARY

Research agenda: statistical or generalization theory for min-norm interpolants

(naive usage of Rademacher complexity, or VC-dim won't explain well)

- Regression: [L. & Rakhlin '18], [L. & Dou '19], [L., Rakhlin & Zhai '19]
- Classification: [L. & Sur '20]

Thank you!

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

   arXiv:2002.01586

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

   arXiv:1908.10292

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

   The Annals of Statistics, to appear

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

   Journal of the American Statistical Association, to appear

PROOF IDEA: RESTRICTED LOWER ISOMETRY

Proof Idea: on a **filtration of spaces** , establish **restricted lower isometry** .

Koltchinskii and Mendelson (2015); Mendelson (2014)

PROOF IDEA: RESTRICTED LOWER ISOMETRY

Proof Idea: on a **filtration of spaces** indexed by polynomial basis, establish **restricted lower isometry** of the empirical kernel.

Define $n\mathbf{K} := [k(x_i, x_j)]_{i,j\in[n]} \in \mathbb{R}^{n\times n}$

$$n\mathbf{K}_{ij} = h\left(\frac{x_i^\top x_j}{d}\right) = \sum_{\iota=0}^{\infty} \alpha_\iota \left(\frac{x_i^\top x_j}{d}\right)^\iota$$
$$= \sum_{r_1,\cdots,r_d\geq 0} c_{r_1\cdots r_d}\, \alpha_{r_1+\cdots+r_d}\, p_{r_1\cdots r_d}(x_i)\, p_{r_1\cdots r_d}(x_j)/d^{r_1+\cdots+r_d}$$

Define **filtrated** empirical kernel

$$n\mathbf{K}_{ij}^{[\leq\iota]} := \sum_{\substack{r_1,\cdots,r_d\geq 0 \\ r_1+\cdots+r_d\leq\iota}} c_{r_1\cdots r_d}\, \alpha_{r_1+\cdots+r_d}\, p_{r_1\cdots r_d}(x_i)\, p_{r_1\cdots r_d}(x_j)/d^{r_1+\cdots+r_d}$$

$c_{r_1\cdots r_d} = \frac{(r_1+\cdots+r_d)!}{r_1!\cdots r_d!}$, $p_{r_1\cdots r_d}(x_i) = (x_i[1])^{r_1}\cdots(x_i[d])^{r_d}$ monomials with multi-index $r_1\cdots r_d$

RESTRICTED LOWER ISOMETRY OF KERNEL

filtrated empirical kernel

$$n\mathbf{K}_{ij}^{[\leq \iota]} := \sum_{\substack{r_1,\cdots,r_d \geq 0 \\ r_1+\cdots+r_d \leq \iota}} c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d} p_{r_1\cdots r_d}(x_i) p_{r_1\cdots r_d}(x_j) / d^{r_1+\cdots+r_d}$$

$$n\mathbf{K}^{[\leq \iota]} = \underbrace{\Phi}_{n\times\binom{\iota+d}{\iota}} \cdot \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota}\times n}$$

filtrated polynomial features

$$\Phi_{i,(r_1\cdots r_d)} = \left(c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d}\right)^{1/2} p_{r_1\cdots r_d}(x_i) / d^{(r_1+\cdots+r_d)/2}$$

filtrated sample covariance operator

$$\Theta^{[\leq \iota]} := \frac{1}{n} \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota}\times n} \cdot \underbrace{\Phi}_{n\times\binom{\iota+d}{\iota}}$$

RESTRICTED LOWER ISOMETRY OF KERNEL

filtrated empirical kernel

$$n\mathbf{K}_{ij}^{[\leq\iota]} := \sum_{\substack{r_1,\cdots,r_d \geq 0 \\ r_1+\cdots+r_d \leq \iota}} c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d} p_{r_1\cdots r_d}(x_i) p_{r_1\cdots r_d}(x_j) / d^{r_1+\cdots+r_d}$$

$$n\mathbf{K}^{[\leq\iota]} = \underbrace{\Phi}_{n\times\binom{\iota+d}{\iota}} \cdot \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota}\times n}$$

filtrated polynomial features

$$\Phi_{i,(r_1\cdots r_d)} = \left(c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d}\right)^{1/2} p_{r_1\cdots r_d}(x_i) / d^{(r_1+\cdots+r_d)/2}$$

filtrated sample covariance operator

$$\Theta^{[\leq\iota]} := \frac{1}{n} \underbrace{\Phi^\top}_{\binom{\iota+d}{\iota}\times n} \cdot \underbrace{\Phi}_{n\times\binom{\iota+d}{\iota}}$$

Restricted Lower Isometry of Kernel:
all non-zero eigenvalues of $\mathbf{K}^{[\leq\iota]}$ is lower bounded by $d^{-\iota}$, i.e.,

$$\lambda_{\min}\left(\Theta^{[\leq\iota]}\right) \gtrsim d^{-\iota}$$

RESTRICTED LOWER ISOMETRY OF KERNEL

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

Assume that Taylor coefficients of h satisfy $\alpha_i > 0 \ \forall i$.

Consider any positive integer ι that satisfy $d^\iota \log d = o(n)$. and $\iota < \nu$. ν is the tail decay of $\mathcal{P}$.

Then with probability at least $1 - \exp(-C \cdot n/d^\iota)$,

$$\text{all non-zero eigenvalues of } \mathbf{K}^{[\leq \iota]} \geq C \cdot d^{-\iota}.$$

RESTRICTED LOWER ISOMETRY OF KERNEL

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

Assume that Taylor coefficients of h satisfy $\alpha_i > 0 \ \forall i$.

Consider any positive integer ι that satisfy $d^{\iota} \log d = o(n)$. and $\iota < \nu$. ν is the tail decay of $\mathcal{P}$.

Then with probability at least $1 - \exp(-C \cdot n/d^{\iota})$,

$$\text{all non-zero eigenvalues of } \mathbf{K}^{[\leq \iota]} \geq C \cdot d^{-\iota}.$$

Some wrong but useful intuition:

- eigenvalues of $\mathbf{K}^{[\leq \iota]}$ equals that of $\Theta^{[\leq \iota]}$

RESTRICTED LOWER ISOMETRY OF KERNEL

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

Assume that Taylor coefficients of h satisfy $\alpha_i > 0\ \forall i$.

Consider any positive integer ι that satisfy $d^\iota \log d = o(n)$. and $\iota < \nu$. ν is the tail decay of $\mathcal{P}$.

Then with probability at least $1 - \exp(-C \cdot n/d^\iota)$,

$$\text{all non-zero eigenvalues of } \mathbf{K}^{[\le\iota]} \ge C \cdot d^{-\iota}.$$

Some wrong but useful intuition:

- eigenvalues of $\mathbf{K}^{[\le\iota]}$ equals that of $\Theta^{[\le\iota]}$
- suppose monomials $\prod_{i=1}^{d}(x[i])^{r_i}$ are orthogonal (wrong), then

$$\mathbb{E}\left[\Theta^{[\le\iota]}\right] = \mathrm{diag}(C(0), \cdots, C(\iota')\cdot d^{-\iota'}, \cdots, \underbrace{C(\iota)\cdot d^{-\iota}}_{\binom{d+\iota-1}{d-1} \text{ such entries}})$$

RESTRICTED LOWER ISOMETRY OF KERNEL

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

Assume that Taylor coefficients of h satisfy $\alpha_i > 0 \ \forall i$.

Consider any positive integer ι that satisfy $d^{\iota} \log d = o(n)$. and $\iota < \nu$. ν is the tail decay of $\mathcal{P}$.

Then with probability at least $1 - \exp(-C \cdot n/d^{\iota})$,

$$\text{all non-zero eigenvalues of } \mathbf{K}^{[\leq \iota]} \geq C \cdot d^{-\iota}.$$

Some wrong but useful intuition:

- eigenvalues of $\mathbf{K}^{[\leq \iota]}$ equals that of $\Theta^{[\leq \iota]}$
- suppose monomials $\prod_{i=1}^{d} (x[i])^{r_i}$ are orthogonal (wrong), then

$$\mathbb{E}\left[\Theta^{[\leq \iota]}\right] = \mathrm{diag}(C(0), \cdots, C(\iota')\cdot d^{-\iota'}, \cdots, \underbrace{C(\iota)\cdot d^{-\iota}}_{\binom{d+\iota-1}{d-1} \text{ such entries}})$$

- even so, standard concentration (fails, at least apply naively)

$$\sup_{u \in B_2^{\binom{d+\iota}{\iota}}} u^{\top} \left(\Theta^{[\leq \iota]} - \mathbb{E}\left[\Theta^{[\leq \iota]}\right] \right) u \leq \frac{1}{\sqrt{n}} \mathrm{Var} \cdots$$

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

Idea 1: Gram-Schimdt process on polynomials, weakly-dependent

$$\{1, t, t^2, \cdots\} \to \{1, q_1(t), q_2(t), \cdots\} \quad q \text{ orthogonal polynomial basis on } L^2_{\mathcal{P}}$$

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

Idea 1: Gram-Schimdt process on polynomials, weakly-dependent

$$\{1, t, t^2, \cdots\} \to \{1, q_1(t), q_2(t), \cdots\} \quad q \text{ orthogonal polynomial basis on } L^2_{\mathcal{P}}$$

$$\Phi_{i,(r_1\cdots r_d)} \to \Psi_{i,(r_1\cdots r_d)} = \left(c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}\right)^{1/2} \prod_{j\in[d]} q_{r_j}(x_i[j])/d^{(r_1+\cdots+r_d)/2}$$

$$\Phi = \Psi\Lambda, \quad \Lambda \in \mathbb{R}^{\binom{\iota+d}{\iota}\times\binom{\iota+d}{\iota}} \quad \text{upper-triangular}$$

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

Idea 1: Gram-Schimdt process on polynomials, weakly-dependent

$$\{1, t, t^2, \cdots\} \to \{1, q_1(t), q_2(t), \cdots\} \quad q \text{ orthogonal polynomial basis on } L^2_{\mathcal{P}}$$

$$\Phi_{i,(r_1 \cdots r_d)} \to \Psi_{i,(r_1 \cdots r_d)} = \left(c_{r_1 \cdots r_d} \alpha_{r_1 + \cdots + r_d}\right)^{1/2} \prod_{j \in [d]} q_{r_j}(x_i[j]) / d^{(r_1 + \cdots + r_d)/2}$$

$$\Phi = \Psi \Lambda, \quad \Lambda \in \mathbb{R}^{\binom{\iota+d}{\iota} \times \binom{\iota+d}{\iota}} \quad \text{upper-triangular}$$

$$\text{Claim: weakly-dependent} \quad \Rightarrow \quad \|\Lambda\|_{\mathrm{op}}, \|\Lambda^{-1}\|_{\mathrm{op}} \le C(\iota)$$

$$u^\top \Theta^{[\le \iota]} u = \frac{1}{n}\|\Phi u\|^2 = \frac{1}{n}\|\Psi \Lambda u\|^2 \ge \lambda_{\min}\left(\frac{1}{n}\Psi^\top \Psi\right)\|\Lambda u\|^2 \asymp \lambda_{\min}\left(\frac{1}{n}\Psi^\top \Psi\right)\|u\|^2$$

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

Idea 2: **small-ball** approach rather than standard concentration

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

Idea 2: **small-ball** approach rather than standard concentration

lower bound $\lambda_{\min}\left(\frac{1}{n}\Psi^\top\Psi\right)$ equiv. $\forall u, \|u\| = 1$, lower bound $\|\Psi u\|^2$

utilize non-negativity

$$\|\Psi u\|^2 = \frac{1}{n}\sum_{i=1}^{n}\langle\Psi(x_i), u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(x_i), u\rangle^2] \cdot \frac{1}{n}\sum_{i=1}^{n} I_{\langle\Psi(x_i),u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(X),u\rangle^2]}$$

small-ball property, $\exists$ constants c_1, c_2

$$\boxed{\mathbb{P}\left(\langle\Psi(x_i), u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(X), u\rangle^2]\right) \geq c_2}$$

which will imply w.p. at least $1 - \exp(-c \cdot n)$

$$\frac{1}{n}\sum_{i=1}^{n} I_{\langle\Psi(x_i),u\rangle^2 \geq c_1\mathbb{E}[\langle\Psi(X),u\rangle^2]} \geq c_2/2$$

Non-trivial: verify **small-ball** property for polynomials (weakly dependent) via Paley-Zygmund

RESTRICTED LOWER ISOMETRY OF KERNEL

Then, how to make it right? **Two Ideas**.

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

all non-zero eigenvalues of $\mathbf{K}^{[\leq \iota]} \geq C \cdot d^{-\iota}$.

Mendelson (2014); Liang et al. (2019); Ghorbani et al. (2019)

INTUITION: WEAKLY DEPENDENT

For any three distinct polynomial features indexed by $(r_1 \cdots r_d), (r'_1 \cdots r'_d), (r''_1 \cdots r''_d)$

$$\prod_{j \in [d]} q_{r_j}(x[j]), \prod_{j \in [d]} q_{r'_j}(x[j]), \prod_{j \in [d]} q_{r''_j}(x[j])$$

Third moment

$$\mathbb{E}\left[q_{r_1 \cdots r_d} q_{r'_1 \cdots r'_d} q_{r''_1 \cdots r''_d}\right] \neq 0$$

only if $\forall j \in [d], r_j + r'_j \geq r''_j$.

Among such triplets, at most $\frac{3^{2\iota}}{d^{\iota}} = O(1/d^{\iota})$ fraction has non-zero third moment.

BACK TO MULTIPLE DESCENT PROOF: SKETCH

Decompose Risk to Bias and Variance. Surprisingly, both terms can be bounded by $\mathbb{E}_{x \sim \mathcal{P}^d} \| k(\mathbf{X}, \mathbf{X})^{-1} k(\mathbf{X}, x) \|^2$.

BACK TO MULTIPLE DESCENT PROOF: SKETCH

Decompose Risk to Bias and Variance. Surprisingly, both terms can be bounded by $\mathbb{E}_{x\sim\mathcal{P}^d}\|k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x)\|^2$.

Sketch:

$$
\begin{aligned}
&\mathbb{E}_x\|k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x)\|^2 \\
&\lesssim \sum_{i=0}^{\iota}\mathbb{E}_x\|\mathbf{K}^{-1}\frac{1}{n}(\mathbf{X}x)^i/d^i\|^2 + \mathbb{E}_x\|\mathbf{K}^{-1}\frac{1}{n}\sum_{i=\iota+1}^{\infty}(\mathbf{X}x)^i/d^i\|^2 \\
&\lesssim \frac{1}{n^2}\sum_{i=0}^{\iota}\mathbb{E}_x\|\mathbf{K}^{-1}(\mathbf{X}x)^i/d^i\|^2 + \|(n\mathbf{K})^{-1}\|_{\mathrm{op}}^2\cdot\mathbb{E}_x\|\sum_{i=\iota+1}^{\infty}(\mathbf{X}x)^i/d^i\|^2 \\
&\lesssim \frac{1}{n^2}\sum_{i=0}^{\iota}\mathbb{E}_x\left[\|(\mathbf{K}^{[\leq i]})^{+}\|_{\mathrm{op}}^2\cdot\|(\mathbf{X}x)^i/d^i\|^2\right] + \frac{n}{d^{\iota+1}} \\
&\lesssim \frac{1}{n^2}\sum_{i=0}^{\iota}\mathbb{E}_x\left[d^{2i}\cdot\|(\mathbf{X}x)^i/d^i\|^2\right] + \frac{n}{d^{\iota+1}} \quad \text{use restricted lower isometry} \\
&\lesssim \frac{d^{\iota}}{n} + \frac{n}{d^{\iota+1}} \quad .
\end{aligned}
$$