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

Tengyuan Liang

CHICAGO BOOTH
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 for over-parametrized models
- Regression: multiple descent of risk for kernels/neural networks
- Classification: precise asymptotics of boosting algorithms

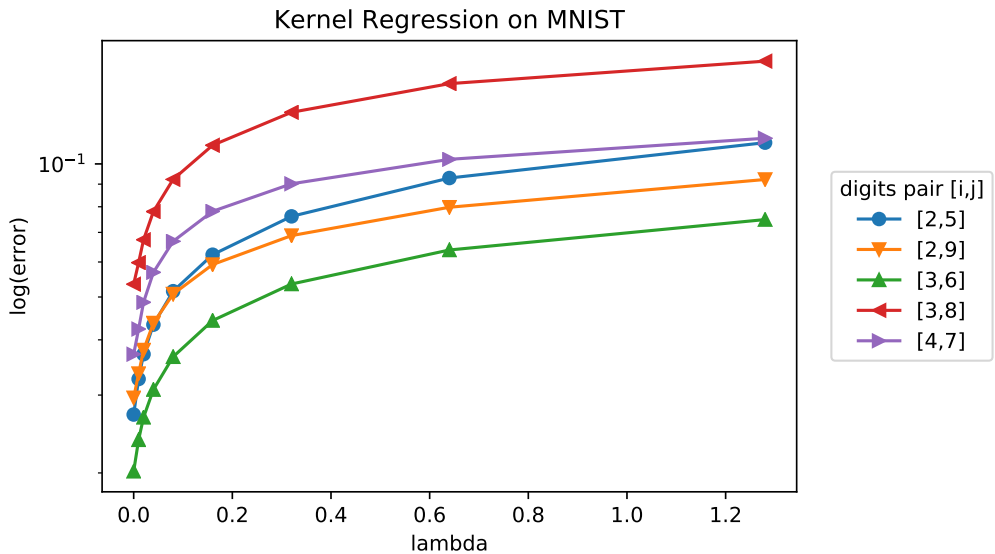
## OVERPARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.
And generalize well on unseen data.

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

Belkin, Hsu, Ma, and Mandal (2018a); Belkin, Hsu, and Mitra (2018b); Belkin, Rakhlin, and Tsybakov (2018c)

Liang and Rakhlin (2018); Bartlett, Long, Lugosi, and Tsigler (2019); Hastie, Montanari, Rosset, and Tibshirani (2019)




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

MNIST data from LeCun, Cortes, and Burges (2010)

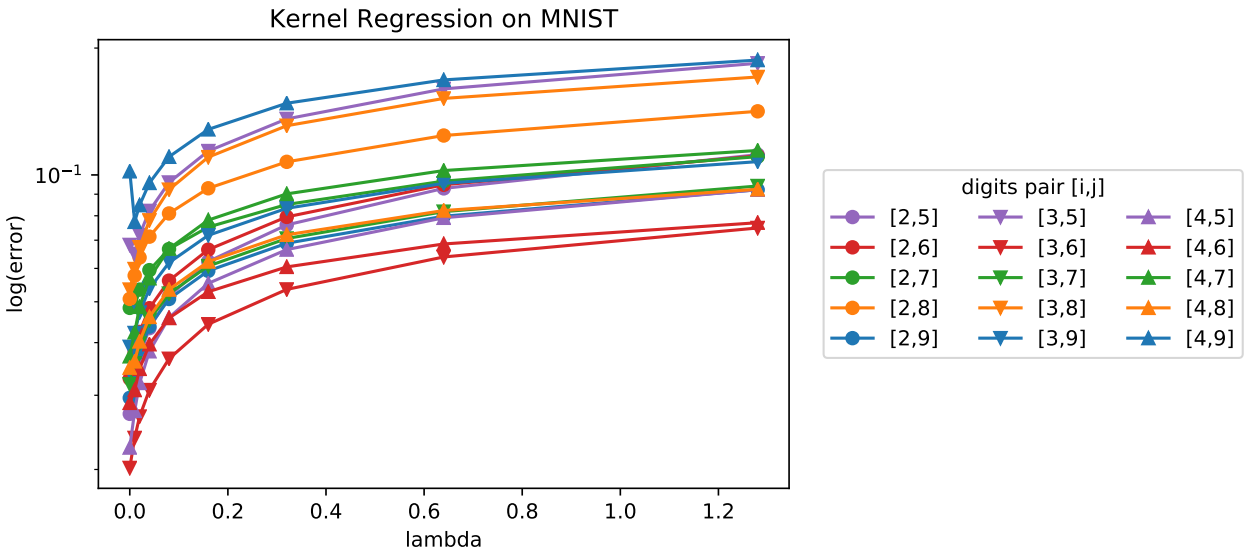
## OVERPARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.
And generalize well on unseen data.

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

Belkin, Hsu, Ma, and Mandal (2018a); Belkin, Hsu, and Mitra (2018b); Belkin, Rakhlin, and Tsybakov (2018c)

Liang and Rakhlin (2018); Bartlett, Long, Lugosi, and Tsigler (2019); Hastie, Montanari, Rosset, and Tibshirani (2019)

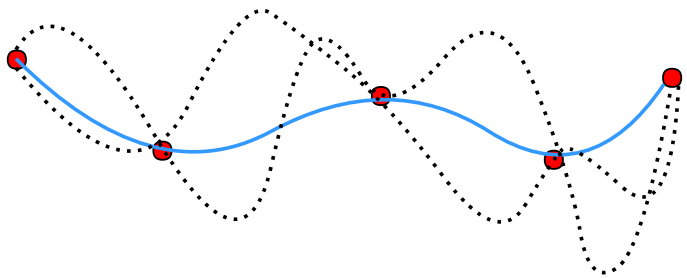



$\lambda = 0$: the interpolants on training data.

MNIST data from LeCun, Cortes, and Burges (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

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

## WHY ERM AND ULLN WOULD NOT SUFFICE

Empirical Risk Minimization (ERM) and Uniform Law of Large Numbers (ULLN) worked nicely to justify classical models

Overparametrized regime: many empirical risk minimizers, yet min-norm interpolator favors one

ERM and ULLN fail to capture the performance of min-norm interpolator

- distribution properties of data
- analytical properties of min-norm interpolator
- growth of the norm of interpolators

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

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

## Regression

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

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

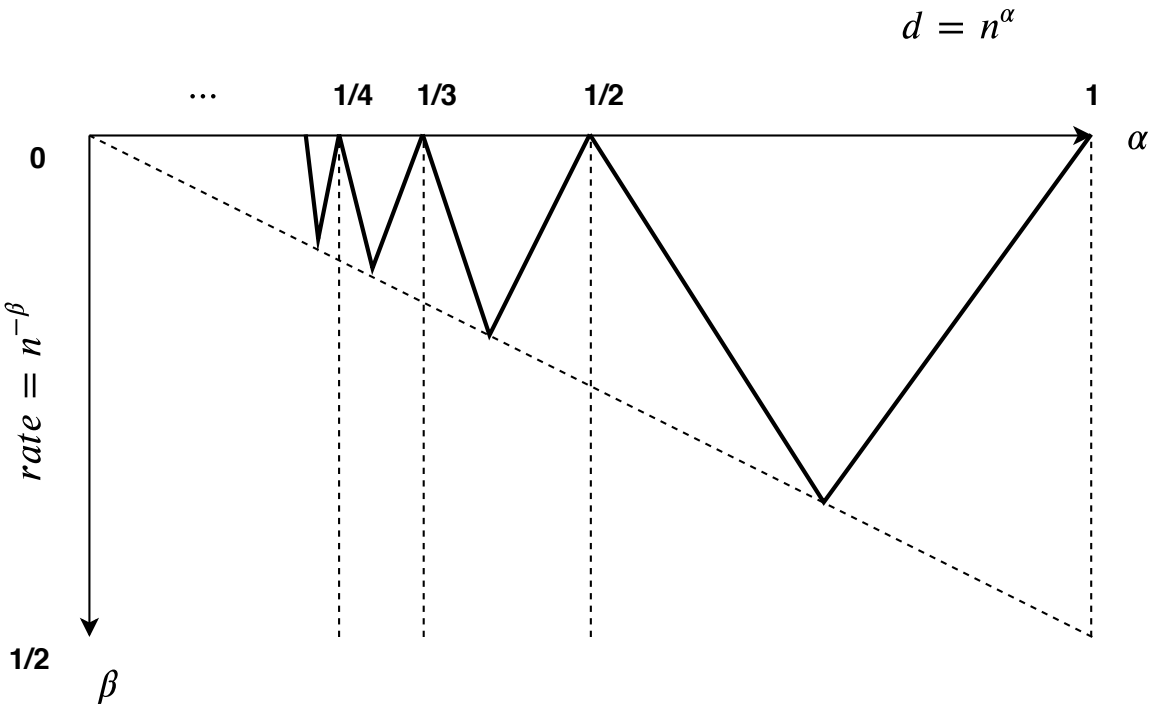
With probability at least $1 - \delta - e^{-n/d^{\iota}}$ on the design $\mathbf{X} \in \mathbb{R}^{n \times d}$,

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

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

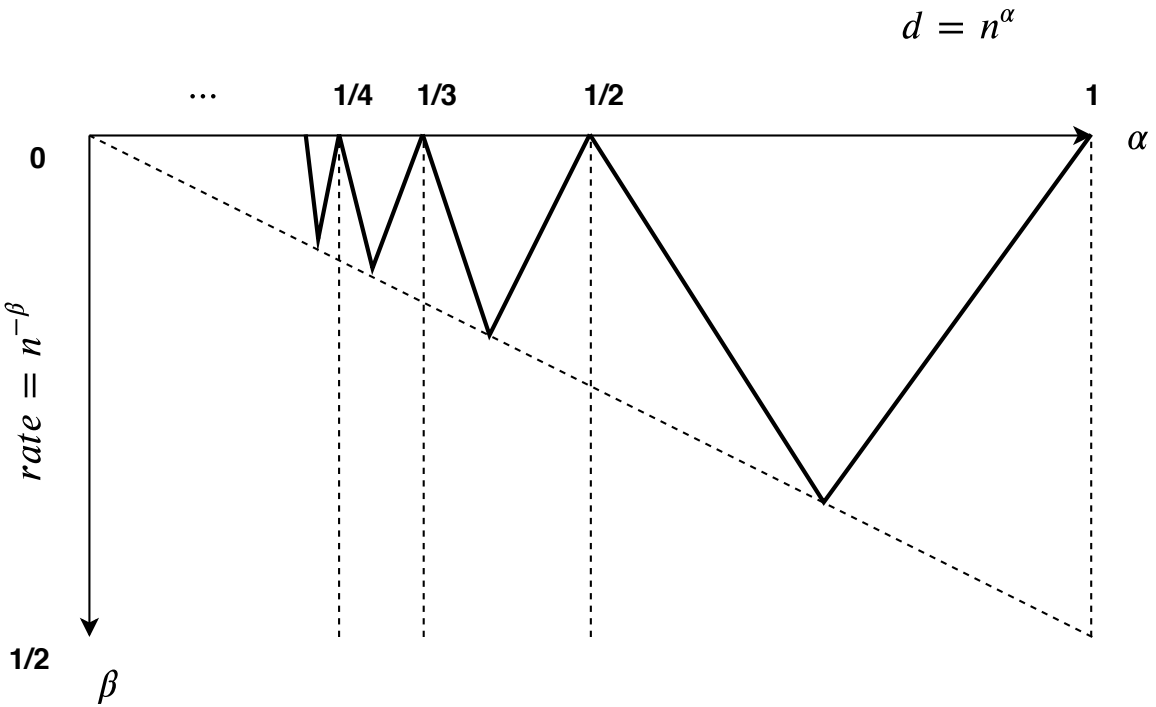
Here the constant $C(\delta, \iota, h, \mathcal{P})$ does not depend on $d, n$.

## MULTIPLE DESCENT




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

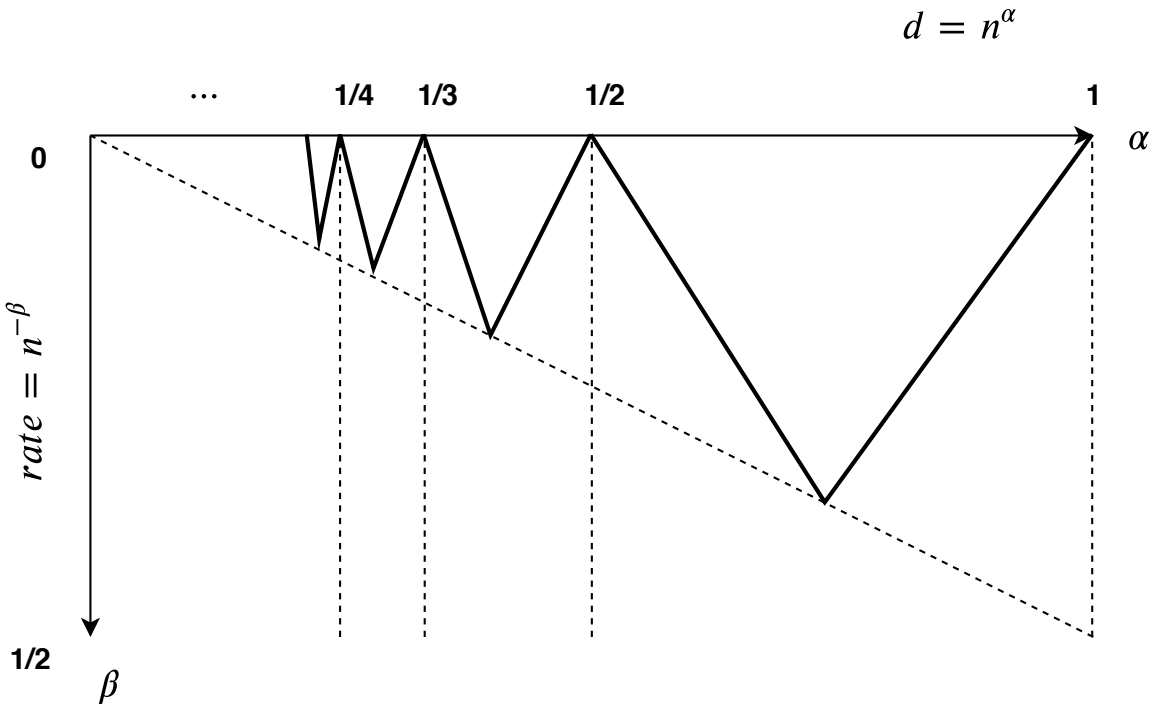
## MULTIPLE DESCENT




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

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

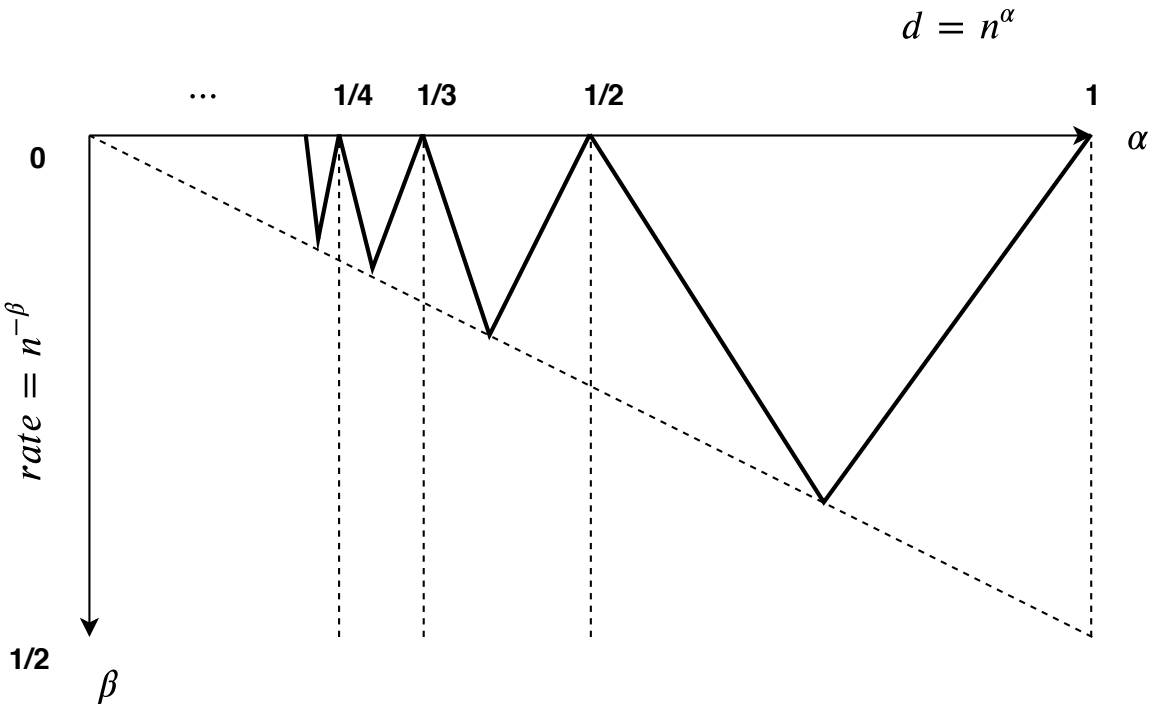
## MULTIPLE DESCENT




**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

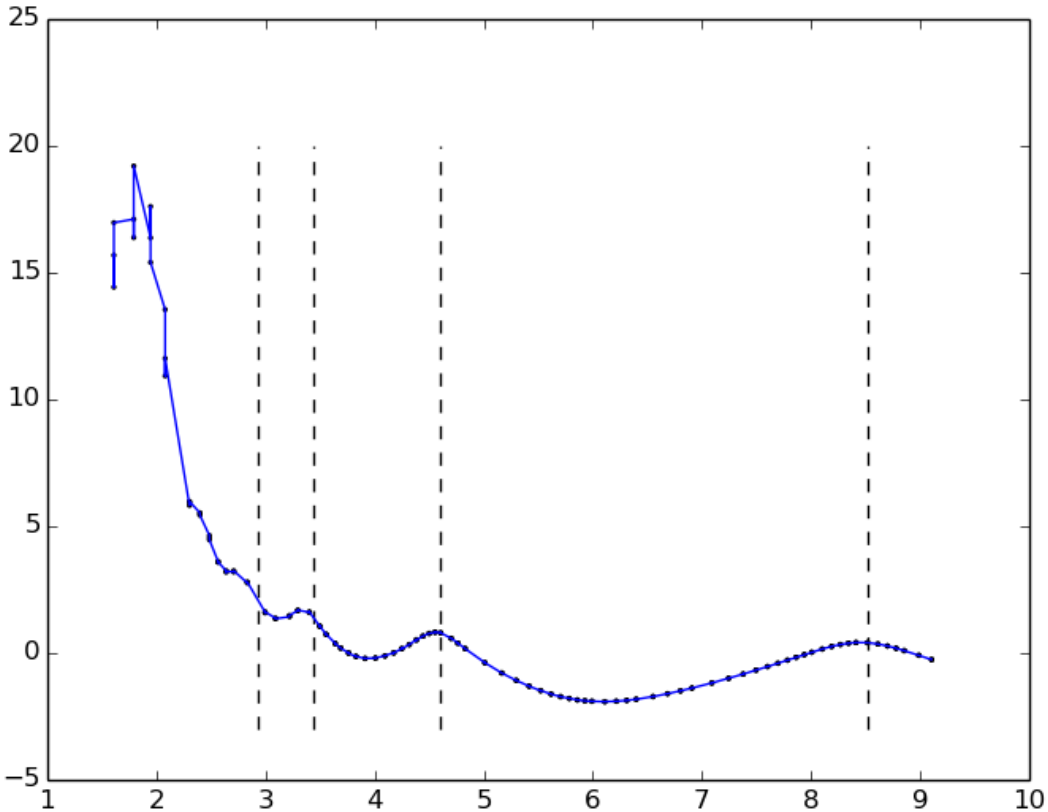



**multiple-descent behavior** of the rates as the scaling $d = n^{\alpha}$ changes.

- **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

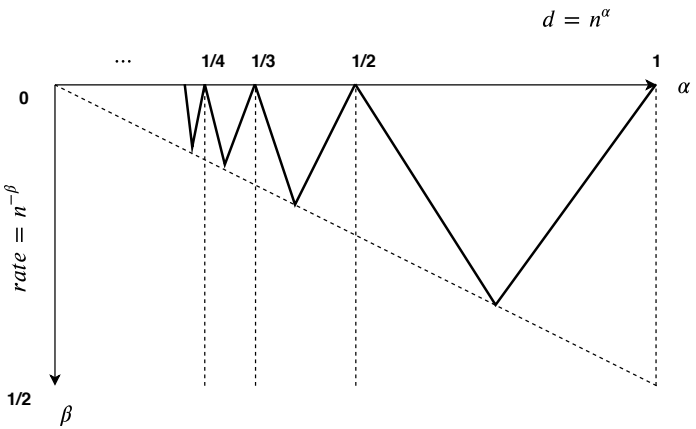
Ghorbani, Mei, Misiakiewicz, and Montanari (2019)

## EMPIRICAL EVIDENCE

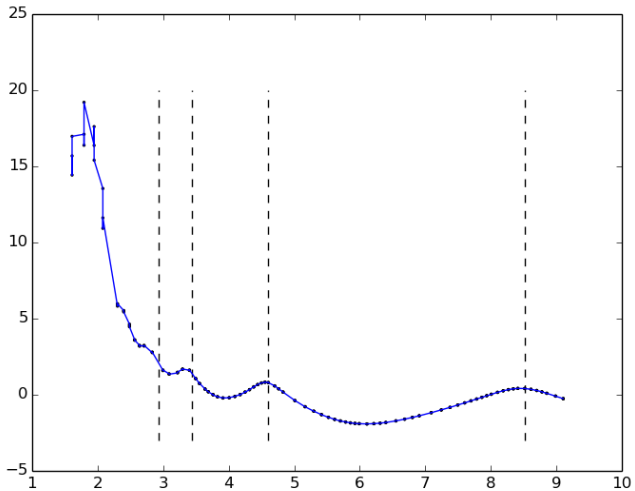



**empirical evidence** of **multiple-descent behavior** as the scaling $d = n^{\alpha}$ changes.

# MULTIPLE DESCENT




theory




empirical

## APPLICATION TO WIDE NEURAL NETWORKS

1. Neural Tangent Kernel (NTK)

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

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

2. Compositional Kernel of Deep Neural Network (DNN)

Rahimi and Recht (2008); Daniely, Frostig, and Singer (2016); Poole, Lahiri, Raghu, Sohl-Dickstein, and Ganguli (2016)
Liang and Tran-Bach (2020); Shankar, Fang, Guo, Fridovich-Keil, Schmidt, Ragan-Kelley, and Recht (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

1. Neural Tangent Kernel (NTK)

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

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

2. Compositional Kernel of Deep Neural Network (DNN)

Rahimi and Recht (2008); Daniely, Frostig, and Singer (2016); Poole, Lahiri, Raghu, Sohl-Dickstein, and Ganguli (2016)
Liang and Tran-Bach (2020); Shankar, Fang, Guo, Fridovich-Keil, Schmidt, Ragan-Kelley, and Recht (2020)

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

**Corollary.**

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

## SOME AFTERTHOUGHTS

1. Index for overparametrization is important

2. Around $d^\iota \approx n$, $d$ could be a poor index of capturing complexity

$$\mathbf{K} \approx \underbrace{\Psi_{\le \iota} \Psi_{\le \iota}^\top}_{\text{low freq.: low degree polynomials}} + \underbrace{\gamma_{>\iota} \cdot \mathbf{I}_n}_{\text{high freq.: implicit regularization}}$$

since the low freq. component $\Psi_{\le \iota} \in \mathbb{R}^{n \times \binom{d+\iota}{\iota}}$ is close to singular, $\|\widehat{f}\|_{\mathcal{H}}$ grows rapidly

3. In practice, multiple descent may be hard to observe, due to high freq. implicit regularization $\gamma_{>\iota}$

Ghorbani, Mei, Misiakiewicz, and Montanari (2019)
d'Ascoli, Sagun, and Biroli (2020); Chen, Min, Belkin, and Karbasi (2020)

Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-$\ell_1$-Norm Interpolated Classifiers
with Pragya Sur (Harvard)

## Classification

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

## 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

Initialize $\theta_0 = \mathbf{0} \in \mathbb{R}^p$, set data weights $\eta_0 = (1/n, \cdots, 1/n) \in \Delta_n$. At time $t \geq 0$:

1. Learner/Feature Selection: $j_t^\star := \arg\max_{j \in [p]} |\eta_t^\top Z \mathbf{e}_j|$, set $\gamma_t = \eta_t^\top Z \mathbf{e}_{j_t^\star}$ ;
2. Adaptive Stepsize: $\alpha_t = \frac{1}{2} \log\left(\frac{1+\gamma_t}{1-\gamma_t}\right)$ ;
3. Coordinate Update: $\theta_{t+1} = \theta_t + \alpha_t \cdot \mathbf{e}_{j_t^\star}$ ;
4. Weight Update: $\eta_{t+1}[i] \propto \eta_t[i] \exp(-\alpha_t y_i x_i^\top \mathbf{e}_{j_t^\star})$, normalized $\eta_{t+1} \in \Delta_n$.

Terminate after $T$ steps, and output the vector $\theta_T$.

Freund and Schapire (1995, 1996)

## 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)

## KEY: EMPIRICAL MARGIN

Empirical margin is **key** to Generalization and Optimization.

Generalization: for all $f(x) = x^\top \theta / \|\theta\|_1$ and $\kappa > 0$,

$$\mathbb{P}\big(\mathbf{y} f(\mathbf{x}) < 0\big) \le \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i f(x_i) < \kappa)}_{\text{empirical margin}} + \underbrace{\sqrt{\frac{\log n \log p}{n\kappa^2}}}_{\text{generalization error}} + \sqrt{\frac{\log(1/\delta)}{n}}, \text{ w.p. } 1-\delta$$

Schapire, Freund, Bartlett, and Lee (1998)

Choose classifier $f$ that maximizes minimal margin $\kappa$

$$\kappa = \max_{\theta \in \mathbb{R}^p} \min_{1 \le i \le n} y_i x_i^\top \theta / \|\theta\|_1$$

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

## KEY: EMPIRICAL MARGIN

Empirical margin is **key** to Generalization and Optimization.

Generalization: for all $f(x) = x^\top \theta / \|\theta\|_1$ and $\kappa > 0$,

$$\mathbb{P}\big(\mathbf{y} f(\mathbf{x}) < 0\big) \le \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i f(x_i) < \kappa)}_{\text{empirical margin}} + \underbrace{\sqrt{\frac{\log n \log p}{n\kappa^2}}}_{\text{generalization error}} + \sqrt{\frac{\log(1/\delta)}{n}}, \text{ w.p. } 1-\delta$$

Schapire, Freund, Bartlett, and Lee (1998)

*"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)

KEY: EMPIRICAL MARGIN

Empirical margin is **key** to Generalization and Optimization.

Optimization: for AdaBoost, $p$-weak learners, $Z := y \circ X \in \mathbb{R}^{n \times p}$

$$\sum_{i=1}^{n} \mathbb{I}(-y_i x_i^\top \theta_T > 0) \le ne \cdot \exp\Big( - \sum_{t=1}^{T} \frac{\gamma_t^2}{2} (1 + o(\gamma_t)) \Big) \ .$$

By Minimax Thm.

$$|\gamma_t| = \|Z^\top \eta_t\|_\infty \ge \min_{\eta \in \Delta_n} \|Z^\top \eta\|_\infty = \min_{\eta \in \Delta_n} \max_{\|\theta\|_1 \le 1} \eta^\top Z \theta = \max_{\|\theta\|_1 \le 1} \min_{1 \le i \le n} \mathbf{e}_i^\top Z \theta \ge \kappa$$

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

Stopping time (zero-training error)

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

## $\ell_1$ GEOMETRY, MARGIN, AND INTERPOLATION

min-$\ell_1$-norm interpolation equiv. max-$\ell_1$-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)

## $\ell_1$ 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 $\ell_1$-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 $f : \mathbb{R} \to [0, 1]$,

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

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

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

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

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

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

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

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)

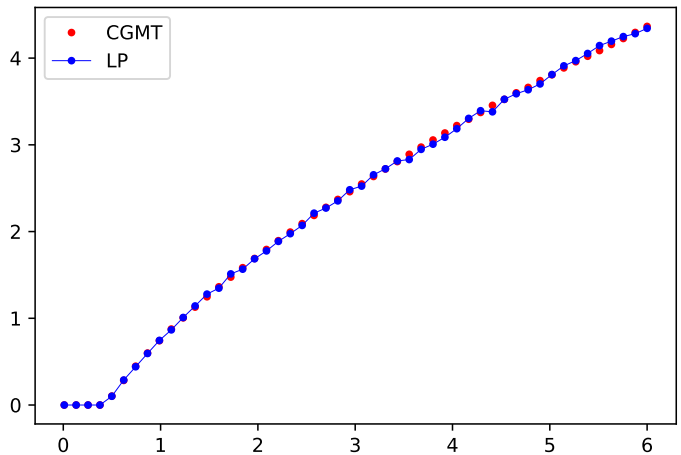
Convex Gaussian comparison: Thrampoulidis, Oymak, and Hassibi (2014, 2015); Thrampoulidis, Abbasi, and Hassibi (2018)

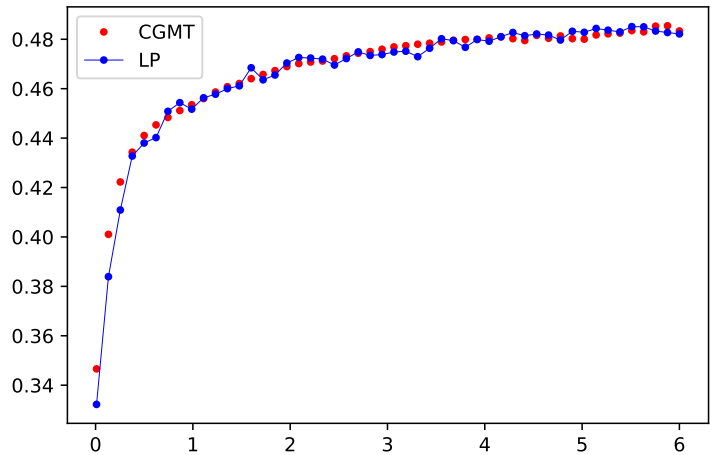
$\ell_2$-margin: Gardner (1988); Shcherbina and Tirozzi (2003)

Montanari, Ruan, Sohn, and Yan (2019); Deng, Kammoun, and Thrampoulidis (2019)

## THEORY VS. EMPIRICAL

$x$-axis, varying $\psi$ overparametrization ratio




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

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

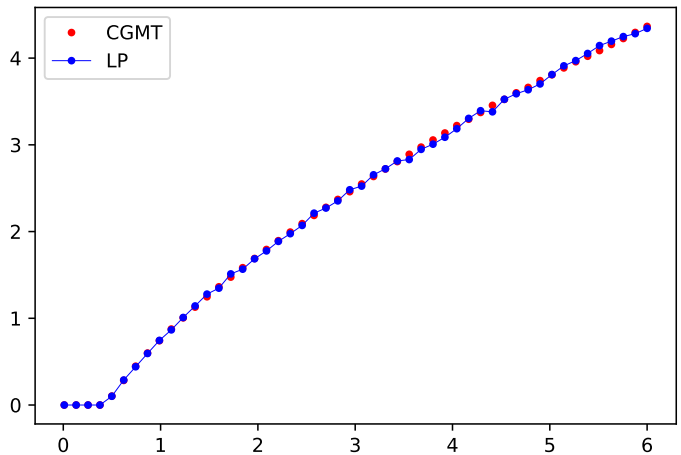
Blue: empirical (numerical solution via linear programming)

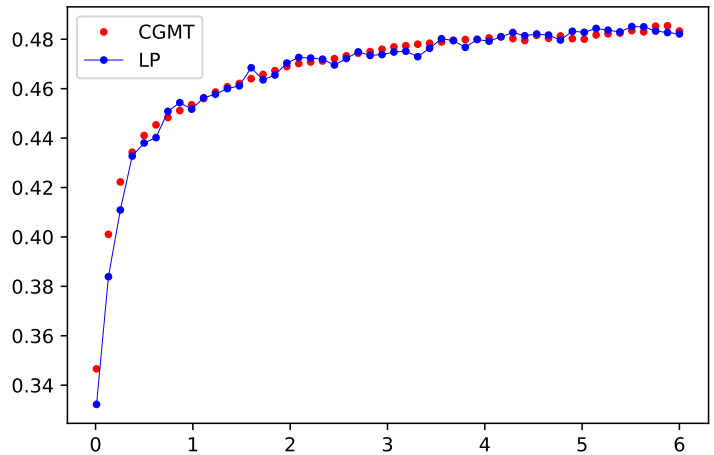
vs.

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

## THEORY VS. EMPIRICAL

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

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

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

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

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

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

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

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

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

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

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

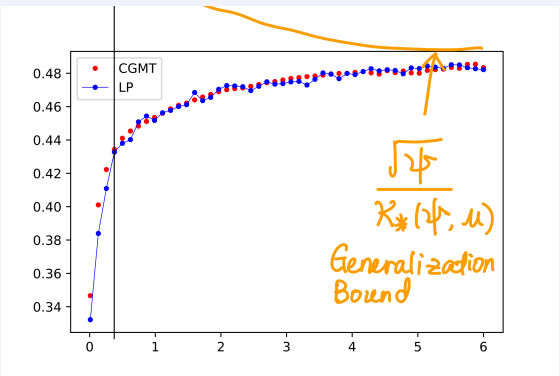
$$\frac{\langle \hat{\theta}_{\ell_1}, \theta_\star \rangle_\Lambda}{\|\hat{\theta}_{\ell_1}\|_\Lambda \|\theta_\star\|_\Lambda} \to \frac{c_1^\star}{\sqrt{(c_1^\star)^2 + (c_2^\star)^2}}$$

Mannor, Meir, and Zhang (2002); Jiang (2004); Lugosi, Vayatis, et al. (2004)
Zhang and Yu (2005); Bartlett and Traskin (2007)

Resolves a question posed in Breiman '99.

Statistical and Algorithmic implications

significantly improves over prior generalization bounds




overparametrization → faster optimization

overparametrization → sparser solution

## BACK TO GENERALIZATION

Known generalization bounds:

$$\text{generalization error} < \frac{1}{\sqrt{n}\,\kappa_{\ell_1}(X,y)} \cdot (\text{log factors, constants})$$

$$= \frac{\sqrt{\psi}}{\kappa_\star(\psi,\mu)} \cdot (\text{log factors, constants})$$

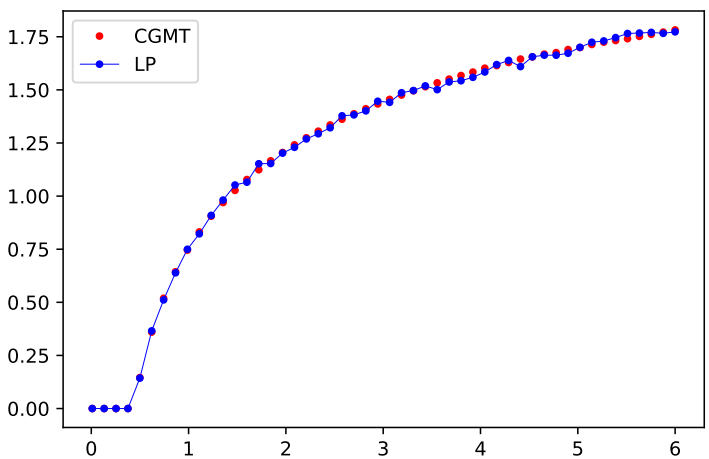
## BACK TO GENERALIZATION

Known generalization bounds:

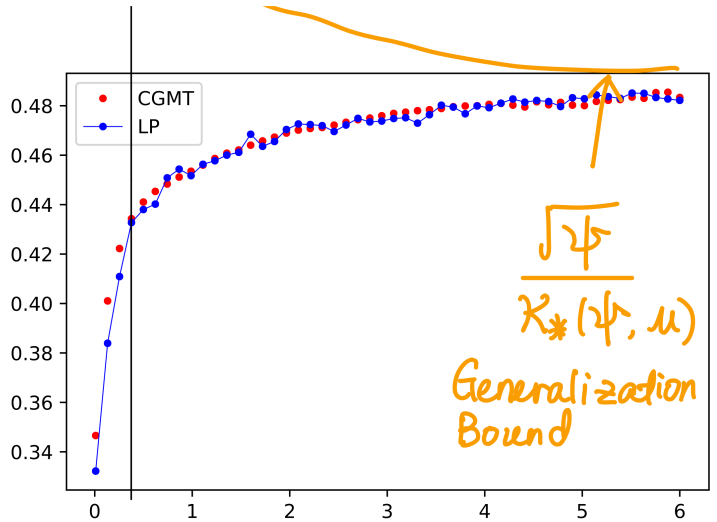
$$\text{generalization error} < \frac{1}{\sqrt{n}\,\kappa_{\ell_1}(X,y)} \cdot (\text{log factors, constants})$$

$$= \frac{\sqrt{\psi}}{\kappa_\star(\psi,\mu)} \cdot (\text{log factors, constants})$$

Let's plot generalization error and $\kappa_\star(\psi,\mu)/\sqrt{\psi}$




$\kappa_\star(\psi,\mu)/\sqrt{\psi}$ against $\psi$




generalization error vs. known bounds

## BACK TO BOOSTING ALGORITHMS

Known computation results:

$$\text{optimization steps} < \frac{1}{\kappa^2_{\ell_1}(X,y)} \cdot (\text{log factors, constants})$$

$$\lim_{s\to 0}\lim_{T\to\infty} \min_{i\in[n]} \frac{y_i x_i^\top \theta^{T,s}_{\text{boost}}}{\|\theta^{T,s}_{\text{boost}}\|_1} = \kappa_{\ell_1}(X,y)$$

## BACK TO BOOSTING ALGORITHMS

Known computation results:

$$\text{optimization steps} < \frac{1}{\kappa^2_{\ell_1}(X,y)} \cdot (\text{log factors, constants})$$

$$\lim_{s\to 0}\lim_{T\to\infty} \min_{i\in[n]} \frac{y_i x_i^\top \theta^{T,s}_{\text{boost}}}{\|\theta^{T,s}_{\text{boost}}\|_1} = \kappa_{\ell_1}(X,y)$$

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

With proper (non-vanishing) stepsize $s$, the sequence $\{\theta^{t,s}_{\text{boost}}\}_{t=0}^{\infty}$ satisfy: for any $0 < \epsilon < 1$, with stopping time

$$t \geq T_\epsilon(p) \quad \text{with} \quad \boxed{\frac{T_\epsilon(p)}{n\log^2 n} \to \frac{12\epsilon^{-2}}{\left(\kappa_\star(\psi,\mu)/\sqrt{\psi}\right)^2}} \ ,$$

the solution approximates the Min-$\ell_1$-Interpolated Classifier

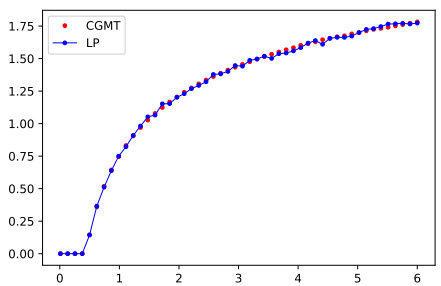
$$p^{1/2} \cdot \min_{i\in[n]} \frac{y_i x_i^\top \theta^{t,s}_{\text{boost}}}{\|\theta^{t,s}_{\text{boost}}\|_1} \in \left[(1-\epsilon)\cdot\kappa_\star(\psi,\mu), \kappa_\star(\psi,\mu)\right] \ .$$

## BACK TO BOOSTING ALGORITHMS

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

With proper (non-vanishing) stepsize $s$, the sequence $\{\theta^{t,s}_{\text{boost}}\}_{t=0}^{\infty}$ satisfy: for any $0 < \epsilon < 1$, with stopping time

$$t \geq T_\epsilon(p) \quad \text{with} \quad \boxed{\frac{T_\epsilon(p)}{n \log^2 n} \to \frac{12\epsilon^{-2}}{\left(\kappa_\star(\psi, \mu)/\sqrt{\psi}\right)^2}} \,,$$




$\kappa_\star(\psi, \mu)/\sqrt{\psi}$ against $\psi$

overparametrization → faster optimization

## ALGORITHMIC: ACTIVATED FEATURES BY BOOSTING

Boosting chooses weak-learner (WL) adaptively. How sparse is $\frac{\text{Selected WL}}{\text{Total WL}}$?

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

Let $S_0(p)$ be the number of weak-learner selected when Boosting hits zero training error $\frac{1}{n}\sum_{i=1}^n \mathbb{I}(y_i x_i^\top \theta^t < 0) = 0$ with initialization $\theta^0 = \mathbf{0}$,

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

We show that

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

## TECHNICAL REMARKS

Our proof build upon Convex Gaussian Minimax Theorem Thrampoulidis, Oymak, and Hassibi (2014, 2015); Thrampoulidis, Abbasi, and Hassibi (2018); Gordon (1988) and is inspired by the work on the $\ell_2$-margin by Montanari, Ruan, Sohn, and Yan (2019).

$\ell_1$-case has technical difficulties to overcome

- we prove a stronger uniform deviation result that suits the $\ell_1$ case, by exploiting a self-normalization property.
- different fixed point equation systems.

(normalized) max $\ell_1$ margin larger than max $\ell_2$ margin

## SOME EXTENSIONS

Our theoretical analysis can be extended to:

1. other geometry:

Max-$\ell_q$-margin, $q \geq 1$, both the statistical theory and algorithmic analysis

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

## SOME EXTENSIONS

Our theoretical analysis can be extended to:

1. other geometry:

Max-$\ell_q$-margin, $q \geq 1$, both the statistical theory and algorithmic analysis

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

2. other models:

- Model misspecification: let $\tilde{x}_i = (x_i, z_i)$,
  $\mathbb{P}(y_i = +1|\tilde{x}_i) = 1 - \mathbb{P}(y_i = -1|\tilde{x}_i) = f\left(\tilde{x}_i^\top \theta_\star\right)$, only $(x_i, y_i)$ is observed
- Gaussian mixture models: $\mathbb{P}(y_i = +1) = 1 - \mathbb{P}(y_i = -1) = \upsilon \in (0, 1)$,
  $x_i|y_i \sim \mathcal{N}(y_i \cdot \theta_\star, \Lambda)$, $\|\theta_\star\| = \rho$
- Models with planted structure in $x$

  Chatterji and Long (2020); Deng, Kammoun, and Thrampoulidis (2019)

Define the following function $F_{\kappa,\sigma}(\cdot,\cdot):\mathbb{R}\times\mathbb{R}_+\to\mathbb{R}$

$$F_\kappa(c_1,c_2):=\left(\mathbb{E}(\kappa-c_1\cdot\mathbf{y}\mathbf{u}-c_2\cdot\mathbf{z})_+^2\right)^{1/2} \tag{2.1}$$

Let us denote two functions related to truncated moment of Gaussian

$$m_0(x):=\mathop{\mathbb{E}}_{\mathbf{G}\sim\mathcal{N}(0,1)}[(x+\mathbf{G})_+^0]=1-\Phi(-x) \tag{2.2}$$

$$m_1(x):=\mathop{\mathbb{E}}_{\mathbf{G}\sim\mathcal{N}(0,1)}[(x+\mathbf{G})_+]=x[1-\Phi(-x)]+\phi(-x) \tag{2.3}$$

$$m_2(x):=\mathop{\mathbb{E}}_{\mathbf{G}\sim\mathcal{N}(0,1)}[(x+\mathbf{G})_+^2]=(1+x^2)[1-\Phi(-x)]+x\phi(-x) \tag{2.4}$$

In the case when $\mathbf{y}=\mathbf{u}$, we have

$$F_\kappa(c_1,c_2)=c_2\sqrt{m_2(\tfrac{\kappa-c_1}{c_2})} \tag{2.5}$$

$$\partial_1F_\kappa(c_1,c_2)=-\frac{m_1(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}} \tag{2.6}$$

$$\partial_2F_\kappa(c_1,c_2)=\sqrt{m_2(\tfrac{\kappa-c_1}{c_2})}-\frac{\kappa-c_1}{c_2}\cdot\frac{m_1(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}}=\frac{m_0(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}} \tag{2.7}$$

It is easy to verify that $\partial_2F_\kappa(c_1,c_2)>0$ for any $(\kappa-c_1)/c_2\neq-\infty$.

The fixed point equations for $(c_1,c_2,s)$ satisfy

$$\mathbb{E}\Big[\theta\eta(\theta,g)\Big]=c_2^{-1}c_1\frac{m_0(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}} \tag{4.5}$$

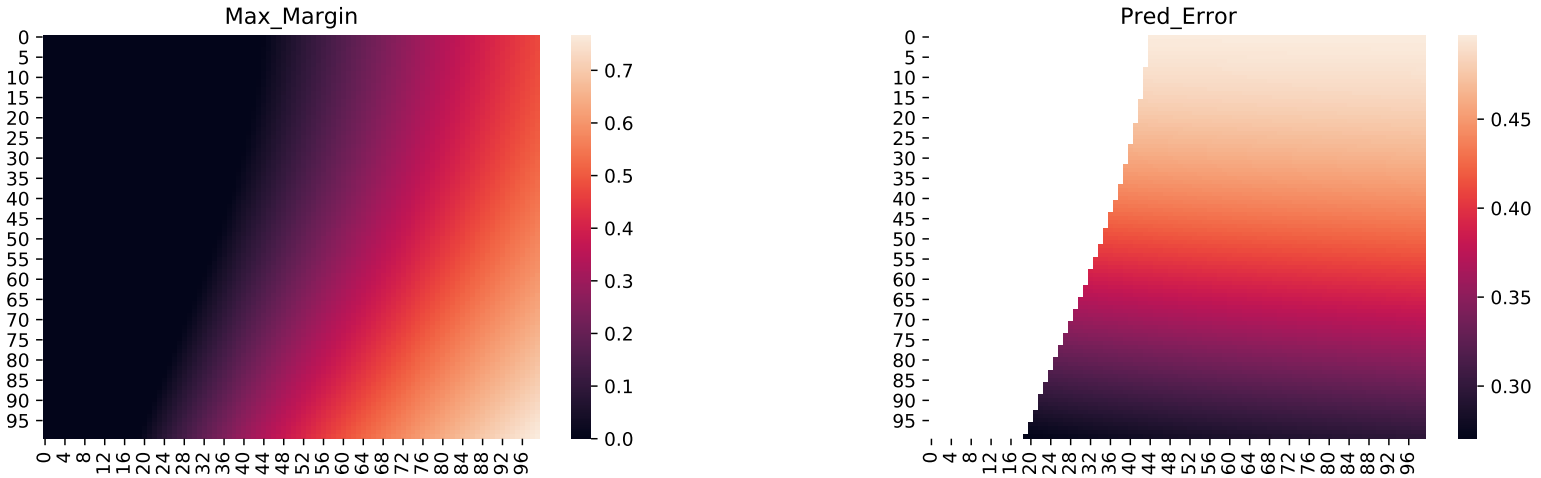
$$\mathbb{E}\Big[|\eta(\theta,g)|\Big]=c_2^{-1}\frac{m_0(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}} \tag{4.6}$$

$$\mathbb{E}\Big[|\eta(\theta,g)|^2\Big]=\left(\frac{m_0(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}}\right)^2 \tag{4.7}$$

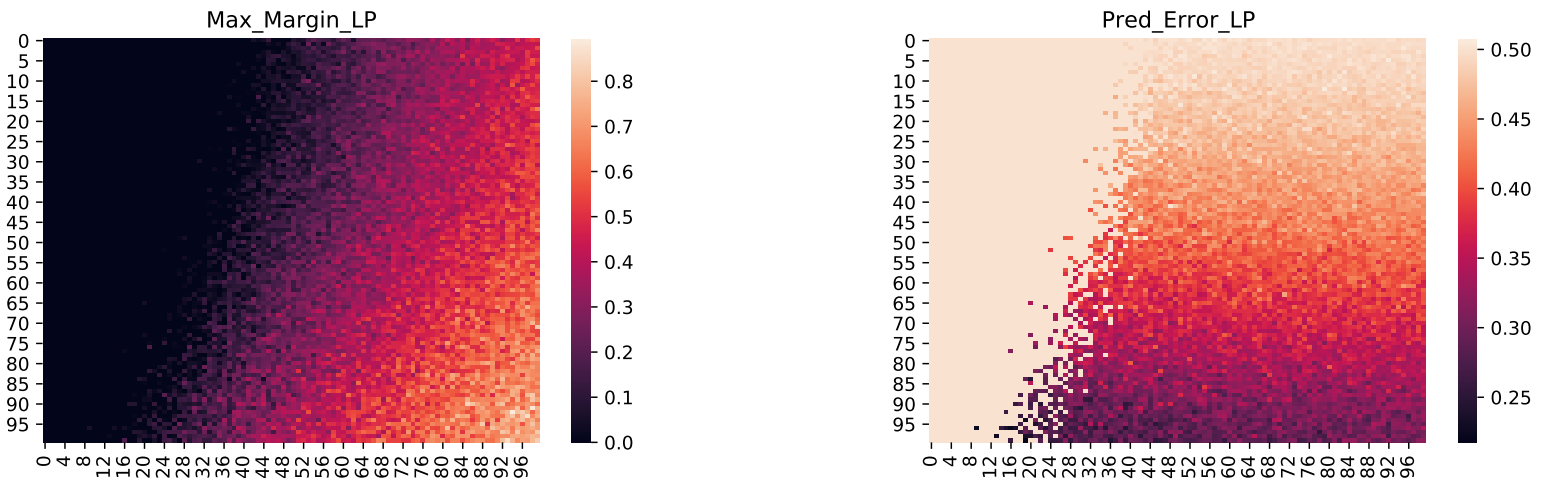
where

$$\eta(\theta,g):=\psi^{1/2}\,\mathrm{prox}_s\left(\psi^{-1/2}\frac{m_1(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}}\cdot\theta+g\right) \tag{4.8}$$

$$s:=\psi^{-1/2}\frac{m_1(\frac{\kappa-c_1}{c_2})}{\sqrt{m_2(\frac{\kappa-c_1}{c_2})}}\kappa \tag{4.9}$$

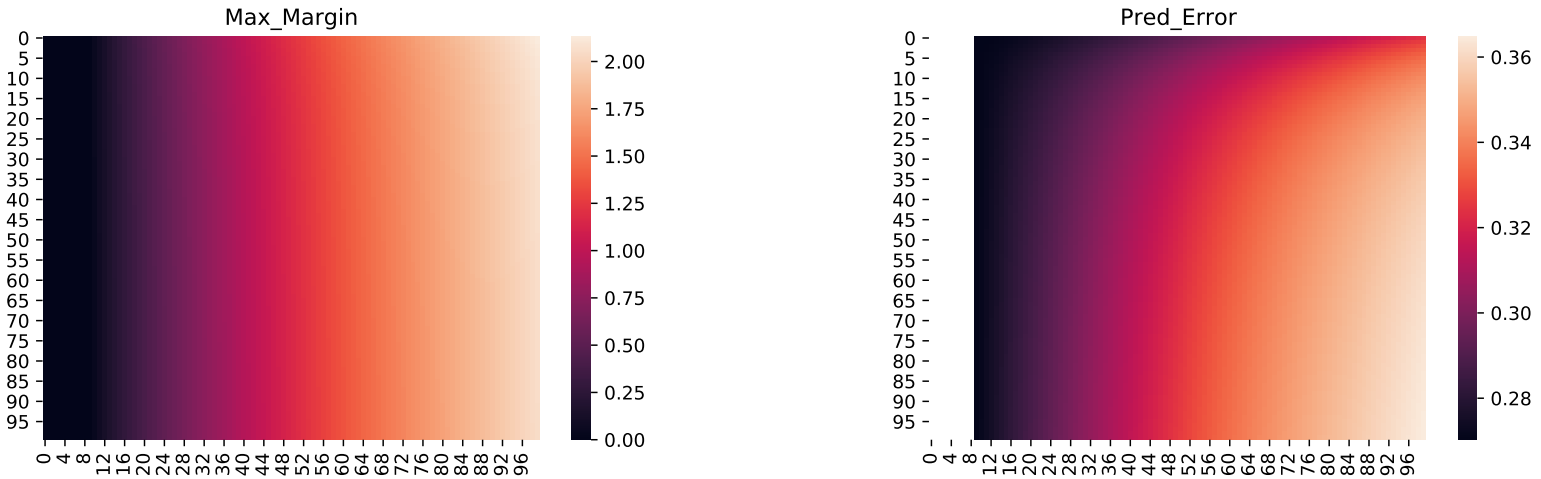



(a) Fixed-point equations.

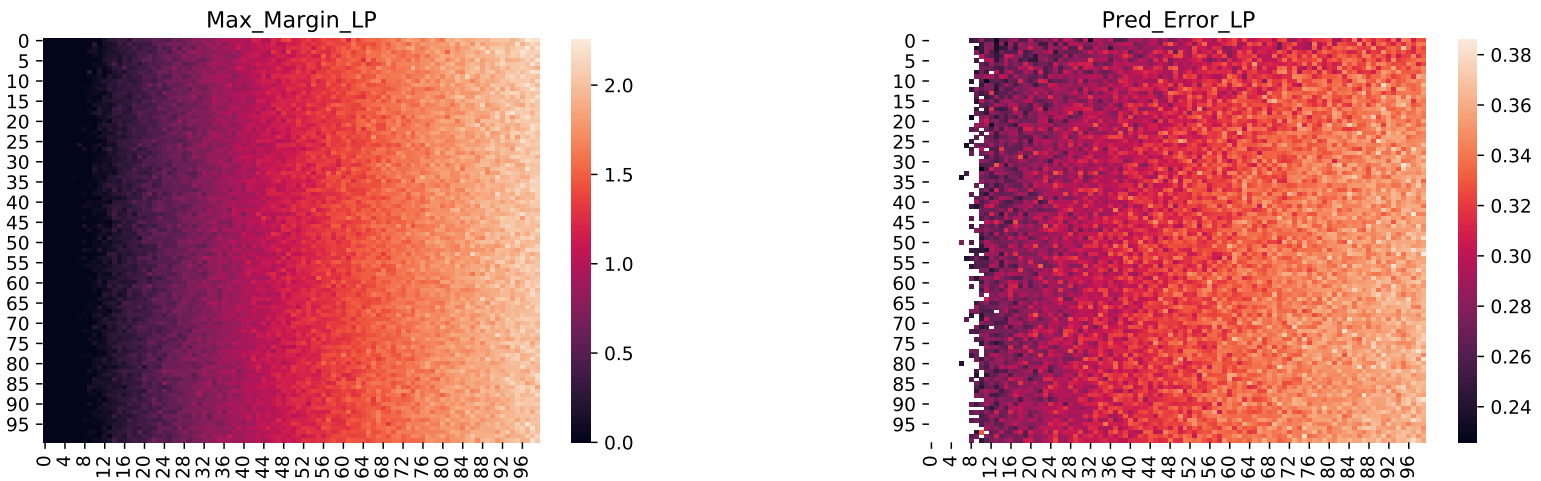



(b) Linear programming.

Diagram for $\ell_2$.



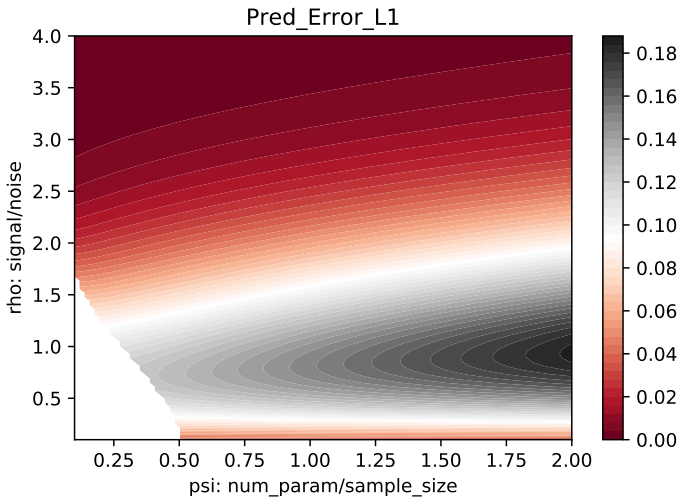

(a) Fixed-point equations.

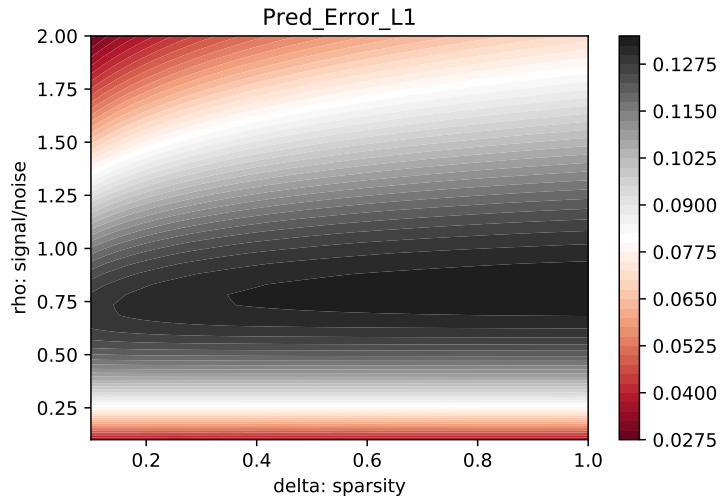



(b) Linear programming.

Diagram for $\ell_1$.

what if Excess Risk, rather than Prediction Error



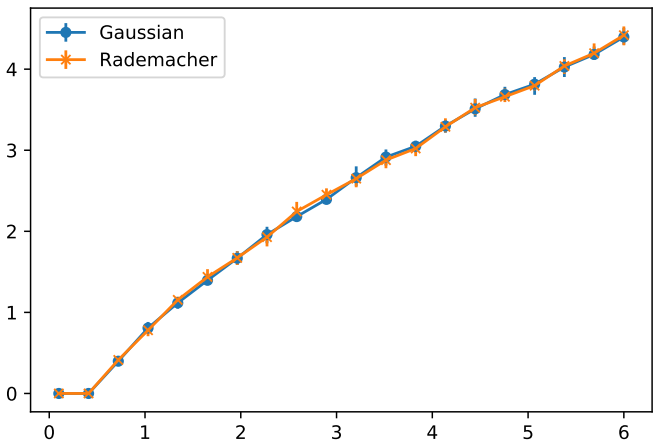



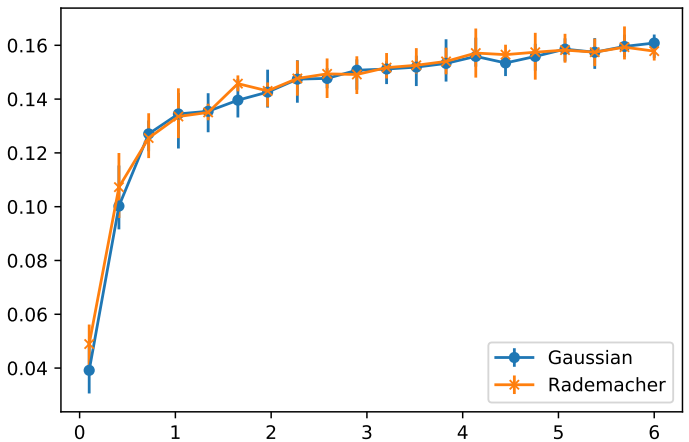

ExcessRisk = PredErr-BayesErr

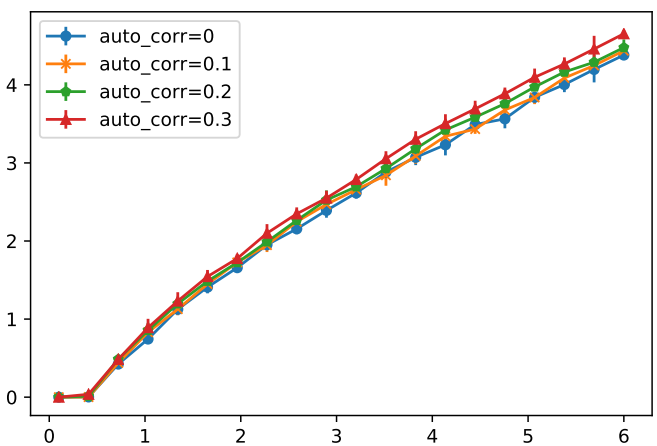
FUTURE WORK

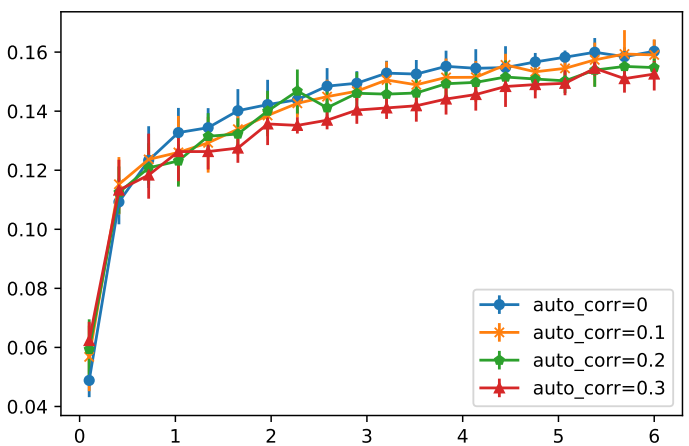
1. Quality of interpolated solution induced by different geometry

2. Beyond Gaussian, universality?



## FUTURE WORK

1. Quality of interpolated solution induced by different geometry

2. Beyond Gaussian, universality?

3. Nonlinear models: kernels, random features

optimization:

- e.g., for 2-homogenous Neural Networks, gradient flow converges to $\ell_1$-margin, feature selection component Bach-Chizat/Amid-Warmuth
- multi-layer Neural networks for classification on separable data, what type of solution gradient descent finds? Srebro-coauthors/Ji-Telgarsky/Chatterji-Long-Bartlett

what are the statistical properties?

SUMMARY

Continuous effort: statistical and computational theory for min-norm interpolants

(naive usage of ERM/ULLN, distribution agnostic VC-theory struggle to explain)

**Many breakthroughs are made by this audience.**

SUMMARY

Continuous effort: statistical and computational theory for min-norm interpolants

(naive usage of ERM/ULLN, distribution agnostic VC-theory struggle to explain)

**Many breakthroughs are made by this audience.**

some of my efforts:

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

# 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.

  *Journal of the American Statistical Association, 2020*

- **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*