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

Tengyuan Liang

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
OVERPARAMETERIZED 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)

\[ \lambda = 0: \text{the interpolants on training data.} \]

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)

\[ \lambda = 0: \text{the interpolants on training data.} \]

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
**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**

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

**Classification**

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

\[ \hat{f} = \arg \min_{f} \|f\|_{\text{norm}}, \quad \text{s.t.} \quad y_i = f(x_i) \ \forall i \in [n]. \]
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 \( f_* (x) := \mathbb{E}[Y|X = x] \), with bounded \( \text{Var}[Y|X = x] \).

**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 \( f_* (x) = \int k(x,z) \rho_* (z) \mu (dz) \) with \( \| \rho_* \|_{\mu} \leq C \).
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 $\tilde{f}$?

$$\tilde{f} = \arg\min_{f} \|f\|_\mathcal{H}, \text{ s.t. } y_i = f(x_i) \forall i \in [n].$$
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 $\alpha \in \left(\frac{1}{\iota+1}, \frac{1}{\iota}\right)$.

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

$$\mathbb{E}\left[\left\|\hat{f} - f^*_\mu\right\|_2^2 | X\right] \leq C \cdot \left(\frac{d^\iota}{n} + \frac{n}{d^{\iota+1}}\right) \asymp n^{-\beta},$$

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

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

\[ d = n^\alpha \]

multiple-descent behavior of the rates as the scaling \( d = n^\alpha \) changes.
MULTIPLE DESCENT

\[ d = n^\alpha \]

multiple-descent behavior of the rates as the scaling \( d = n^\alpha \) changes.

- **valley**: “valley” on the rate curve at \( d = n^{\frac{1}{\ell + 1/2}}, \ell \in \mathbb{N} \)
**MULTIPLE DESCENT**

![Graph showing rates and scaling]

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

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

- **over-parametrization**: towards over-parametrized regime, the good rate at the bottom of the valley is better
**MULTIPLE DESCENT**

\[ d = n^\alpha \]

![Graph showing multiple descent behavior](image)

**multiple-descent behavior** of the rates as the scaling \( d = n^\alpha \) changes.

- **valley**: “valley” on the rate curve at \( d = n^{\frac{1}{i+1/2}}, \ i \in \mathbb{N} \)

- **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

\[
\beta = n^{-\beta}
\]

\[d = n^2\]

---

**theory**

**empirical**
APPLICATION TO WIDE NEURAL NETWORKS

Neural Tangent Kernel (NTK)

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

Compositional Kernel of Deep Neural Network (DNN)

\[ k_{DNN}(x, x') = \sum_{i=0}^{\infty} \alpha_i \cdot \left( \frac{\langle x, x' \rangle}{\|x\| \|x'\|} \right)^i \]

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

Daniely et al. (2016); Poole et al. (2016); Liang and Tran-Bach (2020)
APPLICATION TO WIDE NEURAL NETWORKS

Neural Tangent Kernel (NTK)

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

Compositional Kernel of Deep Neural Network (DNN)

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

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-\(\ell_1\)-Norm Interpolated Classifiers
with Pragya Sur (Harvard)

Classification

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

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

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

Consider when data is **linearly separable**

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

Natural to consider **overparametrized regime**

$$
p/n \rightarrow \psi \in (0, \infty).
$$
**Boosting / AdaBoost**

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


“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)
\[\min \ell_1\text{-norm interpolation equiv. max-}\ell_1\text{-margin}\]

\[
\max \min_{\|\theta\|_1 \leq 1} y_i x_i^T \theta =: \kappa_{\ell_1}(X, y).
\]

Prior understanding:

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

Schapire, Freund, Bartlett, and Lee (1998)

**optimization steps** < \(\frac{1}{\kappa^2}\) \cdot (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 classifier \( \hat{\theta} \) 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 \( f : \mathbb{R} \to [0, 1] \),

\[
\mathbb{P}(y_i = +1|x_i) = 1 - \mathbb{P}(y_i = -1|x_i) = f(x_i^\top \theta^*_\tau),
\]

with some \( \theta^*_\tau \in \mathbb{R}^p \).

Consider high-dim asymptotic regime with overparametrized ratio

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

Signal strength: \( \| \Lambda^{1/2} \theta^*_\tau \| \to \rho \in (0, \infty) \),

Coordinate: \( \bar{w}_j = \sqrt{p} \frac{\lambda_j^{1/2} \theta^*_{\tau,j}}{\rho}, 1 \leq j \leq p \).

Assume

\[
\frac{1}{p} \sum_{j=1}^{p} \delta(\lambda_j, \bar{w}_j) \Rightarrow \mu, \text{ a dist. on } \mathbb{R}_{>0} \times \mathbb{R}
\]
**Theorem (L. & Sur, ’20).**

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

**Margin:** \[
\lim_{n,p \to \infty} \frac{p^{1/2}}{n} \kappa_{\ell_1}(X, y) = \kappa_*(\psi, \mu), \text{ a.s.}
\]

**Generalization error:** \[
\lim_{n,p \to \infty} \mathbb{P}_{x,y} (y \cdot x^\top \hat{\theta}_{\ell_1} < 0) = \text{Err}_* (\psi, \mu), \text{ a.s.}
\]
**Theorem (L. & Sur, ’20).**

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

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

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

precise asymptotics can also be established on

**Angle:**  
$$\frac{\langle \hat{\theta}_{\ell_1}, \theta_\star \rangle \wedge}{\| \hat{\theta}_{\ell_1} \| \wedge \| \theta_\star \| \wedge},$$  

**Loss:**  
$$\sum_{j \in [p]} \ell(\hat{\theta}_{\ell_1,j}, \theta_{\star,j})$$
For $\psi \geq \psi^*$ (separability threshold), sharp asymptotic characterization holds:

Margin: \[
\lim_{n,p \to \infty} p^{1/2} \cdot \kappa_1 (X, y) = \kappa_* (\psi, \mu), \quad a.s.
\]

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

precise asymptotics can also be established on

Angle: \[
\frac{\langle \hat{\theta}_{\ell_1}, \theta_* \rangle \wedge}{\| \hat{\theta}_{\ell_1} \| \wedge \| \theta_* \| \wedge}, \quad \text{Loss: } \sum_{j \in [p]} \ell(\hat{\theta}_{\ell_1,j}, \theta_{*,j})
\]

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

$\ell_2$-margin: Gardner (1988); Shcherbina and Tirozzi (2003); Deng et al. (2019); Montanari et al. (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: $\mathbb{P}_{x,y} \left( y \cdot x^T \hat{\theta}_{\ell_1} < 0 \right) \to \text{Err}_\star(\psi, \mu)$

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

Generalization: $\mathbb{P}_{x,y}(y \cdot x^T \hat{\theta}_{\ell_1} < 0) \rightarrow Err_*(\psi, \mu)$

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!

$$\max_{\|\theta\|_1 \leq 1} \min_{1 \leq i \leq n} y_i x_i^T \theta$$
NON-LINEAR EQUATION SYSTEM: FIXED POINT

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

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

$F_\kappa(c_1, c_2) := \left( \mathbb{E} \left[ (\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2 \right] \right)^{\frac{1}{2}}$ 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]: \( \kappa^* (\psi, \mu) \) 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 \( (\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0, 1) \):

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

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

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

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

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

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

\[
\kappa^* (\psi, \mu) := \inf \{ \kappa \geq 0 : T(\psi, \kappa) \geq 0 \}
\]
GENERALIZATION ERROR, BAYES ERROR, AND ANGLE

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

$$\text{Err}_*(\psi, \mu) = \mathbb{P}(c_1^* Y Z_1 + c_2^* Z_2 < 0)$$

$$\text{BayesErr}(\psi, \mu) = \mathbb{P}(Y Z_1 < 0)$$
GENERALIZATION ERROR, BAYES ERROR, AND ANGLE

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

\[ \text{Err}_*(\psi, \mu) = \mathbb{P}(c_1^* Y Z_1 + c_2^* Z_2 < 0) \]
\[ \text{BayesErr}(\psi, \mu) = \mathbb{P}(YZ_1 < 0) \]

\[ \frac{\langle \hat{\theta}_{\ell_1}, \theta_* \rangle \wedge}{\| \hat{\theta}_{\ell_1} \wedge \theta_* \wedge} \rightarrow \frac{c_1^*}{\sqrt{(c_1^*)^2 + (c_2^*)^2}} \]

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!

  *arXiv:2002.01586*

  *arXiv:2004.04767*

  *Conference on Learning Theory (COLT), 2020*

  *The Annals of Statistics, 2020*

  *Journal of the American Statistical Association, 2020*