Boosting, Min-Norm Interpolated Classifiers, and Overparametrization: a precise asymptotic theory

Tengyuan Liang

The University of Chicago Booth School of Business

joint work with Pragya Sur (Harvard)

OUTLINE

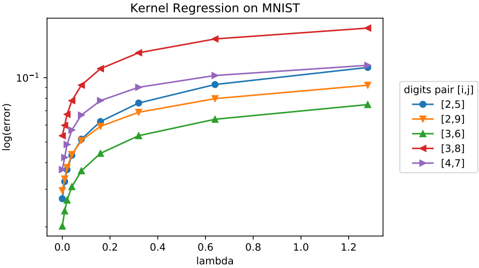
- Motivation: min-norm interpolants under overparametrized regime
- Classification: boosting on separable data
  - precise asymptotics of margin
  - fixed point of a non-linear system of equations
  - statistical and algorithmic implications
- Proof Sketch: Gaussian comparison and convex geometry tools

OVERPARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.

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

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



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

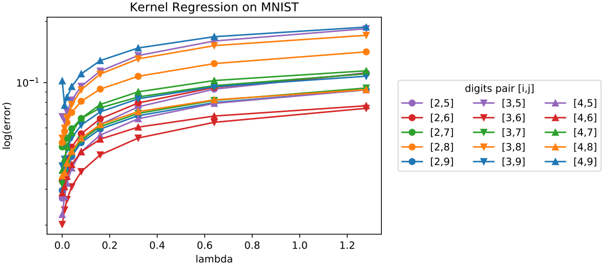
MNIST data from LeCun et al. (2010)

OVERPARAMETRIZED REGIME OF STAT/ML

Model class complex enough to interpolate the training data.

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

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

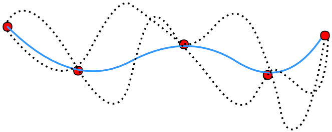


$\lambda = 0$: the interpolants on training data.

MNIST data from LeCun et al. (2010)

OVERPARAMETRIZED REGIME OF STAT/ML

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



Practical methods or algorithms favor certain functions!

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

OVERPARAMETRIZED REGIME OF STAT/ML

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

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

OVERPARAMETRIZED REGIME OF STAT/ML

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

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

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

MIN-NORM INTERPOLANTS

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

Regression

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

Classification

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

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

`tyliang.github.io/Tengyuan.Liang/pdf/Liang-Sur-20.pdf`

Classification

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

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 θ_T.

Freund and Schapire (1995, 1996)

BOOSTING / ADABOOST

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

Wald Lecture, Breiman (2004)

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 κ

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

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

KEY: EMPIRICAL MARGIN

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

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

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

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, $Z := y \circ X \in \mathbb{R}^{n\times p}$

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

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)

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

L_1 GEOMETRY, MARGIN, AND INTERPOLATION

We consider min-L_1-norm interpolated classifier on separable data

$$\hat{\theta}_{\ell_1} = \arg\min_{\theta} \|\theta\|_1, \quad \text{s.t. } y_i x_i^\top \theta \geq 1, \forall i \in [n] \ .$$

Algorithmic: on separable data, Boosting algorithm $\theta_{\text{boost}}^{T,s}$ with infinitesimal step-size s agrees with the min-L_1-norm interpolation asymptotically

$$\lim_{s\to 0}\lim_{T\to\infty} \theta_{\text{boost}}^{T,s} / \|\theta_{\text{boost}}^{T,s}\|_1 = \hat{\theta}_{\ell_1} \ .$$

Freund and Schapire (1995); Rosset et al. (2004); Zhang and Yu (2005)

L_1 GEOMETRY, MARGIN, AND INTERPOLATION

min-L_1-norm interpolation equiv. max-L_1-margin

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

Prior understanding:

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

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

L_1 GEOMETRY, MARGIN, AND INTERPOLATION

Prior understanding:

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

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

However, many questions remain:

Statistical

- how large is the L_1-margin $\kappa_{\ell_1}(X, y)$?
- angle between the interpolated clasifier $\hat{\theta}$ and the truth $\theta_\star$?
- precise generalization error of Boosting? relation to Bayes Error?

Computational

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

Data Generating Process

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

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

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

Consider high-dim asymptotic regime with overparametrized ratio

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

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 some $f : \mathbb{R} \to [0, 1]$,

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

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

Consider high-dim asymptotic regime with overparametrized ratio

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

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

Assume

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Theorem (L. & Sur, '20).

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

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

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

Theorem (L. & Sur, '20).

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

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

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

precise asymptotics can also be established on

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

PRECISE HIGH-DIM ASYMPTOTIC THEORY FOR BOOSTING

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

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

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

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

precise asymptotics can also be established on

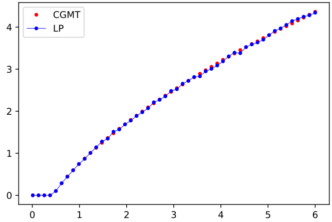
$$\text{Angle:} \quad \frac{\langle \hat{\theta}_{\ell_1}, \theta_\star \rangle_\Lambda}{\|\hat{\theta}_{\ell_1}\|_\Lambda \|\theta_\star\|_\Lambda}, \qquad \text{Loss:} \quad \sum_{j\in[p]} \ell(\hat{\theta}_{\ell_1,j}, \theta_{\star,j})$$

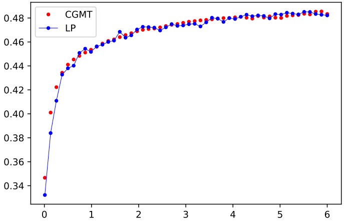
Gaussian comparison: Gordon (1988); Thrampoulidis et al. (2014, 2015, 2018)

L_2-margin: Gardner (1988); Shcherbina and Tirozzi (2003); Deng et al. (2019); Montanari et al. (2019)

THEORY VS. EMPIRICAL

x-axis, varying ψ overparametrization ratio





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

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

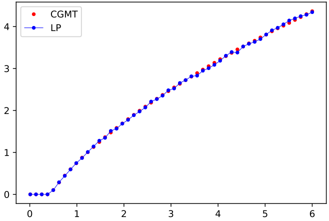
Blue: empirical (numerical solution via linear programming)

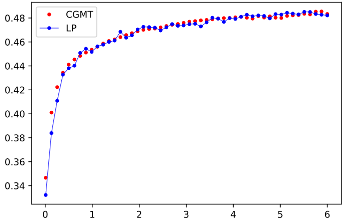
vs.

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

THEORY VS. EMPIRICAL

x-axis, varying ψ overparametrization ratio





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

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

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

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

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

NON-LINEAR EQUATION SYSTEM: FIXED POINT

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

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

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

NON-LINEAR EQUATION SYSTEM: FIXED POINT

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

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

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

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

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

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

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

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

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

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

Statistical and Algorithmic implications

BACK TO GENERALIZATION

Known generalization bounds:

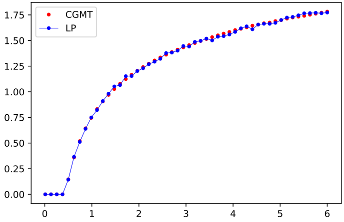
$$\text{generalization error} < \frac{1}{\sqrt{n}\,\kappa_{\ell_1}(X, y)} \cdot (\text{log factors, constants})$$
$$= \frac{\sqrt{\psi}}{\kappa_\star(\psi, \mu)} \cdot (\text{log factors, constants})$$

BACK TO GENERALIZATION

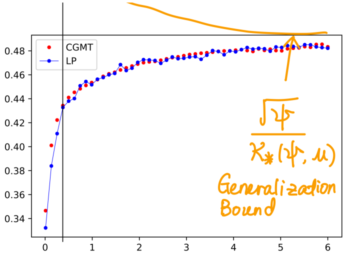
Known generalization bounds:

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

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



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



generalization error vs. known bounds

L_2-margin: Montanari et al. (2019)

BACK TO BOOSTING ALGORITHMS

Known computation results:

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

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

BACK TO BOOSTING ALGORITHMS

Known computation results:

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

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

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

With proper (non-vanishing) stepsize s, the sequence $\{\theta_{\text{boost}}^{t,s}\}_{t=0}^{\infty}$ 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-L_1-Interpolated Classifier

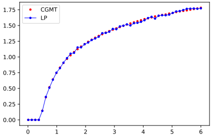
$$p^{1/2} \cdot \min_{i \in [n]} \frac{y_i x_i^\top \theta_{\text{boost}}^{t,s}}{\|\theta_{\text{boost}}^{t,s}\|_1} \in \left[(1 - \epsilon) \cdot \kappa_\star(\psi, \mu), \kappa_\star(\psi, \mu)\right] \ .$$

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

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



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

overparametrization → faster optimization

ALGORITHMIC: ACTIVATED FEATURES BY BOOSTING

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

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 $\theta^0 = \mathbf{0}$,

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

We show that

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

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

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

In the numerical example: overparametrization $\psi > 5$, $\frac{12}{\kappa_\star^2(\psi,\mu)} \ll 1$.

Proof Sketch

Gaussian Comparison + Convex Geometry + New Uniform Convergence

TECHNICAL REMARKS

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

L_1-case has technical difficulties to overcome

- we prove a stronger uniform deviation result that suits the L_1 case, by exploiting a self-normalization property.
- different fixed point equation systems.

(normalized) max L_1 margin much larger than max L_2 margin

PROOF SKETCH

Step 1:

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

It is not hard to see that

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

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

PROOF SKETCH

Step 1:

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

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

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

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

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

Gordon's statement of Slepian-Fernique-Sudakov

Let $\{X_{ij}\}$ and $\{Y_{ij}\}$, $1 \le i \le n$, $1 \le j \le m$, be two centered Gaussian processes which satisfy for all indices:

(i) $\mathbb{E}X_{ij}^2 = \mathbb{E}Y_{ij}^2$,

(ii) $\mathbb{E}(X_{ij}X_{ik}) \ge \mathbb{E}(Y_{ij}Y_{ik})$,

(iii) $\mathbb{E}(X_{ij}X_{\ell k}) \le \mathbb{E}(Y_{ij}Y_{\ell k})$, if $i \neq \ell$.

Then

$$\mathbb{E} \min_i \max_j X_{ij} \le \mathbb{E} \min_i \max_j Y_{ij} .$$

Gordon (1988)

[Backup] Convex Gaussian MinMax Theorem

Let $\Omega_1 \subset \mathbb{R}^n, \Omega_2 \subset \mathbb{R}^p$ be two compact sets and let $U : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a continuous function. Let $Z = (Z_{i,j}) \in \mathbb{R}^{n \times p}, g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_p)$ be independent vectors and matrices with standard Gaussian entries. Define

$$V_1(Z) = \min_{w_1 \in \Omega_1} \max_{w_2 \in \Omega_2} w_1^\top Z w_2 + U(w_1, w_2) \ ,$$

$$V_2(g, h) = \min_{w_1 \in \Omega_1} \max_{w_2 \in \Omega_2} \|w_2\| g^\top w_1 + \|w_1\| h^\top w_2 + U(w_1, w_2) \ .$$

Then

1. For all $t \in \mathbb{R}$,
$$\mathbb{P}(V_1(Z) \leq t) \leq 2\mathbb{P}(V_2(g, h) \leq t) \ .$$
2. Suppose Ω_1 and Ω_2 are both convex, and U is convex concave in (w_1, w_2). Then, for all $t \in \mathbb{R}$,
$$\mathbb{P}(V_1(Z) \geq t) \leq 2\mathbb{P}(V_2(g, h) \geq t) \ .$$

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

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

$$\boxed{\lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \hat{\xi}_{\psi,\kappa}^{(n,p)} \overset{\text{a.s.}}{=} \tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)} \qquad \text{“the a.s. limit”}}$$

TECHNICAL CHALLENGES IN L_1 CASE

Step 3: large n, p limit

The empirical problem (finite-dim optimization)

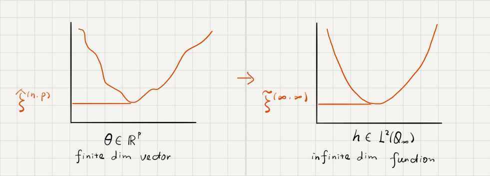
$$\hat{\xi}^{(n,p)}_{\psi,\kappa} = \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa \left(\langle w, \Lambda^{1/2}\theta \rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2 \right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right]$$

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

$$\boxed{\lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \hat{\xi}^{(n,p)}_{\psi,\kappa} \overset{\text{a.s.}}{=} \tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa} \qquad \text{“the a.s. limit”}}$$



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

$$\boxed{\lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \hat{\xi}_{\psi,\kappa}^{(n,p)} \overset{\text{a.s.}}{=} \tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)} \qquad \text{“the a.s. limit”}}$$

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 $\sqrt{p} \to \infty$.

KKT TO SYSTEM OF EQUATIONS

To prove "the a.s. limit", start with the KKT condition

$$\begin{aligned} \Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}\Lambda^{1/2}\left[\partial_1 F_\kappa(c_1,c_2)W + \partial_2 F_\kappa(c_1,c_2)\Pi_{W^\perp}(Z)\right] + s\cdot\partial\|h\|_{L_1(\mathcal{Q}_\infty)} &= 0 \ , \\ s(1-\|h\|_{L_1(\mathcal{Q}_\infty)}) &= 0 \ , \\ s \geq 0, \|h\|_{L_1(\mathcal{Q}_\infty)} &\leq 1 \ . \end{aligned}$$

which implies

$$h^\star = -\frac{\Lambda^{-1}\operatorname{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)} .$$

plugging in the system

$$c_1 = \langle \Lambda^{1/2}h^\star, W\rangle_{L_2(\mathcal{Q}_\infty)}, \qquad c_1^2 + c_2^2 = \|\Lambda^{1/2}h^\star\|^2_{L_2(\mathcal{Q}_\infty)}, \qquad \|h^\star\|_{L_1(\mathcal{Q}_\infty)} = 1$$

UNIFORM DEVIATION ON FIXED POINT EQUATIONS

$$V_1^{(\infty,\infty)}(c_1, c_2, s) :=$$

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

$$V_2^{(\infty,\infty)}(c_1, c_2, s) :=$$

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

$$V_3^{(\infty,\infty)}(c_1, c_2, s) :=$$

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

UNIFORM DEVIATION ON FIXED POINT EQUATIONS

$$V_1^{(\infty,\infty)}(c_1, c_2, s) :=$$
$$c_1 + \mathop{\mathbb{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\left(\frac{\Lambda^{-1/2}W \cdot \mathrm{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)$$
$$V_2^{(\infty,\infty)}(c_1, c_2, s) :=$$
$$c_1^2 + c_2^2 - \mathop{\mathbb{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\left(\frac{\Lambda^{-1/2}\,\mathrm{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)^2$$
$$V_3^{(\infty,\infty)}(c_1, c_2, s) :=$$
$$1 - \mathop{\mathbb{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\left|\frac{\Lambda^{-1}\,\mathrm{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right|,$$

if uniform convergence result holds, in the region $c_1 \in [0, M]$, $c_2 > 0$, $s > 0$

$$\lim_{n\to\infty, p(n)/n=\psi}\ \sup_{c_1\in[0,M], c_2>0, s>0} (c_2 \vee 1)^{-1}|V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,\infty)}(c_1,c_2,s)| = 0$$
$$\lim_{n\to\infty, p(n)/n=\psi}\ \sup_{c_1\in[0,M], c_2>0, s>0} (c_2 \vee 1)^{-2}|V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,\infty)}(c_1,c_2,s)| = 0$$
$$\lim_{n\to\infty, p(n)/n=\psi}\ \sup_{c_1\in[0,M], c_2>0, s>0} (c_2 \vee 1)^{-1}|V_3^{(n,p)}(c_1,c_2,s) - V_3^{(\infty,\infty)}(c_1,c_2,s)| = 0$$

uniform convergence + uniqueness ⇒ "the a.s. limit"

KEY: NEW UNIFORM DEVIATION

We derive uniform deviation over unbounded domain for the fixed-point equations, using a key self-normalization property of $\partial_i F_\kappa(c_1, c_2)$.

[L. & Sur '20] For $i = 1, 2$, we have w.p. at least $1 - n^{-2}$,

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

KEY: NEW UNIFORM DEVIATION

We derive uniform deviation over unbounded domain for the fixed-point equations, using a key self-normalization property of $\partial_i F_\kappa(c_1, c_2)$.

[L. & Sur '20] For $i = 1, 2$, we have w.p. at least $1 - n^{-2}$,

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

$$\partial_1 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbb{E}}_n[YZ_1 \sigma(\kappa - c_1 YZ_1 - c_2 Z_2)]}{(\widehat{\mathbb{E}}_n[\sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)])^{1/2}} = -\frac{\widehat{\mathbb{E}}_n[YZ_1 \sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)]}{(\widehat{\mathbb{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)])^{1/2}}$$

$$\partial_2 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbb{E}}_n[Z_2 \sigma(\kappa - c_1 YZ_1 - c_2 Z_2)]}{(\widehat{\mathbb{E}}_n[\sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)])^{1/2}} = -\frac{\widehat{\mathbb{E}}_n[Z_2 \sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)]}{(\widehat{\mathbb{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)])^{1/2}}$$

where $\sigma(t) := \max(t, 0)$ satisfies the positive homogeneity $\sigma(|c|t) = |c|\sigma(t)$.

- region (i) $(c_1, c_2) \in [-M, M] \times (0, M]$
- region (ii) $(c_1, c_2) \in [-M, M] \times (M, \infty) \Rightarrow (c_2^{-1}, c_1 c_2^{-1}) \in [0, 1/M) \times (-1, 1)$

Large n limit: $\widehat{\mathbb{E}}_n \to \mathbb{E}$, key uniform deviation, self-normalization property.

Large p limit: $\mathcal{Q}_p \to \mathcal{Q}_\infty$, 2-uniform integrability of $\mathcal{Q}_p$ due to W_2.

SOME EXTENSIONS

Our theoretical analysis can be extended to:

1. other geometry:

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

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

SOME EXTENSIONS

Our theoretical analysis can be extended to:

1. other geometry:

Max-L_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: $\mathbb{P}(y_i = +1) = 1 - \mathbb{P}(y_i = -1) = \upsilon \in (0, 1)$, $x_i|y_i \sim \mathcal{N}(y_i \cdot \theta_\star, \Lambda)$
- Models with planted structure in x

FUTURE WORK

1. quality of interpolated solution induced by different geometry
2. beyond Gaussian
3. nonlinear random feature models

SUMMARY

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

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

SUMMARY

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

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

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

Thank you!

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

  `https://tyliang.github.io/Tengyuan.Liang/pdf/Liang-Sur-20.pdf`

- **Liang, T.**, Tran-Bach, H. (2020). — Mehlers Formula, Branching Process, and Compositional Kernels of Deep Neural Networks.

- **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)

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

  The Annals of Statistics

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

  Journal of the American Statistical Association