A precise high-dimensional asymptotic theory for Boosting

Pragya Sur

Harvard University

joint work with Tengyuan Liang (UChicago)

Boosting

- Roots trace back to Valiant ('84)
- Improve generalization weak learning algos. combining them "smartly": Schapire ('90), Freund ('95)
- Adaboost (Freund and Schapire ('95,'96))

BOOSTING

- Roots trace back to Valiant ('84)
- Improve generalization weak learning algos. combining them "smartly": Schapire ('90), Freund ('95)
- Adaboost (Freund and Schapire ('95,'96))

Initialize $\theta_0 = \mathbf{0} \in \mathbb{R}^p$, training examples $\{(x_i, y_i)\}_{i=1}^n$, set data weights $\eta_0 = (1/n, \cdots, 1/n) \in \Delta_n$. $Z = y \circ X$.
At time $t \geq 0$:

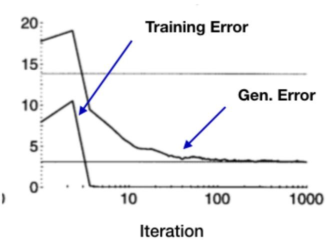
1. Learner/Feature Selection: $j_t^\star := \arg\max_{j \in [p]} |\eta_t^\top Z \mathbf{e}_j|$, set $\gamma_t = \eta_t^\top Z \mathbf{e}_{j_t^\star}$;
2. Adaptive Stepsize: $\alpha_t = \frac{1}{2} \log\left(\frac{1+\gamma_t}{1-\gamma_t}\right)$;
3. Coordinate Update: $\theta_{t+1} = \theta_t + \alpha_t \cdot \mathbf{e}_{j_t^\star}$;
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 ('95,'96)

INTERPOLATION PHENOMENON

Observed long ago for boosting (and bagging, etc)!!

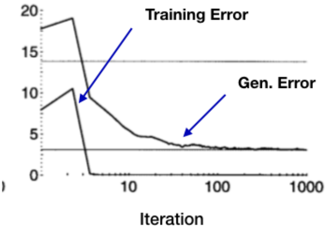


Drucker and Cortes ('96), Quinlan ('96), Breiman ('98),

Interpolation: a common phenomenon now (overparametrized linear regression, kernel regression, neural nets, etc.); Pic. court.:Schapire et al. ('98)

INTERPOLATION PHENOMENON

Observed long ago for boosting (and bagging, etc)!!



Drucker and Cortes ('96), Quinlan ('96), Breiman ('98),

- Search for an explanation—several proposals.

Interpolation: a common phenomenon now (overparametrized linear regression, kernel regression, neural nets, etc.); Pic. court.:Schapire et al. ('98)

INTERPOLATION PHENOMENON

Observed long ago for boosting (and bagging, etc)!!

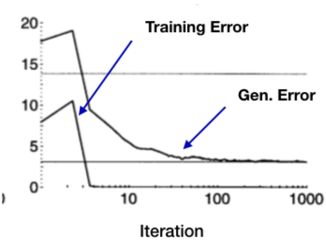
Drucker and Cortes ('96), Quinlan ('96), Breiman ('98), Schapire, Freund, Bartlett and Lee ('98)

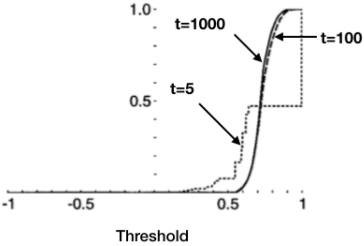
- Search for an explanation—several proposals.
- Key quantity : empirical margin distribution
- Fraction of examples for which $y_i f(x_i)$ is below some threshold

Interpolation: a common phenomenon now (overparametrized linear regression, kernel regression, neural nets, etc.); Pic. court.:Schapire et al. ('98)

Interpolation phenomenon

Observed long ago for boosting (and bagging, etc)!!





Drucker and Cortes ('96), Quinlan ('96), Breiman ('98), Schapire, Freund, Bartlett and Lee ('98)

- Search for an explanation—several proposals.
- Key quantity : empirical margin distribution
- Fraction of examples for which $y_i f(x_i)$ is below some threshold

Interpolation: a common phenomenon now (overparametrized linear regression, kernel regression, neural nets, etc.); Pic. court.:Schapire et al. ('98)

KEY: EMPIRICAL MARGIN

Empirical margin is **key** to Generalization.

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 ('98)

KEY: EMPIRICAL MARGIN

Empirical margin is **key** to Generalization.

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 ('98)

Optimize upper bound: Choose κ to be the max-min ℓ_1 margin:

$$\kappa_{\ell_1} = \max_{\theta \in \mathbb{R}^p} \min_{1 \le i \le n} y_i x_i^\top \theta / \|\theta\|_1$$

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

KEY: EMPIRICAL MARGIN

Empirical margin is **key** to Generalization.

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 ('98)

Optimize upper bound: Choose κ to be the max-min ℓ_1 margin:

$$\kappa_{\ell_1} = \max_{\theta \in \mathbb{R}^p} \min_{1 \le i \le n} y_i x_i^\top \theta / \|\theta\|_1$$

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

- Later improved by Koltchinskii and Panchenko ('02). But, still upper bound!

KEY: THE MAX-MIN-ℓ_1-MARGIN

Margin is **key** to Generalization and Optimization.

Stopping time (zero-training error)

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

Zhang and Yu ('05)

KEY: THE MAX-MIN-ℓ_1-MARGIN

Margin is **key** to Generalization and Optimization.

Stopping time (zero-training error)

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

Zhang and Yu ('05)

- Is this upper bound tight?

An algorithmic insight: Min L_1 norm interpolants

Define the min-L_1-norm interpolated classifier on linearly separable data

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

AN ALGORITHMIC INSIGHT: MIN L_1 NORM INTERPOLANTS

Define the min-L_1-norm interpolated classifier on linearly separable data

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

On linearly separable data, Boosting iterates $\theta_{\text{boost}}^{T,s}$ with infinitesimal stepsize s agrees with the min-L_1-norm interpolant in infinite time limit

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

Rosset et al. ('04), Zhang and Yu ('05)

AN ALGORITHMIC INSIGHT: MIN L_1 NORM INTERPOLANTS

Define the min-L_1-norm interpolated classifier on linearly separable data

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

On linearly separable data, Boosting iterates $\theta_{\text{boost}}^{T,s}$ with infinitesimal stepsize s agrees with the min-L_1-norm interpolant in infinite time limit

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

Rosset et al. ('04), Zhang and Yu ('05)

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

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

An algorithmic insight: Min L_1 norm interpolants

Define the min-L_1-norm interpolated classifier on linearly separable data

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

On linearly separable data, Boosting iterates $\theta_{\text{boost}}^{T,s}$ with infinitesimal stepsize s agrees with the min-L_1-norm interpolant in infinite time limit

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

Rosset et al. ('04), Zhang and Yu ('05)

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

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

An algorithmic insight: Min L_1 norm interpolants

Define the min-L_1-norm interpolated classifier on linearly separable data

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

On linearly separable data, Boosting iterates $\theta_{\text{boost}}^{T,s}$ with infinitesimal stepsize s agrees with the min-L_1-norm interpolant in infinite time limit

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

Rosset et al. ('04), Zhang and Yu ('05)

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

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

- Suggests $\hat{\theta}_{\ell_1}$, κ_{ℓ_1} key players in boosting.

Prior understanding:

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

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

Prior understanding:

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

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

THIS TALK

- Exactly how large is the ℓ_1-margin κ_{ℓ_1}?

Prior understanding:

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

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

THIS TALK

- Exactly how large is the ℓ_1-margin κ_{ℓ_1}?
- What does the limiting object (min norm interpolant) look like?

Prior understanding:

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

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

THIS TALK

- Exactly how large is the ℓ_1-margin κ_{ℓ_1}?
- What does the limiting object (min norm interpolant) look like?
- How long does Boosting take to reach min norm interpolant?

Prior understanding:

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

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

THIS TALK

- Exactly how large is the ℓ_1-margin κ_{ℓ_1}?
- What does the limiting object (min norm interpolant) look like?
- How long does Boosting take to reach min norm interpolant?
- Precise generalization error?

Prior understanding:

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

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

THIS TALK

- Exactly how large is the ℓ_1-margin κ_{ℓ_1}?
- What does the limiting object (min norm interpolant) look like?
- How long does Boosting take to reach min norm interpolant?
- Precise generalization error?
- Other properties: proportion of active weak-learners?
- Understand in a high-dimensional setting?

TOOLS AND INSPIRATION

- Convex Gaussian Minimax Theorem (Gordon('88), Thrampoulidis et al. ('14)),
- max-ℓ_2-margin (Gardner ('88), Shcherbina and Tirozzi ('03), Montanari et al. ('19), Deng et al. ('19)).

Tools and Inspiration

- Convex Gaussian Minimax Theorem (Gordon('88), Thrampoulidis et al. ('14)),
- max-ℓ_2-margin (Gardner ('88), Shcherbina and Tirozzi ('03), Montanari et al. ('19), Deng et al. ('19)).
- But, ℓ_1, ℓ_2 geometries significantly different.
- ℓ_1 lacks important "strong convexity type features" that ℓ_2 has.
- Calls for novel techniques and new uniform convergence arguments.
- different fixed point equation systems

Formal Setting

- High-dim asymptotic regime with overparametrized ratio (No. of samples: n, no. of features: p)
$$p/n \to \psi \in (0, \infty), \quad n, p \to \infty.$$
- Sequence $\{(x_i(n), y_i(n), \theta_\star(n))\}_{i=1}^n$, $x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda(n))$, $\Lambda(n) \in \mathbb{R}^{p \times p}$ diag,
$$\mathbb{P}(y_i = +1 | x_i) = 1 - \mathbb{P}(y_i = -1 | x_i) = f(x_i^\top \theta_\star), \ \theta_\star \in \mathbb{R}^p \ ,$$
- signal strength : $\|\Lambda^{1/2}\theta_\star\| \to \rho \in (0, \infty)$, coordinate : $\bar{w}_j = \sqrt{p}\frac{\lambda_j^{1/2}\theta_{\star,j}}{\rho}, 1 \le i \le p.$
$$\frac{1}{p}\sum_{j=1}^{p} \delta_{(\lambda_j, \bar{w}_j)} \stackrel{\text{Wasserstein-2}}{\Longrightarrow} \mu, \text{ a dist. on } \mathbb{R}_{>0} \times \mathbb{R}$$

FORMAL SETTING

- High-dim asymptotic regime with overparametrized ratio (No. of samples: n, no. of features: p)

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

- Sequence $\{(x_i(n), y_i(n), \theta_\star(n))\}_{i=1}^n$, $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda(n))$, $\Lambda(n) \in \mathbb{R}^{p \times p}$ diag,

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

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

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

- Three problem parameters: ψ, ρ, μ!

Formal Setting

- High-dim asymptotic regime with overparametrized ratio (No. of samples: n, no. of features: p)

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

- Sequence $\{(x_i(n), y_i(n), \theta_\star(n))\}_{i=1}^n$, $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda(n))$, $\Lambda(n) \in \mathbb{R}^{p \times p}$ diag,

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

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

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

- Three problem parameters: ψ, ρ, μ!
- Problem instances linearly separable asymptotically $\leftrightarrow \psi > \psi_\star(\rho)$

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

Logistic regression: Candès and S. ('18); General f: Montanari et al.('19)

The Margin

Recall problem parameters: $p/n \to \psi$; $\|\Lambda^{1/2}\theta_\star\| \to \rho$; $\frac{1}{p}\sum_{j=1}^{p}\delta_{(\lambda_j,\bar{w}_j)} \overset{\text{W-2}}{\Rightarrow} \mu$

Theorem (Liang & S. '20).

For $\psi \geq \psi^\star$ (separability threshold), the max-min-ℓ_1-margin coverges to

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

where

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

THE MARGIN

Recall problem parameters: $p/n \to \psi$; $\|\Lambda^{1/2}\theta_\star\| \to \rho$; $\frac{1}{p}\sum_{j=1}^{p} \delta_{(\lambda_j, \bar{w}_j)} \overset{W\text{-}2}{\Rightarrow} \mu$

Theorem (Liang & S. '20).

For $\psi \geq \psi^\star$ (separability threshold), the max-min-ℓ_1-margin coverges to

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

where

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

THE MARGIN

Recall problem parameters: $p/n \to \psi$; $\|\Lambda^{1/2}\theta_\star\| \to \rho$; $\frac{1}{p}\sum_{j=1}^{p}\delta_{(\lambda_j,\bar{w}_j)} \overset{\text{W-2}}{\Rightarrow} \mu$

Theorem (Liang & S. '20).

For $\psi \geq \psi^\star$ (separability threshold), the max-min-ℓ_1-margin coverges to

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

where

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

- $T_{\psi,\rho,\mu}(\cdot)$ can be explicitly pinned down!
- Related to MLE existence phase transition curve.

THE MARGIN

Recall problem parameters: $p/n \to \psi$; $\|\Lambda^{1/2}\theta_\star\| \to \rho$; $\frac{1}{p}\sum_{j=1}^{p} \delta_{(\lambda_j, \bar{w}_j)} \overset{\text{W-2}}{\Rightarrow} \mu$

Theorem (Liang & S. '20).

For $\psi \geq \psi^\star$ (separability threshold), the max-min-ℓ_1-margin coverges to

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

where

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

- $T_{\psi,\rho,\mu}(\cdot)$ can be explicitly pinned down!
- Related to MLE existence phase transition curve.
- Continuous and non-decreasing.

THE MARGIN LIMIT

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

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

Montanari et al.('19), related to phase transition curve from Candès and S.('18)

THE MARGIN LIMIT

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

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

Montanari et al.('19), related to phase transition curve from Candès and S.('18)

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

with $c_1 \equiv c_1(\psi, \kappa), c_2 \equiv c_2(\psi, \kappa), s \equiv s(\psi, \kappa)$ solves a non-linear system of equations.

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

THE MARGIN LIMIT

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

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

Montanari et al.('19), related to phase transition curve from Candès and S.('18)

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

with $c_1 \equiv c_1(\psi,\kappa), c_2 \equiv c_2(\psi,\kappa), s \equiv s(\psi,\kappa)$ solves a non-linear system of equations.

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

- Proofs for ℓ_1 case require new uniform convergence arguments.
- Discover a key self-normalization property of partial derivatives of $F_\kappa(\cdot,\cdot)$ that fixes the issue.

MIN-ℓ_1-NORM INTERPOLANT

Theorem (Liang and S.('20)).

Min-L_1-norm interpolant admits the gen. error

$$\lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_{\ell_1} < 0\right) = \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- $(Y, Z_1) \perp Z_2; Z_i \sim \mathcal{N}(0,1); \mathbb{P}(Y = +1|Z_1) = 1 - \mathbb{P}(Y = -1|Z_1) = f(\rho \cdot Z_1)$
- $(c_1^\star, c_2^\star, s^\star)$ is unique soln. to non-lin. sys. of eqns parametrized by ψ and the limit of max-min-L_1-margin, $\kappa_\star(\psi, \rho, \mu)$

MIN-ℓ_1-NORM INTERPOLANT

Theorem (Liang and S.('20)).

Min-L_1-norm interpolant admits the gen. error

$$\lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_{\ell_1} < 0\right) = \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- $(Y, Z_1) \perp Z_2; Z_i \sim \mathcal{N}(0,1); \mathbb{P}(Y = +1|Z_1) = 1 - \mathbb{P}(Y = -1|Z_1) = f(\rho \cdot Z_1)$
- $(c_1^\star, c_2^\star, s^\star)$ is unique soln. to non-lin. sys. of eqns parametrized by ψ and the limit of max-min-L_1-margin, $\kappa_\star(\psi, \rho, \mu)$
- $c_1^\star$: Angle, $c_2^\star$: Norm of residual, $s^\star$: Lagrange multipliers for ℓ_1 norm constraints.

MIN-ℓ_1-NORM INTERPOLANT

Theorem (Liang and S.('20)).

Min-L_1-norm interpolant admits the gen. error

$$\lim_{\substack{n,p\to\infty \\ p/n\to\psi}} \mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_{\ell_1}<0\right)=\mathbb{P}\left(c_1^\star YZ_1+c_2^\star Z_2<0\right), \quad a.s.$$

- $(Y,Z_1)\perp Z_2; Z_i\sim\mathcal{N}(0,1); \mathbb{P}(Y=+1|Z_1)=1-\mathbb{P}(Y=-1|Z_1)=f(\rho\cdot Z_1)$
- $(c_1^\star,c_2^\star,s^\star)$ is unique soln. to non-lin. sys. of eqns parametrized by ψ and the limit of max-min-L_1-margin, $\kappa_\star(\psi,\rho,\mu)$
- $c_1^\star$: Angle, $c_2^\star$: Norm of residual, $s^\star$: Lagrange multipliers for ℓ_1 norm constraints.

- The constants $(c_1^\star,c_2^\star,s^\star)$ pin down $\hat{\theta}_{\ell_1}$, e.g. for any convex f_0,

$$\frac{1}{p}\sum_{i=1}^{p} f_0(\hat{\theta}_{\ell_1,1}) \xrightarrow{\text{a.s.}} ??$$

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 (Liang & S. '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 for s.l.n. n, p,

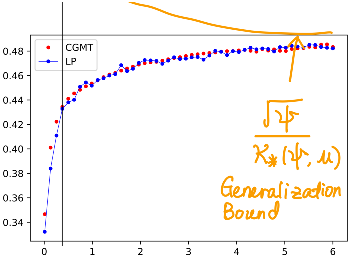
$$p^{1/2} \cdot \min_{i \in [n]} \frac{y_i x_i^\top \theta_{\text{boost}}^{t,s}}{\|\theta_{\text{boost}}^{t,s}\|_1} \in \left[(1-\epsilon) \cdot \kappa_\star(\psi,\mu), \kappa_\star(\psi,\mu)\right] \,.$$

THE GENERALIZATION ERROR

Known generalization bounds:

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

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



generalization error vs. known bounds

Optimization Speed

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 (Liang & S. '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 for s.l.n. n, p,

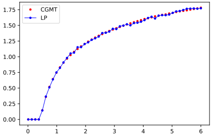
$$p^{1/2}\cdot \min_{i\in[n]} \frac{y_i x_i^\top \theta_{\text{boost}}^{t,s}}{\|\theta_{\text{boost}}^{t,s}\|_1} \in \left[(1-\epsilon)\cdot\kappa_\star(\psi,\mu), \kappa_\star(\psi,\mu)\right]\ .$$

Optimization Speed

Theorem (Liang & S. '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 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 (Liang & S. '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

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

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

ALGORITHMIC: ACTIVATED FEATURES BY BOOSTING

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

Theorem (Liang & S. '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

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

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

- Precise characterizations $\cdots$
- Of the max-min-ℓ_1-margin
- Min-ℓ_1-norm interpolant
- Several insights into boosting — precise gen. error expressions, optimization speed, etc.

- Precise characterizations $\cdots$
- Of the max-min-ℓ_1-margin
- Min-ℓ_1-norm interpolant
- Several insights into boosting — precise gen. error expressions, optimization speed, etc.
- Extensions possible to ℓ_p geometries, certain misspecified models, Gaussian mixture models, $\cdots$
- Many other perspectives in boosting (e.g. effectively minimizes empirical loss functional, fits logistic regression models additively, L_2 boosting, model-based boosting, etc)
- Opens door to further questions!

Thank you!