A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-l1-Norm Classifiers

杨东旭

August 15, 2025

Bias–Variance Balance VS Benign overfitting

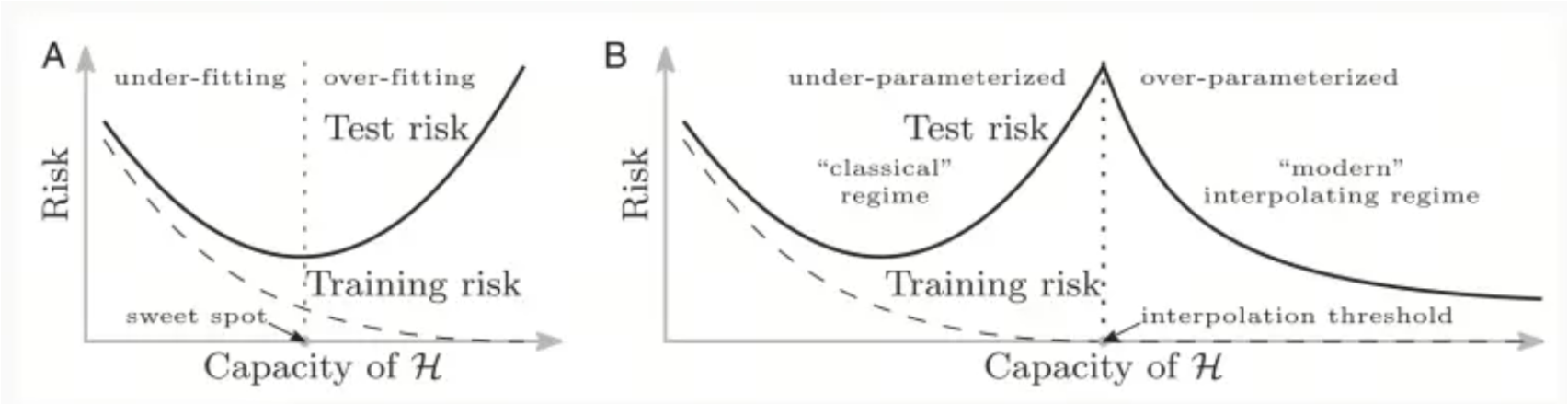

Figure: Bias–Variance Balance VS Benign overfitting

Which algorithms?

minimum-norm interpolation regime

We focus on boosting/AdaBoost

Boosting

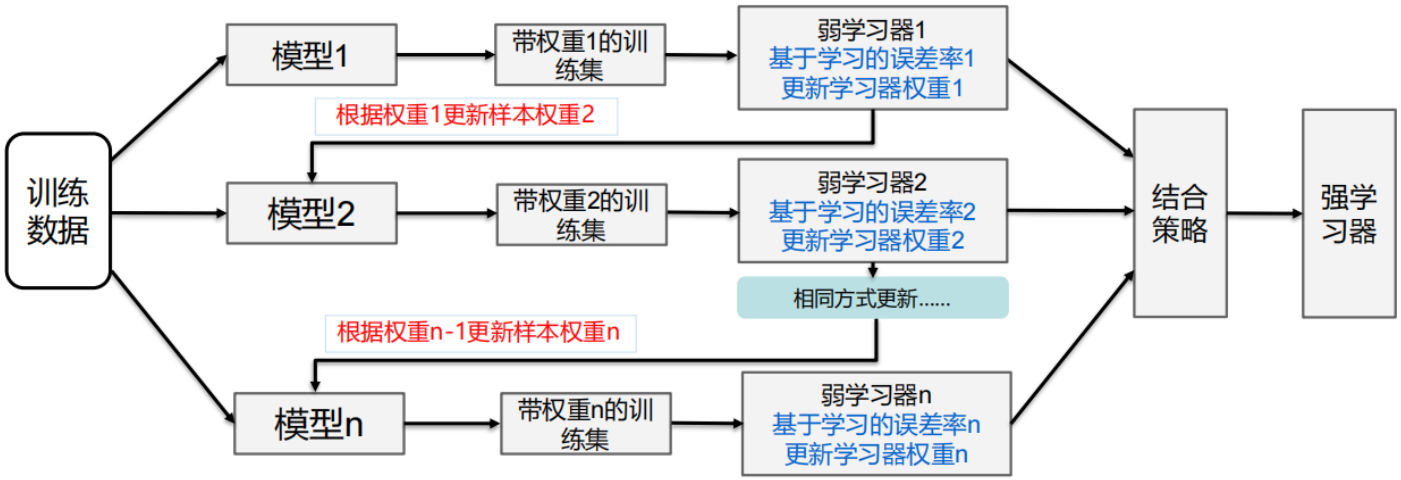

Figure: Boosting

Boosting–Benign Overfitting

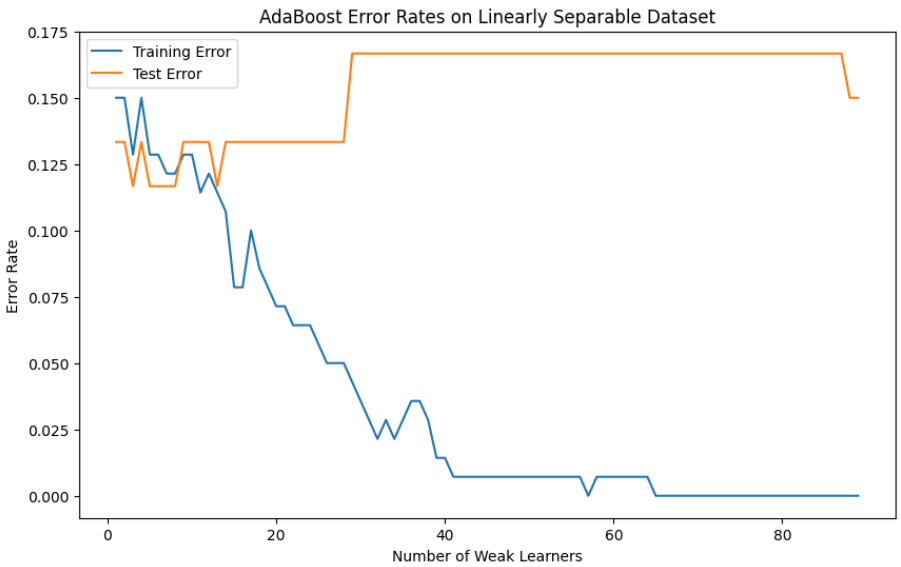

Figure: 1

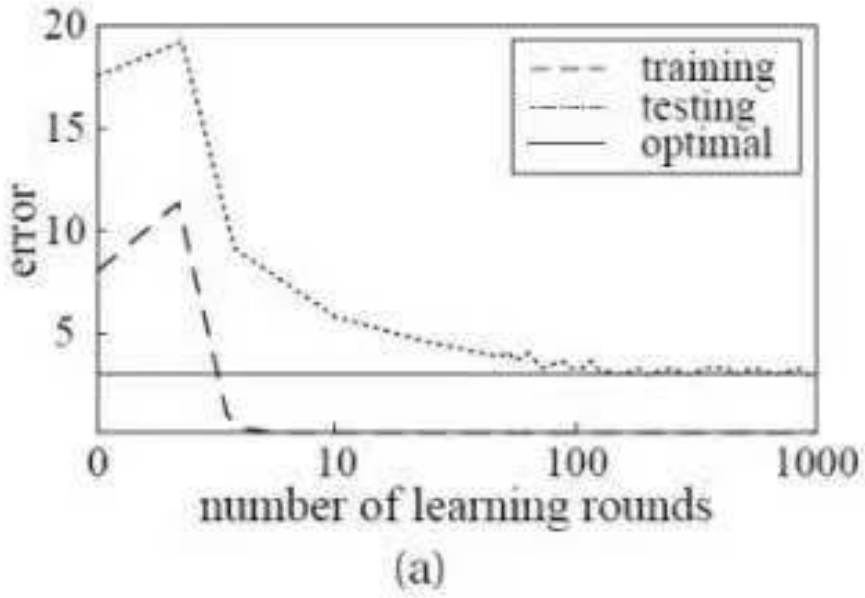

Figure: 2

These results delineate the differences between efficient reduction in training set error and test set accuracy. Arc-fs reaches zero training set error very quickly, after an average of 5 tree constructions (at most). But the accompanying test set error is higher than that of bagging, which takes longer to reach zero training set error. To produce optimum reductions in test set error, arc-fs must be run far past the point of zero training set error.

Background

Margin-based analyses

- AdaBoost increases the fraction of high-margin samples over iterations.
- The empirical margin distribution stabilizes to a limiting one rapidly.
- Key quantities that precisely determine the generalization behavior of AdaBoost remained unanswered.

Consistency and early stopping

- AdaBoost is process consistent
- There exists a stopping time at which the prediction error approximates the optimal Bayes error in the limit of large samples.
- Precise characterization of test error vs. Bayes error in overparametrized regimes still missing

Connections with min-L1-norm interpolation

- For linearly separable data, AdaBoost with infinitesimal step size converges to min-L1-norm interpolant when left to run forever, which equivalent to maximizing L1-margin κ_{n,ℓ_1}
- The number of optimization steps necessary for AdaBoost to reach zero training error can be upper bounded by$T = \mathcal{O}(\kappa_{n,\ell_1}^{-2})$
- Boosting potentially converges (in direction) to a sparse classifier.

This paper

1. Provide a precise characterization of the value of the max-ℓ_1-margin，show that $p^{1/2} \cdot \kappa_{n,\ell_1}$ converges almost surely to a constant $\kappa_\star$
2. Establish a precise formula for the generalization error of the min-ℓ_1-norm interpolant，illuminates that the generalization error is completely governed by the dimensionality parameter $\psi = p/n$ and the limit $\kappa_\star$
3. Develop an exact characterization of a threshold T such that for all iterations $t \geq T$, the boosting iterates. stay arbitrarily close to min-ℓ_1-norm interpolant,in the large n,p limit
4. Keeping other parameters fixed, T decreases with an increase in ψ, suggesting that **overparametrization helps in optimization**
5. Introduce a new class of Boosting Algorithms that converge to the max-ℓ_q-margin direction

Min-ℓ_1-Norm Interpolated Classifier

Definition

Given training data $\{(x_i, y_i)\}_{i=1}^n$ with $y_i \in \{+1, -1\}$:

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

Max-ℓ_1-Margin

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

When $\kappa_{n,\ell_1} > 0$, $\hat{\theta}_{n,\ell_1} / \|\hat{\theta}_{n,\ell_1}\|_1$ solves the margin problem.

Data Generation Model

High-dimensional Setting

- **Dimensionality**: $p/n \to \psi > 0$ (overparametrized regime)
- **Features**: $x_i \sim \mathcal{N}(0, \Lambda(n)) \in \mathbb{R}^p$
  Diagonal covariance $\Lambda(n) = \text{diag}(\lambda_1, ..., \lambda_p)$
- **Labels**: $y_i \in \{+1, -1\}$ via GLM:

$$\mathbb{P}(y_i = +1|x_i) = f(\langle \theta_*(n), x_i \rangle)$$

where $f : \mathbb{R} \to [0, 1]$ is monotonic (e.g., logistic sigmoid)

Key Assumptions

Assumption 1: Eigenvalue Bound

$$c \leq \lambda_i(\Lambda(n)) \leq 1/c \quad \forall i, n \quad (0 < c < 1)$$

Ensures well-conditioned features (no degenerate dimensions)

Key Assumptions

Notation

- $\theta_* \in \mathbb{R}^p$: True parameter vector (ground truth)
- $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$: Diagonal covariance matrix
- e_i: i-th standard basis vector in $\mathbb{R}^p$ (e.g., $e_1 = (1, 0, \ldots, 0)^T$)
- $\delta_{(a,b)}$: Dirac delta measure at point $(a, b) \in \mathbb{R}^2$
- W_2: Wasserstein-2 distance (detailed next page)

Key Assumptions

Wasserstein-2 Distance

For measures μ, ν on $\mathbb{R}^2$:

$$W_2(\mu,\nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int \|x-y\|^2 d\gamma(x,y) \right)^{1/2}$$

where $\Gamma(\mu,\nu)$ is the set of couplings with marginals μ and ν.

Implication

W_2-convergence implies convergence of the second moments and Weak convergence .

Key Assumption

Assumption 2: Signal Structure

$$\rho(n) := (\theta_*^T \Lambda \theta_*)^{1/2}$$

$$\bar{w}_i(n) := \sqrt{p} \cdot \frac{\sqrt{\lambda_i}(\theta_*, e_i)}{\rho(n)}$$

$$\rho(n) \to \rho \geq 0$$

$$\frac{1}{p} \sum_{i=1}^{p} \delta_{(\lambda_i, \bar{w}_i)} \xrightarrow{W_2} \mu$$

remark

Note $\sum_{j=1}^{p} (\theta_{*,j}^{(n)})^2 = \mathcal{O}(1)$, $\theta_{*,i} = \mathcal{O}(1/\sqrt{p})$ and:

$$\int_{\mathbb{R}^2} w^2 d\mu(\lambda, w) = \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \bar{w}_i^2 = 1$$

Key Assumptions

Assumption 3: Uniform Boundedness

$$\|\bar{w}(n)\|_\infty \leq C', \quad \|\bar{w}(n)\|_1 > C'' \quad \forall n, p$$

for constants $C', C'' > 0$, where $\bar{w}(n) \in \mathbb{R}^p$ is defined as:

$$\bar{w}_i(n) := \sqrt{p} \cdot \frac{\sqrt{\lambda_i}\langle \theta_*(n), e_i \rangle}{\rho(n)}$$

- Ensures signal is neither too sparse nor too concentrated
- $\|\cdot\|_\infty$ bound prevents dominance by single feature
- $\|\cdot\|_1$ bound ensures sufficient signal spread

Linear Separability Condition

Linear Separability

$$\lim_{n\to\infty, p(n)\to\infty} \mathbb{P}\left\{\exists \theta \in \mathbb{R}^p : y_i \mathbf{x}_i^\top \theta > 0 \text{ for all } 1 \le i \le n\right\} = 1$$

Phase Transition Threshold

Data is linearly separable *iff*:

$$\psi > \psi^*(\rho, f) = \min_{c\in\mathbb{R}} F_0^2(c, 1)$$

where $F_k(c_1, c_2) := \sqrt{\mathbb{E}[(k - c_1 Y Z_1 - c_2 Z_2)_+^2]}$ with:

- $Z_1, Z_2 \sim \mathcal{N}(0, 1)$ i.i.d., $Z_2 \perp (Y, Z_1)$
- $Y \in \{+1, -1\}$ with $\mathbb{P}(Y = +1|Z_1) = f(\rho Z_1)$

Interpretation

- ψ^* depends on signal strength ρ and link function f
- For logistic model (f =sigmoid), $\psi^* \approx 0.43$

Boosting Algorithm (AdaBoost)

Setup

- Training data $\{(x_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, +1\}$
- Weak learners = coordinate directions: $\{e_j\}_{j=1}^p$
- $Z := [y_1 x_1, \cdots, y_n x_n]^T \in \mathbb{R}^{n \times p}$

Algorithm Steps

1. Initialize weights $\eta_0 = \frac{1}{n}\mathbf{1}_n \in \Delta_n$, $\theta_0 = \mathbf{0}_p$
2. For $t = 0, 1, 2, ...$:
   (a) **Feature selection**: $v_{t+1} = \arg\max_{v \in \{e_j\}} |\eta_t^T Z v|$
   (b) **Adaptive stepsize**: $\alpha_t = \eta_t^T Z v_{t+1}$
   (c) **Parameter update**: $\theta_{t+1} = \theta_t + \alpha_t v_{t+1}$
   (d) **Weight update**: $\eta_{t+1}[i] \propto \eta_t[i] \exp(-\alpha_t y_i x_i^T v_{t+1})$

Boosting Algorithm (AdaBoost)

Remark

- **Weight?**: exponential loss

$$\text{Loss} = \sum_{i=1}^{N} \exp\left(-y_i f_m(x_i)\right)$$

- **Meaning of p**: Overparametrization, weak learners

Max-ℓ_1-Margin Asymptotics

Theorem 3.1: Limit of Scaled Margin

Under Assumptions 1–3 and linear separability ($\psi > \psi^*$):

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

where

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

and

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

- $F_\kappa(c_1, c_2) := \sqrt{\mathbb{E}[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2]}$
- (c_1, c_2, s) solve nonlinear system (details in Following section) parameterized by $(\psi, \rho, \mu, \kappa)$
- the limit is well-defined(details in Following section)

Generalization Error of Min-ℓ_1 Interpolant

Theorem 3.2: Exact Test Error

Under the assumptions of Theorem 3.1, the generalization error of any min-ℓ_1-interpolated classifier,converges almost surely, that is, for a new data point (x,y) drawn from the data-generating distribution

$$\lim_{n\to\infty} \mathbb{P}(y_{\text{new}} x_{\text{new}}^T \hat{\theta}_{n,\ell_1} < 0) \overset{\text{a.s.}}{=} \text{Err}_\star(\psi, \rho, \mu)$$

where

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

with c_1^*, c_2^* = solution to the nonlinear system at $\kappa = \kappa_\star$

Remark

$Err_{\text{Bayes}}(\rho) = \mathbb{P}(YZ_1 < 0)$
$Err_\star(\psi, \rho, \mu) = \mathbb{P}((c_2^*)^{-1}c_1^* YZ_1 + Z_2 < 0)$
$(c_2^*)^{-1}c_1^*$ exactly determines how the test error of $\hat{\theta}_{n,\ell_1}$ differs from the optimal Bayes error.

Asymptotic Geometry of the Interpolant

The angle between the interpolated solution and the target satisfies:

$$\frac{\langle \hat{\theta}_{n,\ell_1}, \theta_* \rangle_\Lambda}{\|\hat{\theta}_{n,\ell_1}\|_\Lambda \|\theta_*\|_\Lambda} \xrightarrow{a.s.} \frac{c_1^*}{\sqrt{(c_1^*)^2 + (c_2^*)^2}}$$

where $\langle \theta_1, \theta_2 \rangle_\Lambda := \theta_1^T \Lambda \theta_2$. Furthermore, c_2^* represents the norm of the orthogonal projection:

$$\left\| \Pi_{(\Lambda^{1/2}\theta_*)^\perp} (\Lambda^{1/2} \hat{\theta}_{n,\ell_1}) \right\| \xrightarrow{a.s.} c_2^*$$

Finite sample result

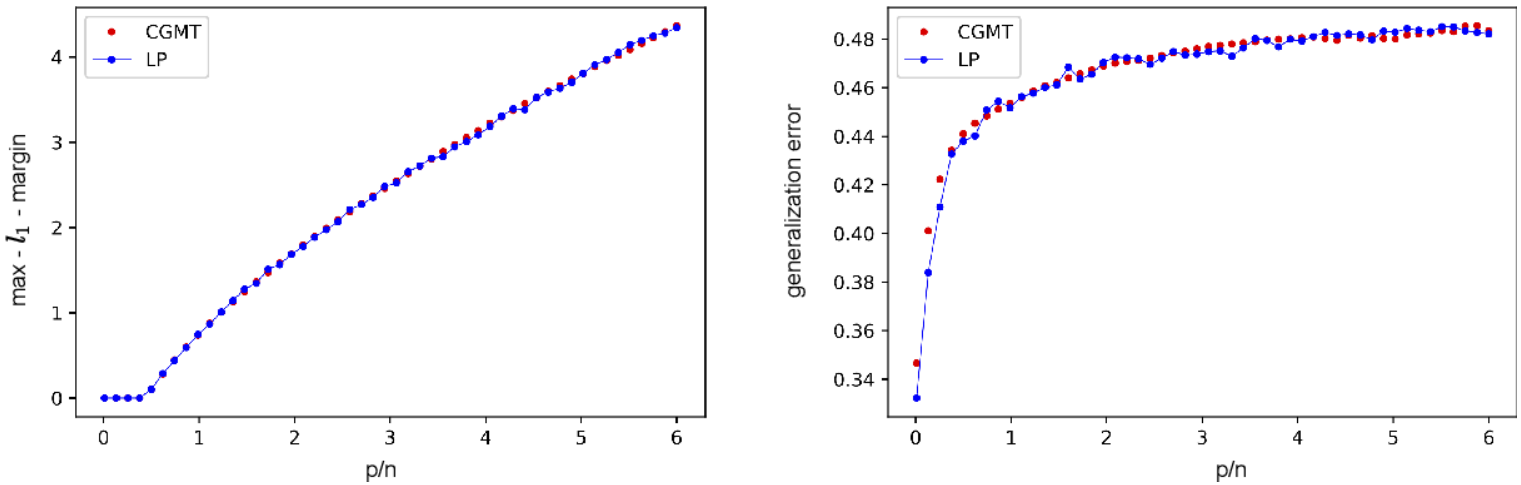

Figure: Finite sample result(n = 400,f:sigmoid)

The Nonlinear System of Equations

Definition 1: Key System for Max-ℓ_1-Margin

For $\psi > 0$, $\kappa \geq 0$, solve for $(c_1, c_2, s) \in \mathbb{R}^3$:

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

$$c_1^2 + c_2^2 = \mathbb{E}_{\mathcal{Q}}\left[\frac{\Lambda^{-1/2}\mathcal{T}}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right]^2$$

$$1 = \mathbb{E}_{\mathcal{Q}}\left[\frac{\Lambda^{-1}|\mathcal{T}|}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right]$$

- $\mathcal{Q} = \mu \otimes \mathcal{N}(0,1)$ with μ from Assumption 2
- $F_\kappa(c_1,c_2) := \sqrt{\mathbb{E}[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2]}$
- Proximal operator:
  $\text{prox}_s(t) = \arg\min_{x\in\mathbb{R}}\{|x| + \frac{1}{2s}(x-t)^2\} = \text{sign}(t)(|t|-s)_+$
- $\mathcal{T} = \text{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}\left[\partial_1 F_\kappa - c_1 c_2^{-1}\partial_2 F_\kappa\right]\Lambda^{1/2}W\right)$

Uniqueness of Solutions

Threshold Functions

Define constants:

$$\zeta := \left(\mathbb{E}_\mu |\Lambda^{-1/2} W|\right)^{-1}, \quad \omega := \left(\mathbb{E}_\mu (W - \zeta \Lambda^{-1/2} \mathsf{sign}(\zeta \Lambda^{-1/2} W))^2\right)^{1/2}$$

and functions:

$$\psi_+(\kappa) := \begin{cases} 0 & \partial_1 F_\kappa(\zeta, 0) > 0 \\ \partial_2^2 F_\kappa(\zeta, 0) - \omega^2 \partial_1^2 F_\kappa(\zeta, 0) & \text{otherwise} \end{cases}$$

$$\psi_-(\kappa) := \begin{cases} 0 & \partial_1 F_\kappa(-\zeta, 0) < 0 \\ \partial_2^2 F_\kappa(-\zeta, 0) - \omega^2 \partial_1^2 F_\kappa(-\zeta, 0) & \text{otherwise} \end{cases}$$

$$\psi^+(\kappa) := \max\{\psi^*(\rho, f), \psi_+(\kappa), \psi_-(\kappa)\}$$

Proposition 3.1 (Uniqueness)

For $\psi > \psi^+(\kappa)$, the system admits a unique solution $(c_1, c_2, s) \in \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$.

Properties Ensuring Well-Definedness

Key Properties of $T(\psi, \kappa)$

1. **Continuity**: $(\psi, \kappa) \mapsto T(\psi, \kappa)$ is continuous on its domain
2. **Monotonicity in** ψ: For fixed $\kappa > 0$, $T(\psi, \kappa)$ is strictly decreasing in ψ
3. **Monotonicity in** κ: For fixed $\psi > 0$, $T(\psi, \kappa)$ is strictly increasing in κ

Asymptotic Behavior

- $\lim_{\psi \to \infty} T(\psi, \kappa) < 0$
- $\lim_{\psi \downarrow \psi^+(\kappa)} T(\psi, \kappa) > 0$
- $\lim_{\kappa \to \infty} T(\psi, \kappa) = \infty$

Implication for $\kappa_\star$

The solution region $\{(\psi, \kappa) : \psi > \psi^+(\kappa)\}$ contains $\{(\psi, \kappa) : T(\psi, \kappa) = 0\}$, ensuring $\kappa_\star$ is well-defined.

Boosting Convergence to Min-ℓ_1 Interpolant

Theorem 3.3: Iteration Threshold

Under the assumptions of Theorem3.1,with suitably chosen learning rate , the boosting iterates $\{\hat{\theta}^t\}$ satisfy:

$$\forall \epsilon > 0, \quad t \geq T_\epsilon(n)$$

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

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

where the threshold scales as:

$$\boxed{\lim_{n\to\infty}\frac{T_\epsilon(n)}{n\log^2 n} \overset{\text{a.s.}}{=} \frac{12\psi}{\kappa_\star^2(\psi,\rho,\mu)}\epsilon^{-2}}$$

Boosting Convergence to Min-ℓ_1 Interpolant

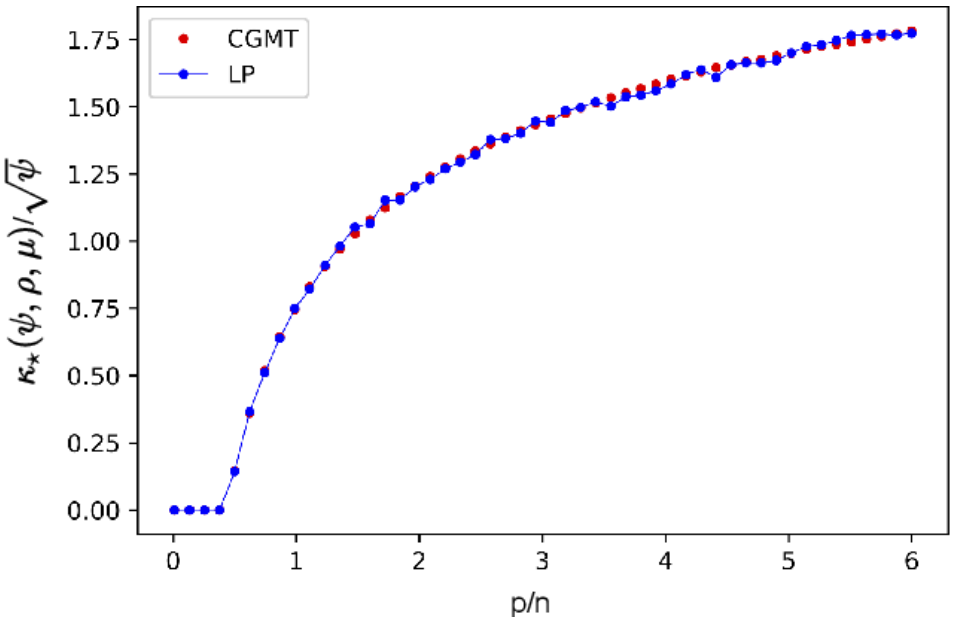

Figure: Finite sample set

the max-ℓ_1-margin is the key quantity governing the generalization error

overparametrization leads to faster optimization

Proportion of Activated Features for AdaBoost

Corollary 3.1: Active Features at Interpolation

Let $S_0(p)$ denote the number of non-zero features when boosting *first* achieves zero training error (with initialization $\hat{\theta}^0 = 0$). Under the assumptions of Theorem 3.3:

$$\limsup_{p\to\infty} \frac{S_0(p)}{p \log^2 p} \leq \frac{12}{\kappa_\star^2(\psi, \rho, \mu)} \quad \text{a.s.}$$

- Larger overparameterization leads to fewer activated coordinates under the assumed data-generating model.

Generalized Boosting for ℓ_q-Geometry

Max-ℓ_q-Margin and Interpolant

For $q \geq 1$, define:

$$\text{Max-}\ell_q\text{-margin:} \quad \kappa_{n,\ell_q} := \max_{\|\theta\|_q \leq 1} \min_{1 \leq i \leq n} y_i x_i^T \theta$$

$$\text{Min-}\ell_q\text{-norm interpolant:} \quad \hat{\theta}_{n,\ell_q} \in \arg\min_{\theta} \|\theta\|_q \quad \text{s.t.} \quad y_i x_i^T \theta \geq 1$$

q_* is the conjugate exponent satisfying $1/q_* + 1/q = 1$

Generalized AdaBoost Algorithm

1. Initialize $\eta_0 = \frac{1}{n}\mathbf{1}_n$, $\theta_0 = \mathbf{0}_p$
2. For $t \geq 0$:
   (a) Update direction: $v_{t+1} = \arg\max_{v:\|v\|_q=1} \langle Z^T \eta_t, v \rangle$
   (b) Adaptive Stepsize: $\alpha_t(\beta) = \beta \cdot \|Z^T \eta_t\|_{q_*}$ $(0 < \beta < 1)$
   (c) Parameter update: $\theta_{t+1} = \theta_t + \alpha_t v_{t+1}$
   (d) Weight update: $\eta_{t+1}[i] \propto \eta_t[i] \exp(-\alpha_t y_i x_i^T v_{t+1})$

Convergence to Max-ℓ_q-Margin

Corollary 3.2: Convergence Guarantee

For any $q \geq 1$ and $0 < \epsilon < 1$, with bounded features $|X_{ij}| \leq M$ and shrinkage $\beta = \epsilon/(p^{2/q_*}M^2)$:

$$\kappa_{n,\ell_q} \geq \min_i \frac{y_i x_i^T \theta_T}{\|\theta_T\|_q} > (1-\epsilon)\kappa_{n,\ell_q}$$

$T \geq \log(1.01ne) \cdot \frac{2p^{2/q_*}M^2\epsilon^{-2}}{\kappa^2_{n,\ell_q}}$ iterations

Corollary 3.3: Limiting Margins ($1 \leq q \leq 2$)

Under model assumptions:

$$p^{\frac{1}{q}-\frac{1}{2}} \kappa_{n,\ell_q} \xrightarrow{\text{a.s.}} \kappa_\star^{(q)}(\psi, \rho, \mu)$$

where $\kappa_\star^{(q)}$ solves:

$$\begin{aligned} c_1 &= \mathbb{E}_{\mathcal{Q}}[\Lambda^{1/2} h^* \cdot W] \\ c_1^2 + c_2^2 &= \mathbb{E}_{\mathcal{Q}} \|\Lambda^{1/2} h^*\|^2 \\ 1 &= \mathbb{E}_{\mathcal{Q}} \|h^*\|_q \end{aligned}$$

with

$$h^* = \text{prox}_{\lambda^*}^{(q)}(t^*)$$

$$t^* = -\frac{\Lambda^{-1/2} G + \psi^{-1/2}[\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)]\Lambda^{-1/2} W}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)}$$

$$\lambda^* = \frac{\Lambda^{-1} s}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)}$$

Robustness to assumptions

1. Going beyond the assumption of independence between the covariates
2. Analyzing sensitivity to the Gaussianity assumption
3. Understanding implications of certain model misspecification(Explore a common source of misspecification that occurs when the model misses a fraction of relevant variables)

Beyond Independent Covariates: Rank-One Perturbation

Gaussian Mixture Model

Data generated via:

- **Class prior**: $\mathbb{P}(y_i = +1) = v \in (0, 1)$
- **Conditional feature distribution**: $x_i | y_i \sim \mathcal{N}(y_i \cdot \theta_\star, \Lambda)$ where $\Lambda \in \mathbb{R}^{p \times p}$ is diagonal

Key Properties

- **Conditional probability**: $\mathbb{P}(y_i = +1 | x_i) = \sigma\left(\log \frac{v}{1-v} + \langle \Lambda^{-1}\theta_\star, x_i \rangle\right)$ where $\sigma(t) = 1/(1 + e^{-t})$
- **Marginal covariance**: $\mathrm{Cov}(x_i) = 4v(1-v)\theta_\star\theta_\star^\top + \Lambda$ (spiked covariance model)

Asymptotics for Rank-One Perturbation

- Maintain $p(n)/n \to \psi$
- Define measure convergence: $\frac{1}{p}\sum_{i=1}^{p} \delta_{(\lambda_i, \sqrt{p}\theta_\star^\top e_i)} \xrightarrow{W_2} \mu$

Define new function $\bar{F}_\kappa : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$:

$$\bar{F}_\kappa(c_1, c_2) := \left(\mathbb{E}\left[(\kappa - c_1 - c_2 Z)_+^2\right]\right)^{1/2}, \quad Z \sim \mathcal{N}(0,1)$$

Equation System for Rank-One Case

Nonlinear System

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

$$c_1 = -\mathbb{E}_{\mathcal{Q}}\left[\frac{\Lambda^{-1}\Theta \cdot \mathcal{T}}{\psi^{-1/2}c_2^{-1}\partial_2 \bar{F}_\kappa}\right]$$

$$c_2^2 = \mathbb{E}_{\mathcal{Q}}\left[\frac{\Lambda^{-1/2}\mathcal{T}}{\psi^{-1/2}c_2^{-1}\partial_2 \bar{F}_\kappa}\right]^2$$

$$1 = \mathbb{E}_{\mathcal{Q}}\left[\frac{\Lambda^{-1}|\mathcal{T}|}{\psi^{-1/2}c_2^{-1}\partial_2 \bar{F}_\kappa}\right]$$

where:

$$\mathcal{T} = \text{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}\partial_1 \bar{F}_\kappa(c_1, c_2)\Theta\right)$$

Result

Theorems 3.1-3.2 hold with F_κ replaced by $\bar{F}_\kappa$ and system replaced

Rank-Two Perturbation Model

Extended Data Generation

$$x_i = y_i \theta_\star + m_i \tilde{\theta} + \tilde{x}_i$$

with:

- $y_i \in \{-1, +1\}$, m_i symmetric around 0
- $\tilde{x}_i \sim \mathcal{N}(0, \Lambda)$ (diagonal covariance)
- Latent m_i not observed

Marginal Covariance

$$\mathrm{Cov}(x_i) = 4v(1-v)\theta_\star \theta_\star^\top + \mathrm{Var}(m_i)\tilde{\theta}\tilde{\theta}^\top + \Lambda$$

(rank-two perturbation of diagonal matrix)

Asymptotics for Rank-Two Case

Measure Convergence

Assume:

$$\frac{1}{p}\sum_{j=1}^{p}\delta_{\left(\lambda_j,\sqrt{p}\theta_\star^\top e_j,\sqrt{p}\tilde{\theta}^\top e_j\right)} \xrightarrow{W_2} \tilde{\mu}$$

Define $\hat{\mathcal{Q}} = \tilde{\mu} \otimes \mathcal{N}(0,1)$ with components $(\Lambda, h_*, \tilde{h}, G)$

Modified Margin Function

$$\tilde{F}_\kappa(c_1,c_2,c_3) = \sqrt{\mathbb{E}\left[(\kappa - c_1 - c_2\tilde{Z} - c_3 M)_+^2\right]}$$

where:

- $M \sim m_i$, $\tilde{Z} \sim \mathcal{N}(0,1)$ independent

Equation System for Rank-Two Case

Extended System (3.30)

$$
\begin{aligned}
c_1 &= \mathbb{E}_{\hat{Q}}[h_* h_{\text{sol}}] \\
c_2 &= \mathbb{E}_{\hat{Q}}[(\Lambda^{1/2} h_{\text{sol}})^2] \\
c_3 &= \mathbb{E}_{\hat{Q}}[\tilde{h} h_{\text{sol}}] \\
1 &= \mathbb{E}_{\hat{Q}}[|h_{\text{sol}}|]
\end{aligned}
$$

where:

$$
h_{\text{sol}} = -\frac{\text{prox}_s\left(\Lambda^{1/2} G + \psi^{-1/2}(\partial_1 \tilde{F}_\kappa h_* + \partial_3 \tilde{F}_\kappa \tilde{h})\right)}{\Lambda \psi^{-1/2} c_2^{-1} \partial_2 \tilde{F}_\kappa}
$$

Pattern

- Each additional spike increases system dimension
- Rank-ℓ perturbation $\Rightarrow$ $(\ell + 2)$-dimensional system
- Core analysis principles remain unchanged

Beyond Independent Covariates

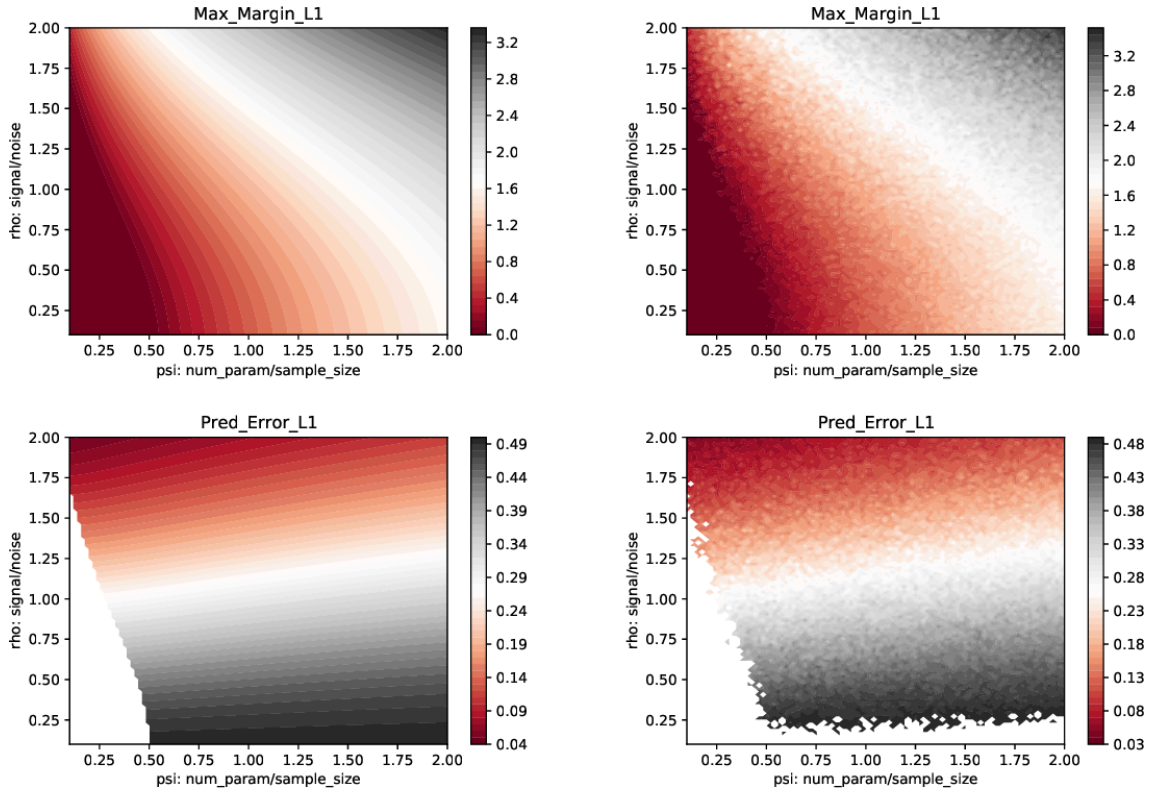

Figure: Rank 1

Beyond Gaussian Covariates: Universality

Nonlinear Random Features in Boosting

Consider:

- **Observed data**: $\{(x_i, y_i)\}_{i=1}^n$ generated as in Section 2
- **Boosting features**: Nonlinear random features $a_i = \sigma(F^\top x_i) \in \mathbb{R}^d$
  - $F \in \mathbb{R}^{p \times d}$: Random weight matrix (e.g., i.i.d. Gaussian entries)
  - $\sigma(\cdot)$: Nonlinear activation (applied element-wise)
- **Max-ℓ_1-margin** is computed using $\{a_i, y_i\}$ instead of $\{x_i, y_i\}$

Beyond Gaussian Covariates: Universality

Linearized Gaussian Counterpart

Define analogous Gaussian features:

$$b_i = \mu_0 \mathbf{1} + \mu_1 F^\top x_i + \mu_2 z_i \in \mathbb{R}^d$$

where:

- $z_i \sim \mathcal{N}(0, I_d)$ independent of everything
- Coefficients match Hermite expansion of σ:

$$\mu_0 = \mathbb{E}[\sigma(Z)], \quad \mu_1 = \mathbb{E}[Z\sigma(Z)], \quad \mu_2 = \sqrt{\mathsf{Var}(\sigma(Z)) - \mu_0^2 - \mu_1^2}$$

with $Z \sim \mathcal{N}(0, 1)$

Universality of Max-ℓ_1-Margin

Theorem 3.4: Universality

Under the setting above, if σ is odd, compactly supported, and has bounded first three derivatives, and $p(n)/n \to \psi$, $d(n)/n \to \phi$ (both $\in (0, \infty)$), then:

$$\boxed{p^{1/2}\left[\kappa_{n,\ell_1}(\{(a_i, y_i)\}) - \kappa_{n,\ell_1}(\{(b_i, y_i)\})\right] \xrightarrow{\mathbb{P}} 0}$$

where $\kappa_{n,\ell_1}(\cdot)$ is computed via:

$$\kappa_{n,\ell_1}(\{r_i, y_i\}) = \max_{\|\theta\|_1 \le 1} \min_{1 \le i \le n} y_i r_i^\top \theta$$

Interpretation

- Asymptotic max-ℓ_1-margin is *insensitive* to higher-order moments of the feature distribution
- Gaussian approximation with matching first/second moments suffices
- Enables extension of Gaussian-based theory to nonlinear random features

Beyond Independent Covariates

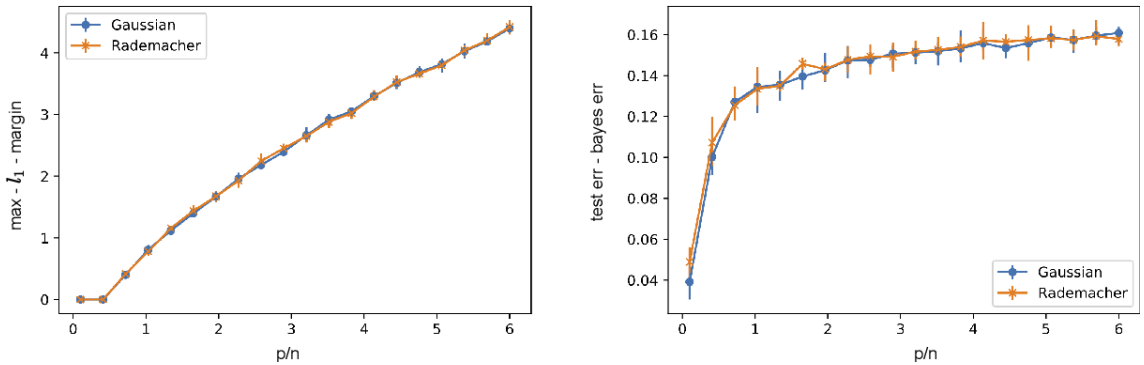

Figure: Sigmoid

Beyond Theory: Covariate Distribution

Test max-ℓ_1-margin and generalization under:

- **Rademacher design**: $X_{ij} = \pm 1$ with prob. $1/2$
- **Gaussian design**: $X_{ij} \sim \mathcal{N}(0, 1)$ (matching 1st/2nd moments)

Model Misspecification: Missing Relevant Variables

Data Generation Process

Observed data: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$
Full feature vector: $\tilde{x}_i = (x_i^T, z_i^T)^T \in \mathbb{R}^{p+q}$

- Observed component: $x_i \sim \mathcal{N}(0, \Lambda_x)$ (diagonal)
- Missing component: $z_i \sim \mathcal{N}(0, \Sigma_z)$
- Label generation: $\mathbb{P}(y_i = +1|\tilde{x}_i) = f(\tilde{x}_i^T \theta_*)$ with $\theta_* := (\theta_{x,*}^T, \theta_{z,*}^T)^T$

High-dimensional Scaling

$$\frac{p(n)}{n} \to \psi > 0, \quad \frac{q(n)}{n} \to \phi > 0$$

Signal strengths:

$$\lim_{n\to\infty} (\theta_{x,*}^T \Lambda_x \theta_{x,*})^{1/2} = \rho$$

$$\lim_{n\to\infty} (\theta_{z,*}^T \Sigma_z \theta_{z,*})^{1/2} = \gamma$$

Asymptotic Theory Under Misspecification

Modified Margin Function

Define $\tilde{F}_\kappa : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$:

$$\tilde{F}_\kappa(c_1, c_2) := \sqrt{\mathbb{E}\left[(\kappa - c_1 Y Z_1 - c_2 Z_3)_+^2\right]}$$

where:

- $Z_1, Z_2, Z_3 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ with $Z_3 \perp (Y, Z_1, Z_2)$
- $Y \in \{+1, -1\}$ with $\mathbb{P}(Y = +1 | Z_1, Z_2) = f(\rho Z_1 + \gamma Z_2)$

Key Result

When $\psi + \phi > \psi^*$ (separability threshold):

- Max-ℓ_1-margin limit: $\sqrt{p}\kappa_{n,\ell_1} \overset{\text{a.s.}}{\to} \kappa_\star$
- Generalization error: $\mathbb{P}(y_{\text{new}} x_{\text{new}}^T \hat{\theta}_{n,\ell_1} < 0) \overset{\text{a.s.}}{\to} \text{Err}_\star$

with identical nonlinear system (3.9) but F_κ replaced by $\tilde{F}_\kappa$

Summary of Contributions

Precise High-Dimensional Asymptotics

- Established exact limit for scaled max-ℓ_1-margin: $\sqrt{p}\kappa_{n,\ell_1} \xrightarrow{a.s.} \kappa_\star(\psi, \rho, \mu)$
- Derived closed-form generalization error for min-ℓ_1-norm interpolant: $\mathbb{P}(y_{\text{new}} x_{\text{new}}^\top \hat{\theta}_{n,\ell_1} < 0) \xrightarrow{a.s.} \text{Err}_\star(\psi, \rho, \mu)$

Boosting Algorithm Characterization

- Quantified iteration threshold $T_\epsilon(n)$ for AdaBoost convergence: $\frac{T_\epsilon(n)}{n \log^2 n} \xrightarrow{a.s.} \frac{12\psi}{\kappa_\star^2} \epsilon^{-2}$
- Revealed **optimization benefit of overparametrization**: $T_\epsilon(n) \downarrow$ as $\psi \uparrow$

Summary of Contributions (Cont.)

Geometric Insights

- Identified key parameters governing Bayes gap: $(c_2^*)^{-1}c_1^*$ controls $\mathsf{Err}_\star - \mathsf{Err}_{\mathsf{Bayes}}$
- Bounded active features at interpolation: $S_0(p)/p\log^2 p \le 12/\kappa_\star^2$

Extensions and Robustness

- Generalized to ℓ_q-geometry ($q \in [1,2]$) with matching asymptotics
- Demonstrated universality: Results hold for spiked covariances and random features
- Verified finite-sample accuracy empirically

Core Tool: Convex Gaussian Min-Max Theorem

Theorem A.1 (CGMT)

For compact sets $\Omega_1 \subset \mathbb{R}^n$, $\Omega_2 \subset \mathbb{R}^p$ and continuous $U : \Omega_1 \times \Omega_2 \to \mathbb{R}$, define:

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

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

where $Z \in \mathbb{R}^{n \times p}$, $g \in \mathbb{R}^n$, $h \in \mathbb{R}^p$ have i.i.d. $\mathcal{N}(0, 1)$ entries. Then:

1. $\forall t \in \mathbb{R} : \mathbb{P}(V_1(Z) \leq t) \leq 2\mathbb{P}(V_2(g, h) \leq t)$
2. If Ω_1, Ω_2 convex and U convex-concave: $\mathbb{P}(V_1(Z) \geq t) \leq 2\mathbb{P}(V_2(g, h) \geq t)$

Significance

Decouples min-max problem with Gaussian matrix Z into simpler min-max problem with Gaussian vectors g, h

Step 1: Basic Reduction

Auxiliary Variable

Define scaled margin deviation:

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

Equivalent to:

$$\xi_{\psi,\kappa}^{(n,p)} = \min_{\|\theta\|_1 \le \sqrt{p}} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot X)\theta)_+\|_2$$

Key Equivalence

$$\xi_{\psi,\kappa}^{(n,p)} = 0 \iff \kappa \le p^{1/2} \kappa_{n,\ell_1}$$

$$\xi_{\psi,\kappa}^{(n,p)} > 0 \iff \kappa > p^{1/2} \kappa_{n,\ell_1}$$

Study ξ to characterize scaled margin

Step 1: Basic Reduction (Cont.)

Whitened Representation

Set $z_i = \Lambda^{-1/2} x_i$, express:

$$x_i^\top \theta_* = \rho(n) z_i^\top w, \quad w := \Lambda^{1/2}\theta_* / \|\Lambda^{1/2}\theta_*\|$$

$$y \odot X = (y \odot Z)\Lambda^{1/2} \overset{d}{=} \left((y \odot z)w^\top + Z\Pi_{w^\perp}\right)\Lambda^{1/2}$$

Rewritten form:

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

where $Z \in \mathbb{R}^{n \times p}$ has i.i.d. $\mathcal{N}(0,1)$ entries

Step 2: Reduction to Gordon's Problem

Decoupled Optimization

Introduce auxiliary Gaussian vectors $g \in \mathbb{R}^p$, $\tilde{z} \in \mathbb{R}^n$:

$$\hat{\xi}_{\psi,\kappa}^{(n,p)} := \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\substack{\|\lambda\|_2 \le 1 \\ \lambda \ge 0}} \frac{1}{\sqrt{p}} \lambda^\top \mathcal{V} + \frac{1}{\sqrt{p}} \|\lambda\|_2 \langle g, \Pi_{w^\perp}(\Lambda^{1/2}\theta) \rangle$$

where $\mathcal{V} = \kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta \rangle - \tilde{z} \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2$

CGMT Application

By Theorem A.1:

$$\mathbb{P}\left(\xi_{\psi,\kappa}^{(n,p)} \le t \,|\, y, z\right) \le 2\mathbb{P}\left(\hat{\xi}_{\psi,\kappa}^{(n,p)} \le t \,|\, y, z\right)$$
$$\mathbb{P}\left(\xi_{\psi,\kappa}^{(n,p)} \ge t \,|\, y, z\right) \le 2\mathbb{P}\left(\hat{\xi}_{\psi,\kappa}^{(n,p)} \ge t \,|\, y, z\right)$$

Sufficient to study the decoupled problem

Step 3: Large n,p Limit & Uniform Deviation

Empirical Function

Define empirical version of F_κ:

$$\hat{F}_\kappa(c_1, c_2) := \left(\hat{\mathbb{E}}_n\left[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2\right]\right)^{1/2}$$

where expectation over empirical distribution

Objective Reformulation

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

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

Key Lemma: Uniform Deviation

Significance

- Let $c_1 = \langle w, \Lambda^{1/2}\theta \rangle$,$c_2 = \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2$
- notice that c_1 is bounded(Assumptiom 3),however， c_2 may potentially grow as $\sqrt{p}$

Lemma 5.1 (Self-Normalization)

For $i = 1, 2$ and any $M > 0$, w.p. $\geq 1 - n^{-2}$:

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

where C is independent of n.

Proof Strategy for Large-n Limit

Proposition A.1

Almost surely:

$$\lim_{\substack{n\to\infty \\ p(n)/n\to\psi}} \hat{\xi}_{\psi,\kappa}^{(n,p)} = \tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)}$$

where $\tilde{\xi}^{(\infty,\infty)}$ is infinite-dimensional optimization

1. Establish Lemma 5.1
2. Analyses the KKT condititon of $\hat{\xi}_{\psi,\kappa}^{(n,p)}$ and $\tilde{\xi}_{\psi,K}^{(\infty,\infty)}$,empirical fixed point (fp) equations converge uniformly to the corresponding fp equations
3. Show convergence of solutions to fixed-point equations

Step 4: Fixed Point Equations & Final Step

KKT Conditions

For infinite-dimensional problem:

$$\Pi_{W^\perp}(G) + \psi^{-1/2}\left[\partial_1 F_\kappa(c_1, c_2)W + c_2^{-1}\partial_2 F_\kappa(c_1, c_2)(\Lambda^{1/2}h - c_1 W)\right] + s \cdot \Lambda^{-1/2}\partial\|h\|_{L_1} = 0$$

with $\|h\|_{L_1} = 1$, $c_1 = \langle \Lambda^{1/2}h, W\rangle$, $c_2 = \|\Pi_{W^\perp}\Lambda^{1/2}h\|$

Proximal Form Solution

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

Completing Theorem 3.1

Fixed Point Characterization

Plug proximal solution into KTT conditions to obtain system:

$$\begin{aligned} c_1 &= -\mathbb{E}\left[\cdots\right] \\ c_1^2 + c_2^2 &= \mathbb{E}\left[\cdots\right]^2 \\ 1 &= \mathbb{E}\left|\cdots\right| \end{aligned}$$

Final Limit

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

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

Proof Sketch for Theorem 3.2

Generalization Error Representation

For Gaussian covariates:

$$\mathbb{P}(y_{\text{new}} x_{\text{new}}^{\top} \hat{\theta}_{n,\ell_1} < 0) = \mathbb{P}(c_{1,n} Y Z_1 + \sqrt{1 - c_{1,n}^2} Z_2 < 0)$$

where:

$$c_{1,n} = \frac{\langle \hat{\theta}_{n,\ell_1}, \theta_* \rangle_\Lambda}{\|\hat{\theta}_{n,\ell_1}\|_\Lambda \|\theta_*\|_\Lambda}$$

then prove $c_{1,n} \xrightarrow{\text{a.s.}} c_1^*$

Boosting as Mirror Descent

Key Insight

AdaBoost can be viewed as a special instance of Mirror Descent:

- **Domain**: Probability Simplex
- **Objective**: Exponential loss minimization
- **Generating Function**:Negative Entropy
- **Bregman divergence**: $D_\phi(\theta\|\theta') = \phi(\theta) - \phi(\theta') - \langle\nabla\phi(\theta'), \theta - \theta'\rangle$ with $\phi(\theta) = \|\theta\|_1^2$

Optimization Problem

Minimize exponential loss:

$$L(\theta) = \sum_{i=1}^{n} \exp(-y_i x_i^\top \theta)$$

with mirror descent update:

$$\theta_{t+1} = \arg\min_{\theta} \left\{\eta_t \langle\nabla L(\theta_t), \theta - \theta_t\rangle + D_\phi(\theta\|\theta_t)\right\}$$

Margin Maximization Rate

Proposition 5.1: Training Error Bound

With $|X_{ij}| \leq M$ and learning rate $\alpha_t(\beta) = \beta \cdot \eta_t^\top Z v_{t+1}$ ($\beta = 1/M^2$), after T iterations:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{y_i x_i^\top \theta_T \leq 0} \leq \epsilon$$

when

$$T \geq \frac{2M^2}{\kappa_{n,\ell_1}^2} \log \frac{ne}{\epsilon}$$

Corollary 5.1: Convergence to $\max$-ℓ_1-margin

Formal Statement

With shrinkage factor $\beta = \epsilon / M^2$, after

$$T \geq \log(1.01ne) \cdot \frac{2M^2\epsilon^{-2}}{\kappa_{n,\ell_1}^2}$$

iterations:

$$\kappa_{n,\ell_1} \geq \min_i \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_1} > (1-\epsilon)\kappa_{n,\ell_1}$$

References

- **Deng, Z., Kammoun, A., & Thrampoulidis, C.** (2019). A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*.
- **Montanari, A., Ruan, F., Sohn, Y., & Yan, J.** (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime.
- **Zhang, T., & Yu, B.** (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4), 1538–1579.
- **Thrampoulidis, C., Oymak, S., & Hassibi, B.** (2014). The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*.
- **Hu, H., & Lu, Y. M.** (2020). Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*.
- **Candès, E. J., & Sur, P.** (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Annals of Statistics*