Active Subsampling for Binary Response Models

Yang Ning

Cornell University, Department of Statistics and Data Science

June, 2026

https://arxiv.org/abs/2411.13763

Joint work with Jingyi Duan (Cornell) and Lehao Fu (Cornell)

Motivation: Why Active Learning/Subsampling?

- **Challenge**: Unlabeled data are abundant; labels are expensive.
  - Image classification
  - Medical diagnosis (expert evaluation or clinical tests)
  - Policy learning in healthcare
  - ...
- **Passive Learning Limitation**:
  - Labeling every data is costly and inefficient
- **Two alternatives**:
  - Semi-supervised learning
  - Active learning/subsampling

Binary Response Models

- **Setup**: $Y \in \{-1, 1\}$, and $W = (X, Z)$
  - Y, binary outcome, (e.g., success of a therapy)
  - $X \in \mathbb{R}$, (e.g., patients' health score)
  - $Z \in \mathbb{R}^d$, $d \gg n$, (e.g., patient-specific features)

- **Problem**: Find a linear decision rule to predict Y
  - Interpretability vs flexibility
  - Estimation vs classification

Example 1: ROC curve and Youden Index

- **Setup**:
  - X: patient biomarker
  - $Y \in \{1, -1\}$: disease status (diseased/healthy)
  - $Z \in \mathbb{R}^d$: patient-specific features

- **Goal**: Est optimal linear threshold $\theta^T Z$ that maximizes

$$\mathbb{P}(X \geq \theta^T Z | Y = 1) + \mathbb{P}(X < \theta^T Z | Y = -1) - 1$$

Example 2: Linear Binary Choice Model

- Linear binary choice model

$$Y = \text{sign}(X - \theta^T Z + \epsilon),$$

  ϵ random noise with $\text{median}(\epsilon|X, Z) = 0$

- θ minimzes

$$\mathbb{P}(X \geq \theta^T Z, Y = -1) + \mathbb{P}(X < \theta^T Z, Y = 1)$$

- Maximum score estimator

Example 3: Individualized Treatment Rule (ITR)

- **Setup**:
  - R: reward variable for patentis
  - $A \in \{1, -1\}$: treatment variable
  - $W = (X, Z)$: baseline subject features
  - Linear ITR: $D(W) = \text{sign}(X - \theta^T Z)$

- **Goal**: Est optimal linear ITR that minimizes

$$\mathbb{E}\Big[\frac{R}{\pi(A|W)} I(A \neq D(W))\Big],$$

where $\pi(A|W)$ is the probability of the treatment assignment.

General Formulation: Weighted Classification

- **Goal**: Est optimal linear threshold

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \left\{ w(-1)\mathbb{P}(X \geq \theta^T Z, Y = -1) + w(1)\mathbb{P}(X < \theta^T Z, Y = 1) \right\},$$

where $w(\cdot)$ a known weight function.

- **Challenges**:
  - Passive learning requires n i.i.d copies of (X, Y, Z)
  - Can we improve the est of θ in active learning framework?

Contributions

- Active subsampling algorithm
- Nonstandard statistical rate of convergence
- Minimax lower bound and adaptive estimation

Passive Learning Algorithm: Smoothed Surrogate Loss

Passive Learning

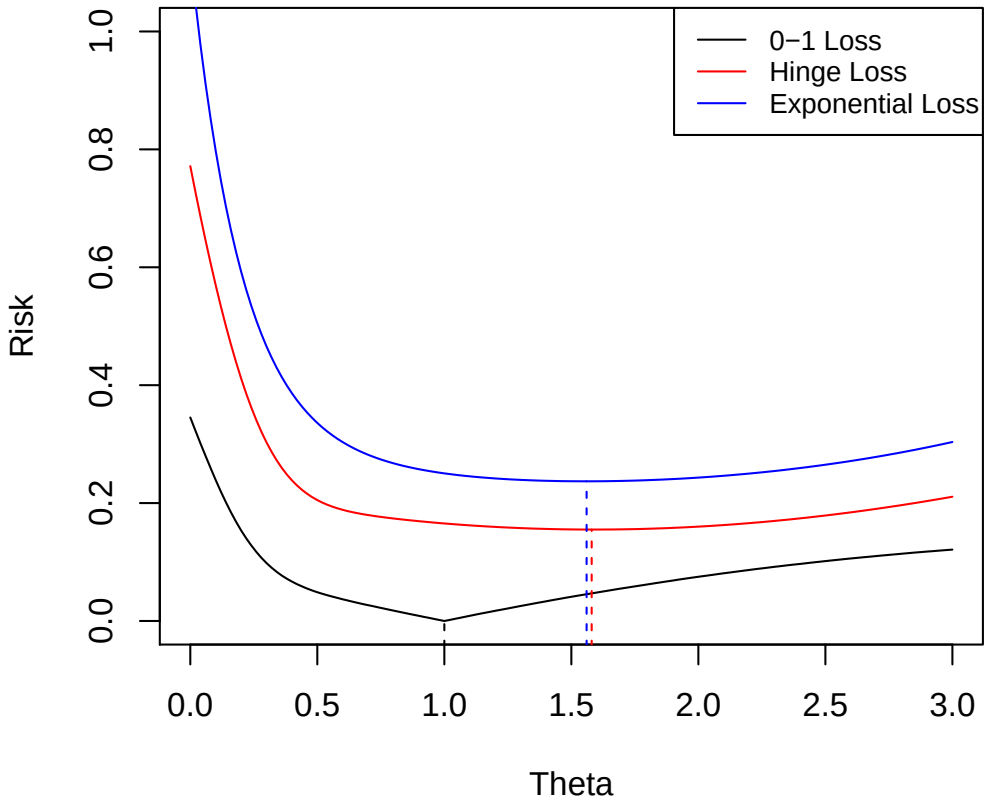
- **Setup**: Observe n i.i.d copies of (X, Y, Z)
- **Goal**: Est optimal linear threshold

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, R(\theta), \quad R(\theta) = \mathbb{E}\Big[L_{01}\Big(Y(X - \theta^T Z)\Big)\Big]$$

- $L_{01}(u) = \frac{1}{2}(1 - \operatorname{sign}(u))$ is the 0-1 loss.
- **Empirical risk minimization** (ERM):
  - ⊗ Empirical risk minimization with 0-1 loss is NP hard
  - ⊗ Convex surrogate loss is not Fisher consistent

Estimation vs Classification

- SVM and Adaboost are not Fisher consistent for est θ
- $X \sim N(0,1)$, $Z \in \{0.5, 5\}$ and $Y = \text{sign}(X - Z)$

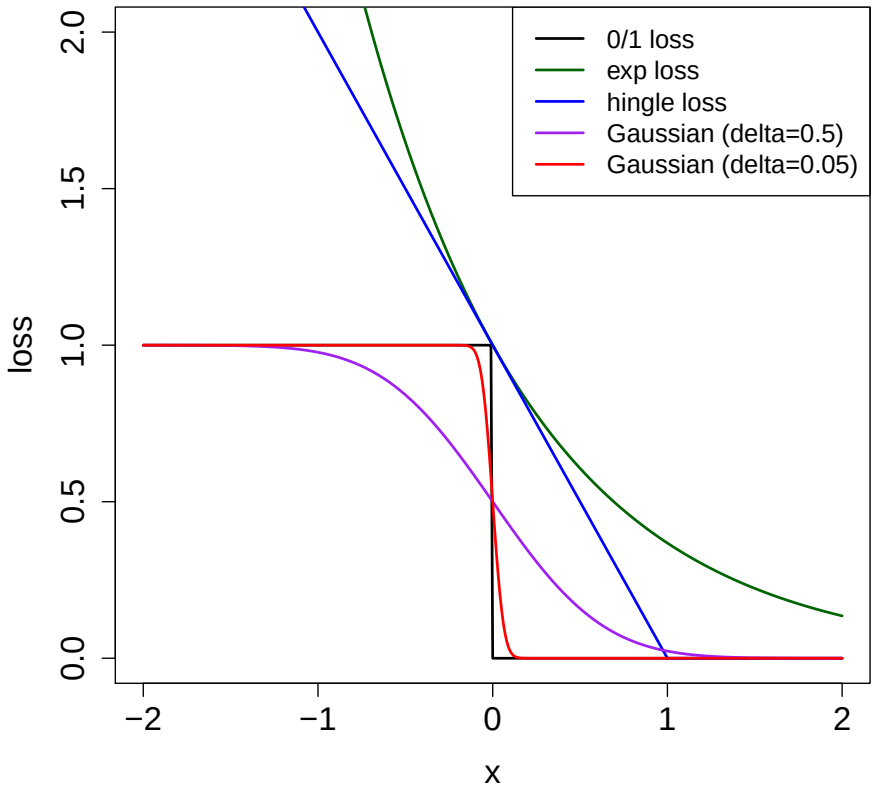


Smoothed Surrogate Loss

- Smoothed surrogate loss

$$L_\delta(u) = \int_{u/\delta}^{\infty} K(t)dt$$

- $K(t)$ kernel function and δ bandwidth parameter



Active Subsampling Algorithm

Active Subsampling Algorithm

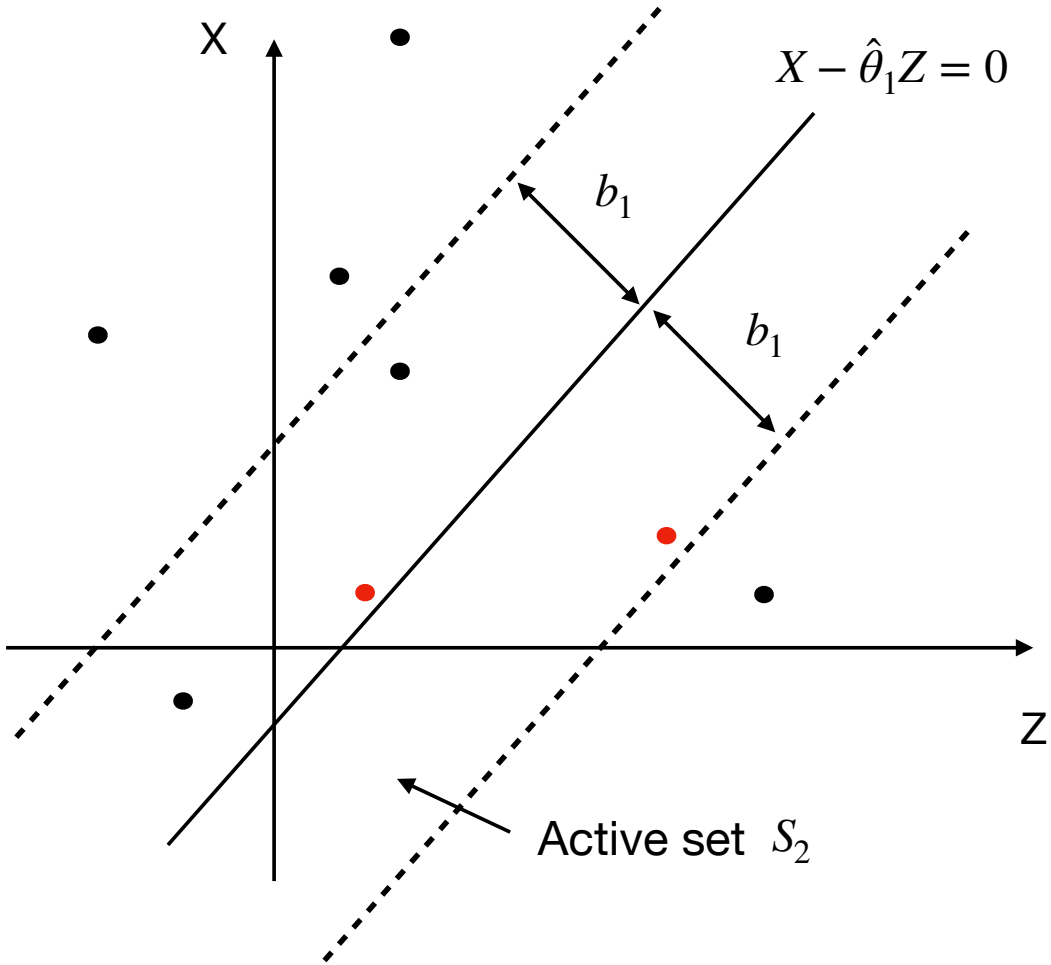
- Unlabeled data: $D = \{X_i, Z_i\}_{i=1}^{N}$ with N i.i.d. samples
- Label budget, $n \ll N$
- Split D into K batches: $D_1, \cdots, D_K$, with equal batch size
- $R_i = 1$ if (X_i, Z_i) is sampled and $R_i = 0$ otherwise
- Estimation via ERM

$$R_{\delta_k}^{D_k}(\theta) = \frac{1}{|D_k|} \sum_{i \in D_k} L_{\delta_k}(Y_i(X_i - \theta^T Z_i)) R_i,$$

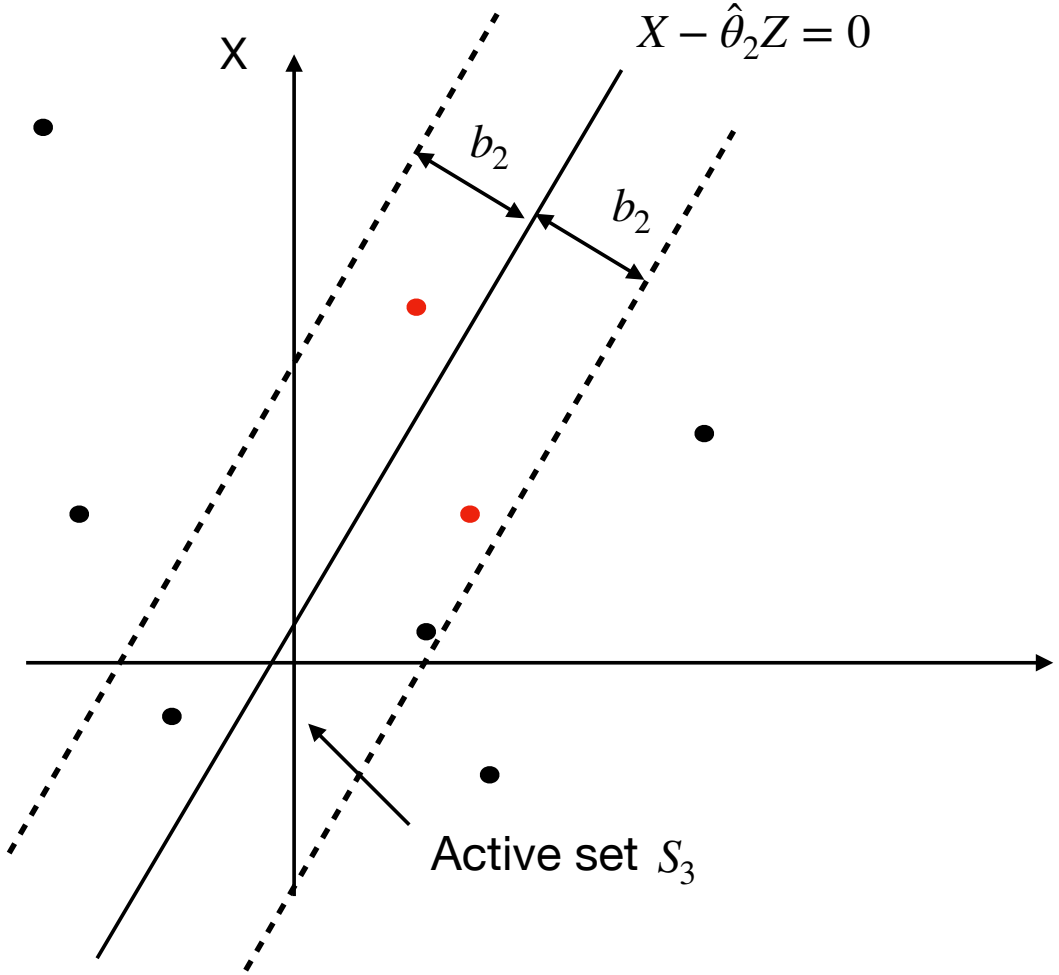
$$\hat{\theta}_k = \underset{\theta}{\operatorname{argmin}} \{R_{\delta_k}^{D_k}(\theta) + \lambda_k \|\theta\|_1\},$$

- Sampling scheme?

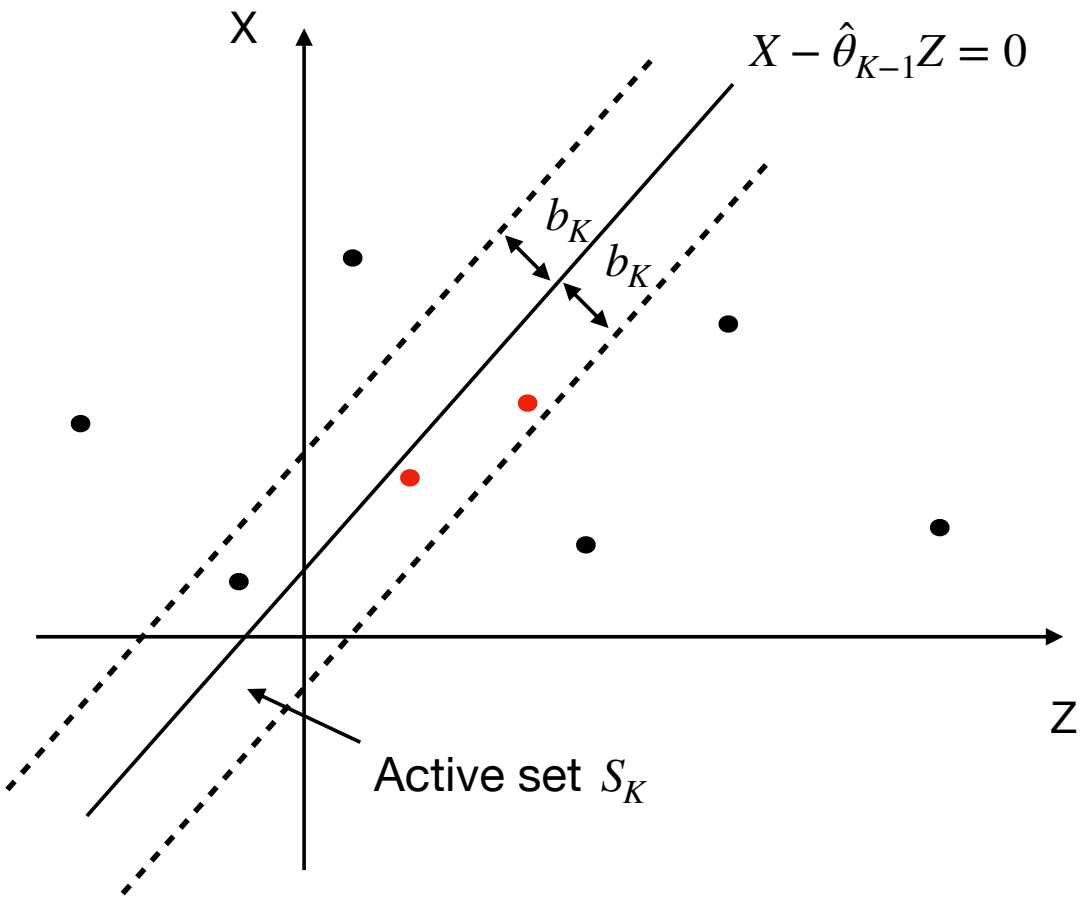
Sampling Scheme



Sampling Scheme



Sampling Scheme



Statistical Theory

Assumptions

- θ is s-sparsity
- $f(x|y,z)$ is β-smooth, ...
- Z given $Y = y$ is a sub-Gaussian vector
- RSC/RSM...

Master Theorem

Master Theorem (informal)

For $1 \leq k \leq K$, with proper choice of λ_k and δ_k,

$$\|\hat{\theta}_1 - \theta\|_2 \lesssim \left(\frac{s \log d}{n_1}\right)^{\frac{\beta \vee 1}{2\beta+1}}.$$

If we further assume

$$b_{k-1} \geq C\delta_k, \quad b_{k-1} \geq C\|\hat{\theta}_{k-1} - \theta\|_2 \sqrt{\log \frac{n_k}{s \log d}},$$

then

$$\|\hat{\theta}_k - \theta^*\|_2 \lesssim \left(\frac{\mathbb{P}\left((X, Z) \in S_k\right) s \log d}{n_k}\right)^{\frac{\beta \vee 1}{2\beta+1}},$$

where n_k is the label budget at the kth iteration.

Master Theorem: Discussion

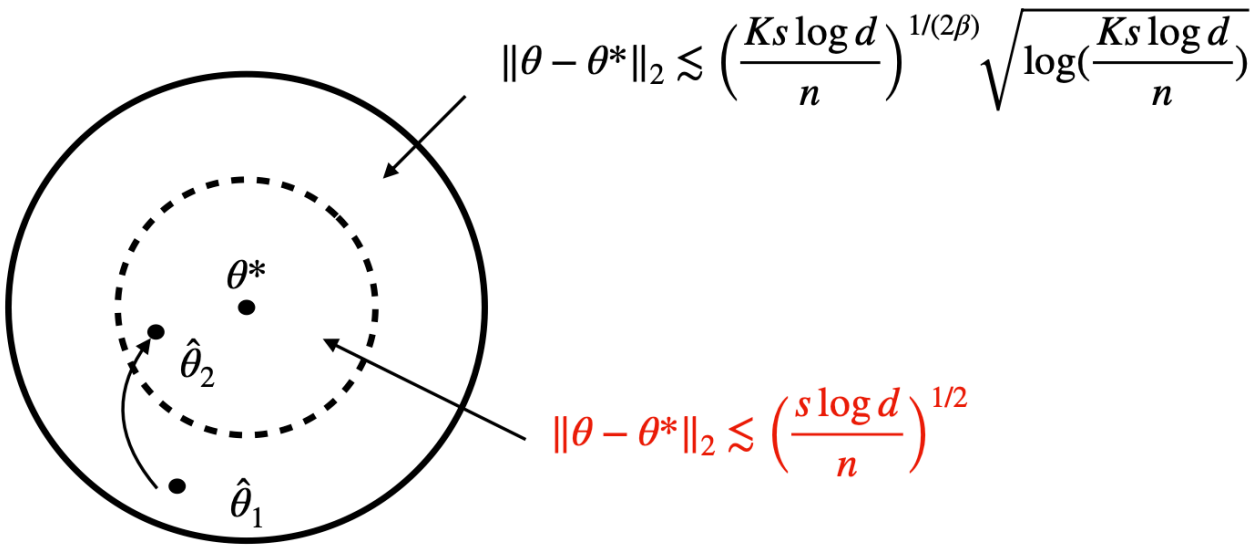
Master Theorem (informal)

$$\|\hat{\theta}_1 - \theta\|_2 \lesssim \left(\frac{s \log d}{n_1}\right)^{\frac{\beta \vee 1}{2\beta+1}}$$

$$\|\hat{\theta}_k - \theta\|_2 \lesssim \left(\frac{\mathbb{P}((X,Z) \in S_k)\, s \log d}{n_k}\right)^{\frac{\beta \vee 1}{2\beta+1}}$$

- Nonstandard rate (due to bias and variance tradeoff)
- $\mathbb{P}((X,Z) \in S_k) \asymp b_{k-1}$, faster rate if $b_{k-1} \to 0$.
- Optimize $\{\lambda_k\}_{k=1}^K, \{b_k\}_{k=1}^{K-1}, \{\delta_k\}_{k=1}^K, \{n_k\}_{k=1}^K$ and K

Optimal rate for $\beta > \frac{1+\sqrt{3}}{2}$



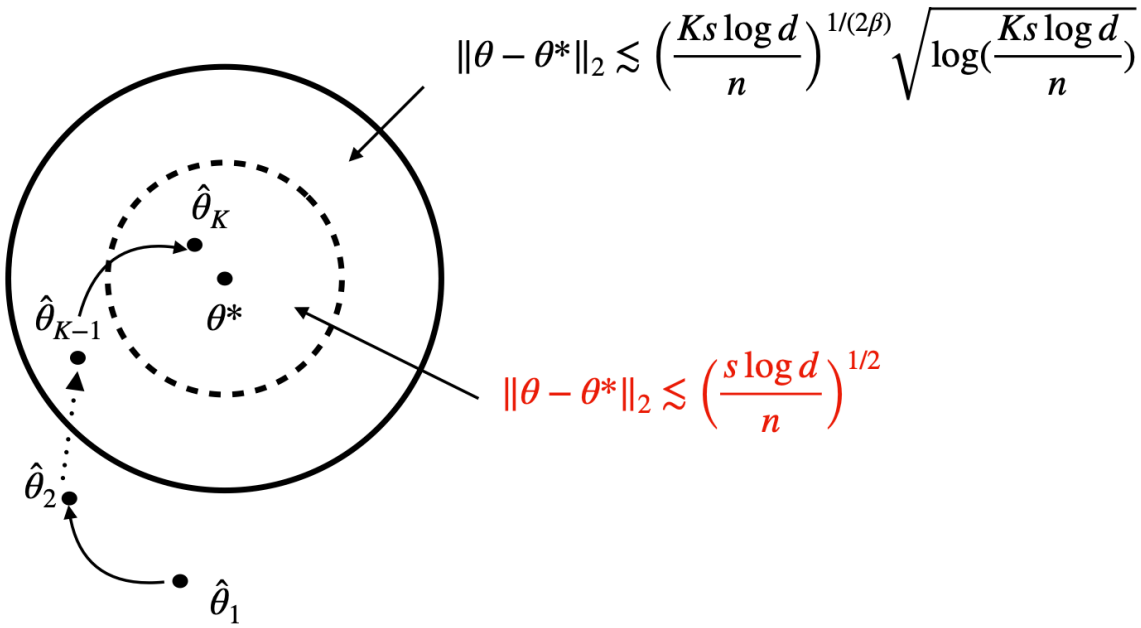
Optimal rate for $\beta > \frac{1+\sqrt{3}}{2}$ (informal)

Uniformly over $2 \le k \le K$,

$$\|\hat{\theta}_k - \theta\|_2 \lesssim \left(\frac{s\log d}{n}\right)^{1/2}$$

faster than the passive rate $\left(\frac{s\log d}{n}\right)^{\frac{\beta}{2\beta+1}}$.

Optimal rate for $1 < \beta \leq \frac{1+\sqrt{3}}{2}$



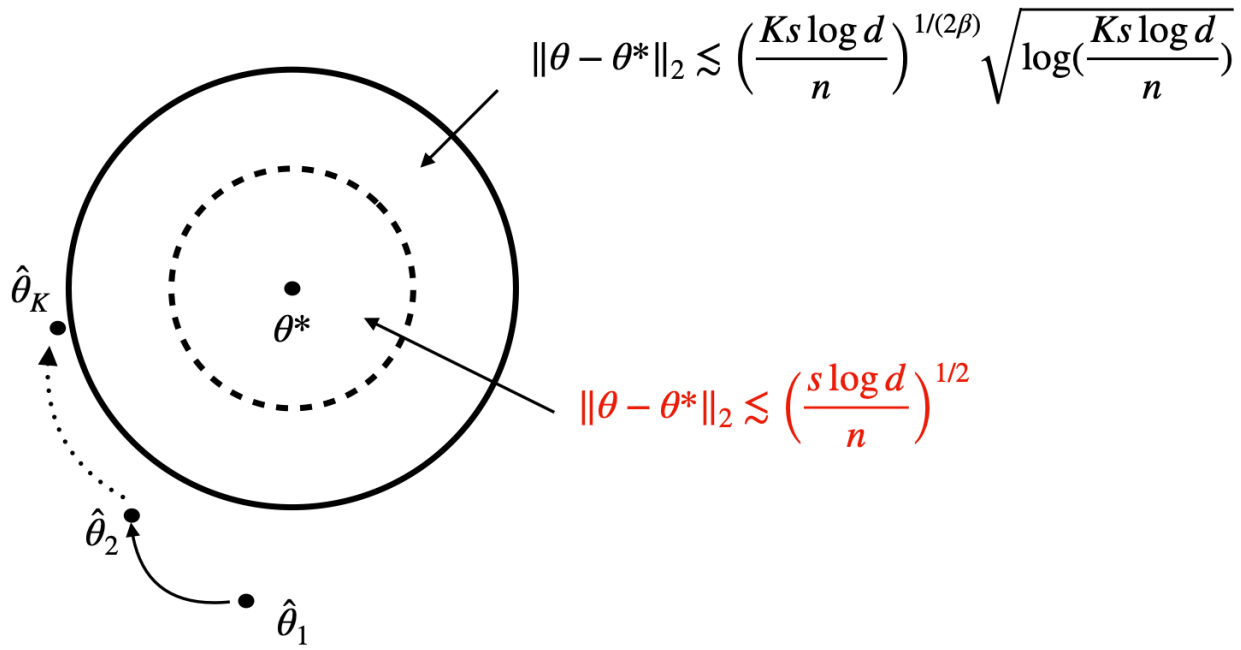
Optimal rate for $1 < \beta \leq \frac{1+\sqrt{3}}{2}$ (informal)

For $K = \lceil \log_{\frac{\beta}{2\beta+1}}(1 - \frac{\beta+1}{2\beta^2}) \rceil + 1$,

$$\|\hat{\theta}_K - \theta\|_2 \lesssim \left(\frac{s\log d}{n}\right)^{1/2}$$

faster than the passive rate $\left(\frac{s\log d}{n}\right)^{\frac{\beta}{2\beta+1}}$.

Optimal rate for $\beta \leq 1$



Optimal rate for $\beta \leq 1$ (informal)

For $K = \lceil \log_{2\beta+1}(\log N) \rceil$,

$$\|\hat{\theta}_K - \theta\|_2 \lesssim \left(\log(\frac{n}{Ks\log d})\right)^{\frac{1}{4\beta}} \left(\frac{Ks\log d}{n}\right)^{\frac{1}{2\beta}}$$

faster than the passive rate $\left(\frac{s\log d}{n}\right)^{\frac{1}{2\beta+1}}$.

Numerical Results

Practical Considerations

- Nonconvex optimization
- We recommend $K = 2$
- Cross-validation for λ_1, λ_2 and b_1
- $\delta_1 = \delta_2 = 1$
- We recommend larger value for n_2 (e.g., 80%)

Simulation Setting

- **Binary response mode**:

$$Y = \text{sign}(X - \theta^T Z + u),$$

u is a random error with Median$(u|X, Z) = 0$.

- **Conditional mean model**: Consider $Y \in \{-1, 1\}$,

$$X = \mu Y + \theta^T Z + u,$$

$u \perp (Y, Z)$ and $\mu > 0$ is a constant.

Simulation Setting

- **Logistic model**:

  $X \sim N(0,1)$, $Z \sim N_d(0, \mathbf{I})$

- **Binary response model**:

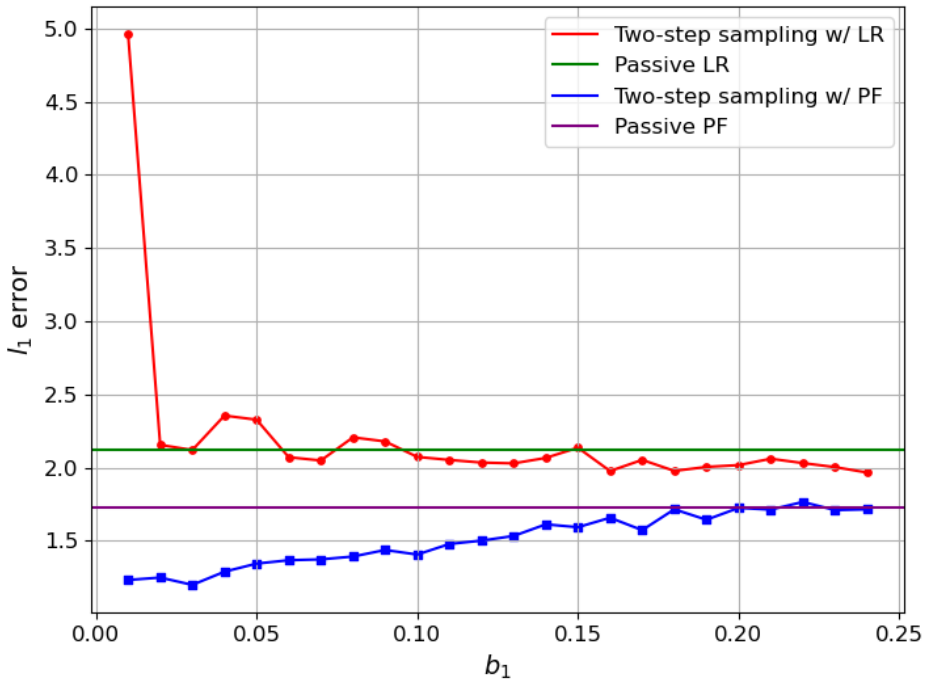
  $X \sim N(0,1)$, $Z \sim N_d(0, \mathbf{I})$, and $u \sim N(0, \sigma^2(1 + 2(X - \theta^T Z)^2))$.

- **Conditional mean model**:

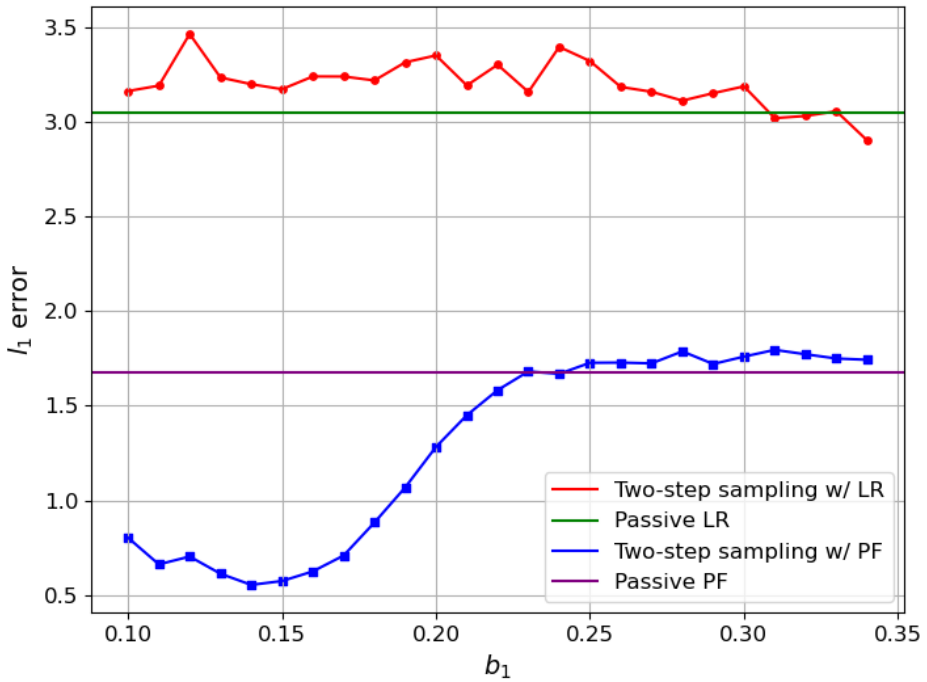
  $Y \sim \text{Uniform}(\{-1, 1\})$, $Z \sim N_d(0, \mathbf{I})$, $u \sim N(0, (0.1)^2)$ and $\mu = 2$.

- $N = 20000$, $n = 2000$, $d = 200$, $s = 10$ and $\|\theta\|_2 = 1$

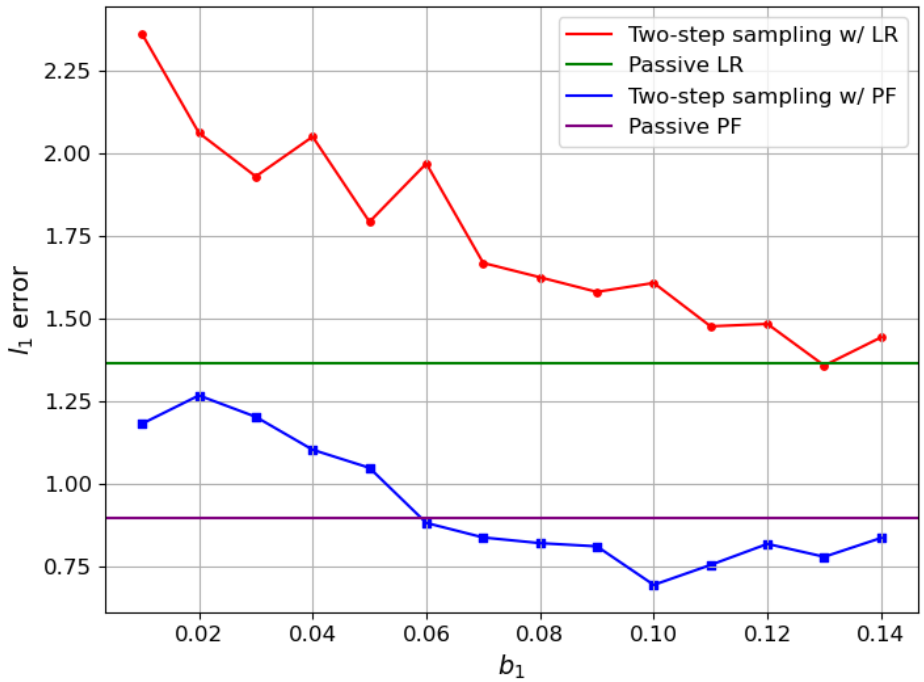
Est Error: Logistic Model



Est Error: Conditional Mean Model



Est Error: Binary Response Model



Est Error via CV

Error	Conditional mean model		Logistic model		Binary response model	
	Two-step	Passive	Two-step	Passive	Two-step	Passive
ℓ_1	0.918	1.648	1.514	1.625	0.835	0.937
ℓ_2	0.313	0.525	0.525	0.559	0.319	0.341
ℓ_∞	0.150	0.196	0.270	0.275	0.192	0.187

Hospital Data

- 101,766 hospital records collected over a decade (1999- 2008) from 130 US hospitals.
- Race, gender, age, admission type, readmission status, length of hospital stay, medical specialty of the admitting physician, number of lab tests performed, Hemoglobin A1c (HbA1c) test results, diagnosis...
- Y: $+1$ if the patient was readmitted within 30 days of discharge, and -1 otherwise
- X: HbA1c test results

Goal: est linear threshold for X to predict patients' readmission.

Est Results

n	3000		4000		5000	
	Two-step	Passive	Two-step	Passive	Two-step	Passive
ℓ_1	2.422	3.192	0.968	1.276	0.379	0.651
ℓ_2	1.403	2.478	1.649	2.048	0.269	0.543
ℓ_∞	0.869	2.369	0.824	2.048	0.203	0.529

Summary

- Active subsampling improves passive learning
- Propose a practically useful active subsampling algorithm
- Establish optimal statistical rate of convergence
- Establish n-budget minimax lower bound and adaptation results

Thank You !