COMPUTATIONAL ASPECTS OF SELECTION OF EXPERIMENTS IN REGRESSION MODELS

Yining Wang

Machine Learning Department, Carnegie Mellon University

EXP SELECTION IN LINEAR REGRESSION

- The linear regression model:

$$y = X\beta_0 + \varepsilon, \qquad X \in \mathbb{R}^{n\times p}$$

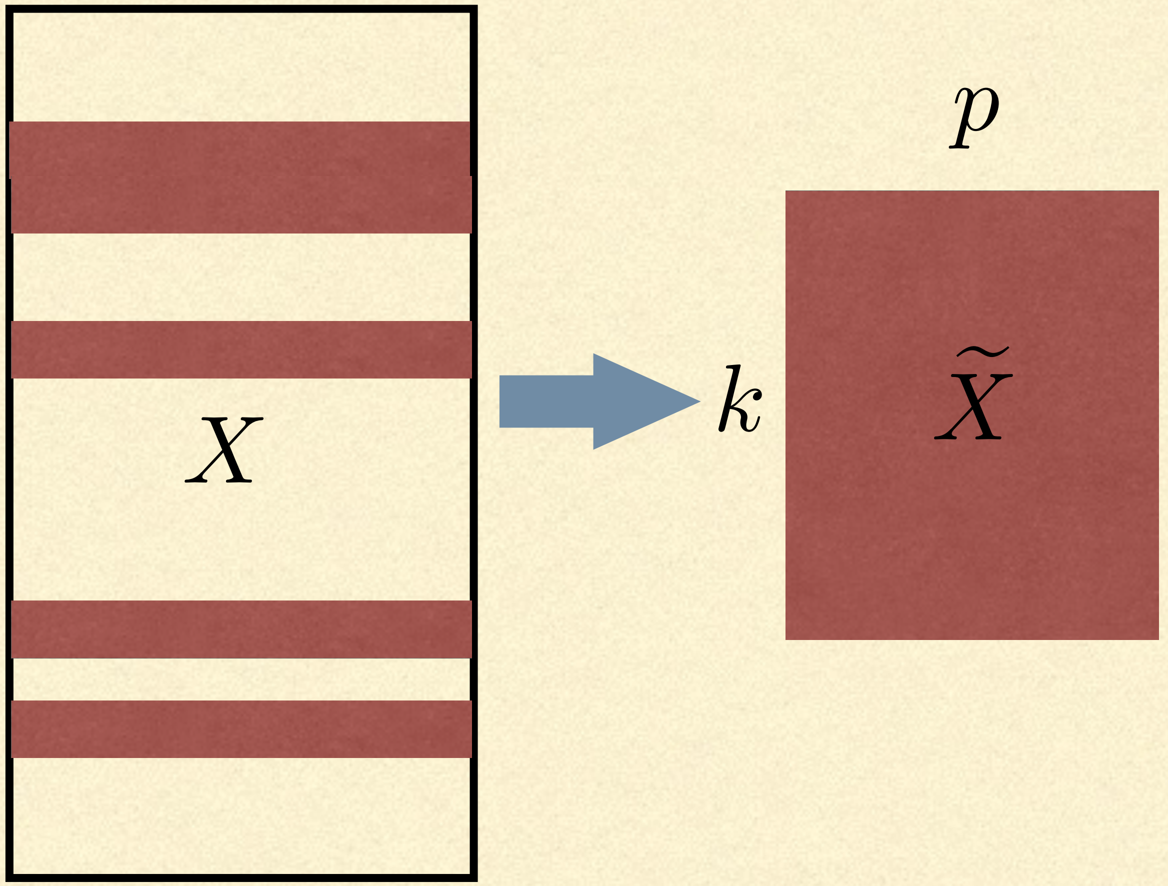
    - We consider the low-dimensional regime: $p < n$

- The subset selection problem: find $p \leq k \ll n$ rows of X that are most "informative" in estimating β_0

- Also known as *experimental design* in statistics literature

EXP SELECTION IN LINEAR REGRESSION

- Motivating examples
  - Material synthesis
    - select "representative" experimental settings
  - Wind speed prediction
    - select "important" locations to measure wind speed



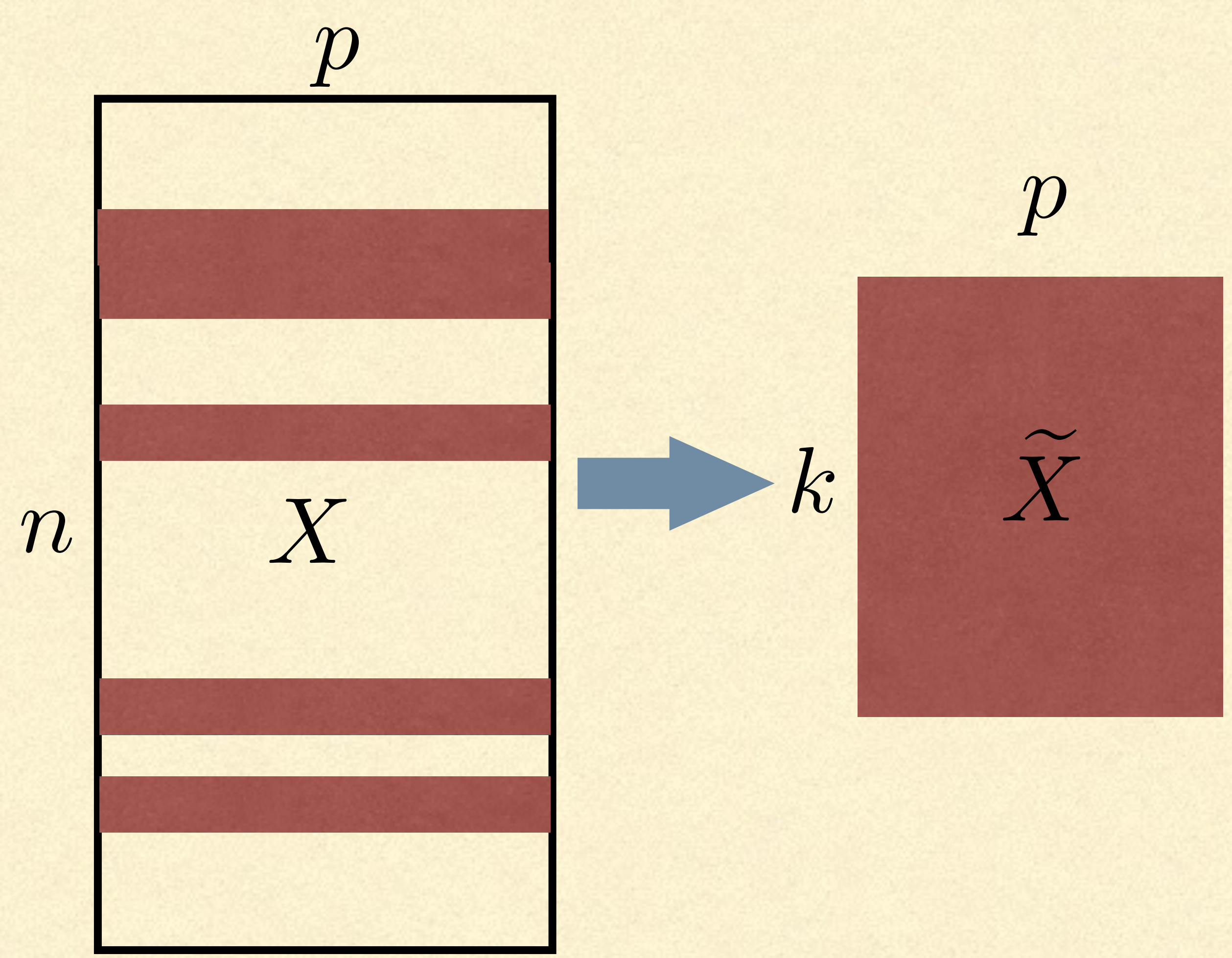
SUBSET SELECTION IN LINEAR REGRESSION

- What do we mean by "informative"?
  - Many criteria exist
  - "Average" Mean-square error:

$$\inf_{\tilde{X}} \mathbb{E}\|\hat{\beta} - \beta_0\|_2^2$$

  - Also known as *A-optimality*

$$\min_{\widetilde{X}} \operatorname{tr}\left[(\tilde{X}^\top \tilde{X})^{-1}\right]$$



COMPUTATIONAL ASPECTS

- The combinatorial A-optimality is difficult to compute:

$$f_{\mathrm{opt}}(k) = \min_{\tilde{X} \in \mathbb{R}^{k \times p}} \mathrm{tr}\left[(\tilde{X}^\top \tilde{X})^{-1} \right]$$

  - Time complexity for brute-force search: $O(n^k)$

- The computational question: polynomial-time algorithm with

$$f(\tilde{X}) \leq C(k,p) \cdot f_{\mathrm{opt}}(k)$$

EXTENSIONS

- Generalized linear model $\eta_i = x_i^\top \beta_0$

$$I(X, \beta_0) = \sum_{i=1}^{n} \mathbb{E} \frac{\partial \log p(y_i | x_i; \beta_0)}{\partial \beta \partial \beta^\top} = \sum_{i=1}^{n} \left(\mathbb{E} \frac{\partial^2 \log p(y_i; \eta_i)}{\partial \eta_i^2} \right) x_i x_i^\top$$

- Reduction to the ordinary linear regression:

$$\tilde{x}_i = \sqrt{\mathbb{E} \frac{\partial^2 \log p(y_i; \eta_i)}{\partial \eta^2}} x_i$$

- Problem: η_i depends on the unknown model parameter
- Solution: *locally* optimal designs.

EXTENSIONS

- Delta's method: estimating $g(\beta_0)$
- Include both in-sample and out-sample predictions

$$\mathrm{tr}\left[\nabla g(\beta_0)(X_S^\top X_S)^{-1}\nabla g(\beta_0)\right] = \mathrm{tr}\left[G_0(X_S^\top X_S)^{-1}\right]$$

$$G_0 = \nabla g(\beta_0)^\top \nabla g(\beta_0) \in \mathbb{R}^{p\times p}$$

- If $G_0 = PP^T$ is invertible:

$$\tilde{x}_i = P^{-1}x_i$$

A CONVEX RELAXATION

- A continuous relaxation: $$f_{\text{opt}}(k) = \min_{\tilde{X} \in \mathbb{R}^{k \times p}} \operatorname{tr}\left[(\tilde{X}^\top \tilde{X})^{-1}\right]$$

$$f^* = \min_{\pi \in \mathbb{R}^p} \operatorname{tr}\left[\left(\sum_{i=1}^n \pi_i x_i x_i^\top\right)^{-1}\right],$$

$$s.t.\;\; \pi \geq 0, \|\pi\|_1 \leq 1, \|\pi\|_\infty \leq 1/k.$$

- The continuous optimization problem is convex.

$$f^* \leq k f_{\text{opt}}(k)$$

A CONVEX RELAXATION

- A few words on how to solve the convex optimization …

$$f^* = \min_{\pi \in \mathbb{R}^p} \operatorname{tr}\left[\left(\sum_{i=1}^n \pi_i x_i x_i^\top\right)^{-1}\right],$$

$$s.t.\ \ \pi \geq 0, \|\pi\|_1 \leq 1, \|\pi\|_\infty \leq 1/k.$$

- *Projective gradient descent:* $\pi^{(t+1)} = \mathcal{P}\left(\pi^{(t)} - \lambda_t \nabla f(\pi^{(t)})\right)$
- Gradient computation: $\frac{\partial f}{\partial \pi_i} = \|\tilde{\Sigma}^{-1} x_i\|_2^2, \tilde{\Sigma} = X^\top \mathrm{diag}(\pi) X.$
- Projection onto $\mathbb{B}_1(1) \cap \mathbb{B}_0(1/k)$: can be done in $O(n \log^2 n)$

SUBSET SELECTION IN LINEAR REGRESSION

- The "only" problem left:

How to turn π^* into a valid subset $\tilde{X}$?

- A simple approach: sampling
  - Sample each row of X with probability $\{\pi_i^*\}_{i=1}^n$
  - Sample *without* replacement

SAMPLING BASED SUBSET SELECTION

- Performance guarantee:

Theorem. Suppose $B = \max_{1 \le i \le n} \|x_i\|_2$ and $\tilde{\Sigma}_* = X^\top \mathrm{diag}(\pi^*) X$. If k satisfies $k \ge \Omega(B^2 \|\tilde{\Sigma}_*^{-1}\|_2 \log n)$ then with probability $1 - O(n^{-1})$

$$f(\tilde{X}) \le O(\log k) \cdot f_{\mathrm{opt}}(k)$$

- Note: $\Omega(B^2 \|\tilde{\Sigma}_*^{-1}\|_2 \log n) \ge \Omega(p \log n)$

- Proof technique: spectral sparsification. [Spielman and Srivastava' 08, *Graph Sparsification by Effective Resistance*]

SAMPLING BASED SUBSET SELECTION

- Spectral approximation:

$$(1-\delta)z^\top \Sigma z \le z^\top \tilde{\Sigma} z \le (1+\delta) z^\top \Sigma z, \qquad \forall z \in \mathbb{R}^p$$

- Goal: find a subset of X such that $\tilde{\Sigma}$ is a spectral approximation of $\tilde{\Sigma}_* = X^\top \mathrm{diag}(\pi^*) X$
  - Immediately yields

$$f(\tilde{X}) \le \frac{k}{1-\delta} f^* \le \frac{1}{1-\delta} f_{\mathrm{opt}}(k)$$

SAMPLING BASED SUBSET SELECTION

- Consider *with replacement* sampling first.
- Define: $$\Pi = \underbrace{\Phi^{1/2}}_{\mathrm{diag}(\pi^*)} X \underbrace{\tilde{\Sigma}_*^{-1}}_{X^\top \mathrm{diag}(\pi^*) X} X^\top \Phi^{1/2}$$
    - (P1). $\lambda_1(\Pi) = \cdots = \lambda_p(\Pi) = 1, \lambda_{p+1}(\Pi) = \cdots = \lambda_n(\Pi) = 0$
    - (P2). $\mathrm{Range}(\Pi) = \mathrm{Range}(\Phi^{1/2} X)$
    - (P3). $\|\Pi_{i\cdot}\|_2^2 = \pi_i^* x_i^\top \tilde{\Sigma}_*^{-1} x_i$

SAMPLING BASED SUBSET SELECTION

Lemma. If $\|\Pi S \Pi - \Pi\|_2 \leq \delta$ then $X^\top \Phi^{1/2} S \Phi^{1/2} X$ is a spectral approximation of $X^\top \Phi X$

- Proof.

$$\frac{z^\top (X^\top \Phi^{1/2} S \Phi^{1/2} X - X^\top \Phi X) z}{z^\top X^\top \Phi X z} = \frac{\tilde{z}^\top (S - I) \tilde{z}}{\tilde{z}^\top \tilde{z}}$$

(Because $\tilde{z}$ is in the range of Π) $$= \frac{\tilde{z}^\top \Pi (S - I) \Pi \tilde{z}}{\tilde{z}^\top \tilde{z}}$$

$$\leq \|\Pi S \Pi - \Pi\|_2$$

SAMPLING BASED SUBSET SELECTION

Lemma. Suppose $S_{ii} = 1/\pi_i^*$ with probability π_i^* and 0 otherwise. Then

$$\Pr\left[\|\Pi S \Pi - \Pi\|_2 > \delta\right] \leq n \exp\left\{-c \cdot \frac{k\delta^2}{B^2 \|\tilde{\Sigma}_*^{-1}\|_2}\right\}$$

- Proof (use matrix Chernoff):
  - Unbiased sub-sampling: $\mathbb{E}[\Pi S \Pi] = \Pi$
  - Recall that $\|\Pi_{i\cdot}\|_2^2 = \pi_i^* x_i^\top \tilde{\Sigma}_*^{-1} x_i$

SAMPLING BASED SUBSET SELECTION

- Taking care of sampling *without replacement*
  - Say we have $\tilde{X}$ sampled *with replacement*
  - $\tilde{X}$ has *O*(log *k*) duplicates because $\|\pi^*\|_\infty \leq 1/k$
  - Remove all duplicates in $\tilde{X}$

$$f(\tilde{X}^{\mathrm{new}}) \leq O(\log k) \cdot f(\tilde{X})$$

SAMPLING BASED SUBSET SELECTION

- Summary

Theorem. Suppose $B = \max_{1 \le i \le n} \|x_i\|_2$ and $\tilde{\Sigma}_* = X^\top \mathrm{diag}(\pi^*) X$. If k satisfies $k \ge \Omega(B^2 \|\tilde{\Sigma}_*^{-1}\|_2 \log n)$ then with probability $1 - O(n^{-1})$

$$f(\tilde{X}) \le O(\log k) \cdot f_{\mathrm{opt}}(k)$$

- Two issues :
  - $O(\log k)$ approximation ratio instead of $(1+\epsilon)$
  - Lower bound of k depends on the conditioning of $\tilde{\Sigma}_*$

GREEDY BASED SUBSET SELECTION

- An interesting *greedy* algorithm presented in Avron and Boutsidis 2012, *Faster Subset Selection For Matrices and Applications*:

1. Start with the full subset $S = \{1, \cdots, n\}$
2. Remove one row in *S* that results in the smallest $f(S')$
3. Repeat step 2 until $|S| = k$

Theorem 3.1 [AB'12]. $\mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right] \leq \frac{n-p+1}{k-p+1}\mathrm{tr}\left[\left(X^\top X\right)^{-1}\right]$

GREEDY BASED SUBSET SELECTION

- Proof idea:

  For $n \times p$ full-rank matrix *A*, there exists a $(n-1) \times p$ matrix *B* such that $$\mathrm{tr}((B^\top B)^{-1}) \leq \frac{n-p+1}{n-p}\mathrm{tr}((A^\top A)^{-1})$$

  - Why?
    - Volume sampling: $\Pr[S] \propto \det(X_S^\top X_S)$
    - Claim: $\mathbb{E}_{|S|=k}\left[\mathrm{tr}((X_S^\top X_S)^{-1}\right] \leq \frac{n-p+1}{k-p+1}\mathrm{tr}((X^\top X)^{-1})$
    - Proof quite complicated.

GREEDY BASED SUBSET SELECTION

Theorem 3.1 [AB'12]. $\mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right] \leq \frac{n-p+1}{k-p+1}\mathrm{tr}\left[\left(X^\top X\right)^{-1}\right]$

- Some notable limitations
  - Additive guarantee: depends on $\mathrm{tr}\left[\left(X^\top X\right)^{-1}\right]$ instead of $f_{\mathrm{opt}}(k)$
  - Computationally heavy: $O(n^2p^2)$ computations at least.
- A better idea: greedy removal on $X^\top \mathrm{diag}(\pi^*)X$

GREEDY BASED SUBSET SELECTION

- Define $S_0 = \mathrm{supp}(\pi^*)$
- Recall that $\|\pi^*\|_\infty \le 1/k$

$$f^* = \min_{\pi \in \mathbb{R}^p} \mathrm{tr}\left[\left(\sum_{i=1}^n \pi_i x_i x_i^\top\right)^{-1}\right],$$
$$s.t. \;\; \pi \ge 0, \|\pi\|_1 \le 1, \|\pi\|_\infty \le 1/k.$$

$$\mathrm{tr}\left[\left(\frac{1}{k} X_{S_0}^\top X_{S_0}\right)^{-1}\right] \le f^* \le k f_{\mathrm{opt}}(k) \implies \mathrm{tr}\left[\left(X_{S_0}^\top X_{S_0}\right)^{-1}\right] \le f_{\mathrm{opt}}(k)$$

- Run the greedy algorithm on X_{S_0} :

$$\mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right] \le \frac{|S_0| - p + 1}{k - p + 1} \mathrm{tr}\left[\left(X_{S_0}^\top X_{S_0}\right)^{-1}\right]$$
$$\le \frac{|S_0| - p + 1}{k - p + 1} \cdot f_{\mathrm{opt}}(k)$$

GREEDY BASED SUBSET SELECTION

$$\mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right] \le \frac{|S_0| - p + 1}{k - p + 1} \cdot f_{\mathrm{opt}}(k)$$

$$\approx 1 + \frac{|S_0|}{k}$$

- Key: upper bound $|S_0| = \|\pi^*\|_0$
- Intuition: the L1 constraint on π should encourage sparsity

GREEDY BASED SUBSET SELECTION

$$f^* = \min_{\pi \in \mathbb{R}^p} \operatorname{tr}\left[\left(\sum_{i=1}^{n} \pi_i x_i x_i^\top\right)^{-1}\right],$$

$$s.t. \;\; \pi \geq 0, \|\pi\|_1 \leq 1, \|\pi\|_\infty \leq 1/k.$$

- The Lagrangian multiplier:

$$\mathcal{L}(\pi; \lambda, \tilde{\lambda}, \mu) = f(\pi) - \sum_{i=1}^{n} \lambda_i \pi_i + \sum_{i=1}^{n} \tilde{\lambda}_i \left(\pi_i - \frac{1}{k}\right) + \mu\left(\sum_{i=1}^{n} \pi_i - 1\right)$$

- KKT condition:

$$x_i^\top \tilde{\Sigma}_*^{-2} x_i = \tilde{\lambda}_i - \lambda_i + \mu$$

GREEDY BASED SUBSET SELECTION

$\lambda_i : \pi_i \geq 0$

$\tilde{\lambda}_i : \pi_i \leq 1/k$

$\mu : \sum_{i=1}^{p} \pi_i \leq 1$

- KKT condition: $x_i^\top \tilde{\Sigma}_*^{-2} x_i = \tilde{\lambda}_i - \lambda_i + \mu$
- Define:
  $A = \{i : \pi_i^* = 1/k\}, B = \{i : 0 < \pi_i^* < 1/k\}, C = \{i : \pi_i^* = 0\}$
- Some facts:
  - *A* has at most *k* elements, and $\|\pi^*\|_0 = |A| + |B| \leq k + |B|$
  - Complementary Slackness: $\forall i \in B, \tilde{\lambda}_i = \lambda_i = 0$

$$x_i^\top \tilde{\Sigma}_*^{-1} x_i = C, \quad \forall i \in B$$

GREEDY BASED SUBSET SELECTION

- $x_i^\top \tilde{\Sigma}_*^{-1} x_i = C, \quad \forall i \in B$

Theorem. Under regularity conditions, the system $x_i^\top A x_i = C$ for all $x_i \in X_S$ has no solution if $|S| > p(p+1)/2$

- Proof. Define

$$\Phi \in \mathbb{R}^{|S| \times p(p+1)/2} \longleftarrow \begin{bmatrix} \phi(x_1) \\ \vdots \\ \phi(x_{|S|}) \end{bmatrix} \begin{bmatrix} \mathrm{vec}(\Sigma) \\ -C \end{bmatrix} = 0$$

GREEDY BASED SUBSET SELECTION

$$\Phi = \begin{bmatrix} \phi(x_1) \\ \vdots \\ \phi(x_{|S|}) \end{bmatrix} \in \mathbb{R}^{|S| \times p(p+1)/2}$$

- If Φ is has full column rank and $|S| > p(p+1)/2$, then $\Phi z = 0$ has no solution except for $z = 0$
- Sufficient if *X* is a random design with an absolutely continuous distribution
- Consequence: $|S_0| \leq k + p^2$

GREEDY BASED SUBSET SELECTION

- Summary:

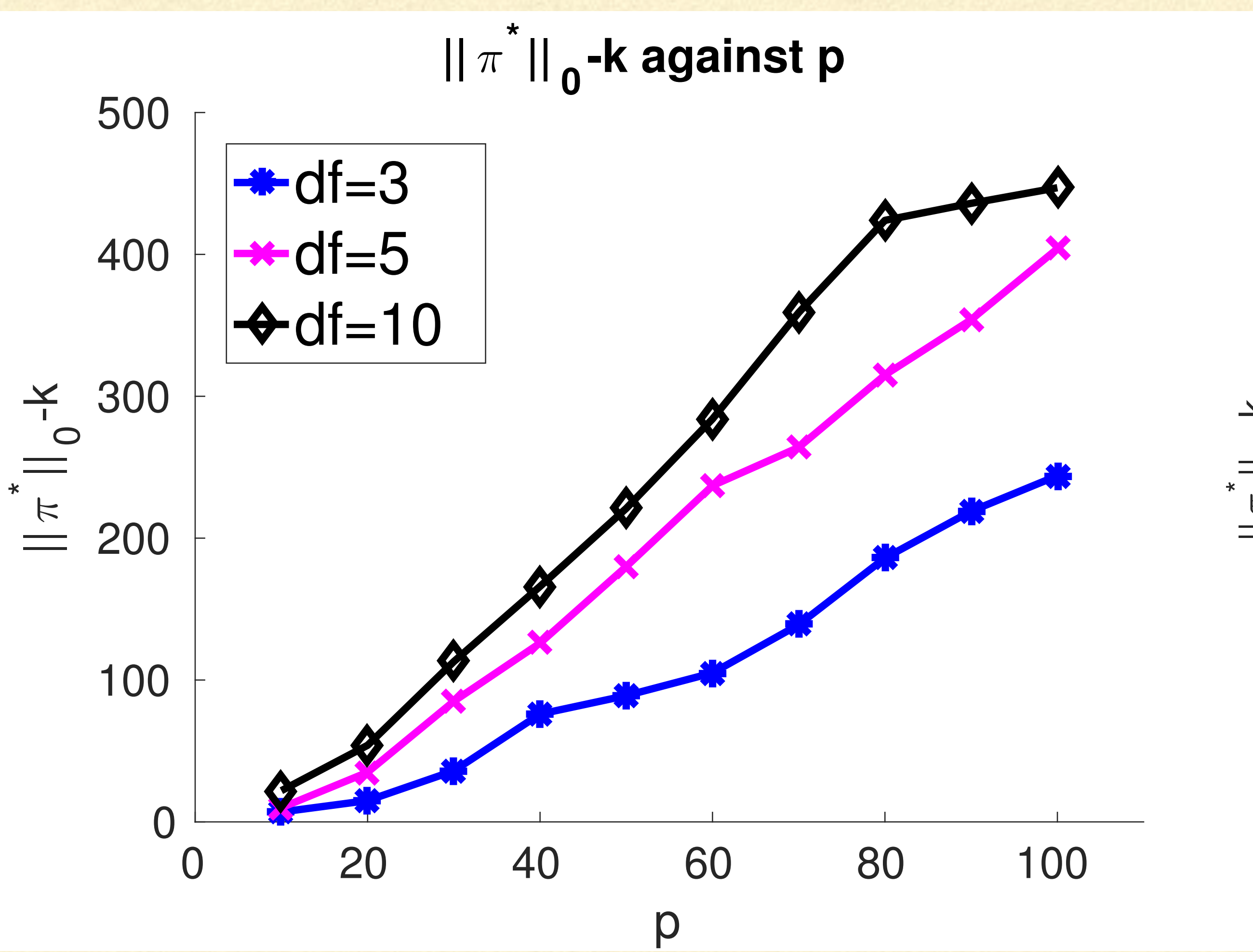
Theorem. Suppose $k \geq 2p$. Under regularity conditions, we have that

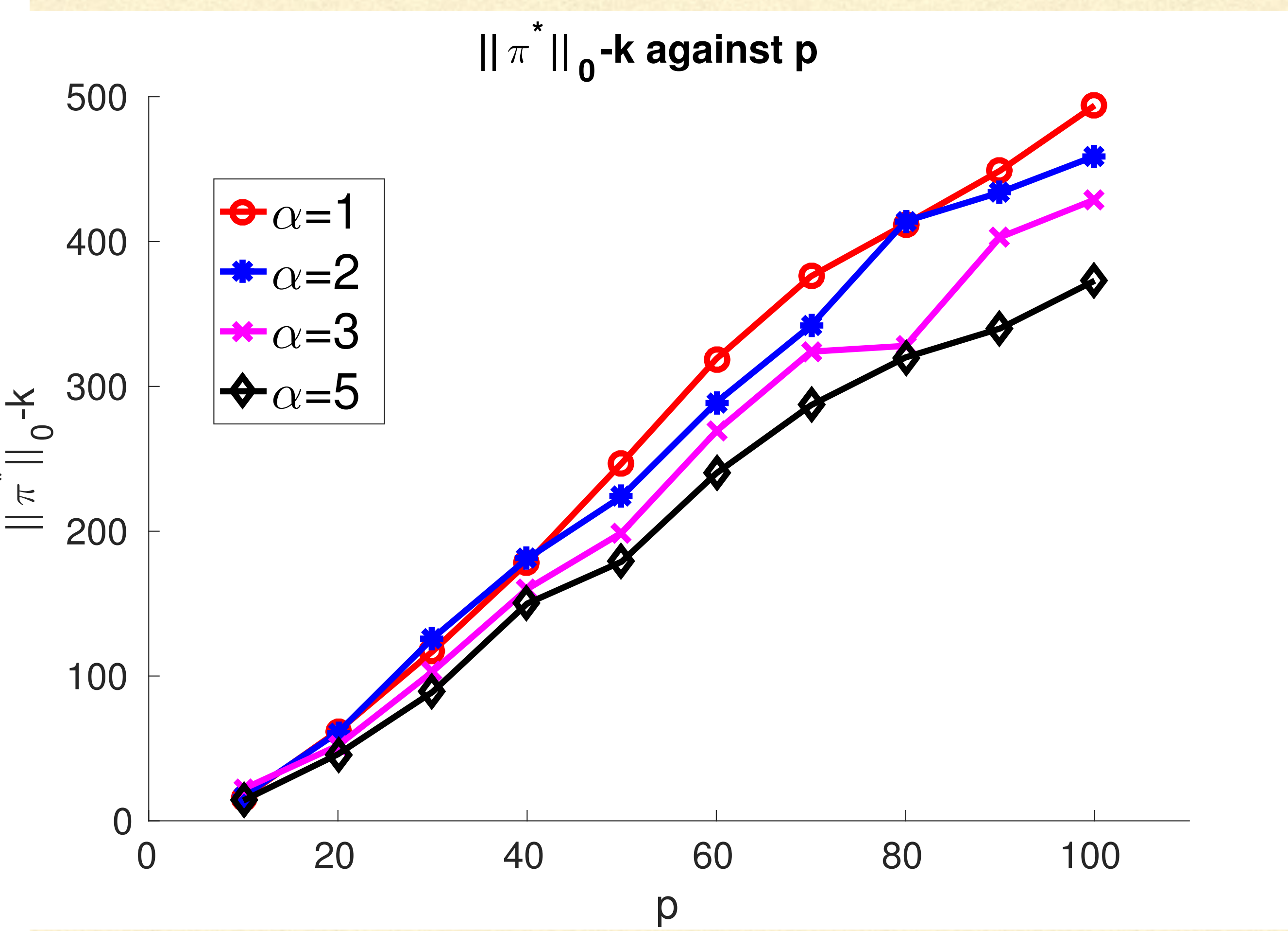
$$\mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right] \leq \left(1 + O(p^2/k)\right) \cdot f_{\mathrm{opt}}(k)$$

- Corollary: if $k = \Omega(\epsilon^{-1} p^2)$ then we achieve $(1+\epsilon)$ approximation of $f_{\mathrm{opt}}(k)$

GREEDY BASED SUBSET SELECTION

- $\|\pi^*\|_0 \leq k + p^2$
- Is this the best we can do?





GREEDY BASED SUBSET SELECTION

Conjecture. $\|\pi^*\|_0 \leq k + O(p)$

- Amazing consequences (near-optimal tractable A-optimality)

Conjecture. There exists a polynomial-time algorithm such that, under regularity conditions, produces $|S| \leq k$ such that if $k \geq \Omega(p/\epsilon)$ then

$$\mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right] \leq (1+\epsilon) \cdot f_{\mathrm{opt}}(k)$$

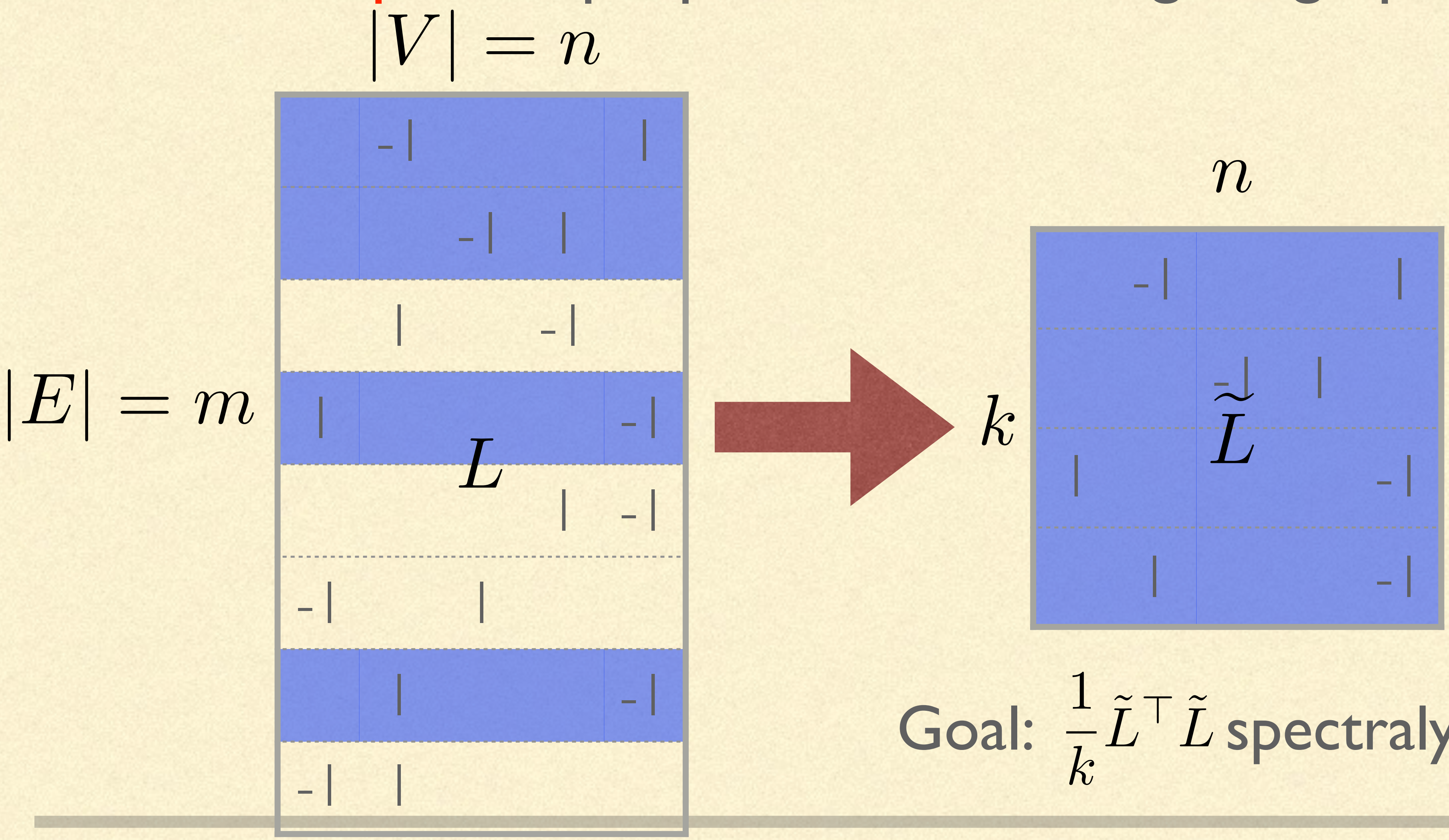
SUMMARY

Algorithm	Model	Bound type	Approx. factor	Condition on k
Leverage score sampling [25]	with replacement	additive	-[3]	asymptotic
Greedy removal [2]	without replacement	additive	$O(n/k)$	$k = \Omega(p)$
Convex A-opt. + sampling	with replacement	relative	$1+\epsilon$	$k = \Omega(\epsilon^{-2} B^2 \\|\boldsymbol{\Sigma}_*^{-1}\\|_2)$
Convex A-opt. + sampling	without replacement	multiplicative	$O(\log k)$	$k = \Omega(B^2 \\|\boldsymbol{\Sigma}_*^{-1}\\|_2)$
Convex A-opt. + greedy	without replacement	relative	$1+\epsilon$	**Rigorous:** $k = \Omega(\epsilon^{-1} p^2)$ **Conjecture:** $k = \Omega(\epsilon^{-1} p)$

- Some open questions:
  - High-dimensional subset selection?
  - Active (feedback-driven) learning

CONNECTIONS TO GRAPH SPARSIFICATION

- Graph sparsification: find a (small) subset of edges in a graph such that the spectral properties of the original graph are preserved.



Goal: $\frac{1}{k}\tilde{L}^\top \tilde{L}$ spectraly approximates $\frac{1}{m}L^\top L$

CONNECTIONS TO GRAPH SPARSIFICATION

- Weighted graph sparsification: new weights allowed to assign to the selected set of edges
  - Spielman and Srivastava'08: *Graph Sparsification by Effective Resistance*
- Unweighted graph sparsification: must keep weights unchanged during sparsification
  - Marcus, Spielman and Srivastava'13: *Interlacing Families and Bipartite Ramanujan Graphs of All Degrees* (Kadison-Singer problem)
  - Anderson, Gu and Melgaard'14: *Efficient Algorithm for Unweighted Spectral Graph Sparsification*

CONNECTIONS TO GRAPH SPARSIFICATION

- One important difference …
    - We don't want to approximate the original design, but rather an optimally-reweighted design.
    - Question: is there an unweighted sparsifier of a weighted graph?