Towards a Zero-One Law for Column Subset Selection

Zhao Song*, David P. Woodruff†, Peilin Zhong‡

*University of Washington, †Carnegie Mellon University, ‡Columbia University

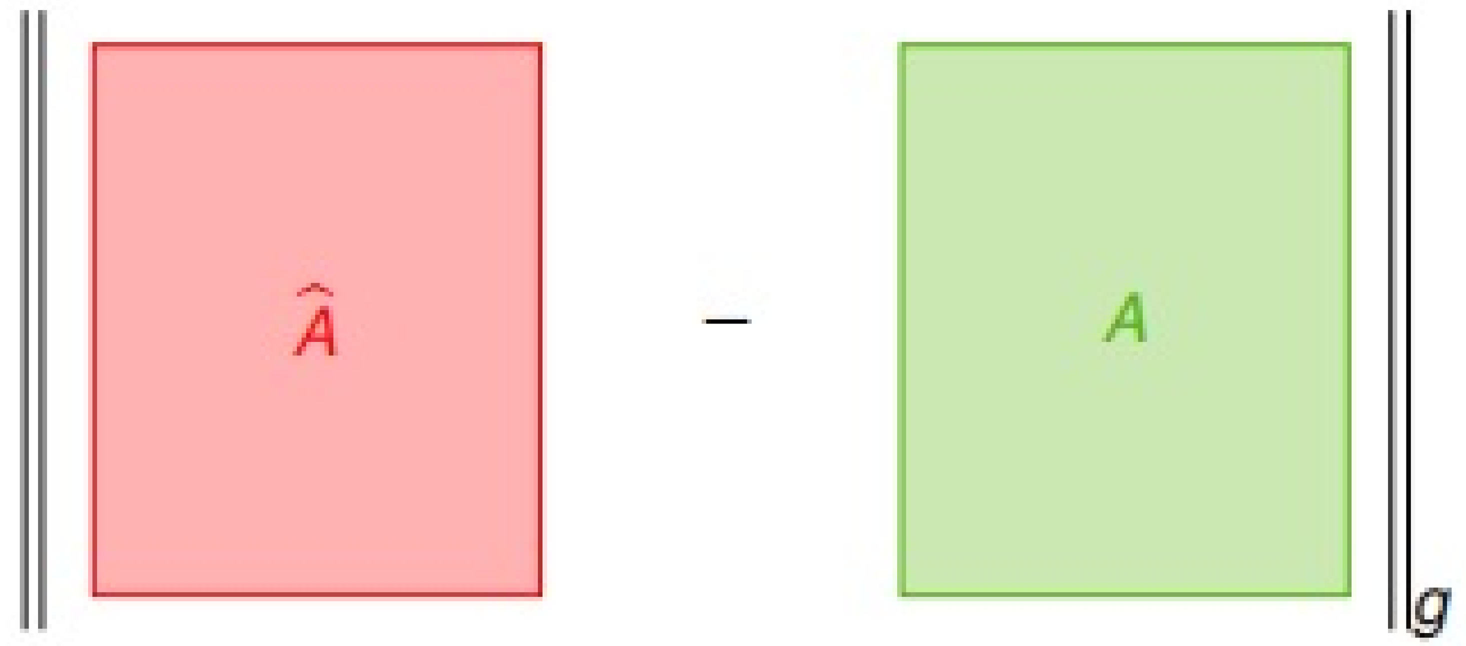
Problem Formulation

Given : $A \in \mathbb{R}^{n\times n}$, $k \ge 1$, $\alpha \ge 1$
Output : a rank-k matrix $\hat{A} \in \mathbb{R}^{n\times n}$ such that

$$\|\hat{A} - A\|_g \le \alpha \cdot \min_{\text{rank}-k\ A'} \|A' - A\|_g,$$

where

$$\|A\|_g = \sum_{i,j} g(A_{i,j}).$$



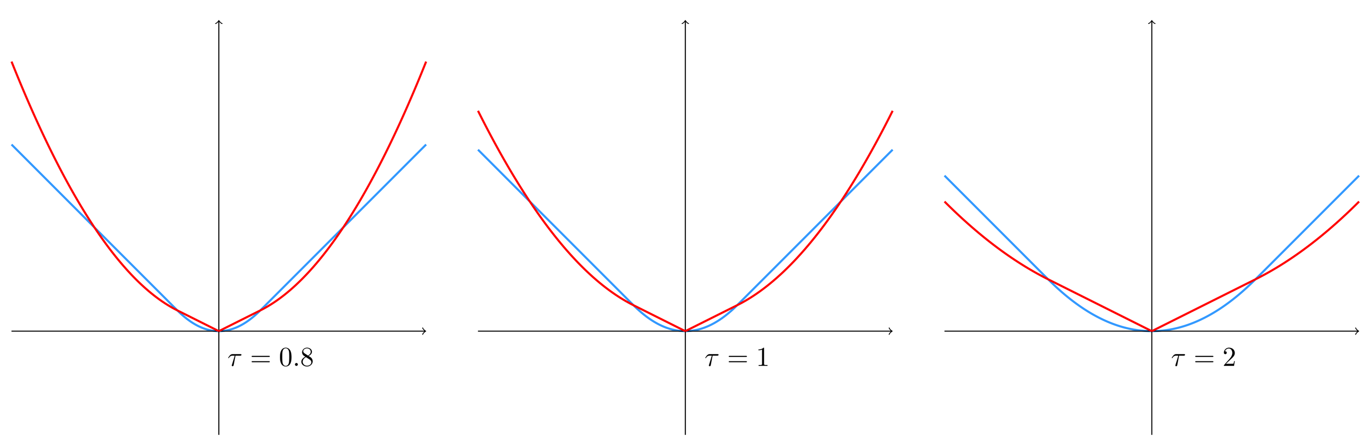
Question: For what kind of g, there could be a fast provable algorithm?

Generalized Low Rank Models

- $g(x) = |x|$ is the maximum likelihood estimator (MLE) for i.i.d. Laplacian noise.
- $g(x) = x^2$ is the MLE for i.i.d. Gaussian noise.
- Huber loss is the MLE for i.i.d. random variables with the Huber density.
- In general, the loss is $g(x)$ when the noise has density $c \cdot e^{-g(t)}$.

Example of functions, Huber and Reverse Huber

- Huber : L2 (when $|x|$ is small) + L1 (when $|x|$ is large)
- Reverse Huber : L1 (when $|x|$ is large) + L2 (when $|x|$ is small)



$$H(x) = \begin{cases} x^2/2\tau & \text{if } |x| \le \tau; \\ |x| - \tau/2 & \text{otherwise.} \end{cases} \qquad \widetilde{H}(x) = \begin{cases} |x|/2 & \text{if } |x| \le \tau; \\ x^2/4\tau + \tau/4 & \text{otherwise.} \end{cases}$$

Huber   Reverse Huber

Column Subset Selection

Given : $A \in \mathbb{R}^{n\times n}$, $k \ge 1$, $\alpha \ge 1$
Output : a subset of k' columns C of A such that

$$\min_X \|CX - A\|_g \le \alpha \cdot \min_{\text{rank}-k\ A'} \|A' - A\|_g.$$

Want k' as small as possible. Column subset selection gives a bicriteria solution for low rank approximation.

Towards a Zero-One Law

- Property 1: approximate triangle inequality
- Property 2: monotonicity
- Property 3: a fast regression algorithm

Approximate Triangle Inequality

For an integer k, we say a function $g(x) : \mathbb{R} \to \mathbb{R}_{\ge 0}$ satisfies the ati$_g$-approximate triangle inequality if for any $x_1, x_2, \cdots, x_k \in \mathbb{R}$ we have

$$g\left(\sum_{i=1}^k x_i\right) \le \mathrm{ati}_{g,k} \cdot \sum_{i=1}^k g(x_i)$$

We allow ati$_{g,k}$ to depend on k, n.

Monotone Property

For any parameter mon$_g \ge 1$, we say function $g(x) : \mathbb{R} \to \mathbb{R}_{\ge 0}$ is mon$_g$-monotone if for any $x, y \in \mathbb{R}$ with $0 \le |x| \le |y|$, we have

$$g(x) \le \mathrm{mon}_g \cdot g(y)$$

Regression Property

We say function $g(x) : \mathbb{R} \to \mathbb{R}_{\ge 0}$ has the $(\mathrm{reg}_{g,d}, T_{g,n,d})$-regression property if given $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times m}$ (where $m \le n$), for each $i \in [m]$, let OPT_i denote $\min_{x\in\mathbb{R}^d} \|Ax - B_i\|_g$. There is an algorithm that runs in $T_{g,n,d}$ time and outputs a matrix $X' \in \mathbb{R}^{d\times m}$ with

$$\|AX'_i - B_i\|_g \le \mathrm{reg}_{g,d} \cdot \mathrm{OPT}_i, \quad \forall i \in [m]$$

with high probability.

Necessity of the First Two Properties

- Approximate triangle inequality:
  - The "jumping function": $g_\tau(x) = |x|$ if $|x| \ge \tau$, and $g_\tau(x) = 0$ otherwise.
  - It has monotone property but not approximate triangle inequality.
  - For the identity matrix I and any $k = \Omega(\log n)$, the Johnson-Lindenstrauss lemma implies one can find a rank-k matrix B for which $\|I - B\|_\infty < 1/2$.
  - If we set $\tau = 1/2$, then $\|I - B\|_{g_\tau} = 0$.
  - But for any subset I_S of columns of the identity matrix we choose, necessarily $\|I - I_S X\|_\infty \ge 1$.
- Monotone property:
  - ReLU function satisfies approximate triangle inequality but not monotone property.
  - The optimal rank-k approximation for any matrix A is 0.
  - Notice though, that there are no good column subset selection algorithms for some matrices A, such as the $n \times n$ identity matrix.

Algorithmic Result

Given $A \in \mathbb{R}^{n\times n}$, let $k \ge 1$, and let $g : \mathbb{R} \to \mathbb{R}_{\ge 0}$ be a function satisfying the ati$_{g,k}$-approximate triangle inequality, the mon$_g$-monotone property, and the $(\mathrm{reg}_{g,d}, T_{g,n,d})$-regression property. Let $\mathrm{OPT} = \min_{\text{rank}-k\ A'} \|A' - A\|_g$.

There is an $O(T_{g,n,k})$ time algorithm for outputting a set $S \subseteq [n]$ with $|S| = O(k \log n)$ for which

$$\min_{X\in\mathbb{R}^{|S|\times n}} \|A_S X - A\|_g \le \mathrm{ati}_{g,k} \cdot \mathrm{mon}_g \cdot \mathrm{reg}_{g,k} \cdot O(k \log k) \cdot \mathrm{OPT},$$

with high probability.

Hardness for Column Subset Selection

Let $H(x)$ denote the Huber function with $\tau = 1$, i.e.,

$$H(x) = \begin{cases} x^2/\tau, & \text{if } |x| < \tau; \\ |x|, & \text{if } |x| \ge \tau. \end{cases}$$

For $k = 1$, there is a matrix $A \in \mathbb{R}^{n\times n}$ such that, if we select $o(\sqrt{\log n})$ columns to fit the entire matrix, there is no $O(1)$-approximation, i.e., for any subset $S \subseteq [n]$ with $|S| = o(\sqrt{\log n})$,

$$\min_{X\in\mathbb{R}^{|S|\times n}} \|A_S X - A\|_F \ge \omega(1) \cdot \min_{\text{rank}-1\ A'} \|A' - A\|_H.$$

Algorithm

Algorithm 1 Column Subset Selection

$r \leftarrow O(\log n)$
$T_0 \leftarrow [n]$
for $i = 1 \to r$ **do**
  $m \leftarrow |T_{i-1}|$
  for $j = 1 \to \log n$ **do**
    Sample $S^{(j)}$ from $\binom{T_{i-1}}{2k}$ uniformly at random
    $m \leftarrow |T_{i-1} \backslash S^{(j)}|$, $d \leftarrow 2k$
    $\{\mathrm{cost}_t\}_{t\in T_{i-1}\backslash S^{(j)}} \leftarrow \textsc{MultipleRegression}(g, n, d, m, A_{S^{(j)}}, A_{T_{i-1}\backslash S^{(j)}})$
    $R^{(j)} \leftarrow \textsc{BottomK}(\textsc{Sort}(\mathrm{cost}), m/20)$
    $v_j \leftarrow \sum_{t\in R^{(j)}} \mathrm{cost}_t$
  end for
  $j^* \leftarrow \min_{j\in[\log n]}\{v_j\}$, $T_i \leftarrow T_{i-1}\backslash(S^{(j^*)} \cup R^{(j^*)})$, $S_i \leftarrow S^{(j^*)}$
end for
$S \leftarrow \cup_i S_i$
Return S

Analysis

- Decompose $A = A^* + \Delta$, where A^* is the optimal rank-k matrix and Δ is the optimal residual matrix, so $\|\Delta\|_g = \mathrm{OPT}$
- Suppose we sample $2k+1$ column indices $i_1, i_2, \cdots, i_{2k+1}$ uniformly at random and consider the submatrix V_S^* which only contains these columns from V^*, where $A^* = U^* \cdot V^*$
  - The k-by-k submatrix of V_S^* with max determinant does not intersect column $V^*_{i_{2k+1}}$ with probability at least $1/2$
  - By Cramer's rule, $V^*_{i_{2k+1}} = \sum_{l=1}^{2k} \alpha_l \cdot V^*_{i_l}$ and $|\alpha_l| \le 1, \forall l \in [2k]$.
  - Can argue $A^*_{i_{2k+1}} = \sum_{l=1}^{2k} \alpha_l \cdot A^*_{i_l}$ and $|\alpha_l| \le 1, \forall l \in [2k]$.
- If we sample $2k$ columns from A^* uniformly at random, a constant fraction of columns of A^* can be interpolated with fitting coefficients of absolute value at most 1
- Suppose the sampled columns have indices $i_1, \cdots, i_{2k}$ and $A_j^* = \sum_{l=1}^{2k} \alpha_l A^*_{i_l}$ where $|\alpha_l| \le 1, \forall l \in [2k]$
- The optimal cost of using $A_{i_1}, \cdots, A_{i_{2k}}$ to fit A_j is at most

$$\begin{aligned} \|A_j - \sum_{l\in[2k]} \alpha_l \cdot A_{i_l}\|_g &= \|\Delta_j - \sum_{l\in[2k]} \alpha_l \cdot \Delta_{i_l}\|_g \\ &\le \mathrm{ati}_g \cdot (\|\Delta_j\|_g + \sum_{l\in[2k]} \|\alpha_l \cdot \Delta_{i_l}\|_g) \\ &\le \mathrm{ati}_g \cdot \mathrm{mon}_g \cdot (\|\Delta_j\|_g + \sum_{l\in[2k]} \|\Delta_{i_l}\|_g) \end{aligned}$$

- Notice that $\mathbf{E}[\sum_{l=1}^{2k} \|\Delta_{i_l}\|_g] = (2k/n) \cdot \mathrm{OPT}$
- Since each time we can fit a constant fraction of the remaining columns with small cost, we can recurse $\log n$ times to fit all columns
- To improve the analysis of the approximation ratio from $k \log n$ to $k \log k$, we condition on each time not sampling the n/k columns with the largest cost