Column Subset Selection

Joel A. Tropp

Applied & Computational Mathematics
California Institute of Technology
jtropp@acm.caltech.edu

Thanks to B. Recht (Caltech, IST)
Column Subset Selection

\[ A = \begin{bmatrix} \text{\cellcolor{orange}} & \text{\cellcolor{green}} & \text{\cellcolor{blue}} & \text{\cellcolor{brown}} & \text{\cellcolor{black}} & \text{\cellcolor{red}} \end{bmatrix} \]

\[ \tau = \{ \text{\cellcolor{orange}}, \text{\cellcolor{green}}, \text{\cellcolor{blue}}, \text{\cellcolor{brown}}, \text{\cellcolor{black}}, \text{\cellcolor{red}} \} \]

\[ A_\tau = \begin{bmatrix} \text{\cellcolor{orange}} & \text{\cellcolor{green}} & \text{\cellcolor{blue}} & \text{\cellcolor{brown}} & \text{\cellcolor{red}} \end{bmatrix} \]
**Spectral Norm Reduction**

**Theorem 1. [Kashin–Tzafriri]**  Suppose the $n$ columns of $A$ have unit $\ell_2$ norm. There is a set $\tau$ of column indices for which

$$|\tau| \geq \frac{n}{\|A\|^2} \quad \text{and} \quad \|A_\tau\| \leq C.$$

**Examples:**

- $A$ has identical columns. Then $|\tau| \geq 1$.

- $A$ has orthonormal columns. Then $|\tau| \geq n$. 

*Column Subset Selection, MMDS, Stanford, June 2008*
Spectral Norm Reduction

**Theorem 1. [Kashin–Tzafriri]** Suppose the \( n \) columns of \( A \) have unit \( \ell_2 \) norm. There is a set \( \tau \) of column indices for which

\[
|\tau| \geq \frac{n}{\|A\|^2} \quad \text{and} \quad \|A_{\tau}\| \leq C.
\]

**Theorem 2. [T 2007]** There is a randomized, polynomial-time algorithm that produces the set \( \tau \).

**Overview:**
- Randomly select columns
- Remove redundant columns
Random Column Selection: Intuitions

- Random column selection reduces norms
- A random submatrix gets “its share” of the total norm
- Submatrices with small norm are ubiquitous
- Random selection is a form of regularization
- Added benefit: Dimension reduction
Example: What Can Go Wrong

\[ A = \begin{bmatrix} \begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array} \end{bmatrix} \]

\[ A_\mathbf{T} = \begin{bmatrix} \begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array} \end{bmatrix} \]

\[ \| A \| = \| A_\mathbf{T} \| = \sqrt{2} \quad \Longrightarrow \quad No \ \text{reduction!} \]
The $(\infty, 2)$ Operator Norm

**Definition 3.** The $(\infty, 2)$ operator norm of a matrix $B$ is

$$\|B\|_{\infty, 2} = \max \{\|Bx\|_2 : \|x\|_{\infty} = 1\}.$$ 

**Proposition 4.** If $B$ has $s$ columns, then the best general bound is

$$\|B\|_{\infty, 2} \leq \sqrt{s} \|B\|.$$
Lemma 5. Suppose the $n$ columns of $A$ have unit $\ell_2$ norm. Draw a uniformly random subset $\sigma$ of columns whose cardinality

$$|\sigma| = \frac{2n}{\|A\|^2}.$$ 

Then

$$\mathbb{E} \|A_\sigma\|_{\infty,2} \leq C\sqrt{|\sigma|}.$$ 

Problem: How can we use this information?
Theorem 6. [Pietsch, Grothendieck] Every matrix $B$ can be factorized as $B = TD$ where

- $D$ is diagonal and nonnegative with $\text{trace}(D^2) = 1$, and
- $\|B\|_{\infty,2} \leq \|T\| \leq \sqrt{\pi/2} \|B\|_{\infty,2}$
Lemma 7. Suppose $B$ has $s$ columns. There is a set $\tau$ of column indices for which

$$|\tau| \geq \frac{s}{2} \quad \text{and} \quad \|B_\tau\| \leq \sqrt{\pi} \cdot \frac{1}{\sqrt{s}} \|B\|_{\infty,2}.$$ 

Proof. Consider a Pietsch factorization $B = TD$. Select

$$\tau = \{j : d_{jj}^2 \leq 2/s\}.$$

Since $\sum d_{jj}^2 = 1$, Markov’s inequality implies $|\tau| \geq s/2$. Calculate

$$\|B_\tau\| = \|TD_\tau\| \leq \|T\| \cdot \|D_\tau\| \leq \sqrt{\pi/2} \|B\|_{\infty,2} \cdot \sqrt{2/s}.$$
Proof of Kashin–Tzafriri

- Suppose the $n$ columns of $A$ have unit $\ell_2$ norm
- Lemma 5 provides (random) $\sigma$ for which
  \[
  |\sigma| = \frac{2n}{\|A\|^2} \quad \text{and} \quad \|A_\sigma\|_{\infty,2} \leq C\sqrt{|\sigma|}
  \]
- Lemma 7 applied to $B = A_\sigma$ yields a subset $\tau \subset \sigma$ for which
  \[
  |\tau| \geq \frac{|\sigma|}{2} \quad \text{and} \quad \|B_\tau\| \leq \sqrt{\pi} \cdot \frac{1}{\sqrt{|\sigma|}} \cdot \|B\|_{\infty,2}
  \]
- Simplify
  \[
  |\tau| \geq \frac{n}{\|A\|^2} \quad \text{and} \quad \|A_\tau\| \leq C\sqrt{\pi}
  \]
- Note: This is almost an algorithm
Consider a matrix $B$ with Pietsch factorization $B = TD$

Suppose $\|T\| \leq \alpha$

Calculate

$$B = TD \implies \|Bx\|_2^2 = \|TDx\|_2^2 \quad \forall x$$

$$\implies \|Bx\|_2^2 \leq \alpha^2 \|Dx\|_2^2 \quad \forall x$$

$$\implies x^*(B^*B)x \leq \alpha^2 \cdot x^*D^2x \quad \forall x$$

$$\implies x^* [B^*B - \alpha^2 D^2] x \leq 0 \quad \forall x$$

$$\implies \lambda_{\text{max}}(B^*B - \alpha^2 D^2) \leq 0$$
Pietsch is Convex

Key new idea: Can find Pietsch factorizations by convex programming

\[
\min \lambda_{\text{max}}(B^*B - \alpha^2F) \\
\text{subject to } F \text{ diagonal, } F \geq 0, \text{ trace}(F) = 1
\]

If value at $F_\star$ is nonpositive, then we have a factorization

\[B = (BF_\star^{-1/2}) \cdot F_\star^{1/2} \] with \[\|BF_\star^{-1/2}\| \leq \alpha\]

Proof of Kashin–Tzafriri offers target value for $\alpha$

Can also perform binary search to approximate minimal value of $\alpha$
An Optimization over the Simplex

Express $F = \text{diag}(f)$

Constraints delineate the probability simplex:

$$\Delta = \{ f : \text{trace}(f) = 1 \text{ and } f \geq 0 \}$$

Objective function and its subdifferential:

$$J(f) = \lambda_{\text{max}}(B^*B - \alpha^2 \text{diag}(f))$$

$$\partial J(f) = \text{conv} \left\{ -\alpha^2 |u|^2 : u \text{ top evec. } B^*B - \alpha^2 \text{diag}(f), \|u\|_2 = 1 \right\}$$

Obtain

$$\min J(f) \text{ subject to } f \in \Delta$$
Entropic Mirror Descent

1. Intialize $f^{(1)} \leftarrow s^{-1}e$ and $k \leftarrow 1$

2. Compute a subgradient: $\theta \in \partial J(f^{(k)})$

3. Determine step size:

$$\beta_k \leftarrow \sqrt{\frac{2 \log s}{k \|\theta\|_\infty^2}}$$

4. Update variable:

$$f^{(k+1)} \leftarrow \frac{f^{(k)} \circ \exp\{-\beta_k \theta\}}{\text{trace}(f^{(k)} \circ \exp\{-\beta_k \theta\})}$$

5. Increment $k \leftarrow k + 1$, and return to 2.

Other Formulations

- Modified primal to simultaneously identify $\alpha$

$$\min \lambda_{\max}(B^* B - \alpha^2 F) + \alpha^2$$
subject to $\quad F$ diagonal, $\quad F \succeq 0$, $\quad \text{trace}(F) = 1$, $\quad \alpha \geq 0$

- Dual problem is the famous MAXCUT SDP:

$$\max \langle B^* B, Z \rangle \quad \text{subject to} \quad \text{diag}(Z) = e, \quad Z \succeq 0$$
Related Results

**Theorem 8. [Bourgain–Tzafriri 1991]** Suppose the $n$ columns of $A$ have unit $\ell_2$ norm. There is a set $\tau$ of column indices for which

$$|\tau| \geq \frac{cn}{\|A\|^2} \quad \text{and} \quad \kappa(A_{\tau}) \leq \sqrt{3}.$$ 

Examples:

- $A$ has identical columns. Then $|\tau| \geq 1$.

- $A$ has orthonormal columns. Then $|\tau| \geq cn$. 
Related Results

Theorem 8. [Bourgain–Tzafriri 1991]  Suppose the $n$ columns of $A$ have unit $\ell_2$ norm. There is a set $\tau$ of column indices for which

$$|\tau| \geq \frac{cn}{\|A\|^2} \quad \text{and} \quad \kappa(A_\tau) \leq \sqrt{3}.$$

Theorem 9. [T 2007]  There is a randomized, polynomial-time algorithm that produces the set $\tau$. 
To learn more...

E-mail:

jtropp@acm.caltech.edu

Web: http://www.acm.caltech.edu/~jtropp

Papers in Preparation:

- T, “Column subset selection, matrix factorization, and eigenvalue optimization”
- T, “Paved with good intentions: Computational applications of matrix column partitions”
- . . .