Column Subset Selection

Joel A. Tropp

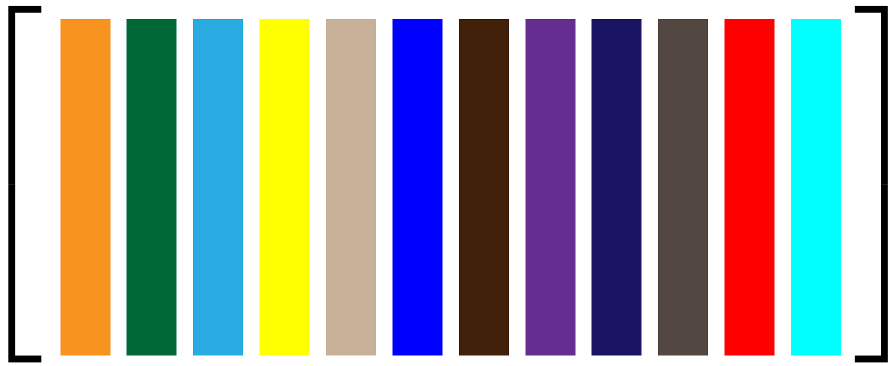
Applied & Computational Mathematics
California Institute of Technology
jtropp@acm.caltech.edu

Thanks to B. Recht (Caltech, IST)

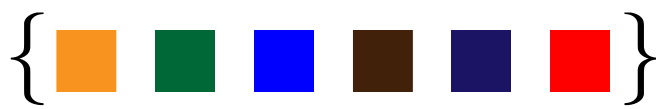
Research supported in part by NSF, DARPA, and ONR

Column Subset Selection

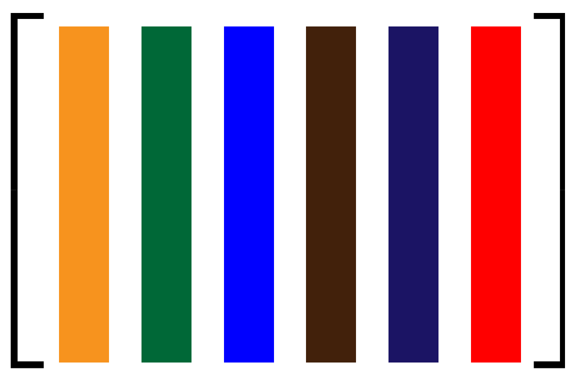
$$A =$$



$$\tau =$$



$$A_\tau =$$



Spectral Norm Reduction

Theorem 1. [Kashin–Tzafriri] *Suppose the n columns of $\boldsymbol{A}$ have unit ℓ_2 norm. There is a set τ of column indices for which*

$$|\tau| \geq \frac{n}{\|\boldsymbol{A}\|^2} \qquad \textit{and} \qquad \|\boldsymbol{A}_\tau\| \leq \mathrm{C}.$$

Examples:

- $\boldsymbol{A}$ has identical columns. Then $|\tau| \geq 1$.
- $\boldsymbol{A}$ has orthonormal columns. Then $|\tau| \geq n$.

Spectral Norm Reduction

Theorem 1. [Kashin–Tzafriri] *Suppose the n columns of $\boldsymbol{A}$ have unit ℓ_2 norm. There is a set τ of column indices for which*

$$|\tau| \geq \frac{n}{\|\boldsymbol{A}\|^2} \qquad \textit{and} \qquad \|\boldsymbol{A}_\tau\| \leq \mathrm{C}.$$

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

Overview:

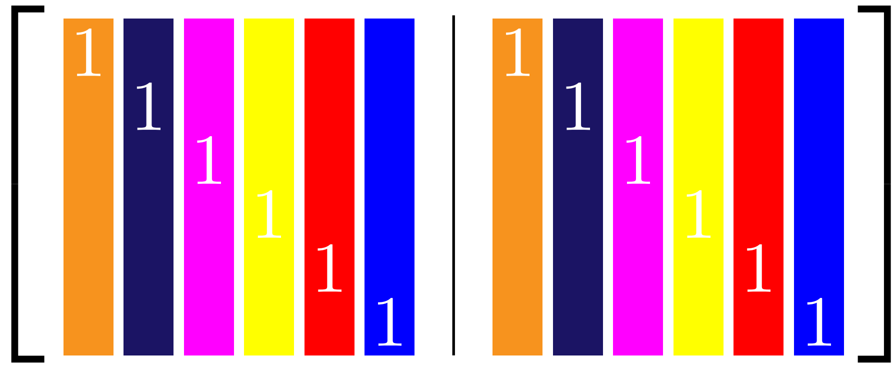
- Randomly select columns
- Remove redundant columns

Random Column Selection: Intuitions

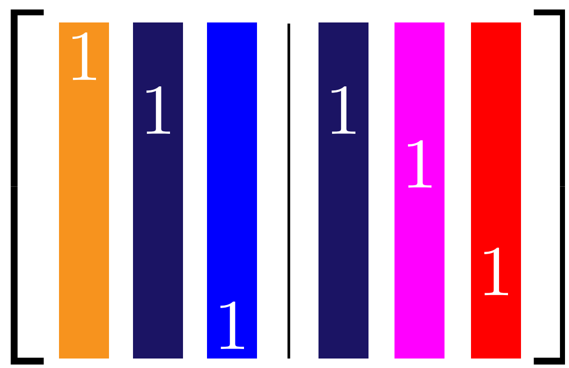
- Random column selection reduces norms
- A random submatrix gets "its share" of the total norm
- Submatrices with small norm are ubiquitous
- Random selection is a form of regularization
- Added benefit: Dimension reduction

Example: What Can Go Wrong

$A =$



$A_\tau =$

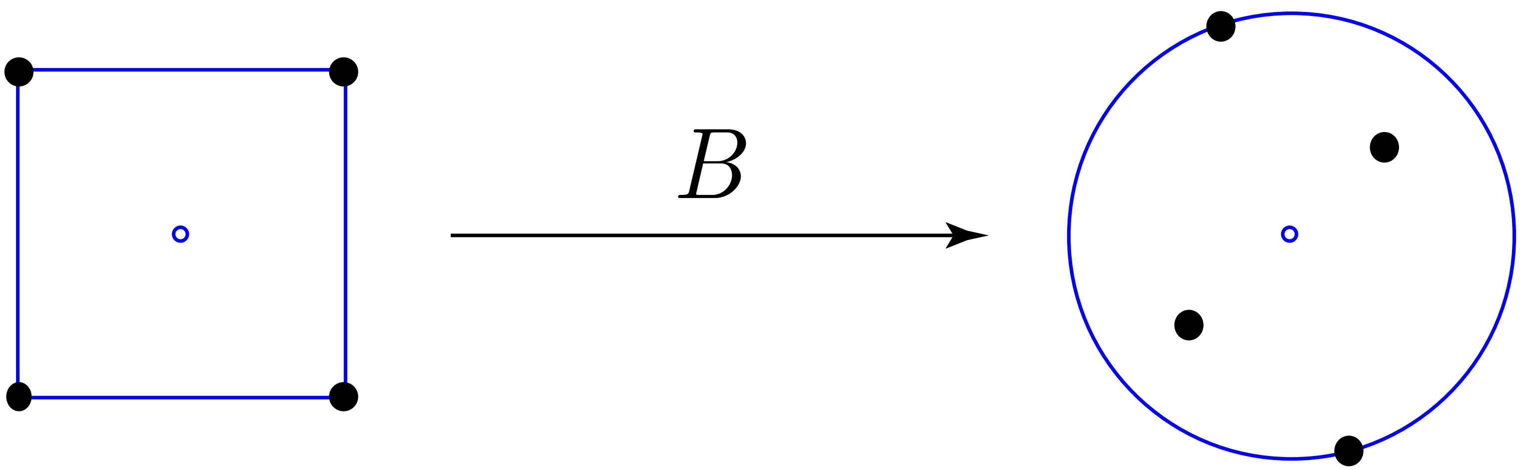


$\|\boldsymbol{A}\| = \|\boldsymbol{A}_\tau\| = \sqrt{2} \qquad \Longrightarrow \qquad$ *No reduction!*

The $(\infty, 2)$ Operator Norm

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

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



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

$$\|\boldsymbol{B}\|_{\infty,2} \le \sqrt{s}\,\|\boldsymbol{B}\|.$$

Random Reduction of $(\infty, 2)$ Norm

Lemma 5. *Suppose the n columns of $\boldsymbol{A}$ have unit ℓ_2 norm. Draw a uniformly random subset σ of columns whose cardinality*

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

Then

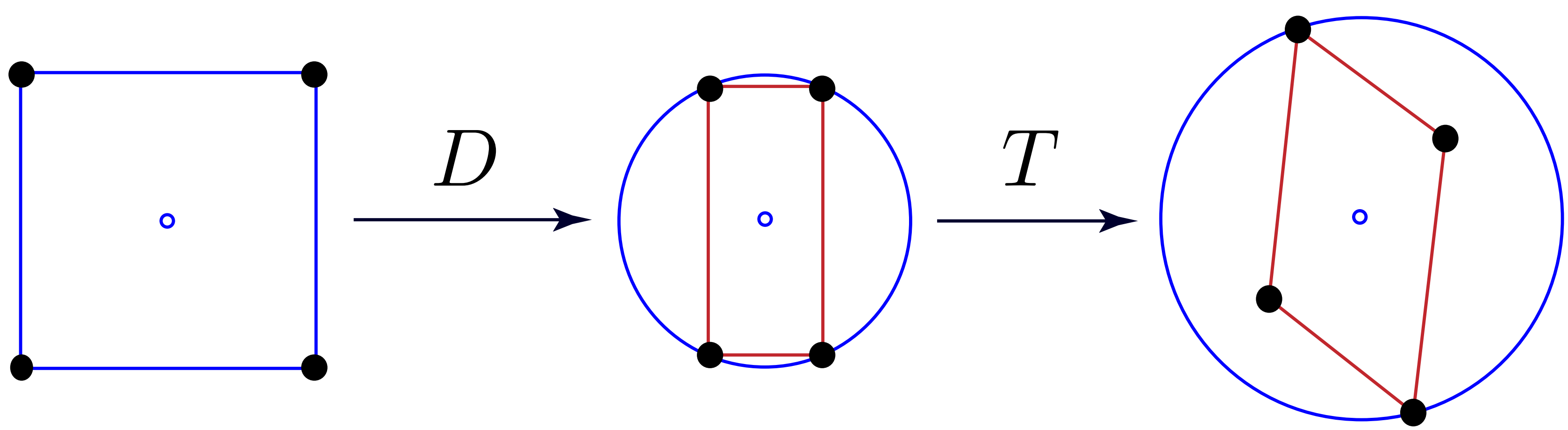
$$\mathbb{E}\,\|\boldsymbol{A}_\sigma\|_{\infty,2} \le \mathrm{C}\sqrt{|\sigma|}.$$

❧ *Problem:* How can we use this information?

Pietsch Factorization

Theorem 6. [Pietsch, Grothendieck] *Every matrix $\boldsymbol{B}$ can be factorized as $\boldsymbol{B} = \boldsymbol{T}\boldsymbol{D}$ where*

- $\boldsymbol{D}$ *is diagonal and nonnegative with* $\operatorname{trace}(\boldsymbol{D}^2) = 1$, *and*
- $\|\boldsymbol{B}\|_{\infty,2} \le \|\boldsymbol{T}\| \le \sqrt{\pi/2}\,\|\boldsymbol{B}\|_{\infty,2}$



Pietsch and Norm Reduction

Lemma 7. *Suppose $\boldsymbol{B}$ has s columns. There is a set τ of column indices for which*

$$|\tau| \geq \frac{s}{2} \qquad \textit{and} \qquad \|\boldsymbol{B}_\tau\| \leq \sqrt{\pi} \cdot \frac{1}{\sqrt{s}} \, \|\boldsymbol{B}\|_{\infty,2} \, .$$

Proof. Consider a Pietsch factorization $\boldsymbol{B} = \boldsymbol{T}\boldsymbol{D}$. Select

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

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

$$\|\boldsymbol{B}_\tau\| = \|\boldsymbol{T}\boldsymbol{D}_\tau\| \leq \|\boldsymbol{T}\| \cdot \|\boldsymbol{D}_\tau\| \leq \sqrt{\pi/2} \, \|\boldsymbol{B}\|_{\infty,2} \cdot \sqrt{2/s}.$$

Proof of Kashin–Tzafriri

- Suppose the n columns of $\boldsymbol{A}$ have unit ℓ_2 norm
- Lemma 5 provides (random) σ for which

$$|\sigma| = \frac{2n}{\|\boldsymbol{A}\|^2} \qquad \text{and} \qquad \|\boldsymbol{A}_\sigma\|_{\infty,2} \leq \mathrm{C}\sqrt{|\sigma|}$$

- Lemma 7 applied to $\boldsymbol{B} = \boldsymbol{A}_\sigma$ yields a subset $\tau \subset \sigma$ for which

$$|\tau| \geq \frac{|\sigma|}{2} \qquad \text{and} \qquad \|\boldsymbol{B}_\tau\| \leq \sqrt{\pi} \cdot \frac{1}{\sqrt{|\sigma|}} \cdot \|\boldsymbol{B}\|_{\infty,2}$$

- Simplify

$$|\tau| \geq \frac{n}{\|\boldsymbol{A}\|^2} \qquad \text{and} \qquad \|\boldsymbol{A}_\tau\| \leq \mathrm{C}\sqrt{\pi}$$

- *Note:* This is almost an algorithm

Pietsch and Eigenvalues

- Consider a matrix $\boldsymbol{B}$ with Pietsch factorization $\boldsymbol{B} = \boldsymbol{TD}$
- Suppose $\|\boldsymbol{T}\| \leq \alpha$
- Calculate

$$
\begin{aligned}
\boldsymbol{B} = \boldsymbol{TD} \qquad &\Longrightarrow \qquad \|\boldsymbol{Bx}\|_2^2 = \|\boldsymbol{TDx}\|_2^2 && \forall \boldsymbol{x} \\
&\Longrightarrow \qquad \|\boldsymbol{Bx}\|_2^2 \leq \alpha^2 \|\boldsymbol{Dx}\|_2^2 && \forall \boldsymbol{x} \\
&\Longrightarrow \qquad \boldsymbol{x}^*(\boldsymbol{B}^*\boldsymbol{B})\boldsymbol{x} \leq \alpha^2 \cdot \boldsymbol{x}^*\boldsymbol{D}^2\boldsymbol{x} && \forall \boldsymbol{x} \\
&\Longrightarrow \qquad \boldsymbol{x}^* \left[\boldsymbol{B}^*\boldsymbol{B} - \alpha^2\boldsymbol{D}^2\right] \boldsymbol{x} \leq 0 && \forall \boldsymbol{x} \\
&\Longrightarrow \qquad \lambda_{\max}(\boldsymbol{B}^*\boldsymbol{B} - \alpha^2\boldsymbol{D}^2) \leq 0
\end{aligned}
$$

Pietsch is Convex

- *Key new idea:* Can find Pietsch factorizations by convex programming

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

- If value at $\boldsymbol{F}_\star$ is nonpositive, then we have a factorization

$$\boldsymbol{B} = (\boldsymbol{B}\boldsymbol{F}_\star^{-1/2}) \cdot \boldsymbol{F}_\star^{1/2} \qquad \text{with} \qquad \left\|\boldsymbol{B}\boldsymbol{F}_\star^{-1/2}\right\| \leq \alpha$$

- Proof of Kashin–Tzafriri offers target value for α
- Can also perform binary search to approximate minimal value of α

An Optimization over the Simplex

- Express $\boldsymbol{F} = \operatorname{diag}(\boldsymbol{f})$
- Constraints delineate the probability simplex:

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

- Objective function and its subdifferential:

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

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

- Obtain

$$\min \ J(\boldsymbol{f}) \qquad \text{subject to} \qquad \boldsymbol{f} \in \Delta$$

Entropic Mirror Descent

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

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

3. Determine step size:

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

4. Update variable:

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

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

References: [Eggermont 1991, Beck–Teboulle 2003]

Other Formulations

- Modified primal to simultaneously identify α

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

- Dual problem is the famous MAXCUT SDP:

$$\max \ \langle \boldsymbol{B}^*\boldsymbol{B},\ \boldsymbol{Z} \rangle \qquad \text{subject to} \qquad \operatorname{diag}(\boldsymbol{Z}) = \mathbf{e}, \quad \boldsymbol{Z} \succcurlyeq \mathbf{0}$$

Related Results

Theorem 8. [Bourgain–Tzafriri 1991] *Suppose the n columns of $\boldsymbol{A}$ have unit ℓ_2 norm. There is a set τ of column indices for which*

$$|\tau| \geq \frac{cn}{\|\boldsymbol{A}\|^2} \qquad \textit{and} \qquad \kappa(\boldsymbol{A}_\tau) \leq \sqrt{3}.$$

Examples:

- $\boldsymbol{A}$ has identical columns. Then $|\tau| \geq 1$.
- $\boldsymbol{A}$ has orthonormal columns. Then $|\tau| \geq cn$.

Related Results

Theorem 8. [Bourgain–Tzafriri 1991] *Suppose the n columns of $\boldsymbol{A}$ have unit ℓ_2 norm. There is a set τ of column indices for which*

$$|\tau| \geq \frac{cn}{\|\boldsymbol{A}\|^2} \qquad \textit{and} \qquad \kappa(\boldsymbol{A}_\tau) \leq \sqrt{3}.$$

Theorem 9. [T 2007] *There is a randomized, polynomial-time algorithm that produces the set τ.*

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"
❧ . . .