Randomized low-rank approximations

in theory and practice

Alex Gittens

ILAS 2013
Randomized Matrix Algorithms Minisymposium
June 4, 2013

Motivation

Our basic task

$\mathbf{A} \in \mathbb{R}^{m \times n}$ is a *huge* matrix. Given $k \ll \min\{m, n\}$, we would like a low-rank approximation to $\mathbf{A}$ with rank about k

1. This abstract problem is ubiquitous in data processing tasks: machine learning, image processing, statistical analysis, optimization, ...
2. Traditional deterministic approaches (via truncated SVD, rank-revealing QR, Krylov space methods) cost at least $O(mnk \log \min\{m, n\})$ operations, and can have high communications costs.

The question arises: can we use randomness to assist in the design of algorithms for finding low-rank approximations of large matrices? We consider two schemes for low-rank approximation:

- Projection-based approximation schemes using fast randomized projections (joint work with C. Boutsidis)
- "Sketching" schemes for positive semidefinite matrices (joint work with M. Mahoney)

Our objective

Determine how the errors of these randomized approximations in the spectral, Frobenius, and trace norms compare with the errors of $\mathbf{A}_k$, the best rank-k approximation to $\mathbf{A}$.

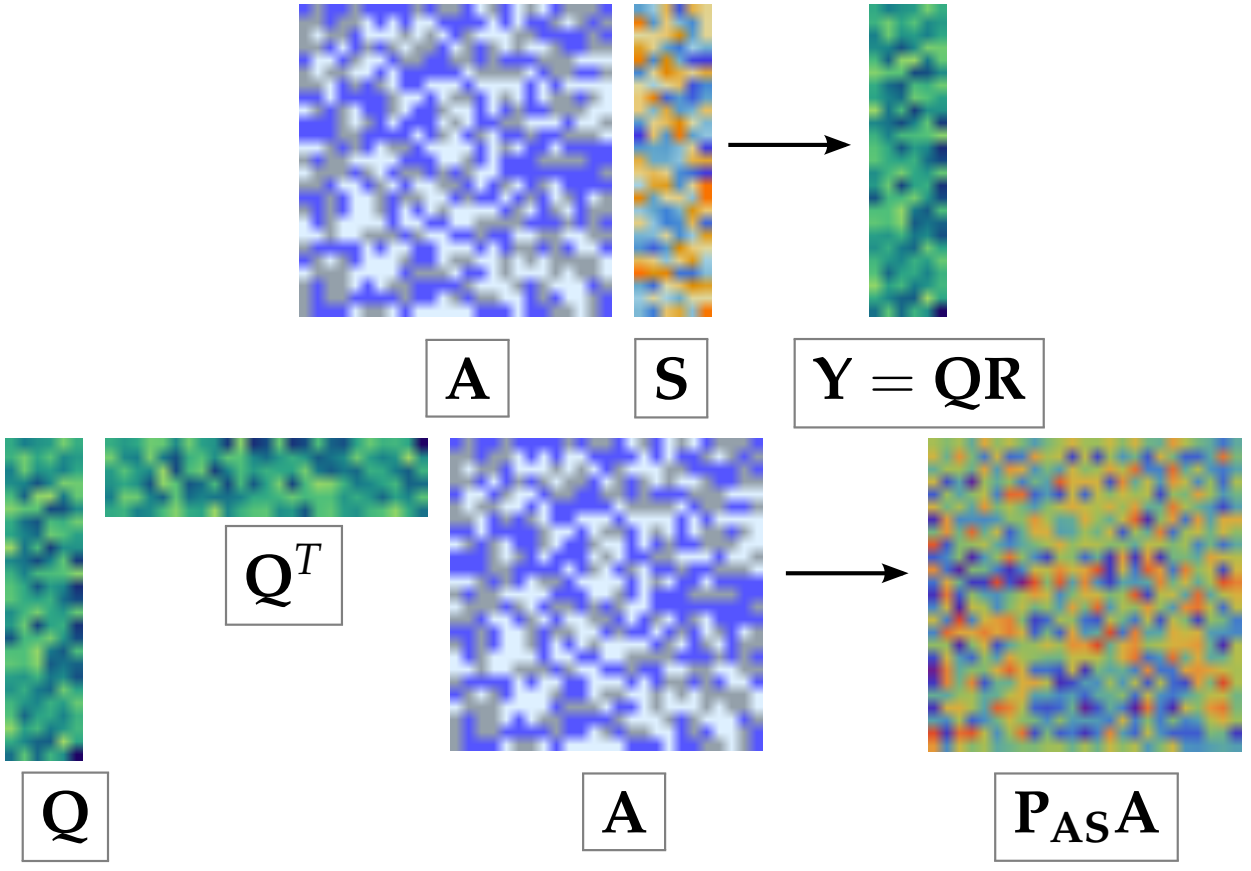
The target audience

Who is interested in these approximations and our guarantees?

- The numerical linear algebra community wants high quality approximations with very low failure rates and low communication cost.
- The machine learning community wants approximations whose errors are on par with modeling inaccuracies and the imprecision of the data
- The optimization community is interested in varying levels of quality.
- The theoretical computer science community is interested in understanding the behavior of these algorithms, e.g. what is the optimal tradeoff between the error, failure rate, and the amount of arithmetic operations involved? How can communication cost be minimized?

The Random Projection Methodology

Capture the range of the "important" part of **A** using a sampling matrix **S**, then project **A** onto this space to reduce rank.



The quality of the approximation depends on how well the range of the dominant k-dimensional left singular space of $\mathbf{A}$ is approximated by the range of $\mathbf{Y}$.

We can use the "power" method to increase the accuracy of the approximation: approximate $\mathbf{A}$ with $\mathbf{P}_{(\mathbf{A}\mathbf{A}^T)^p\mathbf{A}\mathbf{S}}\mathbf{A}$:

1. Form $\mathbf{Y} = (\mathbf{A}\mathbf{A}^T)^p\mathbf{A}\mathbf{S}$.
2. Take the QR decomposition $\mathbf{Y} = \mathbf{Q}\mathbf{R}$.
3. Form the low rank approximation $\mathbf{Q}(\mathbf{Q}^T\mathbf{A})$.

Requires only $2(p+1)$ passes over $\mathbf{A}$.

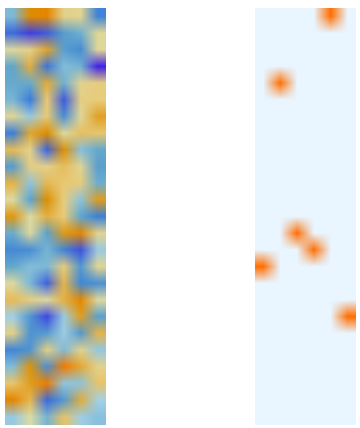
This methodology was popularized by (Papadimitriou et al. 2000), (Sarlós 2006), and (Martinsson et al. 2006).

Our design parameters:

- ℓ, the number of samples ($k \leq \ell \ll \min\{m, n\}$)
- $\mathbf{S} \in \mathbb{R}^{n\times\ell}$, the random sampling matrix.
- p, the number of iterations.

Three factors determine probability of getting a good approximation:

- Spectral decay of $\mathbf{A}$, e.g. the multiplicative gap $\sigma_{k+1}(\mathbf{A})/\sigma_k(\mathbf{A})$, or $(\sigma_{k+1}(\mathbf{A})/\sigma_k(\mathbf{A}))^p$.
- Type of randomness used to generate $\mathbf{S}$.



- Amount of oversampling (as $\ell \to n, \mathbf{P_{AS}A} \to \mathbf{A}$).

Choice of sampling matrix

Dominant arithmetic cost of forming these low-rank approximations is the matrix multiply $\mathbf{AS}$.

A natural choice for $\mathbf{S}$ is a matrix of i.i.d. $\mathcal{N}(0,1)$ Gaussians, proposed in (Martinsson et al. 2006).

- Computation of $\mathbf{AS}$ takes $O(mn\ell)$ time for general $\mathbf{A}$.
- The columns of $\mathbf{A}$ are well-mixed.

(Woolfe et al. 2008) proposed using *structured* random projections.

- Computation of $\mathbf{AS}$ takes reduced time $O(mn\log(\ell))$.
- Mixing not as uniform, so potential accuracy loss.

Assume n is a power of 2. We consider the case where $\mathbf{S}$ is a subsampled randomized Hadamard transform (SRHT):

$$\mathbf{S} = \sqrt{\frac{n}{\ell}}\mathbf{DHR} \in \mathbb{R}^{n\times\ell}.$$

Here:

- $\mathbf{D}$ is a diagonal matrix of random signs,
- $\mathbf{R}$ selects ℓ columns at random, and
- $\mathbf{H} = n^{-1/2}\mathbf{H}_n \in \mathbb{R}^{n\times n}$ is the normalized Walsh–Hadamard matrix. The matrices $\mathbf{H}_n$ are defined recursively by

$$\mathbf{H}_n = \begin{bmatrix} \mathbf{H}_{n/2} & \mathbf{H}_{n/2} \\ \mathbf{H}_{n/2} & -\mathbf{H}_{n/2} \end{bmatrix}, \qquad \text{with} \qquad \mathbf{H}_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}.$$

The matrix–matrix product $\mathbf{AS}$ can be computed in time $\mathrm{O}(mn \log \ell)$.

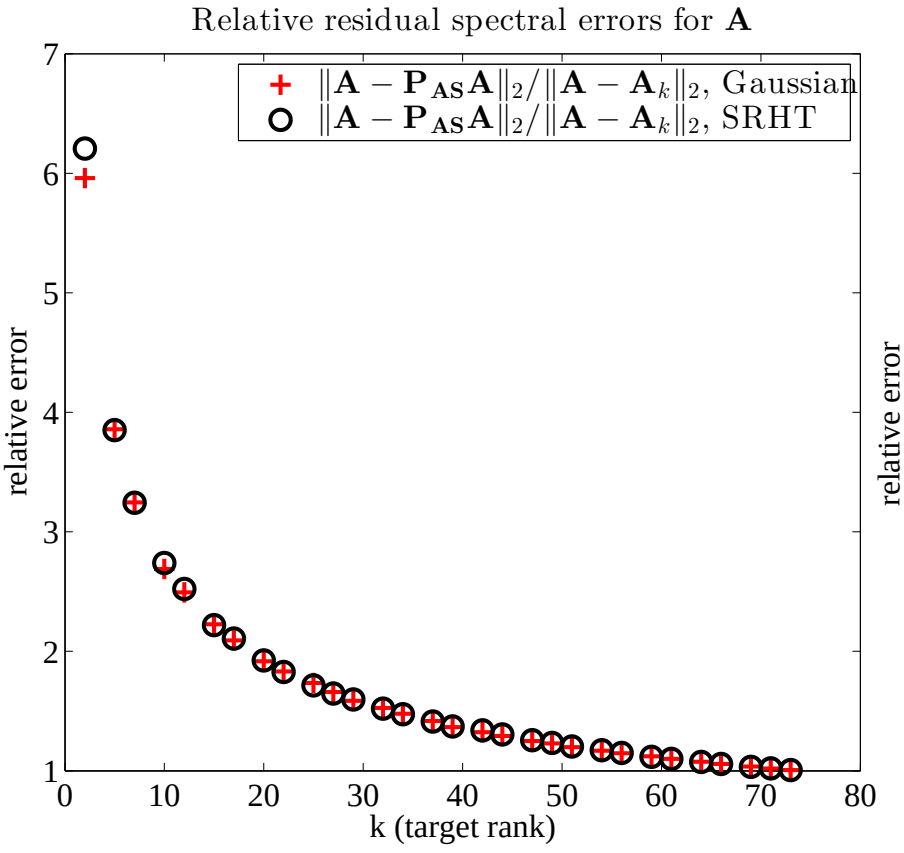
Empirical performance

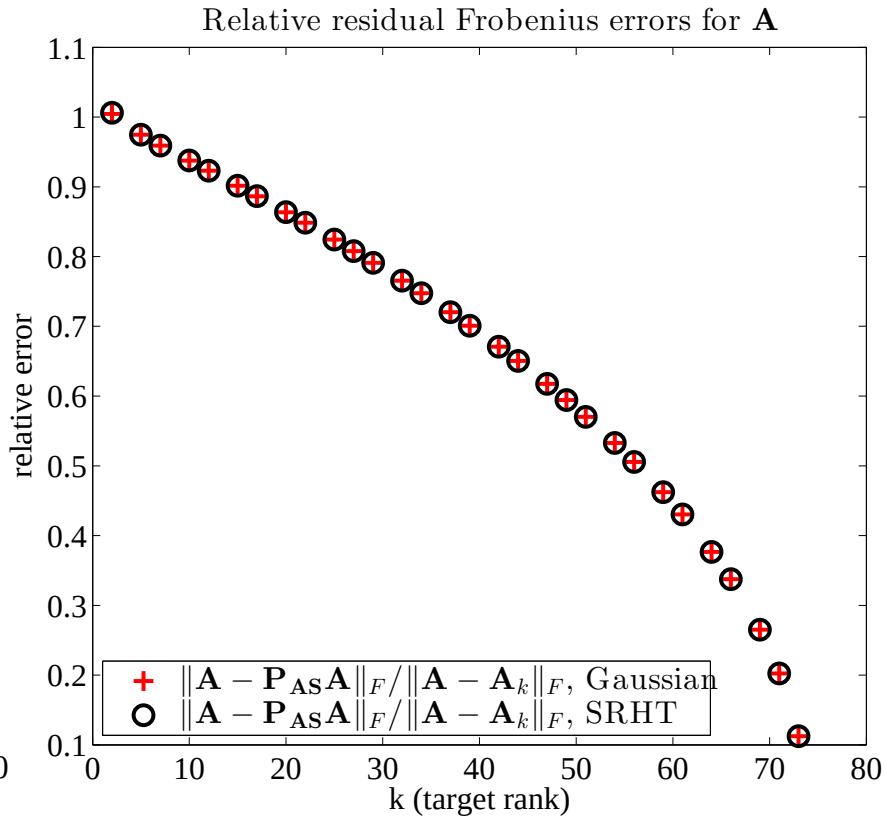
Let $n = 1024$; consider the following test matrix in $\mathbb{R}^{(n+1)\times n}$:

$$\mathbf{A} = [100\mathbf{e}_1 + \mathbf{e}_2, 100\mathbf{e}_1 + \mathbf{e}_3, \ldots, 100\mathbf{e}_1 + \mathbf{e}_{n+1}],$$

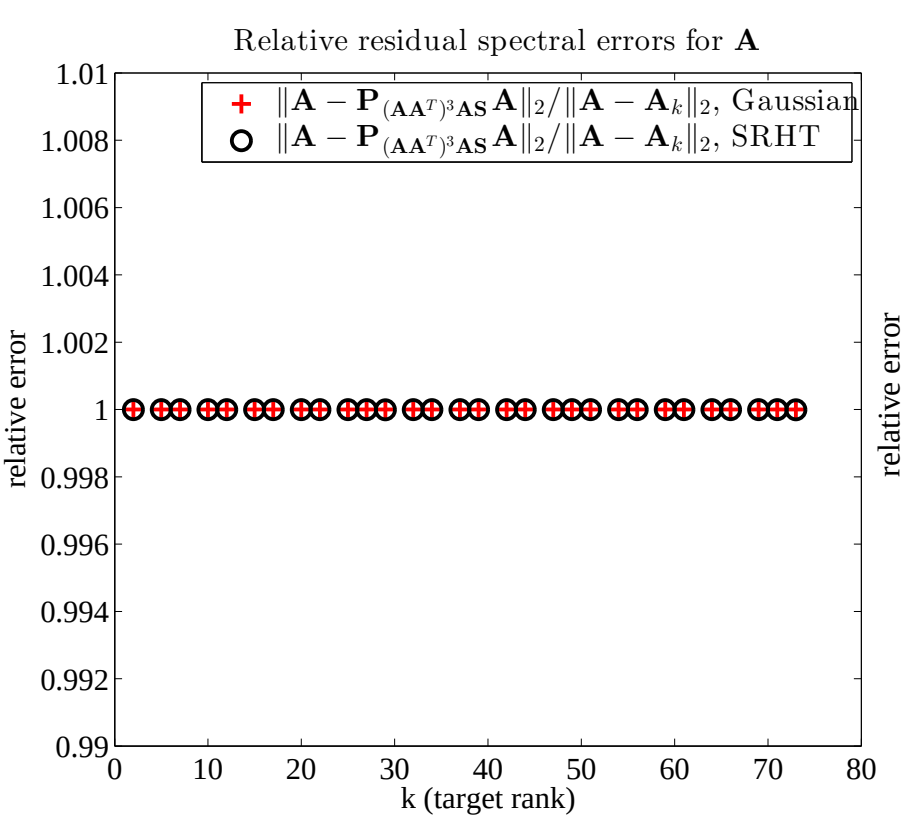
where $\mathbf{e}_i \in \mathbb{R}^{n+1}$ are the standard basis vectors.

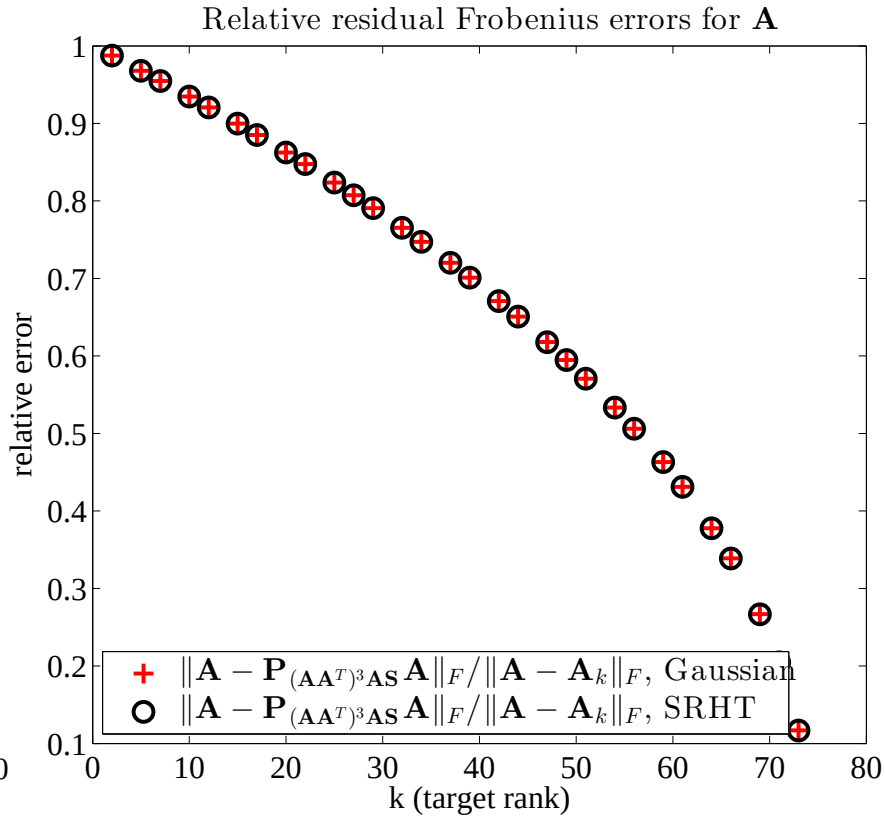
$\mathbf{A}$ is approximately rank-one, and all its columns are biased toward the dominant left singular-vector $\mathbf{e}_1$.





Each point is the average of the errors observed over 30 trials, where each approximation was constructed using $\ell = \lceil 2k \log n \rceil$.



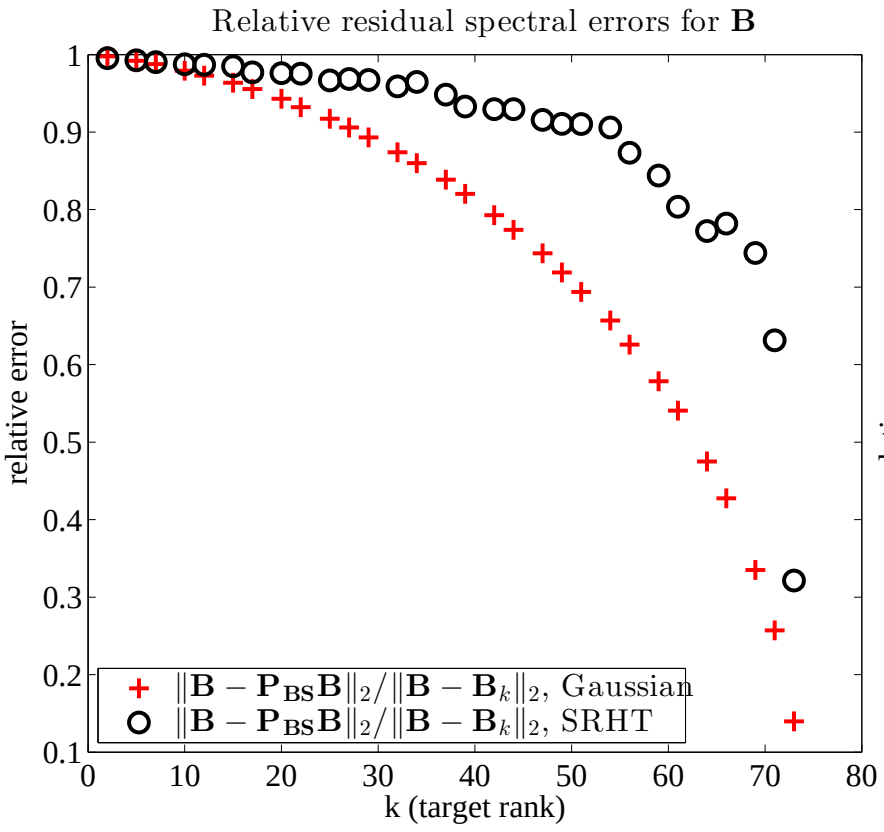


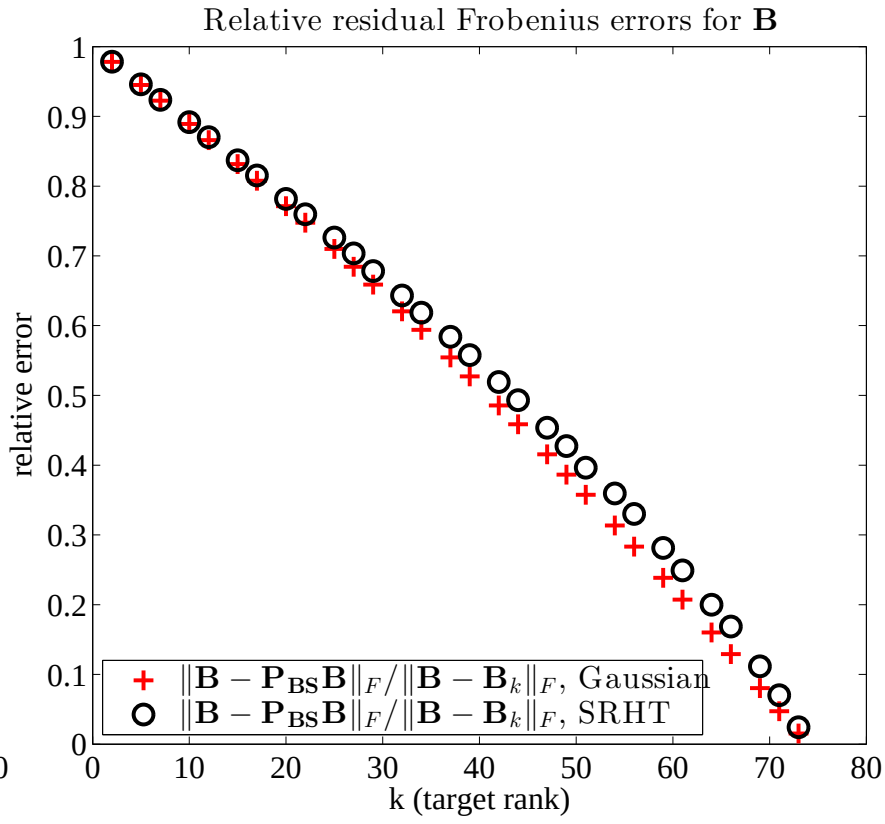
We apply the power method with $p = 3$, while keeping $\ell = \lceil 2k \log n \rceil$.

Let $n = 1024$; consider the diagonal matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ with entries $(\mathbf{B})_{ii} = 100(1 - (i-1)/n)$.

$$\mathbf{B} = \begin{bmatrix} 100 & 0 & 0 & \dots \\ 0 & 99.902 & 0 & \dots \\ 0 & 0 & \ddots & \dots \\ 0 & \dots & 0 & 0.098 \end{bmatrix}$$

$\mathbf{B}$ is full-rank, has slowly decaying spectrum, and only k columns of $\mathbf{B}$ provide any information on the dominant k-dimensional left singular space of $\mathbf{B}$.



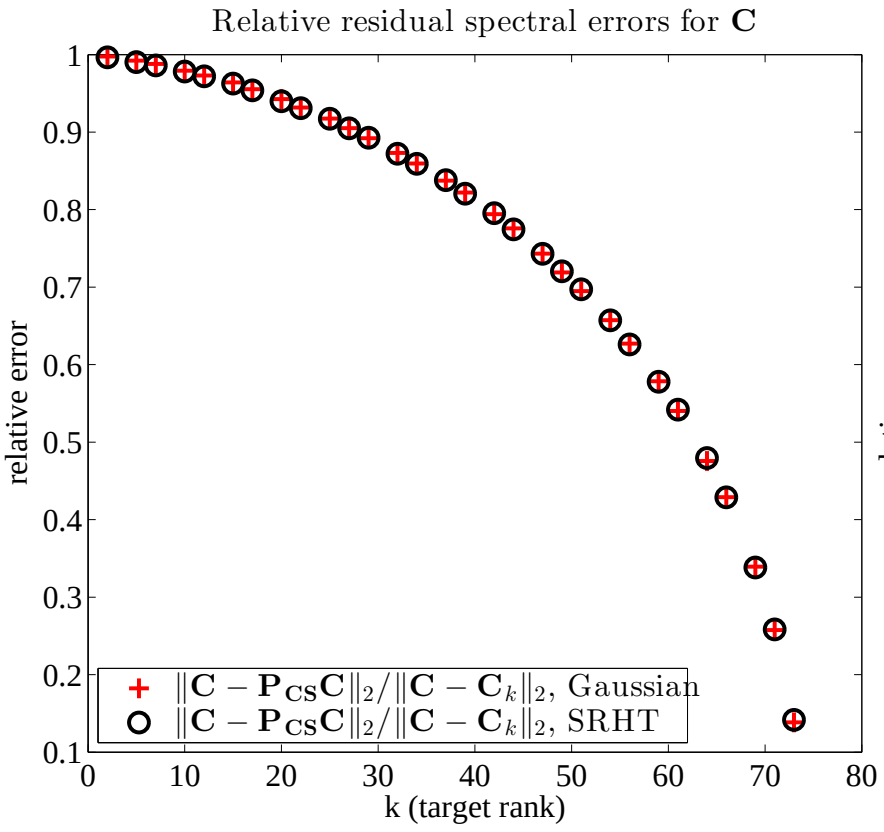


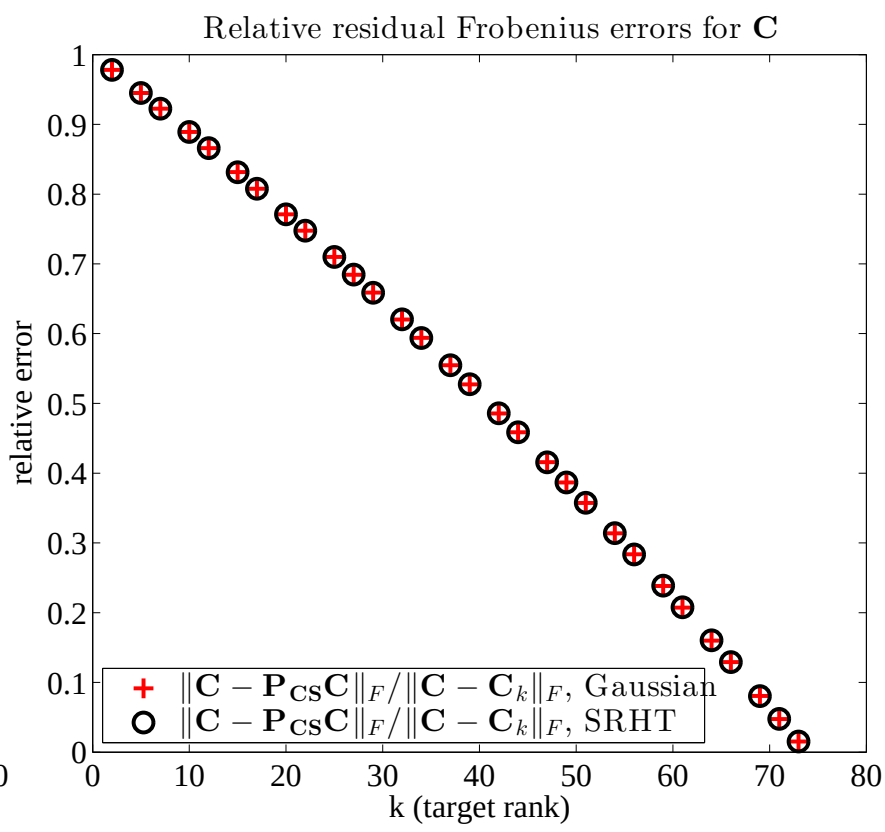
Each point is the average of the errors observed over 30 trials, where each approximation was constructed using $\ell = \lceil 2k \log n \rceil$.

Consider $\mathbf{C} = \mathbf{U}\mathbf{B}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are obtained by taking the SVD of an $n \times n$ matrix of i.i.d. $\mathcal{N}(0,1)$ random variables, $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

$$\mathbf{C} = \mathbf{U} \begin{bmatrix} 100 & 0 & 0 & \dots \\ 0 & 99.902 & 0 & \dots \\ 0 & 0 & \ddots & \dots \\ 0 & \dots & 0 & 0.098 \end{bmatrix} \mathbf{V}^T$$

$\mathbf{C}$ is also full-rank and has slowly decaying spectrum, but every column of $\mathbf{C}$ contains information on every singular space of $\mathbf{C}$.





Each point is the average of the errors observed over 30 trials, where each approximation was constructed using $\ell = \lceil 2k \log n \rceil$.

Observations:

1. When $\mathbf{A}$ is approximately low-rank, the SRHT and Gaussian low-rank approximations exhibit about the same accuracy.
2. When $\mathbf{A}$ is full rank the structures of the singular spaces are important.
   - If the singular vectors are "flat", then SRHT and Gaussian approximations have comparable accuracy.
   - If the singular vectors are axis-aligned, then Gaussian approximations outperform SRHT approximations.
3. Empirically, $\ell = \Omega(k \log n)$ seems to ensure SRHT approximations achieve relative-error bounds.

Prior work

Similar randomized structured projection schemes:

- (Woolfe et al. 2008): if $\ell = \mathrm{O}(k^2)$, then

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 \leq \sqrt{\max\{m, n\}}\|\mathbf{A} - \mathbf{A}_k\|_2$$

- (Nguyen et al. 2009): if $\ell = \mathrm{O}(\varepsilon^{-1} k \log k)$, then

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 \leq (2 + \sqrt{n/\ell})\|\mathbf{A} - \mathbf{A}_k\|_2$$
$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathrm{F}} \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$$

Exactly the same SRHT scheme:

- (Halko et al. 2011): if $\ell = \mathrm{O}(k \log k)$, then for $\xi \in \{2, \mathrm{F}\}$,

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\|_\xi \leq (1 + \sqrt{n/\ell})\|\mathbf{A} - \mathbf{A}_k\|_\xi.$$

For comparison

It follows from (Halko et al. 2011) that when $\mathbf{S}$ is Gaussian and $\ell = \Omega(\epsilon^{-2} k \log n)$,

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2 \leq \left(1 + \frac{\epsilon}{\sqrt{\log n}}\right) \|\mathbf{A} - \mathbf{A}_k\|_2 + \frac{\epsilon}{\sqrt{k \log n}} \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$$

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_F \leq \left(1 + \frac{\epsilon}{\sqrt{\log n}}\right) \|\mathbf{A} - \mathbf{A}_k\|_F$$

simultaneously with probability at least $1 - \frac{2}{n}$.

Gaussians and SRHTs behave quite similar empirically, yet prior analyses for SRHTs are qualitatively poorer than this analysis.

Improved Error Bounds

(Boutsidis and G. 2012)

If $k = \Omega(\log n)$ and $\ell = \Omega(\epsilon^{-2} k \log n)$, then

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2 \leq (4+\epsilon) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2 + \frac{\epsilon}{\sqrt{k}} \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$$

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_{\mathrm{F}} \leq (1+11\epsilon^2) \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}$$

simultaneously with probability at least $1 - \delta$.

This result essentially holds when $\mathbf{S}$ is any subsampled orthogonal transformation,

$$\mathbf{S} = \sqrt{\frac{n}{\ell}} \mathbf{DTR},$$

where $\mathbf{T}$ is an orthogonal transformation matrix with entries on the order of $n^{-1/2}$.

Notation

- Partition the SVD of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \overset{\begin{matrix} k & n-k \end{matrix}}{\begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix}} \begin{bmatrix} \overset{k}{\mathbf{\Sigma}_1} & \\ & \overset{n-k}{\mathbf{\Sigma}_2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}.$$

  Note that $\mathbf{A}_k = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$.

- Define

$$\mathbf{\Omega}_1 = \mathbf{V}_1^T\mathbf{S} \quad \text{and } \mathbf{\Omega}_2 = \mathbf{V}_2^T\mathbf{S},$$

  to capture the interaction of $\mathbf{S}$ with the dominant and residual right singular spaces of $\mathbf{A}$.

Proof sketch

(Boutsidis et al. 2011), (Halko et al. 2011)

If $\mathbf{\Omega}_1 = \mathbf{V}_1^T \mathbf{S}$ has full row rank, then

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_\xi^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_\xi^2 + \left\|\mathbf{\Sigma}_2 \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_\xi^2$$

for $\xi \in \{2, \mathrm{F}\}$.

Geometrical interpretation:

- $\mathbf{V}_1^T \mathbf{S}$ has full row-rank $\Leftrightarrow \tan(\mathbf{V}_1, \mathbf{S}) \neq \infty$.
- $\left\|\mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_2 = \tan(\mathbf{V}_1, \mathbf{S})$.

We bound the additional errors $\left\|\mathbf{\Sigma}_2 \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_2^2$ and $\left\|\mathbf{\Sigma}_2 \mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_\mathrm{F}^2$ when $\mathbf{S}$ is an SRHT.

Key tools

- Matrix Chernoff inequalities that show that if the energy of $\mathbf{M}$ (i.e. its Frobenius norm) is evenly distributed throughout its columns, then matrices consisting of randomly sampled columns of $\mathbf{M}$ have similar extreme singular values to $\mathbf{M}$.
- An extension of a result in (Tropp 2011) to show right multiplication by $\mathbf{DH}$ distributes the energy of any matrix $\mathbf{M}$ evenly over its columns.

Sketch of spectral norm proof

1. By the structural result,

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2^2 + \|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{DHR}\|_2^2 \cdot \left\|(\mathbf{V}_1^T\mathbf{DHR})^\dagger\right\|_2^2.$$

2. **DH** spreads the energy of $\mathbf{V}_1^T$ throughout its columns sufficiently that when enough columns are selected by $\mathbf{R}$, the spectrum of the resulting matrix is close to that of $\mathbf{V}_1^T$. Consequently,

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2^2 + (1 - \sqrt{\epsilon})^{-1}\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{DHR}\|_2^2.$$

3. **DH** spreads the energy of $\mathbf{\Sigma}_2\mathbf{V}_2^T$ throughout its columns sufficiently that when enough columns are selected by $\mathbf{R}$, the norm of the resulting matrix is not much larger than that of $\mathbf{\Sigma}_2$:

$$\mathbb{P}\left\{\|\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{DHR}\|_2^2 \leq \left(5+\frac{\log(n/\delta)}{\ell}\right)\|\mathbf{\Sigma}_2\|_2^2 + \mathrm{O}\left(\frac{\log(n/\delta)^{3/2}}{\ell}\right)\|\mathbf{\Sigma}_2\|_F^2\right\} \geq 1-\delta.$$

4. Combine these pieces and use our lower bounds on k and ℓ to simplify:

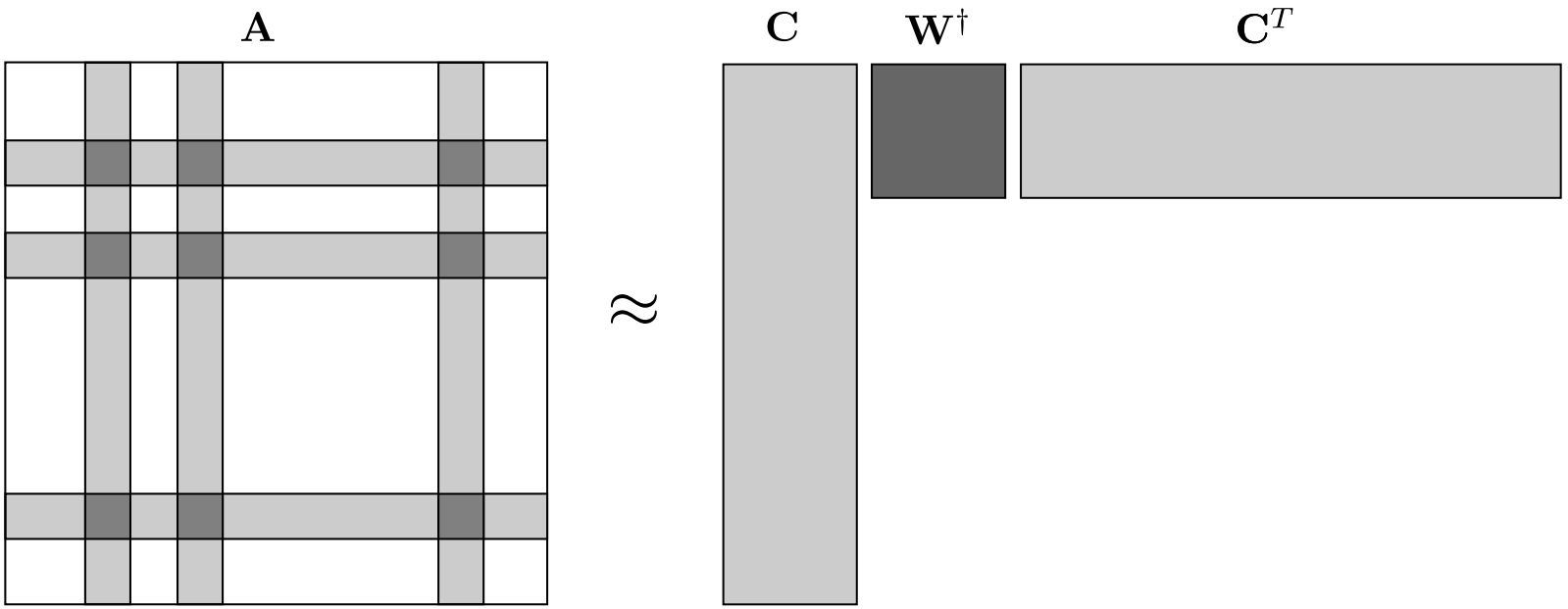
$$\|\mathbf{A}-\mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_2 \leq (4+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_2 + \frac{\epsilon}{\sqrt{k}}\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}.$$

SPSD Sketches

If $\mathbf{A}$ is a positive semidefinite matrix, one may want to *preserve positivity*. SPSD sketches do so:

$$\mathbf{A} \approx \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}$$

where $\mathbf{C} = \mathbf{A}^{p}\mathbf{S}$ and $\mathbf{W} = \mathbf{S}^{T}\mathbf{A}^{2p-1}\mathbf{S}$.
Consider, e.g., if $\mathbf{C}$ corresponds to randomly selected columns of $\mathbf{A}$.



It takes $O(\ell^3 + n\ell^2)$ operations to form this approximation.

The class of SPSD sketches is wide. We consider the following specific sketches:

- When $\mathbf{S}$ selects columns uniformly at random without replacement from $\mathbf{A}$, we call $\mathbf{M}$ a *Nyström extension*.
- When $\mathbf{S}$ consists of i.i.d. $\mathcal{N}(0,1)$ Gaussians, $\mathbf{M}$ is a Gaussian sketch.
- When $\mathbf{S}$ is a subsampled randomized Fourier transform (SRFT), $\mathbf{M}$ is an SRFT sketch.
- When $\mathbf{S}$ selects columns from $\mathbf{A}$ randomly with replacement with probability proportional to their *leverage scores*, $\mathbf{M}$ is a leverage sketch.

Gaussian and SRFT sketches

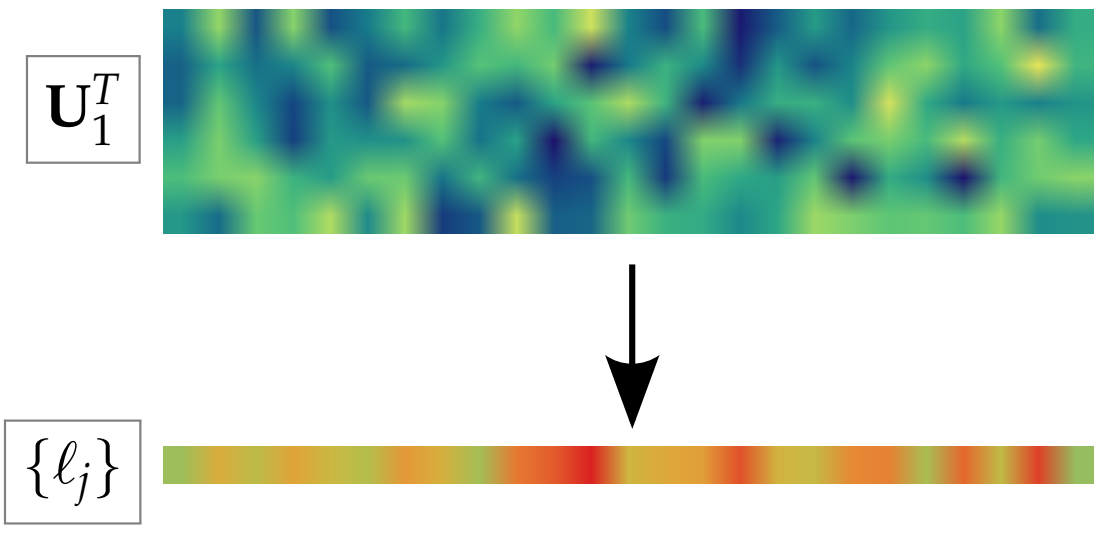
Gaussian and SRFT sketches:

- SRFT sketching suggested in (Chiu and Demanet 2012)
- **S** mixes the columns of **A** together before sampling.
- Mixing process ensures that no columns are ignored.
- Gaussian sketches cost $O(\ell^3 + n^2\ell)$ operations to form.
- SRFT sketches cost $O(\ell^3 + n^2 \log \ell)$ operations to form.

Leverage scores

The statistical leverage scores of the columns of $\mathbf{A}$ (with respect to rank k), are the scaled column norms of $\mathbf{U}_1^T$:

$$\left\{ \ell_j := \frac{n}{k} \left\| (\mathbf{U}_1^T)_j \right\|_2^2, j = 1, \ldots, n \right\}.$$

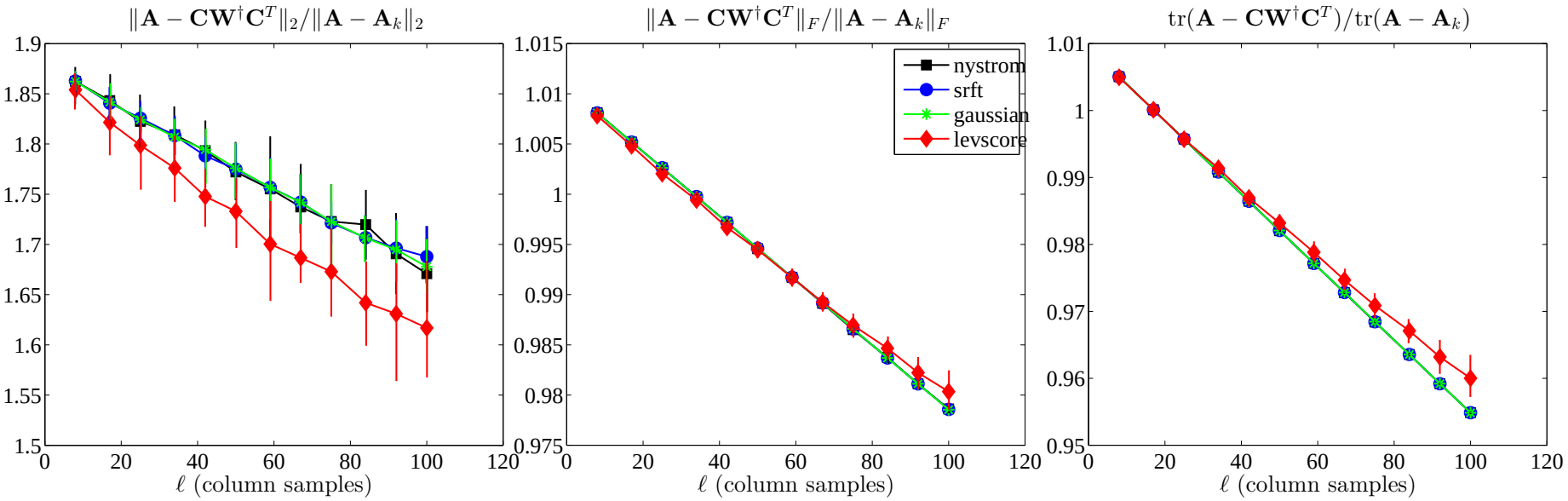


Leverage sketches

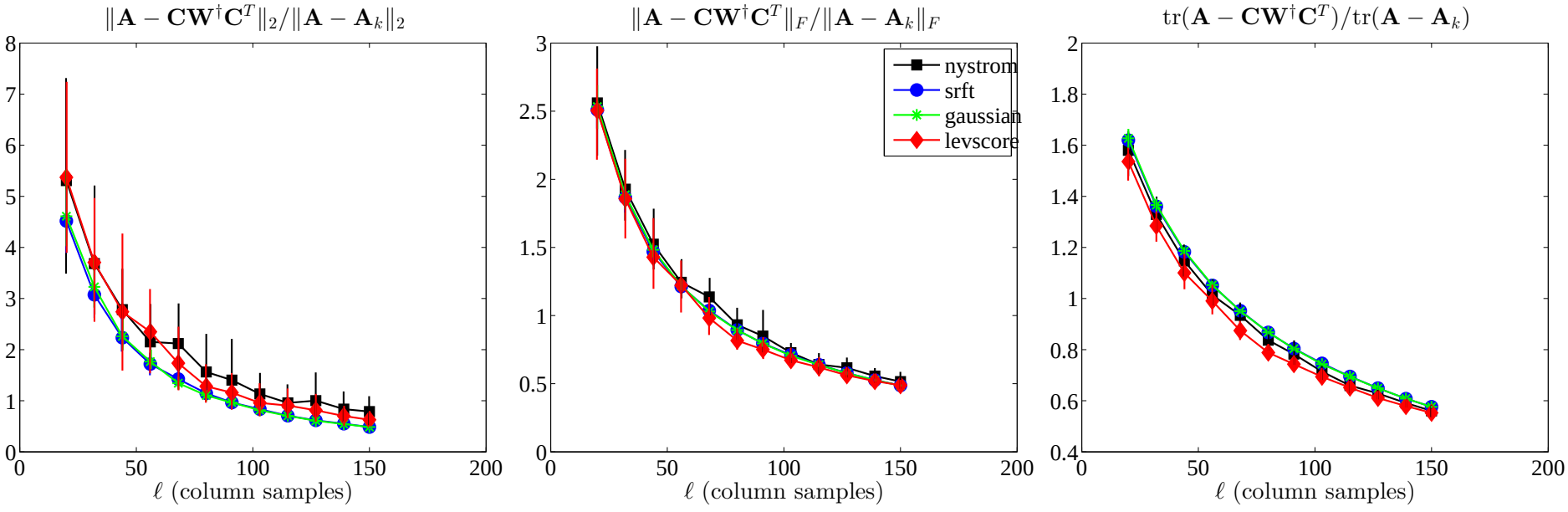
Leverage sketches:

- The idea of leverage score sampling for forming column-sampling based low-rank approximations due to (Drineas et al. 2008).
- Columns are sampled random from $\mathbf{A}$ with probability proportional to their leverage scores.
- Intuitively, leverage score sampling ensures that no important columns are ignored.
- Assuming the leverage scores as given, costs $O(\ell^3 + n^2\ell)$ operations to form.
- The leverage scores can be approximated.

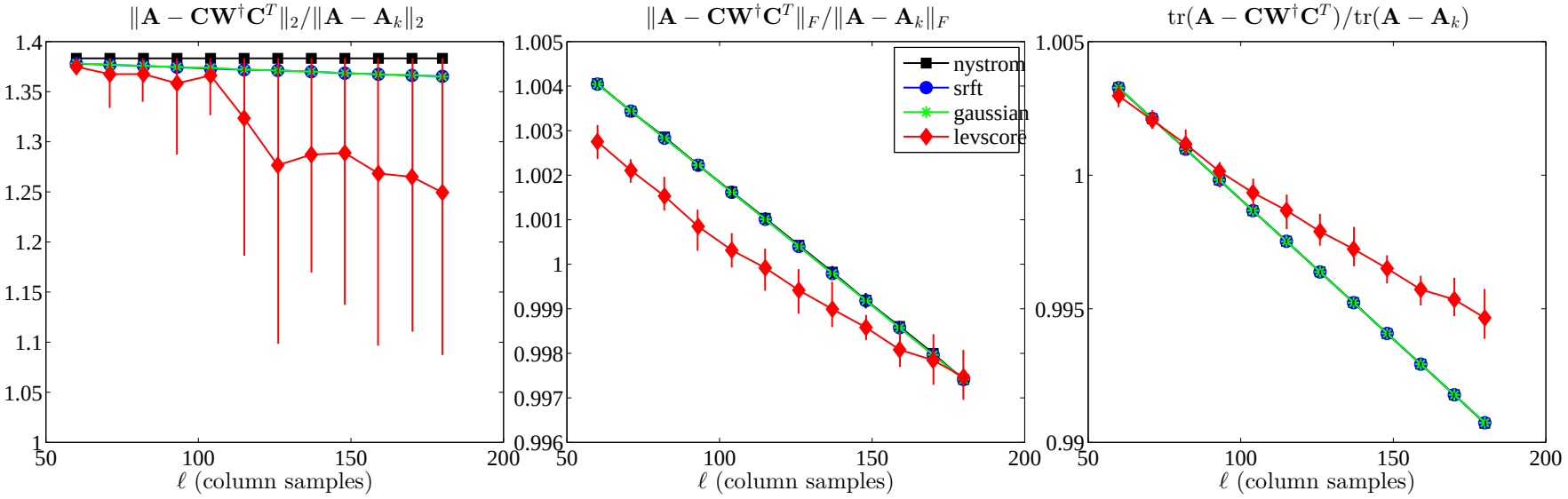
Empirical Performance



Dexter, a 2000×2000 Gram matrix from the UCI Machine Learning Repository. Target rank $k = 8$.



Abalone, a 4898×4898 Radial Basis Kernel matrix from the UCI Machine Learning Repository. Target rank $k = 20$.



Enron, a $10K \times 10K$ Graph Laplacian matrix from the Stanford SNAP collection. Target rank $k = 60$.

Observations:

- The leverage sketches are often the most accurate on average, especially when ℓ is small. However, variance depends on leverage score distribution and can be high.
- The relative trace-norm errors are all smaller than the relative Frobenius-norm errors, which are in turn smaller than the relative spectral-norm errors.
- The sketches are more distinguished by their behavior in the spectral norm than the Frobenius or trace norms.

Prior work ($p = 1$)

Source, sketch	ℓ	$\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_2$
(Drineas and Mahoney 2005), column-sampling	$\Omega(\epsilon^{-4}k)$	$\|\mathbf{A} - \mathbf{A}_k\|_2 + \epsilon \sum_{i=1}^{n} A_{ii}^2$
(Talwalkar and Rostamizadeh 2010), Nyström	$\Omega(\mu k \log k)$	0, if $\text{rank}(\mathbf{A}) = k$
(Kumar et al. 2012), Nyström	$\Omega(1)$	$\|\mathbf{A} - \mathbf{A}_k\|_2 + (n/\sqrt{\ell}) \max_{ii} A_{ii}$
(Chiu and Demanet 2012), Nyström	$\Omega(\mu k \log n)$	$(1 + n/\ell)\|\mathbf{A} - \mathbf{A}_k\|_2$
(Chiu and Demanet 2012), SRFT sketch	$\Omega(k \log^2 n)$	$(1 + n/\ell)\|\mathbf{A} - \mathbf{A}_k\|_2$

- Here $\mu \in [1, \frac{n}{k}]$ is the "coherence" of the matrix.
- The estimated additional error in (Drineas and Mahoney 2005) can be on the order of $\epsilon \operatorname{tr}(\mathbf{A})$.
- The (Talwalkar and Rostamizadeh 2010) exact recovery result requires $\mathbf{A}$ to be exactly low-rank.
- The (Chiu and Demanet 2012) results require $\Omega(k \log n)$ samples as opposed to $\Omega(k \log k)$. The factor n/ℓ is optimal in the Nyström bound, but unnecessary in the SRFT bound.

(G. and Mahoney 2013) provides a framework for deriving significantly improved asymptotic error bounds.

Structural error bounds for SPSD sketches

Recall the partitioned eigendecomposition of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T = [\mathbf{U}_1 \;\; \mathbf{U}_2] \begin{bmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{bmatrix} [\mathbf{U}_1 \;\; \mathbf{U}_2]^T$$

and that

$$\boldsymbol{\Omega}_1 = \mathbf{U}_1^T\mathbf{S} \quad \text{and } \boldsymbol{\Omega}_2 = \mathbf{U}_2^T\mathbf{S}$$

capture the interactions of the sketching matrices with the dominant and residual eigenspaces of $\mathbf{A}$.

If $\mathbf{S}$ has orthonormal columns, then

$$\left\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{\dagger}\right\|_2 = \tan(\mathbf{S}, \mathbf{U}_1).$$

Our error bounds for SPSD sketches follow from the key observation that

SPSD sketches approximate $\mathbf{A}^{1/2}$, (G. 2011)

$$\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T} = (\mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}})(\mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}}\mathbf{A}^{1/2}).$$

Thus the errors of SPSD sketches satisfy

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_{\xi} = \|\mathbf{A} - \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{p-1/2}\mathbf{S}}\mathbf{A}^{1/2}\|_{\xi}$$

for $\xi \in \{2, \mathrm{F}, \mathrm{tr}\}$.

We extend the framework provided in (Halko et al. 2011) for projection-based low-rank approximations to find *deterministic* error bounds for SPSD sketches.

Simplified versions of these bounds (for $p = 1$):

(G. and Mahoney, 2013)

If $\mathbf{\Omega}_1 = \mathbf{U}_1^T \mathbf{S}$ has full row rank, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T\|_2 \leq \left(1 + \left\|\mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_2^2\right) \|\mathbf{A} - \mathbf{A}_k\|_2,$$

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T\|_\mathrm{F} \leq \|\mathbf{A} - \mathbf{A}_k\|_\mathrm{F} + 2\sqrt{2}\left\|\mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_2^2 \cdot \mathrm{tr}(\mathbf{A} - \mathbf{A}_k), \text{ and}$$

$$\mathrm{tr}(\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T) \leq \left(1 + \left\|\mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger\right\|_2^2\right) \cdot \mathrm{tr}(\mathbf{A} - \mathbf{A}_k)$$

- The randomness of $\mathbf{S}$ enters only through the sketching interaction matrix $\mathbf{\Omega}_2 \mathbf{\Omega}_1^\dagger$.
- The spectral-norm and trace-norm additional errors of sketches are proportional to the optimal errors.
- The Frobenius-norm additional errors of sketches are proportional to the optimal trace-norm errors.

Spectral error bounds

Leverage sketches

If $\ell = \Omega(\epsilon^{-1} k \log(k/\delta))$, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}\mathbf{C}^T\|_2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2 + \epsilon \operatorname{tr}(\mathbf{A} - \mathbf{A}_k)$$

with probability at least $1 - \delta - 0.4$.

SRFT sketches

If $k = \Omega(\log n)$ and $\ell = \Omega(\epsilon^{-2} k \log(n/\delta))$, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}\mathbf{C}^T\|_2 \leq \left(5 + \frac{\epsilon}{\sqrt{k}}\right) \|\mathbf{A} - \mathbf{A}_k\|_2 + \frac{\epsilon^2}{k} \operatorname{tr}(\mathbf{A} - \mathbf{A}_k)$$

with probability at least $1 - \delta$.

Gaussian sketches

If $\ell = \Omega((1 + \epsilon^{-1})k)$, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}\mathbf{C}^T\|_2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2 + \frac{\epsilon}{k}\operatorname{tr}(\mathbf{A} - \mathbf{A}_k)$$

with probability at least $1 - k^{-1} - \mathrm{e}^{-k\epsilon^{-1}}$.

NYSTRÖM EXTENSIONS

Nyström extensions perform well when the information in its top k-dimensional eigenspace is spread throughout $\mathbf{A}$:

$$\mathbf{A} = 20 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

versus

$$\mathbf{A} = 20 \begin{bmatrix} 1/2 \\ -1/2 \\ -1/2 \\ 1/2 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 & -1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 5 & -5 & -5 & -5 \\ -5 & 5 & 5 & -5 \\ -5 & 5 & 5 & -5 \\ 5 & -5 & -5 & 5 \end{bmatrix}$$

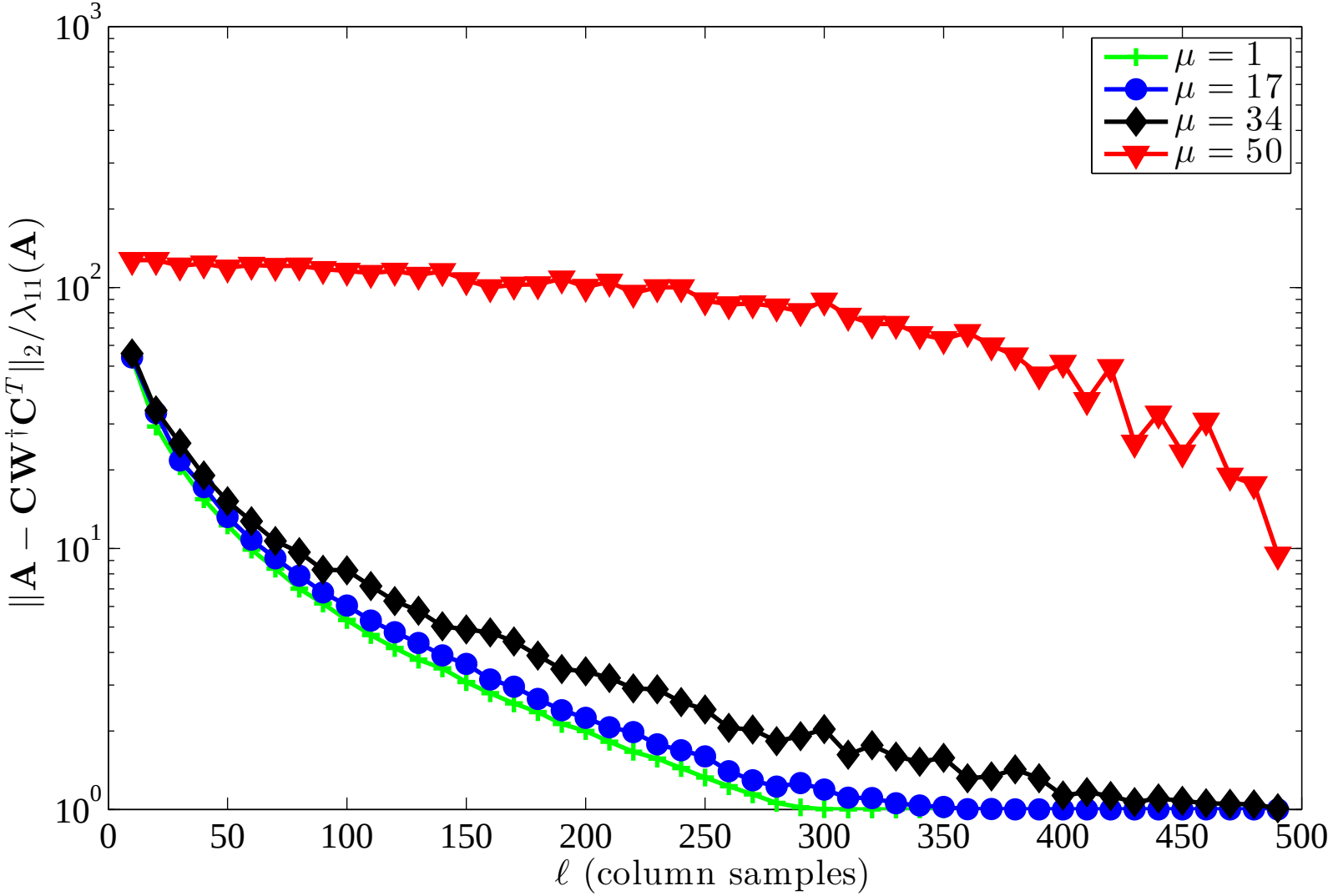
key point: we need the support of the top k eigenvectors to be spread out.

A measure of the "spreadness" of the eigenvectors in $\mathbf{U}_1$ is given by the *coherence* of $\mathbf{U}_1$:

$$\mu := \frac{n}{k} \max_j \|(\mathbf{U}_1)_j\|_2^2.$$

- coherence is the largest of the leverage score of the columns of $\mathbf{A}$.
- μ is between 1 (best case) and n/k (worst case)

Empirical impact of coherence



$\mathbf{A} \in \mathbb{R}^{500\times500}$ is full-rank, but numerically rank 20. The target rank $k = 10$. Each point is the average of 60 trials.

The best previous error bound on the error of Nyström extensions in terms of the coherence is

(Talwalkar and Rostamizadeh 2010)

Let $\mathbf{A}$ be exactly rank-k. If $\ell = \Omega(\mu k \log(k/\delta))$, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\| = 0$$

with probability at least $1 - \delta$.

What about $\mathbf{A}$ that are not exactly low-rank?

Error bounds for Nystrom extensions

The approximation errors can be bounded when ℓ is proportional to the coherence.

Spectral-norm error bound (G. 2011)

If $\ell \geq 8\mu k \log(k/\delta)$, then

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2 \left(1 + \frac{2n}{\ell}\right)$$

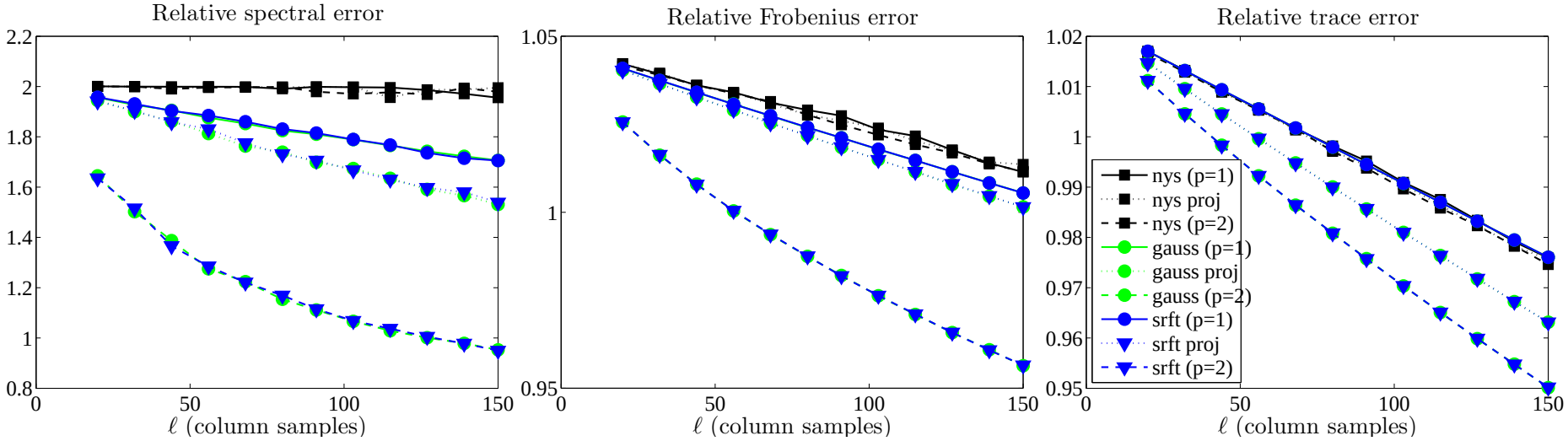
with probability at least $1 - \delta$.

This bound is **tight**: there are matrices for which the relative spectral-norm error is on the order of n/ℓ.

Random projections vs SPSD sketches

Could approximate SPSD matrices with $\mathbf{P_{AS}AP_{AS}}$.

- $\mathbf{P_{AS}AP_{AS}}$ is $(\mathbf{P_{AS}A^{1/2}})(\mathbf{A^{1/2}P_{AS}})$.
- The two-pass sketch $(\mathbf{A^2S})(\mathbf{S}^T\mathbf{A}^3\mathbf{S})^\dagger(\mathbf{S}^T\mathbf{A}^2)$ is $(\mathbf{A^{1/2}P_{A^{3/2}S}})(\mathbf{P_{A^{3/2}S}A^{1/2}})$.



Wine, a 4898×4898 sparse Radial Basis Kernel matrix from the UCI Machine Learning Respository. Target rank $k = 20$. Each point is the average relative error over 30 trials.

Conclusion

Considered two classes of low-rank approximation:

1. A fast projection-based scheme for arbitrary matrices.
   - Established a relative-error Frobenius-norm error bound and an improved additive-error spectral-norm bound.
   - Provided empirical evidence that $\Omega(k \log n)$ samples suffice to obtain small spectral- and Frobenius-norm errors in practice.
2. A new class of SPSD sketches.
   - Introduced leverage score sketches and provided empirical evidence they outperform alternative sketches.
   - Provided theoretical error guarantees for several types of SPSD sketches.
   - Established an optimal relative-error spectral-norm bound for Nyström extensions.
   - Provided empirical evidence that SPSD sketches perform well on a wide range of matrices that arise in machine learning.

References: (preprints on arXiv)

- "The spectral norm error of the naïve Nyström extension", (Gittens 2011).
- "Improved matrix algorithms via the Subsampled Randomized Hadamard Transform", (Boutsidis and G. 2012). SIMAX to appear.
- "Revisiting the Nyström Method for Improved Large-Scale Machine Learning", (G. and Mahoney 2013)
- Nyström Bestiary (Matlab code for experiments with SPSD sketches) `http://users.cms.caltech.edu/~gittens/nystrombestiary/`