Linear and Nonlinear Low-Rank Approximations

in theory and practice

Alex Gittens
International Computer Science Institute (ICSI)
Department of Statistics and AMPLab
University of California, Berkeley

RPI CS Colloquium
May 5, 2016

Motivation

- The matrices arising in modern scientific computation and data analysis applications are massive
- Low-rank matrix approximation is a fundamental tool, used for e.g. model simplification, noise elimination.
- Classical low-rank matrix approximation algorithms (based on e.g. singular value or rank-revealing QR decompositions, or Krylov space methods) can be prohibitively expensive in terms of arithmetic and communication costs.
- Randomized low-rank approximations often have lower arithmetic and communication costs.

What guarantees are available for randomized low-rank approximations? How well do they perform in practice?

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?

Overview

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: statistical analysis, spectral clustering, kernel methods, optimization, ...
2. Traditional deterministic approaches (via truncated SVD, rank-revealing QR, Krylov methods) cost at least $O(mnk \log \min\{m, n\})$ operations, and can have high communications costs.

Naturally, one can consider using randomness to assist in the design of time and communication efficient algorithms for finding low-rank approximations of large matrices.

This talk presents new error bounds for "sketching" schemes for symmetric positive semidefinite matrices.

Our objective

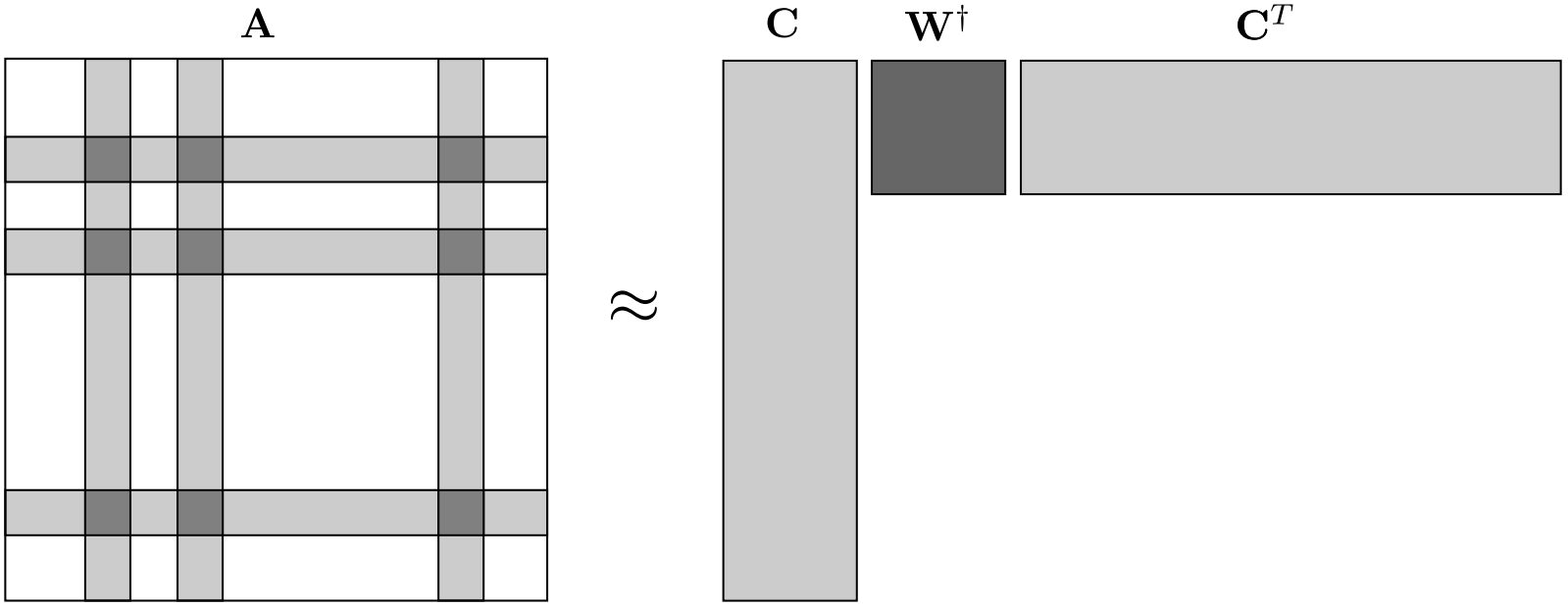
Determine how the errors of SPSD sketches in the spectral, Frobenius, and trace norms compare with the errors of $\mathbf{A}_k$, the best rank-k approximation to $\mathbf{A}$.

NYSTRÖM EXTENSIONS

The simplest SPSD sketch is the Nyström extension:

$$\mathbf{A} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T},$$

where $\mathbf{C}$ consists of uniformly randomly chosen columns of $\mathbf{A}$ and $\mathbf{W}$ consists of $\mathbf{C}$ restricted to the corresponding rows.



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

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

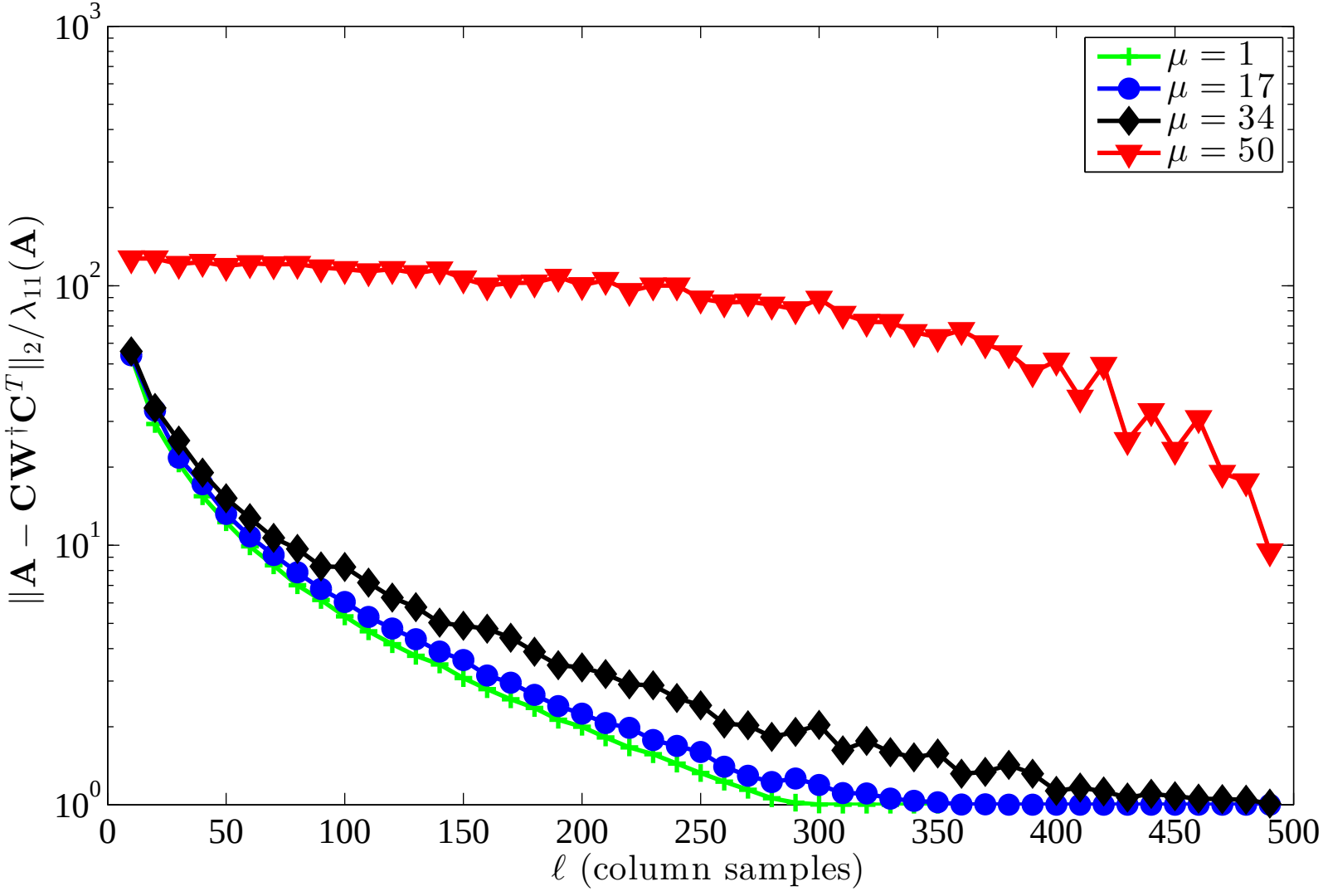
key point: we need the support of the top k eigenvectors to be spread out.

Let $\mathbf{U}_1$ be an orthonormal basis for the dominant k-dimensional eigenspace of $\mathbf{A}$.

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

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

- μ is between 1 (best case) and n/k (worst case)
- μ both theoretically *and empirically* determines the feasibility of Nyström extensions...



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

SPSD Sketches

SPSD sketches generalize the Nyström extension, by potentially mixing information across the columns of $\mathbf{A}$, thereby removing the sensitivity to μ.

Fix an arbitrary "sketching matrix" $\mathbf{S} \in \mathbb{R}^{n \times \ell}$. The corresponding sketch of the SPSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is

$$\hat{\mathbf{A}} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T},$$

where $\mathbf{C} = \mathbf{A}\mathbf{S}$ and $\mathbf{W} = \mathbf{S}^{T}\mathbf{A}\mathbf{S}$.

- The sketch $\hat{\mathbf{A}}$ is also SPSD.
- This model allows for both column-sampling based approximations (e.g. Nyström extensions) and column-mixture based approximations.
- To form the sketch requires only one pass over $\mathbf{A}$.

Choice of sketching matrices

Dominant arithmetic cost of forming the sketch is the matrix–matrix multiply **AS**.

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

- Computation of **AS** takes $O(n^2\ell)$ time for general **A**.
- The columns of **A** are well-mixed.

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

- Computation of **AS** takes reduced time $O(n^2 \log(\ell))$.
- Mixing not as uniform, so potential accuracy loss.

Subsampled randomized orthogonal transforms are one class of structured random projections:

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

Here:

- $\mathbf{D}$ is a diagonal matrix of random signs,
- $\mathbf{R}$ selects ℓ columns at random, and
- $\mathbf{T}$ is the matrix of an orthogonal transformation that is associated with a fast transform.

Examples: if $\mathbf{T}$ is the real Fourier transformation or $\mathbf{T}$ is the normalized Walsh–Hadamard matrix, then the matrix–matrix product $\mathbf{AS}$ can be computed in time $\mathrm{O}(n^2 \log \ell)$.

In this talk, we consider the following specific randomized sketches, corresponding to different distributions on $\mathbf{S}$:

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

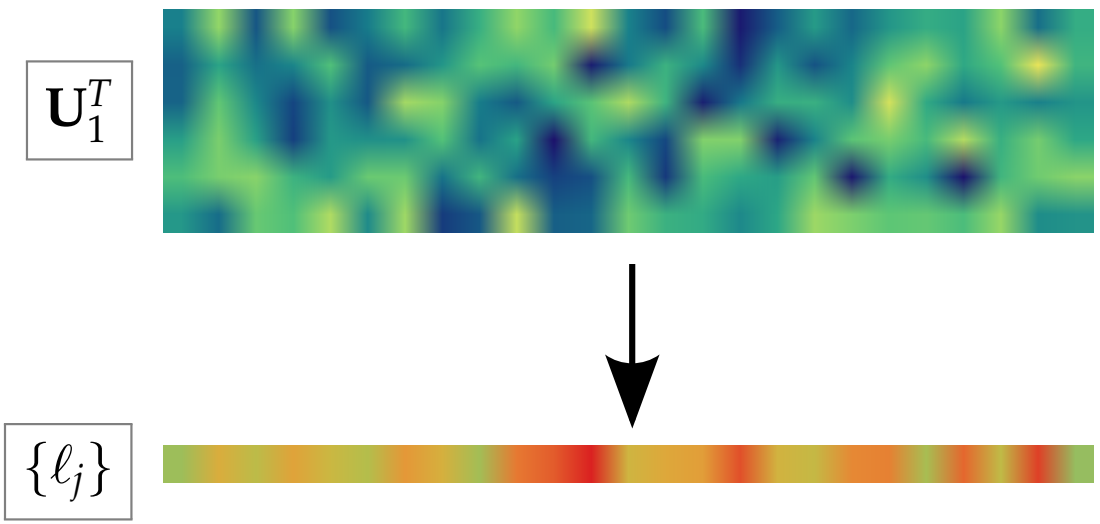
Gaussian and SRFT sketches:

- 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

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

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

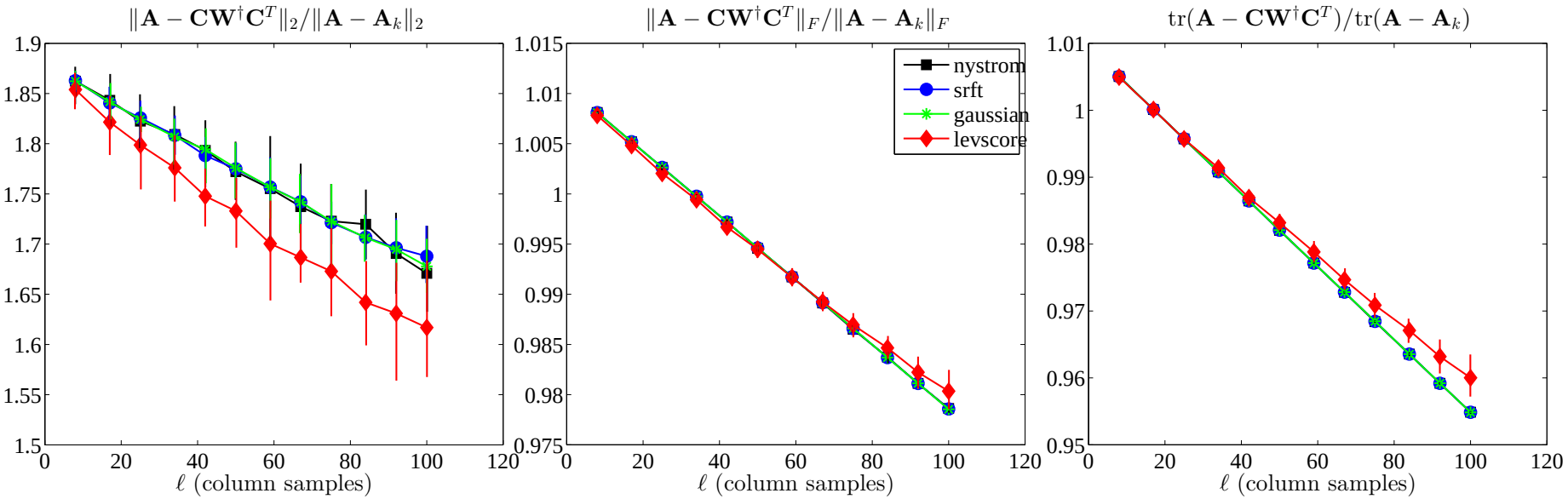


Note that μ, the coherence, is the largest of the leverage scores of the columns of $\mathbf{A}$.

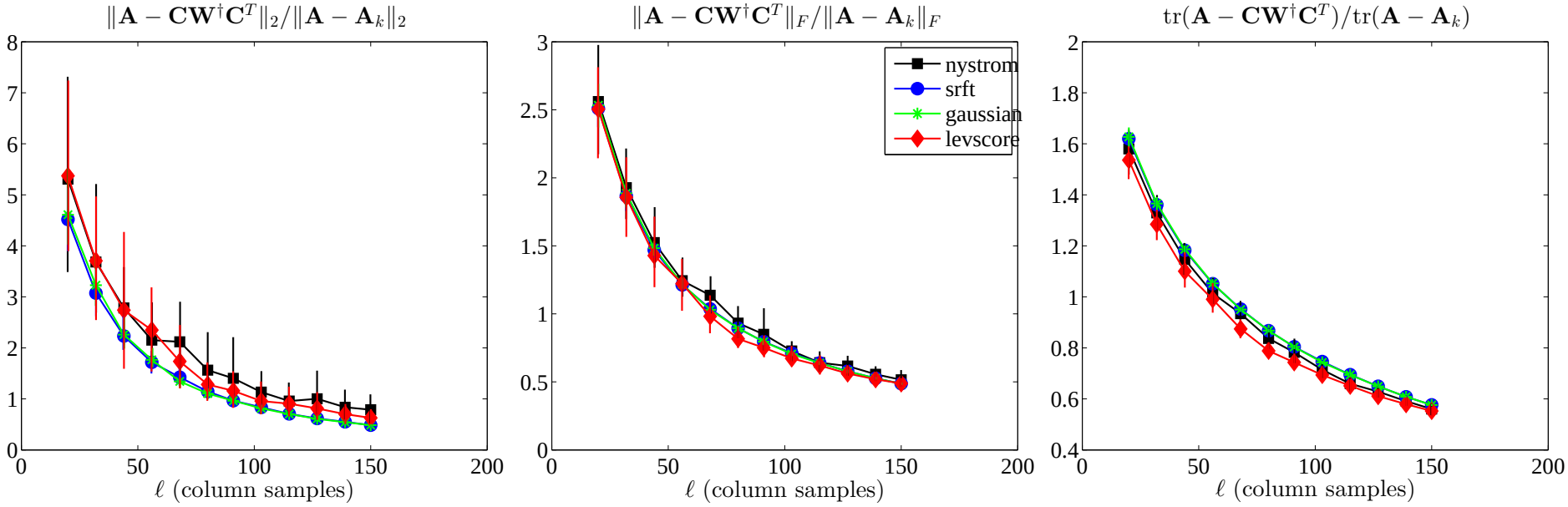
Leverage sketches:

- The idea of leverage score sampling for forming column-sampling based low-rank approximations due to (Drineas et al. 2008).
- Columns are sampled randomly 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.

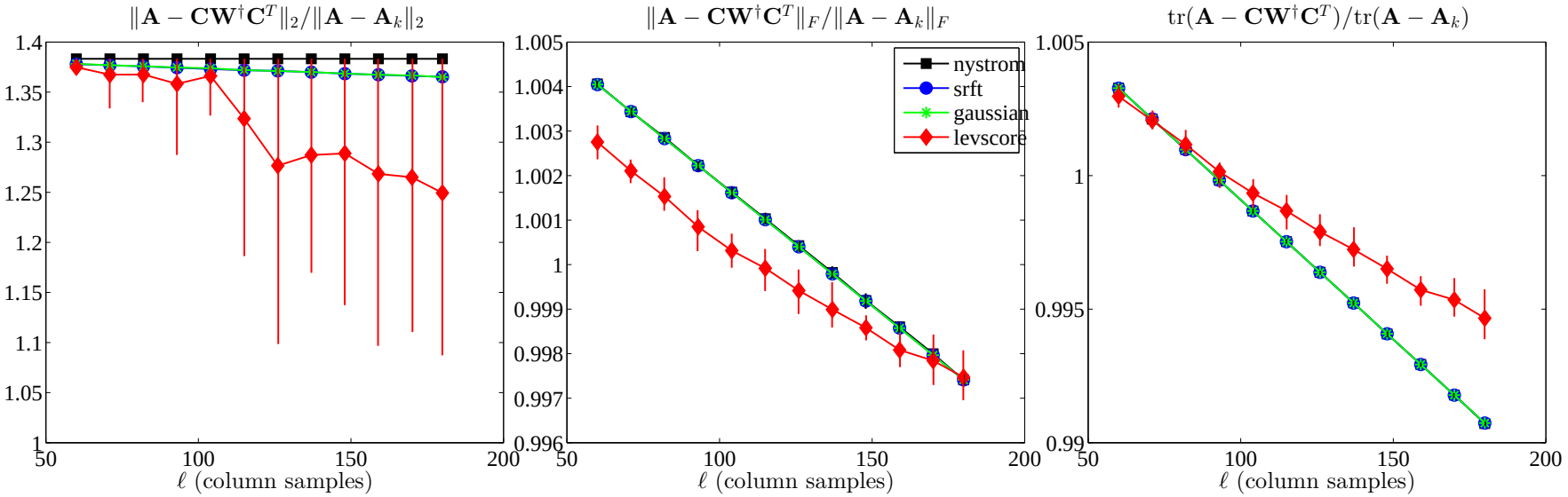
Empirical Performance (Exact schemes)



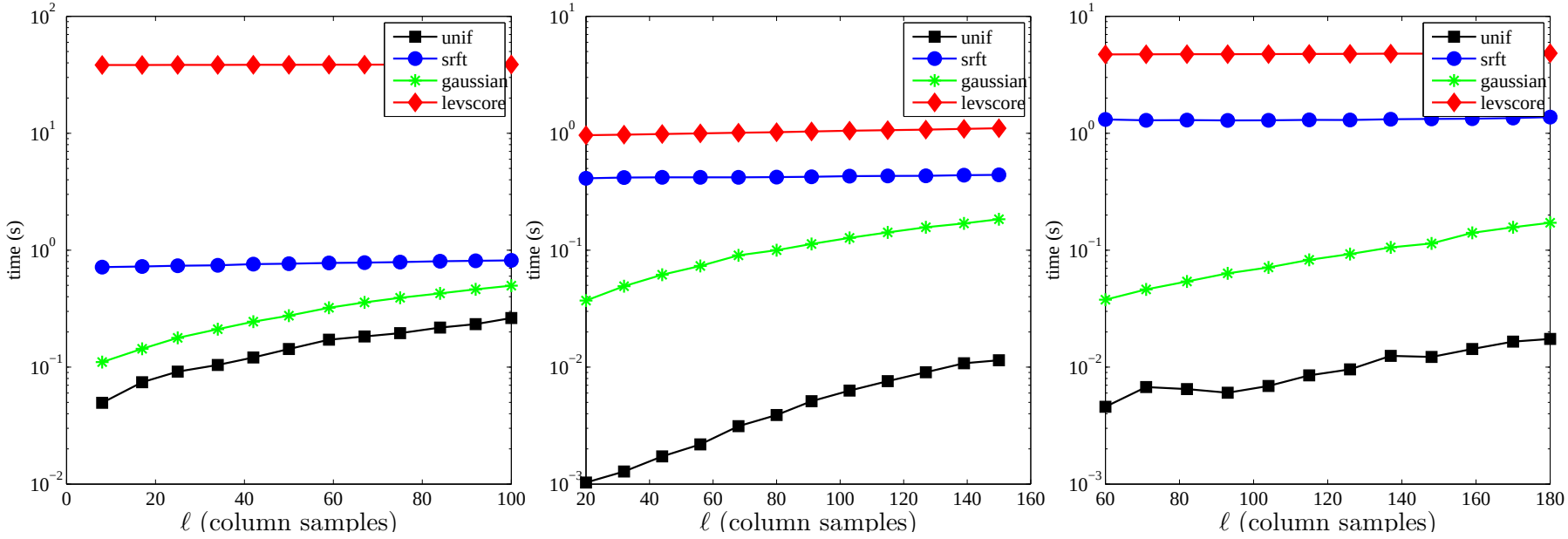
Dexter, a 2000×2000 Gram matrix from the UCI Machine Learning Repository. Target rank $k = 8$.



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 $k = 60$.



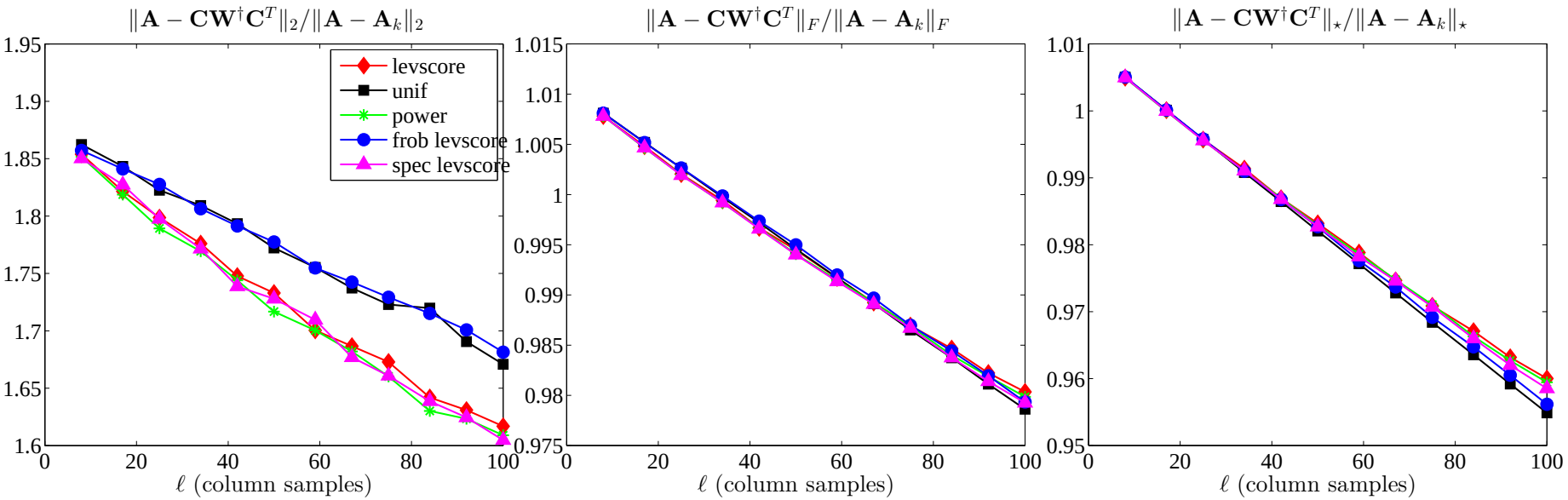
Left to right, computation times for the sketches of the Dexter, Abalone, and Enron matrices.

- The fact that Dexter is a linear kernel (i.e. $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ for some $\mathbf{X}^T$) allows us to speed up the computation of $\mathbf{AS}$.
- We used the $\mathrm{O}(n \log n)$ implementation of the SRFT available in Matlab, as opposed to the $\mathrm{O}(n \log \ell)$ theoretically possible.
- Gaussian multiplication is more efficient on sparse datasets than SRFT multiplication.

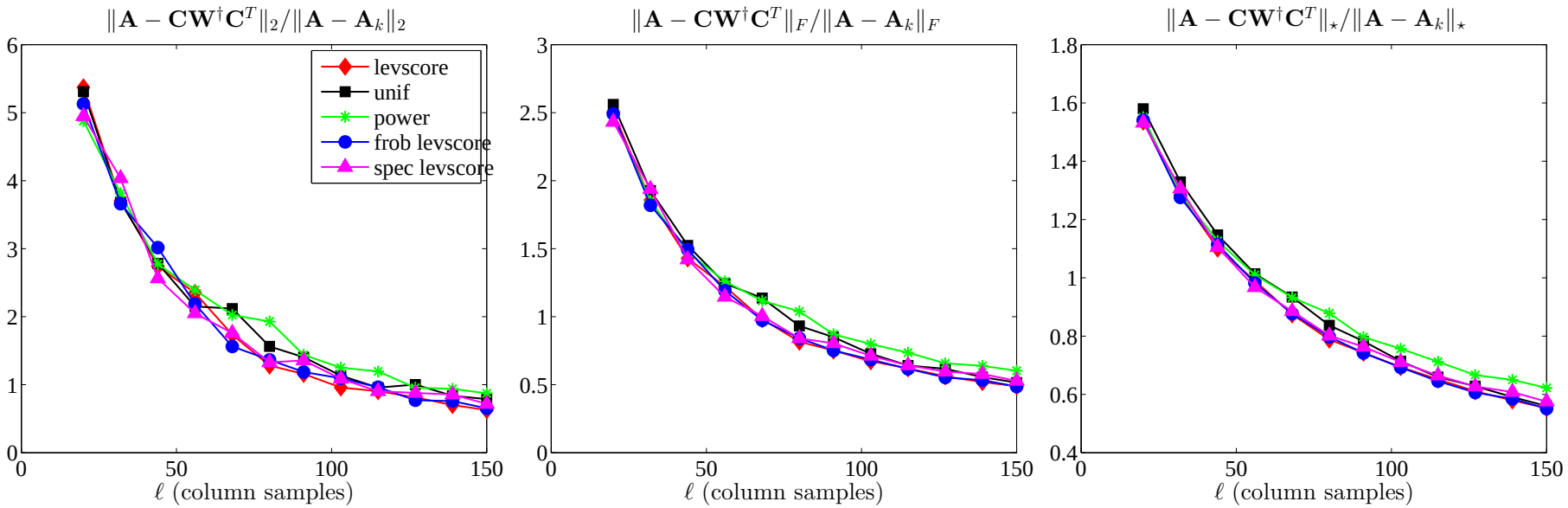
Empirical performance (inexact schemes)

Obtaining the exact leverage scores (via QR or SVD) is expensive. We substitute the following approximations:

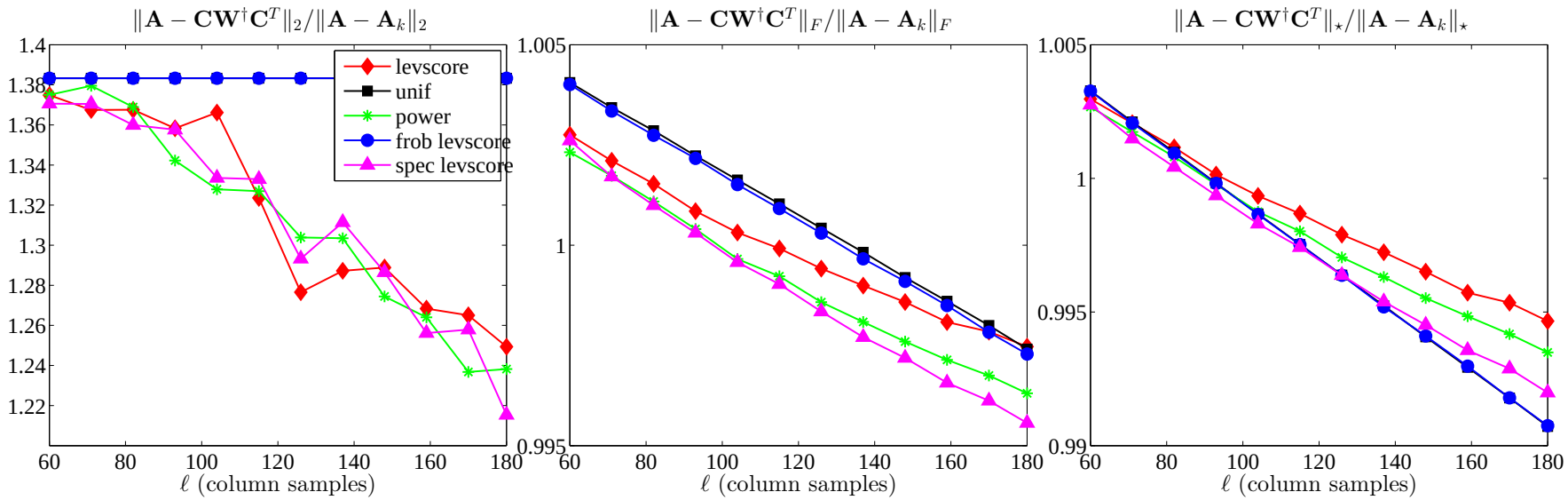
- Power method approximations. Use the power method to obtain approximate bases for the dominant k-dimensional eigenspace of $\mathbf{A}$. Terminate when the leverage scores converge.
- Frobenius-norm approximations. Use a random projection to quickly construct a matrix close to $\mathbf{A}$ in the Frobenius norm, and use the leverage scores of this matrix (Drineas et al. 2012).
- Spectral-norm approximations. Use a random projection to quickly construct a matrix close to $\mathbf{A}$ in the spectral norm, and use the leverage scores of this matrix (Drineas et al. 2012).



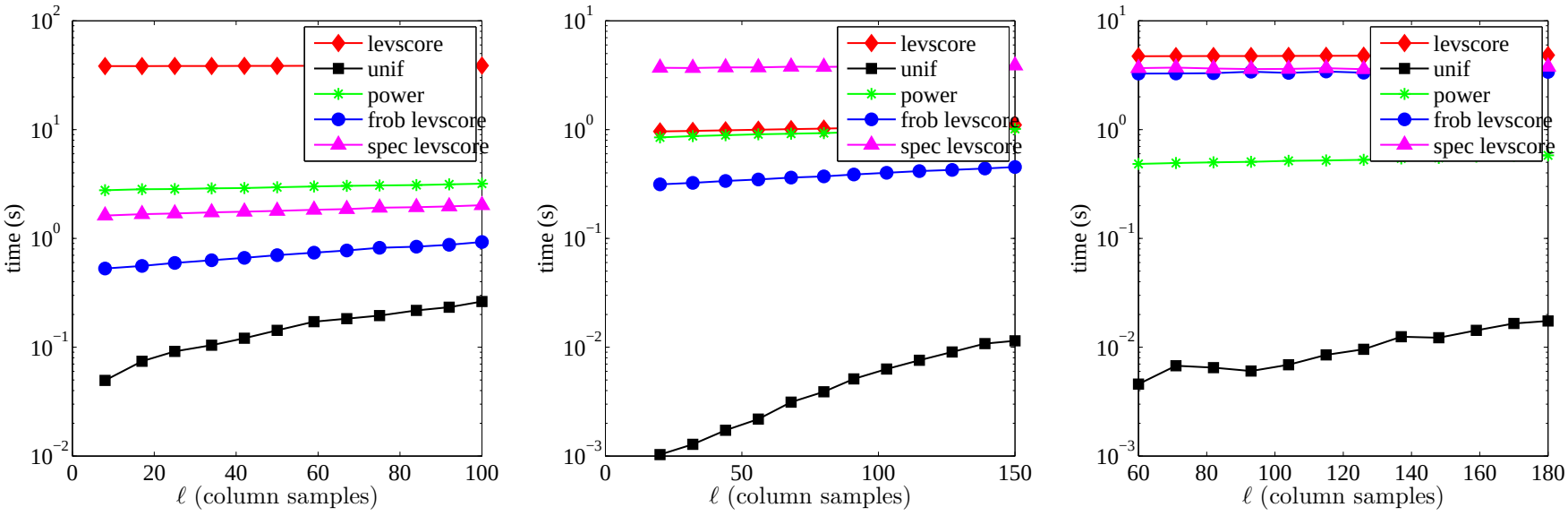
Dexter, a 2000×2000 Gram matrix from the UCI Machine Learning Repository. Target rank $k = 8$.



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 $k = 60$.



Left to right, computation times for the sketches of the Dexter, Abalone, and Enron matrices.

Which approximate leverage score algorithm is appropriate (more efficient than exact leverage score computation) depends on the properties of the matrix.

Prior work

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 $\operatorname{rank}(\mathbf{A}) = k$
(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	$\Omega(k \log^2 n)$	$(1 + n/\ell)\|\mathbf{A} - \mathbf{A}_k\|_2$

- 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.

Application 1: Optimal bounds for Nystrom extensions

The approximation errors can be bounded when ℓ is proportional to the coherence; our framework gives the following result.

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 $1 - \delta$.

A matrix constructed in (Boutsidis et al. 2011) shows that this bound is **tight**: there are matrices for which the relative spectral-norm error is on the order of n/ℓ.

Application 2: Gaussian sketches

Gaussian sketches (G. and Mahoney 2013)

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

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

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_{\mathrm{F}} \leq \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}} + \sqrt{\epsilon\|\mathbf{A} - \mathbf{A}_k\|_2 \operatorname{tr}(\mathbf{A} - \mathbf{A}_k)} + \sqrt{\frac{\epsilon}{k}}\operatorname{tr}(\mathbf{A} - \mathbf{A}_k), \text{ and}$$

$$\operatorname{tr}(\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}) \leq (1+\epsilon)\operatorname{tr}(\mathbf{A} - \mathbf{A}_k).$$

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

This and similarly improved bounds for the other SPSD sketches follow from our deterministic bounds and analyses of the sketching interaction matrices available for the different choices of $\mathbf{S}$.

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

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

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

Corollaries:

- $\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T} \succeq \mathbf{0}$.
- $\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T} \approx \mathbf{A}_k$ when the range of $\mathbf{A}^{1/2}\mathbf{S}$ is close to the range of the dominant k-dimensional eigenspace of $\mathbf{A}^{1/2}$, spanned by $\mathbf{U}_1$.

Deterministic 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 let

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

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

The errors of SPSD sketches are given by:

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

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

The latter expression is similar to the approximation error $\|\mathbf{A} - \mathbf{P}_{\mathbf{AS}}\mathbf{A}\|_{\xi}$, for which deterministic error bounds are available.

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

If $\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}$ 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 $\xi \in \{2, \mathrm{F}\}$.

By extending the perturbation arguments of (Halko et al. 2011),

Simplified error bounds from (G. and Mahoney 2013)

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

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

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

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

Geometrical interpretation (when $\mathbf{S}$ has orthogonal columns):

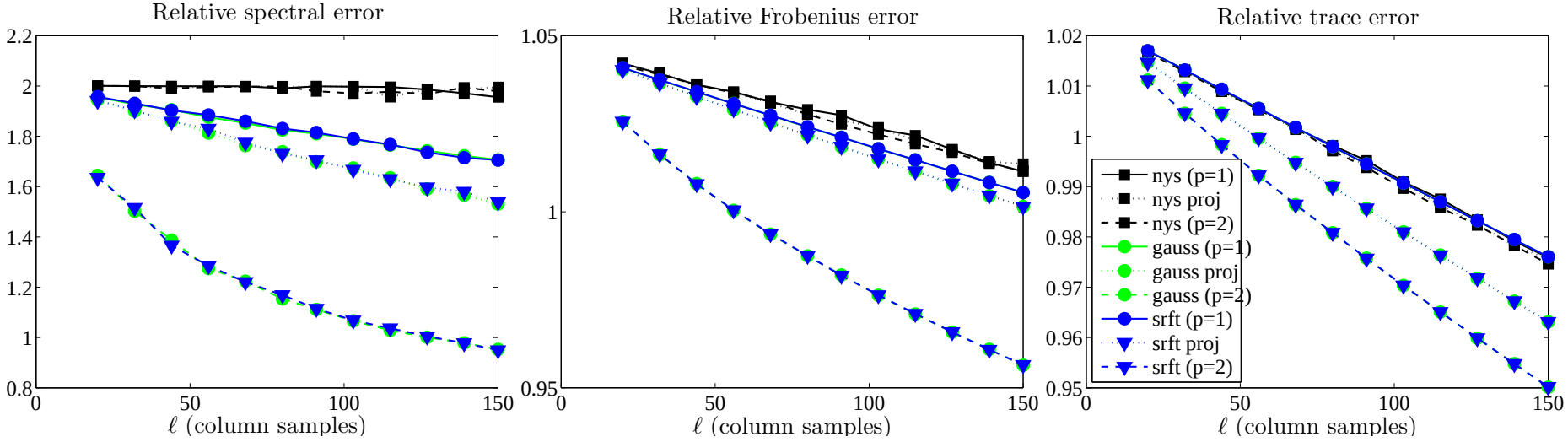
- $\mathbf{U}_1^T\mathbf{S}$ has full row-rank $\Leftrightarrow \tan(\mathbf{U}_1, \mathbf{S}) \neq \infty$.
- $\left\|\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\right\|_2 = \tan(\mathbf{U}_1, \mathbf{S})$.

Note that the randomness of $\mathbf{S}$ enters only through the sketching interaction matrix $\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger$.

Random projections vs SPSD sketches

Could approximate SPSD matrices with $\mathbf{P_{AS}AP_{AS}}$, which requires two passes over $\mathbf{A}$.

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



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

Identified and analyzed a class of low-rank SPSD approximations generalizing the Nyström extension.

- Provided deterministic error bounds and theoretical error guarantees for several types of randomized 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.
- Demonstrated empirically that two-pass SPSD sketches have lower approximation error than the projection-based approximations considered in (Halko et al. 2011).

Paper ID: 785.

Additional 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). Submitted.
- Nyström Bestiary (Matlab code for experiments with SPSD sketches) `http://users.cms.caltech.edu/~gittens/nystrombestiary/`