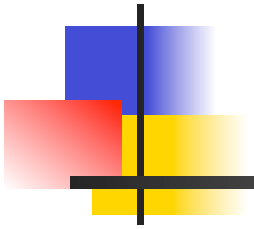
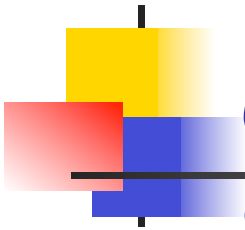


Revisiting the Nyström Method for Improved Large-scale Machine Learning

Michael W. Mahoney

Stanford University
Dept. of Mathematics
http://cs.stanford.edu/people/mmahoney

April 2013



Overview of main results

Gittens and Mahoney (2013)

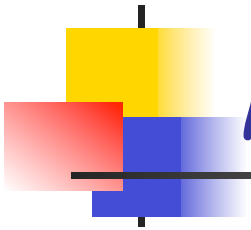
Detailed empirical evaluation:

- On a wide range of SPSD matrices from ML and data analysis
- Considered both random projections and random sampling
- Considered both running time and reconstruction quality
- Many tradeoffs, but prior existing theory was *extremely* weak

Qualitatively-improved theoretical results:

- For spectral, Frobenius, and trace norm reconstruction error
- Structural results (decoupling randomness from the vector space structure) and algorithmic results (for both sampling and projections)

Points to many future extensions (theory, ML, and implementational) ...



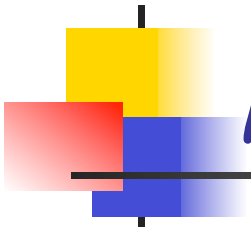
Motivation (1 of 2)

Methods to extract linear structure from the data:

- Support Vector Machines (SVMs).
- Gaussian Processes (GPs).
- Singular Value Decomposition (SVD) and the related PCA.

Kernel-based learning methods to extract non-linear structure:

- Choose features to define a (dot product) space F.
- Map the data, X, to F by ϕ: X$\rightarrow$F.
- Do classification, regression, and clustering in F with linear methods.



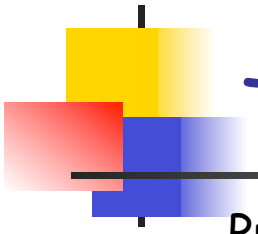
Motivation (2 of 2)

- Use dot products for information about mutual positions.
- Define the kernel or Gram matrix: $G_{ij}=k_{ij}=(\phi(X^{(i)}), \phi(X^{(j)}))$.
- Algorithms that are expressed in terms of dot products can be given the Gram matrix G instead of the data covariance matrix X^TX.

If the Gram matrix G -- $G_{ij}=k_{ij}=(\phi(X^{(i)}), \phi(X^{(j)}))$ -- is dense but (nearly) low-rank, then calculations of interest still need $O(n^2)$ space and $O(n^3)$ time:

- matrix inversion in GP prediction,
- quadratic programming problems in SVMs,
- computation of eigendecomposition of G.

Idea: use random sampling/projections to speed up these computations!



This "revisiting" is particularly timely …

Prior existing theory was *extremely* weak:

- Especially compared with very strong 1±ε results for low-rank approximation, least-squares approximation, etc. of general matrices
- In spite of the empirical success of Nystrom-based and related randomized low-rank methods

Conflicting claims about uniform versus leverage-based sampling:

- Some claim "ML matrices have low coherence" based on one ML paper
- Contrasts with proven importance of leverage scores is genetics, astronomy, and internet applications

High-quality numerical implementations of random projection and random sampling algorithms now exist:

- For L2 regression, L1 regression, low-rank matrix approximation, etc. in RAM, parallel environments, distributed environments, etc.

$$\begin{pmatrix} G \end{pmatrix} \approx \begin{pmatrix} \tilde{G} \end{pmatrix} = \begin{pmatrix} C \end{pmatrix} \begin{pmatrix} W \end{pmatrix}^{+} \begin{pmatrix} C^T \end{pmatrix}$$

Some basics

Leverage scores:

- Diagonal elements of projection matrix onto the best rank-k space
- Key structural property needed to get 1±ε approximation of general matrices

Spectral, Frobenius, and Trace norms:

- Matrix norms that equal {∞,2,1}-norm on the vector of singular values

$$||\mathbf{A}||_2 \leq ||\mathbf{A}||_F \leq ||\mathbf{A}||_* \leq \sqrt{n}||\mathbf{A}||_F \leq n||\mathbf{A}||_2$$

Basic SPSD Sketching Model:

- *SPSD Sketching Model.* Let $\mathbf{A}$ be an $n \times n$ positive semi-definite matrix, and let $\mathbf{S}$ be a matrix of size $n \times \ell$, where $\ell \ll n$. Take

$$\mathbf{C} = \mathbf{AS} \quad \text{and } \mathbf{W} = \mathbf{S^T AS}.$$

Then $\mathbf{CW^+C^T}$ is a low-rank approximation to $\mathbf{A}$ with rank at most ℓ.

Data considered (1 of 2)

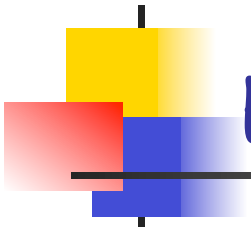
Name	Description	n	d	%nnz
	Laplacians			
HEP	arXiv High Energy Physics collaboration graph	9877	NA	0.06
GR	arXiv General Relativity collaboration graph	5242	NA	0.12
Enron	subgraph of the Enron email graph	10000	NA	0.22
Gnutella	Gnutella peer to peer network on Aug. 6, 2002	8717	NA	0.09
	Linear Kernels			
Dexter	bag of words	2000	20000	83.8
Protein	derived feature matrix for S. cerevisiae	6621	357	99.7
SNPs	DNA microarray data from cancer patients	5520	43	100
Gisette	images of handwritten digits	6000	5000	100
	Dense RBF Kernels			
AbaloneD	physical measurements of abalones	4177	8	100
WineD	chemical measurements of wine	4898	12	100
	Sparse RBF Kernels			
AbaloneS	physical measurements of abalones	4177	8	82.9/48.1
WineS	chemical measurements of wine	4898	12	11.1/88.0

Table 1: The data sets used in our empirical evaluation. The %nnz for the Sparse RBF Kernels depends on the σ parameter.

Data considered (2 of 2)

Name	%nnz	$\lceil \frac{\|\|\mathbf{A}\|\|_F^2}{\|\|\mathbf{A}\|\|_2^2} \rceil$	k	$\frac{\lambda_{k+1}}{\lambda_k}$	$100\frac{\|\|\mathbf{A}-\mathbf{A}_k\|\|_F}{\|\|\mathbf{A}\|\|_F}$	$100\frac{\|\|\mathbf{A}-\mathbf{A}_k\|\|_*}{\|\|\mathbf{A}\|\|_*}$	k^{th}-lrgst lev
HEP	0.06	3078	20	0.998	7.8	0.4	0.261
HEP	0.06	3078	60	0.998	13.2	1.1	0.278
GR	0.12	1679	20	0.999	10.5	0.74	0.286
GR	0.12	1679	60	1	17.9	2.16	0.289
Enron	0.22	2588	20	0.997	7.77	0.352	0.492
Enron	0.22	2588	60	0.999	12.0	0.94	0.298
Gnutella	0.09	2757	20	1	8.1	0.41	0.381
Gnutella	0.09	2757	60	0.999	13.7	1.20	0.340
Dexter	83.8	176	8	0.963	14.5	.934	0.067
Protein	99.7	24	10	0.987	42.6	7.66	0.008
SNPs	100	3	5	0.928	85.5	37.6	0.002
Gisette	100	4	12	0.90	90.1	14.6	0.005
AbaloneD (dense, $\sigma = .15$)	100	41	20	0.992	42.1	3.21	0.087
AbaloneD (dense, $\sigma = 1$)	100	4	20	0.935	97.8	59	0.012
WineD (dense, $\sigma = 1$)	100	31	20	0.99	43.1	3.89	0.107
WineD (dense, $\sigma = 2.1$)	100	3	20	0.936	94.8	31.2	0.009
AbaloneS (sparse, $\sigma = .15$)	82.9	400	20	0.989	15.4	1.06	0.232
AbaloneS (sparse, $\sigma = 1$)	48.1	5	20	0.982	90.6	21.8	0.017
WineS (sparse, $\sigma = 1$)	11.1	116	20	0.995	29.5	2.29	0.2
WineS (sparse, $\sigma = 2.1$)	88.0	39	20	0.992	41.6	3.53	0.098

Table 1: Summary statistics for data sets used in our empirical evaluation.



Effects of "Preprocessing" Decisions

Whitening the input data:

- (mean centering, normalizing variances, etc. to put data points on same scale)
- Tends to homogenize the leverage scores (a little, for fixed rank parameter k)
- Tends to decrease the effective rank & to decrease the spectral gap

Increasing the rank parameter k:

- (leverage scores are defined relative to a given k)
- Tends to uniformize the leverage scores (usually a little, sometimes a lot, but sometimes it increases their nonuniformity)

Increasing the rbf σ scale parameter:

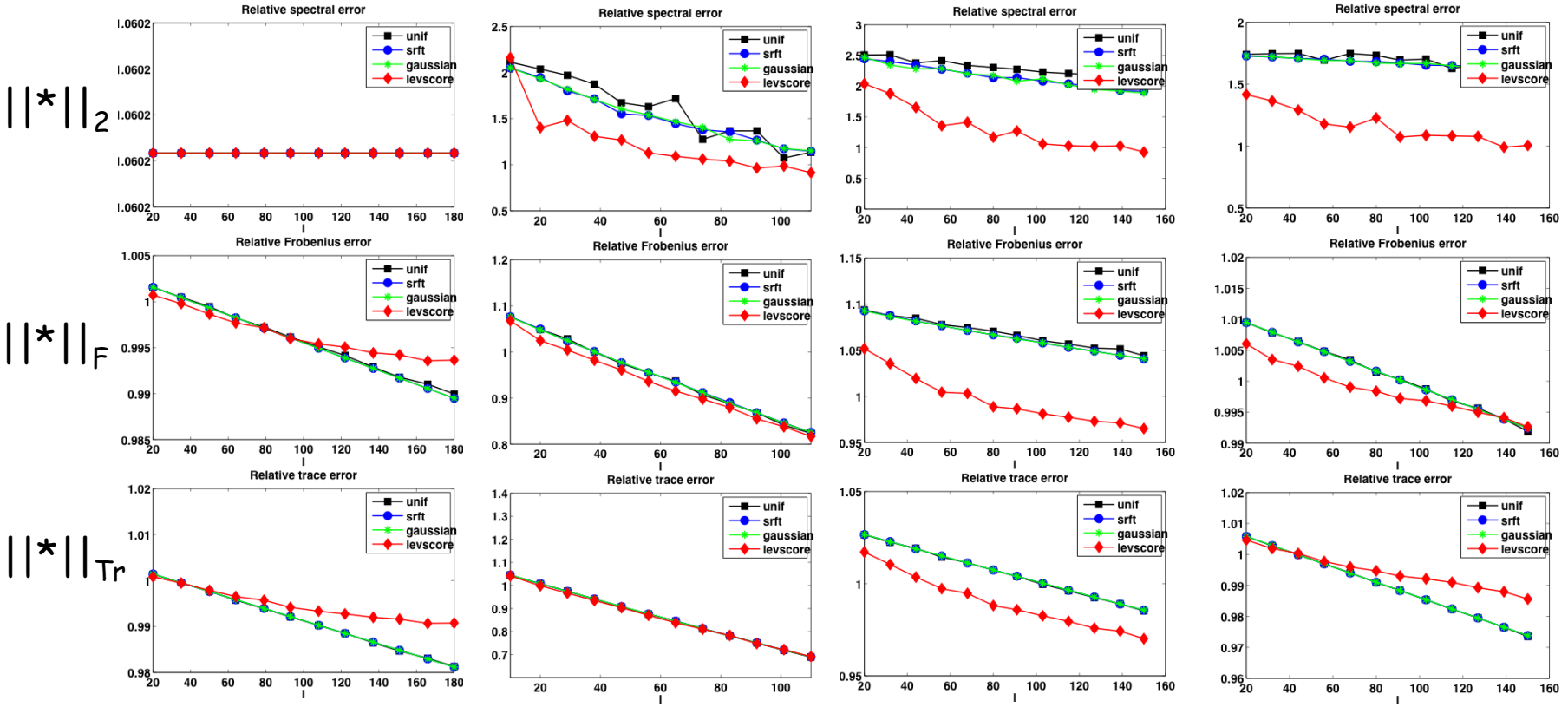
- (defines "size scale" over which a data point sees other data points)
- Tends to uniformize the leverage scores

Zeroing our small matrix entries:

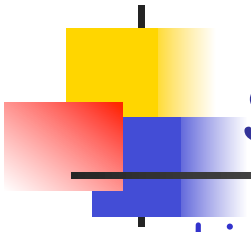
- (replace dense n x n SPSD matrix with a similar sparse matrix)
- Tends to increase effective rank & make leverage scores more nonuniform

Examples of reconstruction error for sampling and projection algorithms

Gittens and Mahoney (2013)



HEP, k=20; Protein k=10; AbaloneD(σ=.15,k=20); AbaloneS(σ=.15,k=20)



Summary of Sampling versus Projection

Linear Kernels & Dense RBF Kernels with larger σ:

- have relatively low rank and relatively uniform leverage scores
- correspond most closely to what is usually studied in ML

Sparsifying RBF Kernels &/or choosing smaller σ :

- tends to make data less low-rank and more heterogeneous leverage scores

Dense RBF Kernels with smaller σ & sparse RBF Kernels:

- leverage score sampling tends to do better than other methods
- Sparse RBF Kernels have many properties of sparse Laplacians corresponding to unstructured social graphs

Choosing more samples l in the approximation:

- Reconstruction quality saturates with leverage score sampling

Restricting the rank of the approximation:

- Rank-restricted approximations (like Tikhonov, not ridge-based) are choppier as a function of increasing l

All methods perform *much* better than theory would suggest!

Approximating the leverage scores (1 of 2, for very rectangular matrices)

Drineas, Magdon-Ismail, Mahoney, and Woodruff (2012)

Input: $\mathbf{A} \in R^{n \times d}$ (with $n \gg d$ and SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$), and $\epsilon \in (0, 1/2]$.

- Let $\mathbf{\Pi}_1 \in R^{r_1 \times n}$ be an SRFT with $r_1 = \Omega(\epsilon^{-2}(\sqrt{d} + \sqrt{\ln n})^2 \ln d)$.
- Compute $\mathbf{\Pi}_1\mathbf{A} \in R^{r_1 \times d}$ and its QR factorization $\mathbf{\Pi}_1\mathbf{A} = \mathbf{QR}$.
- Let $\mathbf{\Pi}_2 \in R^{d \times r_2}$ be a matrix of i.i.d. standard Gaussian random variables, where $r_2 = \Omega\left(\epsilon^{-2} \ln n\right)$.
- Construct the product $\mathbf{\Omega} = \mathbf{AR}^{-1}\mathbf{\Pi}_2$.
- For $i = 1, \ldots, n$ compute $\tilde{\ell}_i = ||\Omega_{(i)}||_2^2$.

Output: $\tilde{\ell}_i, i = 1, \ldots, n$, approximations to the leverage scores of $\mathbf{A}$.

• This algorithm returns relative-error (1±ε) approximations to all the leverage scores of an arbitrary tall matrix in time

$$\mathrm{O}(nd \ln(\sqrt{d} + \sqrt{\ln n}) + nd\epsilon^{-2} \ln n + d^2\epsilon^{-2}(\sqrt{d} + \sqrt{\ln n})^2 \ln d).$$

Approximating the leverage scores (2 of 2, for general matrices)

Drineas, Magdon-Ismail, Mahoney, and Woodruff (2012)

Input: $\mathbf{A} \in R^{n \times d}$, a rank parameter k, and an error parameter $\epsilon \in (0, 1/2]$.

- Construct $\mathbf{\Pi} \in R^{d \times 2k}$ with i.i.d. standard Gaussian entries.
- Compute $\mathbf{B} = \left(\mathbf{A}\mathbf{A}^T\right)^q \mathbf{A}\mathbf{\Pi} \in R^{n \times 2k}$ with

$$q \geq \left\lceil \frac{\ln\left(1 + \sqrt{\frac{k}{k-1}} + e\sqrt{\frac{2}{k}}\sqrt{\min\{n,d\} - k}\right)}{2\ln\left(1 + \epsilon/10\right) - 1/2} \right\rceil,$$

- Approximate the leverage scores of $\mathbf{B}$ by calling the "rectangular" algorithm with inputs $\mathbf{B}$ and ϵ; let $\hat{\ell}_i$ for $i = 1, \ldots, n$ be the outputs.

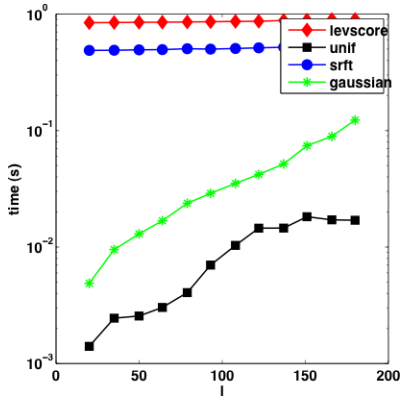
Output: $\hat{\ell}_i, i = 1, \ldots, n$, approximations to the leverage scores of $\mathbf{A}$ filtered through its dominant dimension-k subspace.

• Output is relative-error (1±ε) approximation to all leverage scores of an arbitrary matrix (i.e., the leverage scores of a nearby--in Frobenius norm, q=0, or spectral norm, q>0--matrix) in time $O(ndkq) + T_{\text{RECTANGULAR}}$.

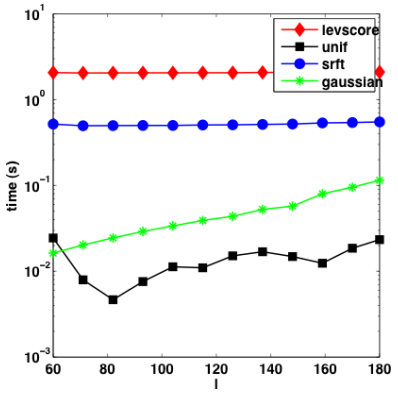
Examples of running times for SLOW low-rank SPSD approximations

Gittens and Mahoney (2013)

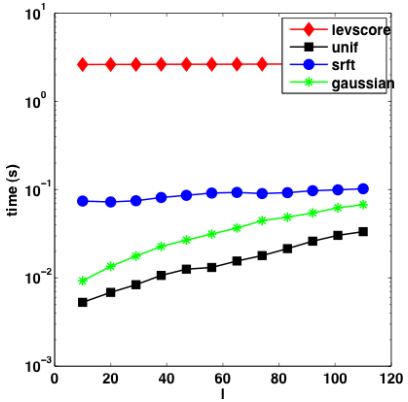
Time



GR(k=20)

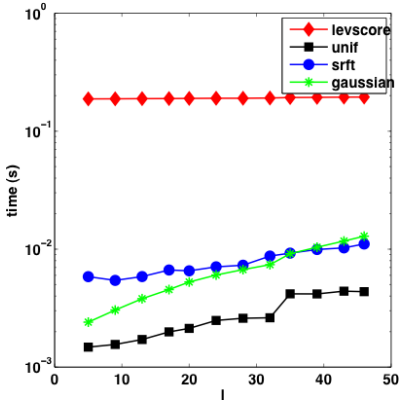


GR(k=60)

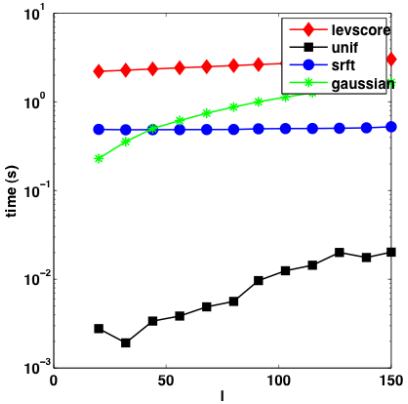


Protein(k=10)

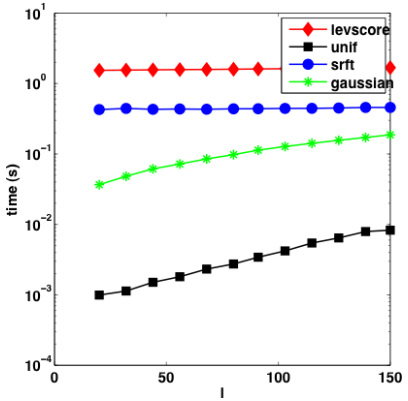
Time



SNPs(k=5)



AbaloneD(σ=.15,k=20)

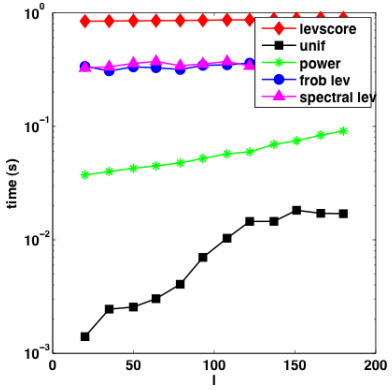


AbaloneS(σ=.15,k=20)

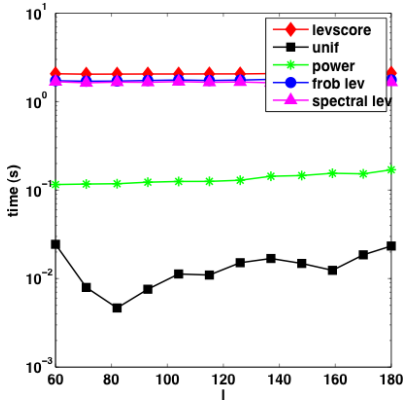
Examples of running times for FAST low-rank SPSD approximations

Gittens and Mahoney (2013)

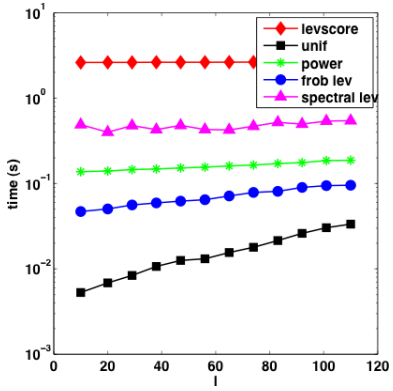
Time



GR(k=20)

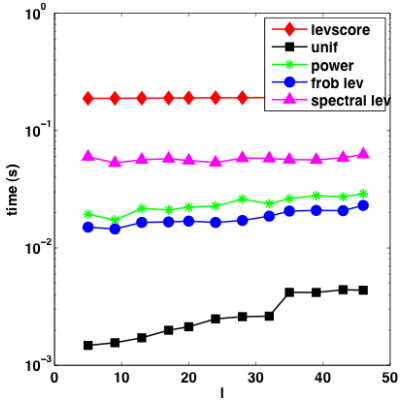


GR(k=60)

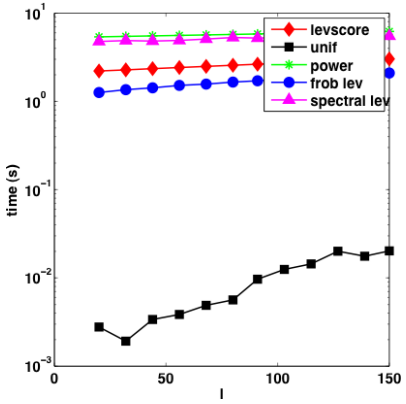


Protein(k=10)

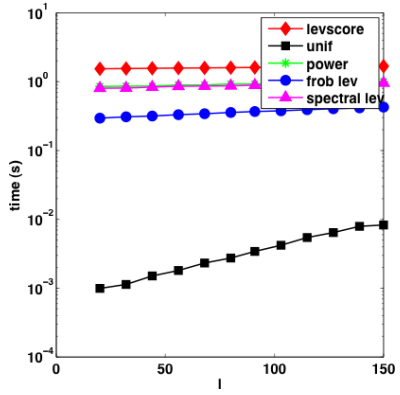
Time



SNPs(k=5)



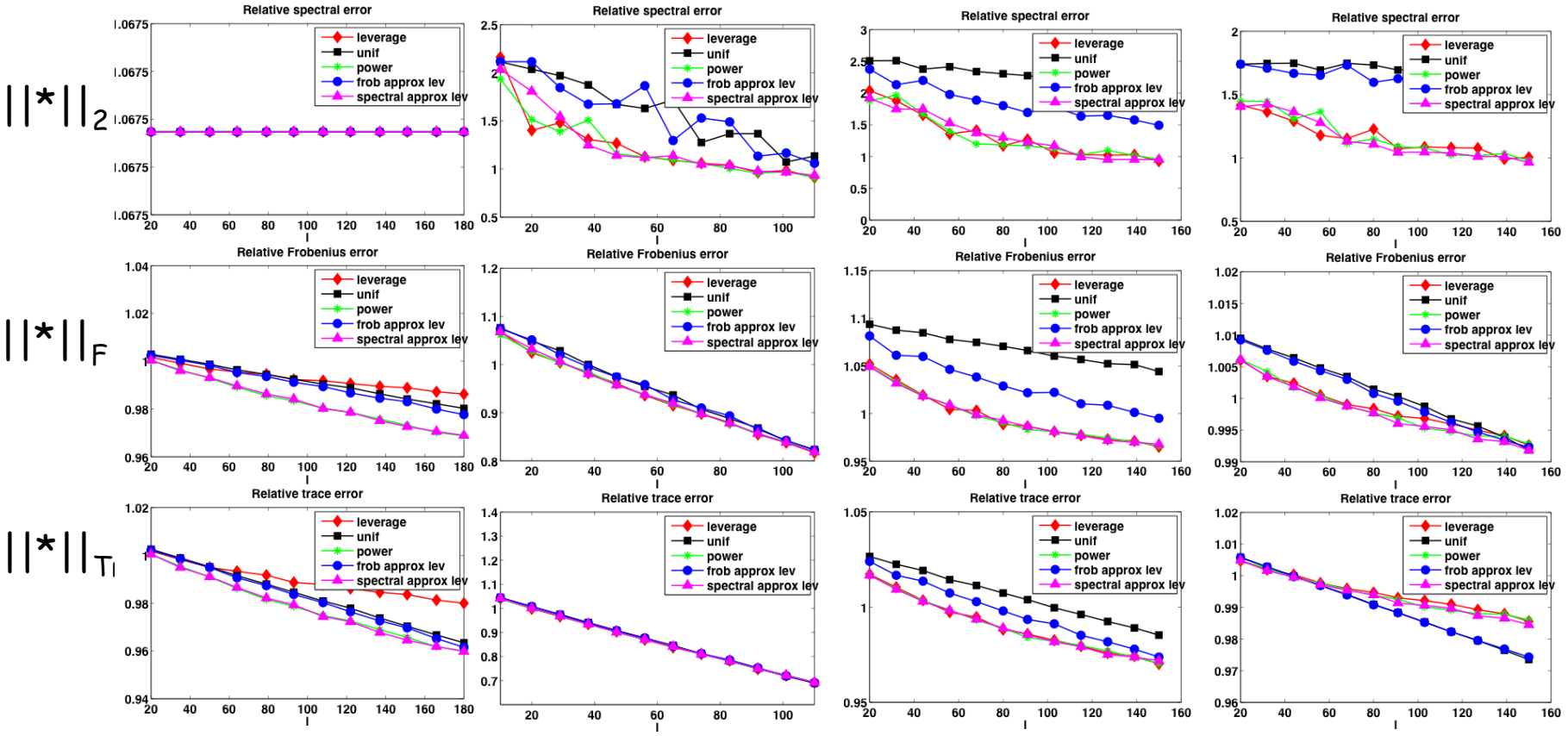
AbaloneD(σ=.15,k=20)



AbaloneS(σ=.15,k=20)

Examples of reconstruction error for FAST low-rank SPSD approximations

Gittens and Mahoney (2013)



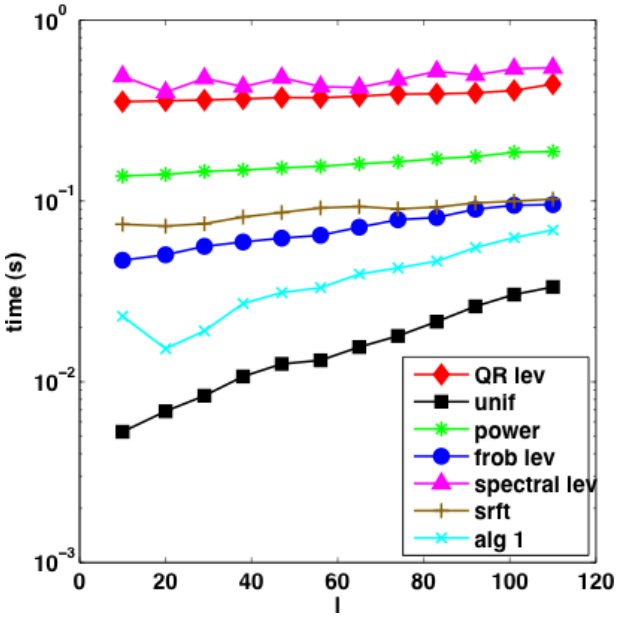
HEP, k=20; Protein k=10; AbaloneD(σ=.15,k=20); AbaloneS(σ=.15,k=20)

An aside: Timing for fast approximating leverage scores of rectangular matrices

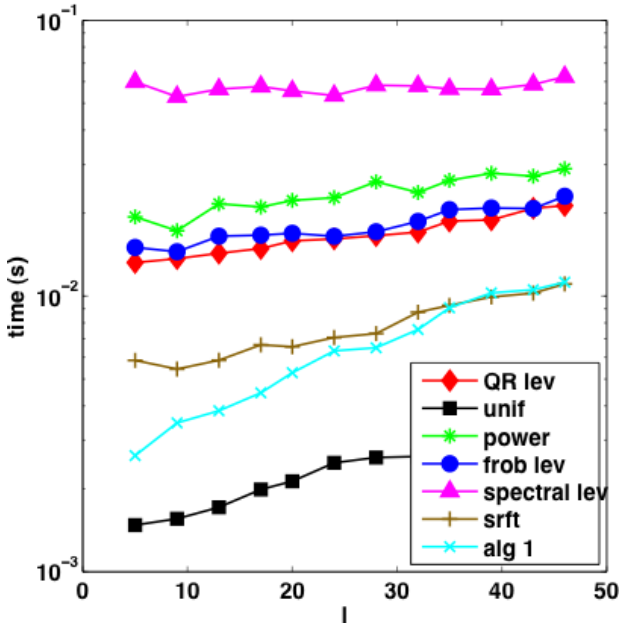
Gittens and Mahoney (2013)

Running time is comparable to underlying random projection

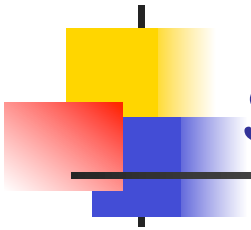
• (Can solve the subproblem directly; or, as with Blendenpik, use it to precondition to solve LS problems of size ≥ thousands-by-hundreds faster than LAPACK.)



Protein k=10;



SNPs(k=5)



Summary of running time issues

Running time of exact leverage scores:

- worse than uniform sampling, SRFT-based, & Gaussian-based projections

Running time of approximate leverage scores:

- can be much faster than exact computation
- with $q=0$ iterations, time comparable to SRFT or Gaussian projection time
- with $q>0$ iterations, time depends on details of stopping condition

The leverage scores:

- with $q=0$ iterations, the actual leverage scores are poorly approximated
- with $q>0$ iterations, the actual leverage scores are better approximated
- reconstruction quality is often no worse, and is often *better*, when using approximate leverage scores

On "tall" matrices:

- running time is comparable to underlying random projection
- can use the coordinate-biased sketch thereby obtained as preconditioner for overconstrained L2 regression, as with Blendenpik or LSRN

Weakness of previous theory (1 of 2)

Drineas and Mahoney (COLT 2005, JMLR 2005):

- If sample $\Omega(k\,\varepsilon^{-4}\log(1/\delta))$ columns according to diagonal elements of A, then

$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^{+}\mathbf{C}^{T}\|_{2,F}\leq\|\mathbf{A}-\mathbf{A}_{k}\|_{2,F}+\varepsilon\sum_{k=1}^{n}(\mathbf{A})_{ii}^{2}$$

Kumar, Mohri, and Talwalker (ICML 2009, JMLR 2012):

- If sample $\Omega(\tau\,k\log(k/\delta))$ columns uniformly, where $\tau \approx$ coherence and A has exactly rank k, then can reconstruct A, i.e.,

$$\mathbf{A}=\mathbf{C}\mathbf{W}^{+}\mathbf{C}^{T}$$

Gittens (arXiv, 2011):

- If sample $\Omega(\mu\,k\log(k/\delta))$ columns uniformly, where μ = coherence, then

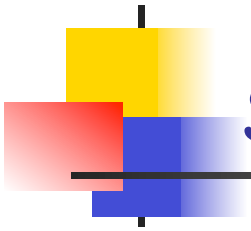
$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^{+}\mathbf{C}^{T}\|_{2}\;\leq\;\|\mathbf{A}-\mathbf{A}_{k}\|_{2}\left(1+\frac{2n}{\ell}\right)$$

$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^{+}\mathbf{C}^{T}\|_{2}\;\leq\;\|\mathbf{A}-\mathbf{A}_{k}\|_{2}+\frac{2}{\delta}\cdot\|\mathbf{A}-\mathbf{A}_{k}\|_{Tr}$$

So weak that these results aren't even a qualitative guide to practice

Weakness of previous theory (2 of 2)

source, sketch	pred./obs. spectral error	pred./obs. Frobenius error	pred./obs. trace error
Protein, $k = 10$			
DM05 nonuniform Nyström	119.2	18.6	–
BW09 uniform Nyström	–	–	3.6
KMT12 uniform Nyström	33.4	20.5	–
GM13 Leverage-based Lemma	42.5	6.9	2.0
GM13 Fourier-based Lemma	297.5	21.7	3.1
GM13 Gaussian-based Lemma	3.8	3.3	1.8
GM13 uniform Nyström Lemma	86.3	91.3	8
AbaloneD, $\sigma = .15, k = 20$			
DM05 nonuniform Nyström	349.9	42.5	–
BW09 uniform Nyström	–	–	2.0
KMT12 uniform Nyström	62.9	46.7	–
GM13 Leverage-based Lemma	235.3	14.6	1.3
GM13 Fourier-based Lemma	139.4	36.9	1.7
GM13 Gaussian-based Lemma	5.2	4.7	1.1
GM13 uniform Nyström Lemma	12.9	228.3	5.1
WineS, $\sigma = 1, k = 20$			
DM05 nonuniform Nyström	422.5	41.0	–
BW09 uniform Nyström	–	–	2.1
KMT12 uniform Nyström	72.8	44.2	–
GM13 Leverage-based Lemma	244.9	13.4	1.2
GM13 Fourier-based Lemma	186.7	36.8	1.7
GM13 Gaussian-based Lemma	6.6	4.7	1.2
GM13 uniform Nyström Lemma	13.7	222.6	5.1



Strategy for improved theory

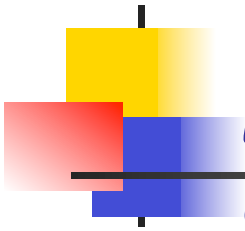
Decouple the randomness from the vector space structure

- This used previously with least-squares and low-rank CSSP approximation

This permits much finer control in the application of randomization

- Much better worst-case theory
- Easier to map to ML and statistical ideas
- Has led to high-quality numerical implementations of LS and low-rank algorithms
- Much easier to parameterize problems in ways that are more natural to numerical analysts, scientific computers, and software developers

This implicitly looks at the "square root" of the SPSD matrix



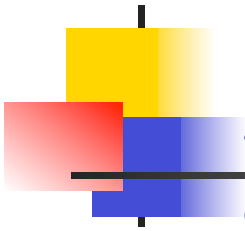
Main structural result

Gittens and Mahoney (2013)

Theorem. Let $\mathbf{A}$ be an $n \times n$ SPSD matrix with eigenvalue decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{U}_1$ is top k eigenvalues, $\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}$ etc., and let $\mathbf{S}$ be a sampling matrix of size $n \times \ell$. Then when $\mathbf{C} = \mathbf{AS}$ and $\mathbf{W} = \mathbf{S}^T\mathbf{AS}$, the corresponding low-rank SPSD approximation satisfies

$$
\begin{aligned}
\|\mathbf{A} - \mathbf{C}\mathbf{W}^+\mathbf{C}^T\|\|_2 &\leq \|\mathbf{\Sigma}_2\|_2 + \|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^+\|_2^2 \\
\|\mathbf{A} - \mathbf{C}\mathbf{W}^+\mathbf{C}^T\|_F &\leq \|\mathbf{\Sigma}_2\|_F + \sqrt{2}\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^+\|_F + \|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^+\|_F^2 \\
\|\mathbf{A} - \mathbf{C}\mathbf{W}^+\mathbf{C}^T\|_{Tr} &\leq \mathrm{Tr}\mathbf{\Sigma}_2 + \|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^+\|_F^2,
\end{aligned}
$$

assuming $\mathbf{\Omega}_1$ has full row rank.



Algorithmic applications (1 of 2)

Gittens and Mahoney (2013)

Lemma. Let $\mathbf{S}$ be a sampling matrix of size $n \times \ell$ corresponding to a leverage-based probability distribution derived from the top k-dimensional eigenspace of $\mathbf{A}$ s.t. for some $\beta \in (0, 1]$. If $\ell \geq 3200(\beta\varepsilon^2)^{-1}k\ln(4k/(\beta\delta))$, then w.p. $1-\delta$ the corresponding low-rank SPSD approximation satisfies

$$\begin{aligned}
\|\mathbf{A} - \mathbf{C}\mathbf{W}^{+}\mathbf{C}^T\|_2 &\leq \|\mathbf{A} - \mathbf{A}_k\|_2 + \varepsilon^2\|\mathbf{A} - \mathbf{A}_k\|_{Tr}, \\
\|\mathbf{A} - \mathbf{C}\mathbf{W}^{+}\mathbf{C}^T\|_F &\leq (1+\sqrt{2}\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F + \varepsilon^2\|\mathbf{A} - \mathbf{A}_k\|_{Tr}, \\
\|\mathbf{A} - \mathbf{C}\mathbf{W}^{+}\mathbf{C}^T\|_{Tr} &\leq (1+\varepsilon^2)\|\mathbf{A} - \mathbf{A}_k\|_{Tr}.
\end{aligned}$$

Similar bounds for uniform sampling, except that need to sample proportional to the coherence (the largest leverage score).

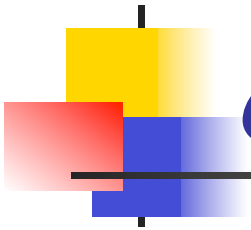
Algorithmic applications (2 of 2)

Gittens and Mahoney (2013)

Lemma. Let $\mathbf{S} = \sqrt{\frac{n}{\ell}}\mathbf{DFR}$ be a structured random projection of size $n \times \ell$. If $\ell \geq 24\varepsilon^{-1}[\sqrt{k} + \sqrt{8\ln(8n/\delta)}]^2 \ln(8k/\delta)$, then w.p. $1-\delta$ the corresponding low-rank SPSD approximation satisfies

$$\begin{aligned}
\|\mathbf{A} - \mathbf{C}\mathbf{W}^+\mathbf{C}^T\|_2 &\leq \left(1 + \frac{1}{1-\sqrt{\varepsilon}} \cdot \left(5 + \frac{16\ln(n/\delta)^2}{\ell}\right)\right)\|\mathbf{A} - \mathbf{A}_k\|_2 \\
&+ \frac{2\ln(n/\delta)}{(1-\sqrt{\varepsilon})\ell}\|\mathbf{A} - \mathbf{A}_k\|_{Tr}, \\
\|\mathbf{A} - \mathbf{C}\mathbf{W}^+\mathbf{C}^T\|_F &\leq (1 + \sqrt{44\varepsilon})\|\mathbf{A} - \mathbf{A}_k\|_F + 22\varepsilon\|\mathbf{A} - \mathbf{A}_k\|_{Tr}, \\
|\mathbf{A} - \mathbf{C}\mathbf{W}^+\mathbf{C}^T\|_{Tr} &\leq (1 + 22\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{Tr}.
\end{aligned}$$

Similar bounds for Gaussian-based random projections.



Conclusions ...

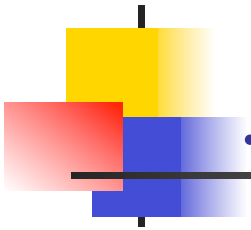
Detailed empirical evaluation:

- On a wide range of SPSD matrices from ML and data analysis
- Considered both random projections and random sampling
- Considered both running time and reconstruction quality
- Many tradeoffs, but prior existing theory was *extremely* weak

Qualitatively-improved theoretical results:

- For spectral, Frobenius, and trace norm reconstruction error
- Structural results (decoupling randomness from the vector space structure) and algorithmic results (for both sampling and projections)

Points to many (theory, ML, and implementational) future directions ...



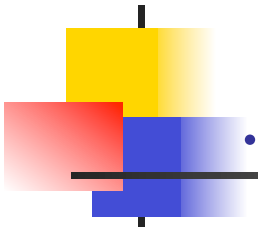
... and Extensions (1 of 2)

More-immediate extension:

Do this on real data 100X or 1000X larger:

- Design the stack to make this possible and relate to related work of Smola et al '13 Fastfood; Rahimi-Recht '07-'08 construction; etc.
- Use Bekas et al '07-'08 "filtering" methods for evaluating matrix functions in DFT and scientific computing
- Focus on robustness and sensitivity issues
- Tighten upper bounds in light of Wang-Zhang-'13 lower bounds
- Extensions of this & related prior work to SVM, CCA, and other ML problems

For software development, concentrate on use cases where theory is well-understood and usefulness has been established.



... and Extensions (2 of 2)

Less-immediate extension:

Relate to recent theory and make it more useful

- Evaluate sparse embedding methods and extend to sparse SPSD matrices
- Apply to solving linear equations (effective resistances are leverage scores)
- Compute the elements of the inverse covariance matrix (localized eigenvectors and implicit regularization)
- Relate to Kumar-Mohri-Talwalkar-'09 Ensemble Nystrom method
- Relate to Bach-'13 use of leverage scores can be used to control generalization