Augmented leverage score sampling

Daniel J. Perry

Scientific Computing and Imaging Institute
University of Utah

September 22, 2016

Joint work with Ross T. Whitaker

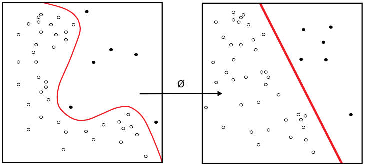
www.sci.utah.edu

Problem statement

Computational expensive analysis of large datasets

- Large n which limits available computation per item
- Desired analysis is computationally expensive for n
- A subset of size $m \ll n$ is representative of the entire dataset

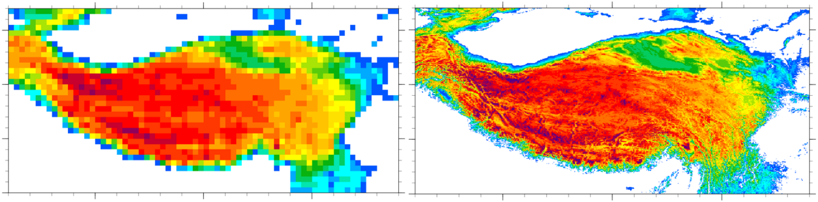
Example: kernel learning on large datasets



- Naive kernel based learning requires kernel (Gram) matrix $\mathbf{A^T A}$
  - $\mathcal{O}(n^2)$ + "analysis cost"
- Nyström approximation to the kernel matrix $\mathbf{A^T A}$
  - Select a subset $\mathbf{C}$ of size $m \ll n$
  - $\mathbf{A^T A}_{i,j} \approx \mathbf{A_i^T C (C^T C)^{-1} C^T A_j}$
  - $\mathcal{O}(m^2)$ + "analysis cost"
- Use on *resource limited machines* requires careful subset selection

[0] image from wikipedia.org

Example: multifidelity simulation



- Uncertainty analysis
  - Random parameter(s) of interest, $\eta \in \mathcal{D}$
  - n samples of $\eta \in \mathcal{D}$ to quantify uncertainty
- Multifidelity uncertainty analysis
  - High fidelity model: only $m \ll n$ simulations
  - Low fidelity model: run all n simulations
  - Representative subset of size m from n for high fidelity simulation
- Careful selection of subset important!

[0] images from www.ncl.ucar.edu

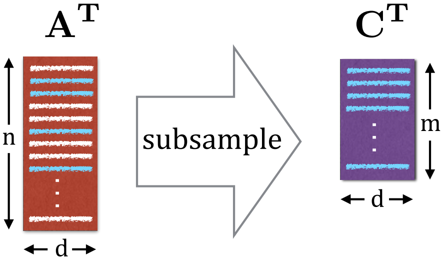
Problem statement

Definition

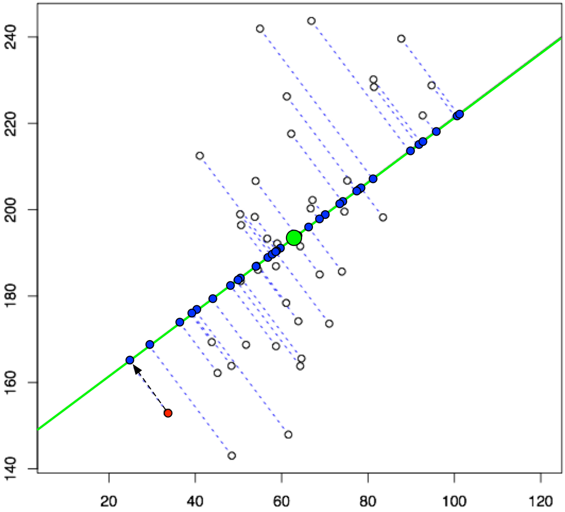
Column Subset Selection Problem. Let $\mathbf{A} \in \mathbb{R}^{d \times n}$. Find $m < n$ columns for $\mathbf{A}$ - denoted as $\mathbf{C} \in \mathbb{R}^{d \times m}$ that minimize

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_{\eta}, \tag{1}$$

for $\eta \in \{F, 2\}$, and where $\mathbf{C}^{\dagger}$ denotes the Moore-Penrose pseudo-inverse.



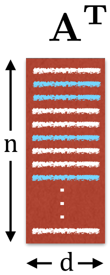
Problem statement: visually



Related work

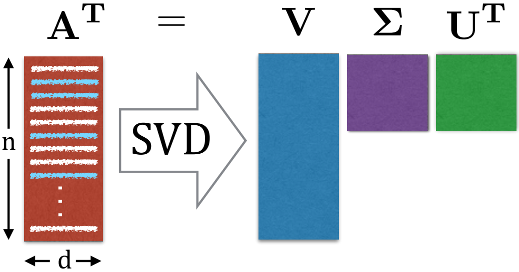
- Leverage sampling [Mahoney, 2011, Mahoney, 2010, Papailiopoulos et al., 2014, Boutsidis et al., 2014]
  - [Papailiopoulos et al., 2014] obtained a very similar bound for deterministic leverage sampling
- Greedy subset selection [Altschuler et al., 2016]
  - obtained a bound in a different form (relative to the optimal subset) for deterministic greedy subset selection
- CUR [Drineas et al., 2008]
  - [Drineas et al., 2008] uses leverage sampling
- Nyström approximation [Drineas and Mahoney, 2005, Gittens and Mahoney, 2013, Gittens, 2011]
  - [Drineas and Mahoney, 2005] uses uniform sampling

Deterministic leverage score sampling



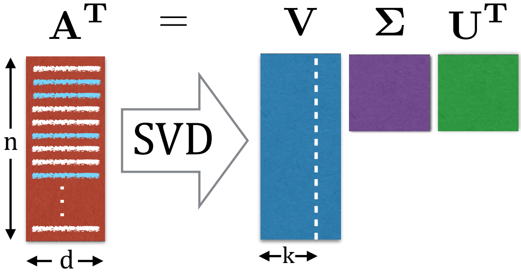
- data matrix $\mathbf{A} \in \mathbb{R}^{d \times n}$

Deterministic leverage score sampling



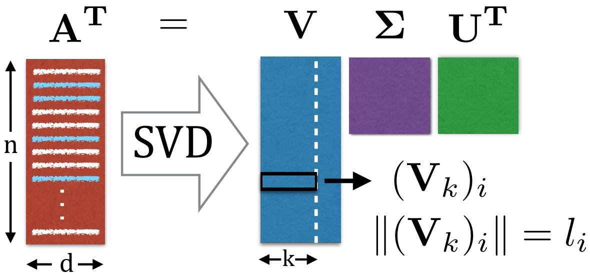
- singular value decomposition

Deterministic leverage score sampling



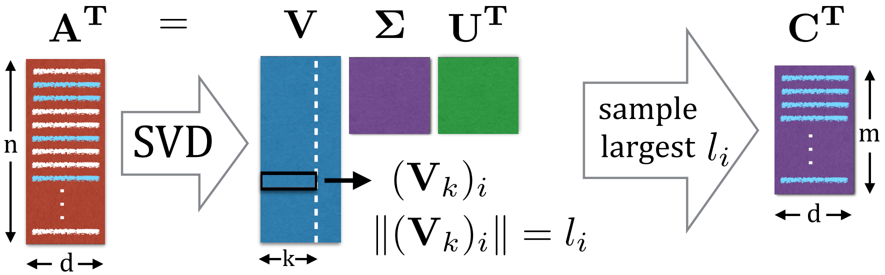
- best rank-k approximation

Deterministic leverage score sampling



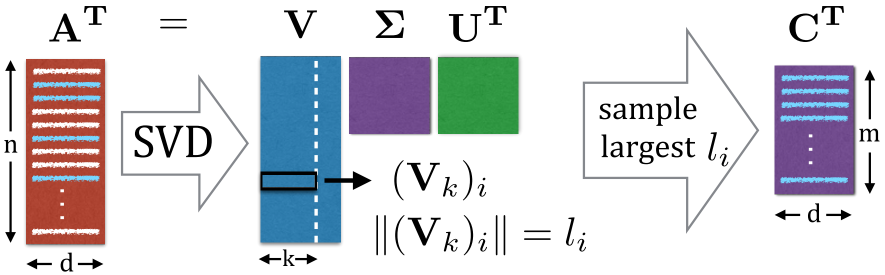
- norm of each row $\mathbf{V_k}$

Deterministic leverage score sampling



- Deterministic samples largest l_i values first

Deterministic leverage score sampling



- Probabilistic version samples with probability l_i

Leverage scores

Definition

Let $\mathbf{V_k} \in \mathbb{R}^{n \times k}$ contain the top k right singular vectors of a $d \times n$ matrix A with rank $\rho = \mathbf{rank}(\mathbf{A}) \geq k$. Then the (rank-k) leverage score of the i-th column of $\mathbf{A}$ is defined as

$$l_i^{(k)} = \|[\mathbf{V_k}]_{i,:}\|_2^2, \;\; i = 1, 2, \ldots, n. \tag{2}$$

Here, $[\mathbf{V_k}]_{i,:}$ denotes the i-th row of $\mathbf{V_k}$.

Leverage score bound

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_{\zeta}^2 < (1 + 2\epsilon) \cdot \|\mathbf{A} - \mathbf{A_k}\|_{\zeta}^2. \tag{3}$$

for $\zeta \in \{2, F\}$, $\epsilon \in (0, .5)$, where $\mathbf{A_k}$ is the best rank-k approximation to $\mathbf{A}$ [Papailiopoulos et al., 2014].

CSSP Objective: data scale also important

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_{\eta}$$
$$\|\mathbf{A} - \mathbf{C}(\mathbf{C}^{\mathbf{T}}(\mathbf{C}\mathbf{C}^{\mathbf{T}})^{-1})\mathbf{A}\|_{\eta}$$
$$\|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}} - \mathbf{W}_{\mathbf{d}}\mathbf{W}_{\mathbf{d}}^{\mathbf{T}}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}\|_{\eta}$$
$$\|(\mathbf{U} - \mathbf{W}_{\mathbf{d}}\mathbf{W}_{\mathbf{d}}^{\mathbf{T}}\mathbf{U})\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}\|_{\eta}$$

where $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$ and $\mathbf{C} = \mathbf{W}\boldsymbol{\Psi}\mathbf{H}^{\mathbf{T}}$ are the respective SVDs.

- Indictates the "unwhitened" data points $\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$, not $\mathbf{V}^{\mathbf{T}}$
- This informs the core idea of an *augmented leverage score*

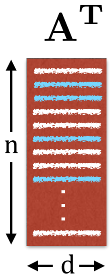
Proposed: augmented leverage score

Definition

Let $\mathbf{Y_k} = \mathbf{V_k \Sigma_k} \in \mathbb{R}^{n \times k}$ contain the top k singular values multiplied with the right singular vectors of a $d \times n$ matrix $\mathbf{A}$ with rank $\rho = \mathbf{rank}(\mathbf{A}) \geq k$. Then the (rank-k) augmented leverage score of the i–th column of $\mathbf{A}$ is defined as

$$\hat{l}_i^{(k)} = \|[\mathbf{Y_k}]_{i,:}\|_2^2, \;\; i = 1, 2, \ldots, n.$$

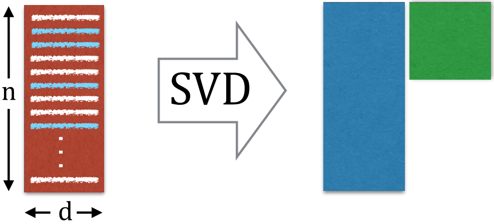
Augmented leverage score sampling



- data matrix $\mathbf{A} \in \mathbb{R}^{d \times n}$

Augmented leverage score sampling

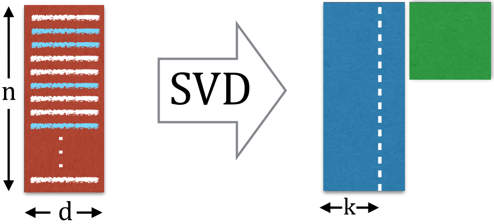
$$\mathbf{A^T} = \mathbf{V\Sigma U^T} = \mathbf{Y^T U^T}$$



- singular value decomposition, combine $\mathbf{V^T\Sigma} = \mathbf{Y^T}$

Augmented leverage score sampling

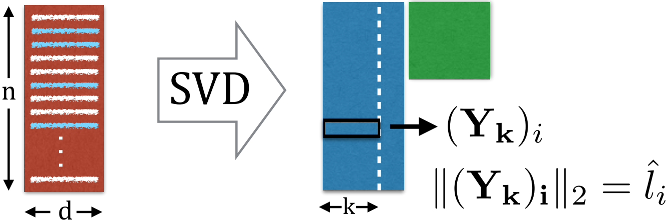
$$\mathbf{A}^{\mathbf{T}} = \mathbf{V\Sigma U}^{\mathbf{T}} = \mathbf{Y}^{\mathbf{T}}\mathbf{U}^{\mathbf{T}}$$



- best rank-k approximation

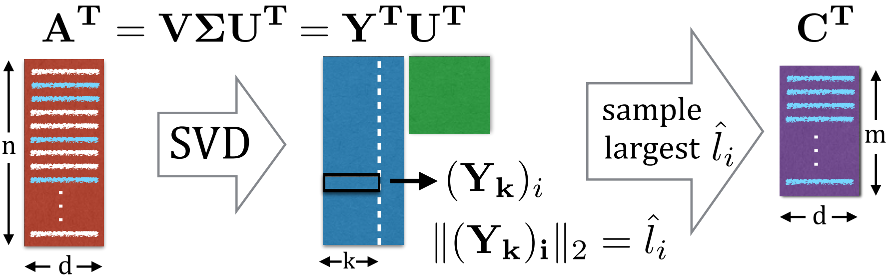
Augmented leverage score sampling

$$\mathbf{A^T} = \mathbf{V \Sigma U^T} = \mathbf{Y^T U^T}$$



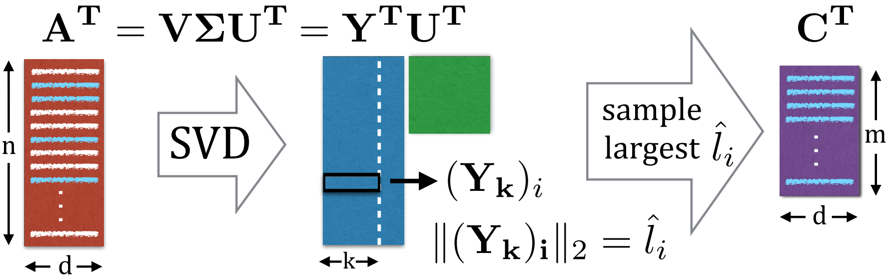
- norm of each row $\mathbf{Y_k}$

Augmented leverage score sampling



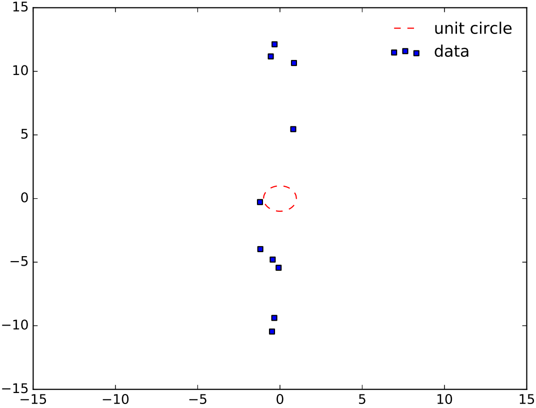
- Augmented samples largest $\hat{l}_i$ values first

Augmented leverage score sampling



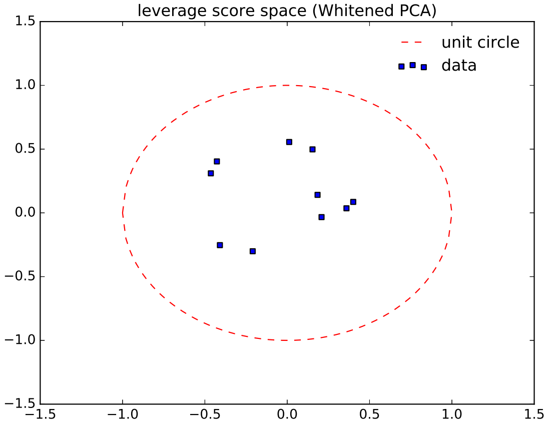
- Probabilistic version samples with probability $\hat{l}_i$

Leverage score: visual example



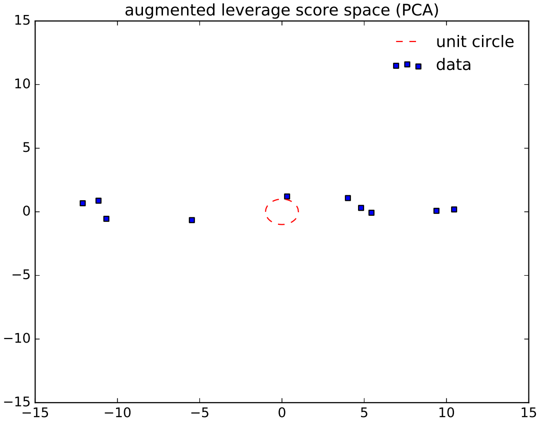
- samples from Gaussian with covariance $\begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$

Leverage score: visual example



- $\mathbf{V}_{\mathbf{k}}^{\mathbf{T}} = \mathbf{\Sigma}_{\mathbf{k}}^{-1}\mathbf{U}_{\mathbf{k}}^{\mathbf{T}}\mathbf{A}$
- Norm sampling based on a whitened rank-k PCA projection

Augmented leverage score: visual example



- $\mathbf{\Sigma_k V_k^T = U_k^T A}$
- Norm sampling based on an (unwhitened) rank-k PCA projection

Augmented Leverage Score Sampling Algorithm

Input $\mathbf{A} \in \mathbb{R}^{d \times n}, k, \theta$
Compute $\mathbf{Y_k} = \mathbf{V_k}\hat{\mathbf{\Sigma}}_\mathbf{k} \in \mathbb{R}^{n \times k}$
for $i = 1, 2, \ldots, n$
    $\hat{l}_i^{(k)} = \|[\mathbf{Y_k}]_{i,:}\|_2^2$
end for
Sort $\hat{l}_i^{(k)}$ **in place**
Find index $m \in \{1, \ldots, n\}$ **such that:**

$$m = \arg\min_m \left(\sum_{i=1}^{m} \hat{l}_i^{(k)} > \theta \right).$$

If $m < k$**, set** $m = k$**.**
Output $\mathbf{S} \in \mathbb{R}^{n \times m}$**, s.t. AS has the top** m **columns of A.**

Bounds of augmented leverage score sampling

Theorem

Let $\theta = k \cdot \frac{\sigma_1^2}{\sigma_k^2} - \epsilon$ *for some* $\epsilon \in (0, 0.5)$, *and let* $\mathbf{S} \in \mathbb{R}^{n \times m}$ *be the sampling matrix from Augmented Leverage Score Sampling Algorithm, then, for* $\mathbf{C} = \mathbf{AS}$ *and* $\zeta = \{2, F\}$

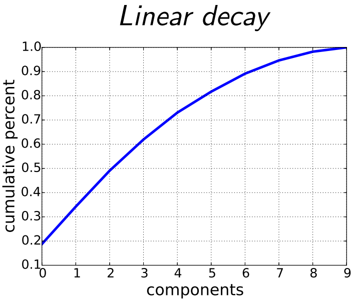
$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_{\zeta}^{2} < \frac{\sigma_1^2}{\sigma_k^2}(1 + 2\epsilon) \cdot \|\mathbf{A} - \mathbf{A_k}\|_{\zeta}^{2}. \tag{4}$$

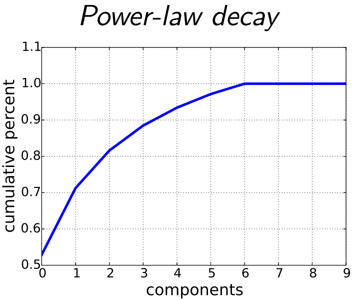
Experiments: synthetic data attributes

name	n	d	k
Synthetic linear decay	1000	10	9
Synthetic power law decay	1000	10	6

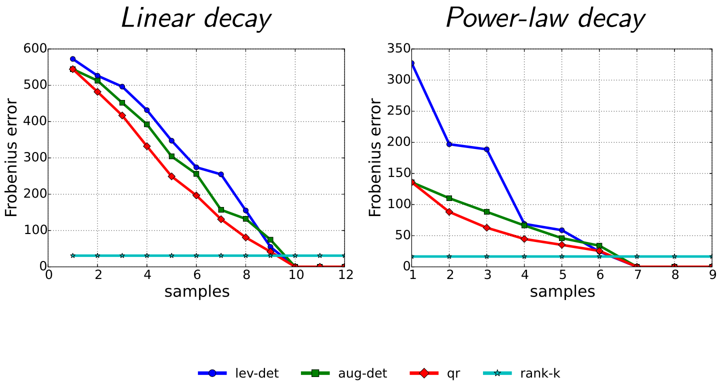
- n - number of columns
- d - dimension of each column
- k - number of PCA dimensions to preserve 90% of spectral energy

Experiments: synthetic data spectra

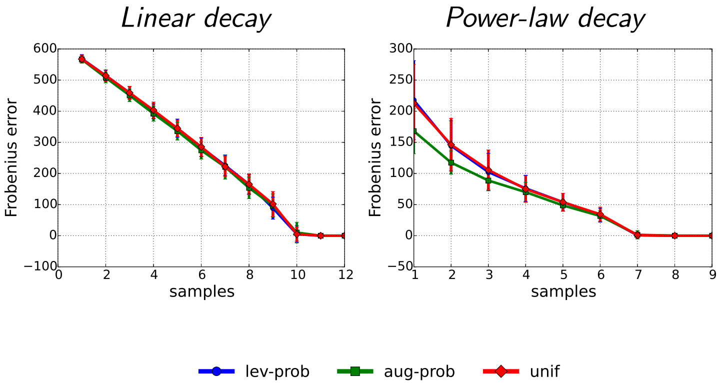




Experiments: synthetic data deterministic results



Experiments: synthetic data probabilistic results

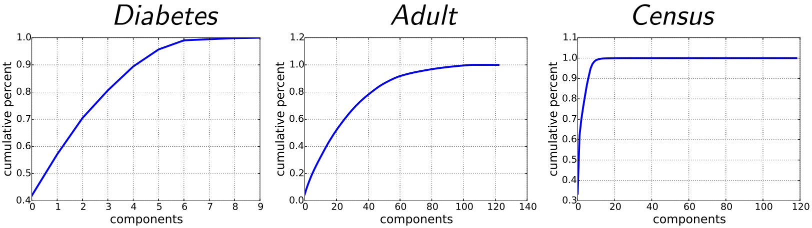


Experiments: real data attributes

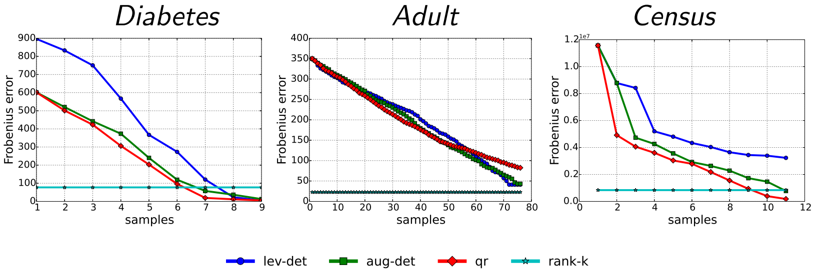
name	n	d	k
Diabetes	442	9	6
Adult	16281	123	73
Census	2273	119	8

- n - number of columns
- d - dimension of each column
- k - number of PCA dimensions to preserve 90% of spectral energy

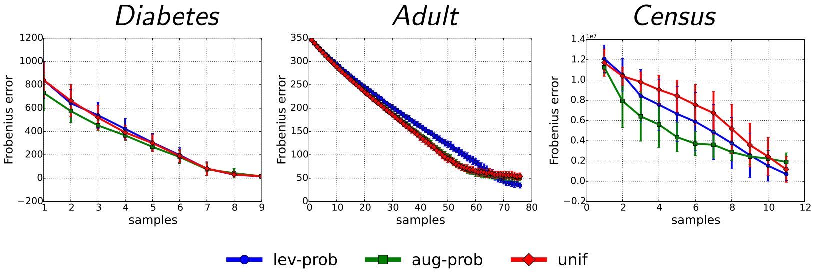
Experiments: real data spectra



Experiments: real data results



Experiments: real data results



Conclusion

- Proposed the *augmented leverage score* motivated by the CSSP objective
- Provided an initial error bound for deterministic augmented leverage score sampling
- Shown empricial results comparing the method on a variety of data sets
  - advantages for data sets with sharp spectral decay
  - shown advantages in both deterministic and probabilistic settings

Acknowledgments

- Thanks to ExxonMobil for funding

Thank you!

Questions?

Bibliography

Altschuler, J., Bhaskara, A., Mirrokni, V., Rostamizadeh, A., Zadimoghaddam, M., et al. (2016).
Greedy column subset selection: New bounds and distributed algorithms.
arXiv preprint arXiv:1605.08795.

Boutsidis, C., Drineas, P., and Magdon-Ismail, M. (2014).
Near-optimal column-based matrix reconstruction.
SIAM Journal on Computing, 10598(i):1–27.

Drineas, P., Mahoney, M., and Muthukrishnan, S. (2008).
relative-error cur matrix decompositions.
SIAM Journal of Matrix Analysis and Applications2, 30(2):844–881.

Drineas, P. and Mahoney, M. W. (2005).
On the Nystrom method for Approximating a gram matrix for improved kernel-based learning (Extended abstract).
Learning Theory, Proceedings, 3559:323–337.