On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning

(Petros Drineas & Michael Mahoney, JMLR, 2005)

Presented by Ameet Talwalkar

Motivation

- Kernel-based algorithms
  - rely on inner-product between data points
  - e.g., SVMs, Kernel PCA, Gaussian Processes
- Introduce non-linearity via PDS kernels
  - $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$
  - $k(x_i, x_j) = k(x_j, x_i)$
- Resulting Gram (Kernel) matrix is positive semidefinite
  - non-negative eigenvalues

Motivation

- Kernel-based algorithms are sometimes costly
  - Kernel PCA: eigendecomposition
  - GP: matrix inversion
  - SVM: $O(n^3)$ instance of Quadratic Programming

- Low-rank approximation of G (dense)
  - approximate eigenvectors/values
  - approximate inverse (matrix inversion lemma)

- Goal: Find low-rank approximation of Gram matrix to improve efficiency of Kernel-based algorithms

Outline

- Terminology
- Basic Idea of Algorithm
- Main Theorem
- Connection to Nyström Method
- Experiments

Terminology – SVD

- Full SVD: $A = U\Sigma V^T,\ A \in \mathbb{R}^{m\,x\,n}$

- Singular Values: $\Sigma \in \mathbb{R}^{m\,x\,n};\ \Sigma = \boldsymbol{diag}(\sigma_1, \ldots, \sigma_\rho)$

  $\rho = min(m, n);\ \sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_\rho \geqslant 0$

- Singular Vectors: $U \in \mathbb{R}^{m\,x\,m},\ U^T U = I_m$

  $V \in \mathbb{R}^{n\,x\,n},\ V^T V = I_n$

- “Best” rank-k Approx: $A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{i=k} \rho_i U^i (V^i)^T$

- Pseudoinverse: $A^+ = V\Sigma^{-1} U^T$

Terminology – Gram Matrix

- Dataset of n points: $X \in \mathbb{R}^{m\,x\,n}$, $X = U\,\Sigma\,V^T$
- Gram matrix: $G \in \mathbb{R}^{n\,x\,n}$, $G = X^T X = V\,\Sigma^2 V^T$
- Partition of G: $G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$
- W: $W \in \mathbb{R}^{c\,x\,c}$, $W = G_{11}$
- C: $C \in \mathbb{R}^{n\,x\,c}$; $C = \begin{bmatrix} W \\ G_{21} \end{bmatrix} = \begin{bmatrix} W & G_{12} \end{bmatrix}^T$

Basic Idea of Algorithm

- From last slide: $G = \begin{bmatrix} W & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$

- Nyström Approximation: $G \approx \tilde{G} = CW^{+}C^{T} = \bar{U}\,\Sigma_{W}\,\bar{U}^{T}$

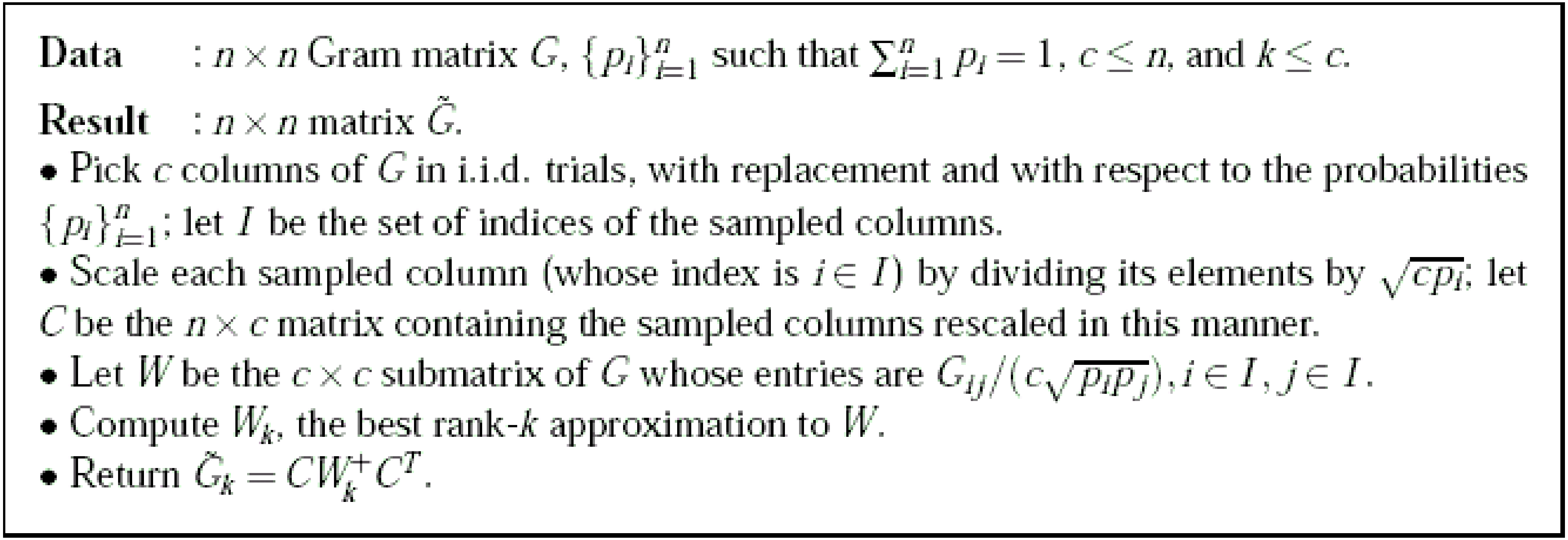
- Multiplying:

$$\tilde{G} = \begin{bmatrix} W \\ G_{21} \end{bmatrix} W^{+} \begin{bmatrix} W & G_{12} \end{bmatrix} = \begin{bmatrix} W & G_{12} \\ G_{21} & G_{21}W^{+}G_{12} \end{bmatrix}$$

- Estimate matrix by:
  - exact decomposition of small piece (W)
  - interpolate by relating sampled points to full dataset

- Runtime: $O(c^{3} + nc^{2} + s)$, where s based on sampling method

Main Algorithm



Data : $n \times n$ Gram matrix G, $\{p_i\}_{i=1}^{n}$ such that $\sum_{i=1}^{n} p_i = 1$, $c \le n$, and $k \le c$.

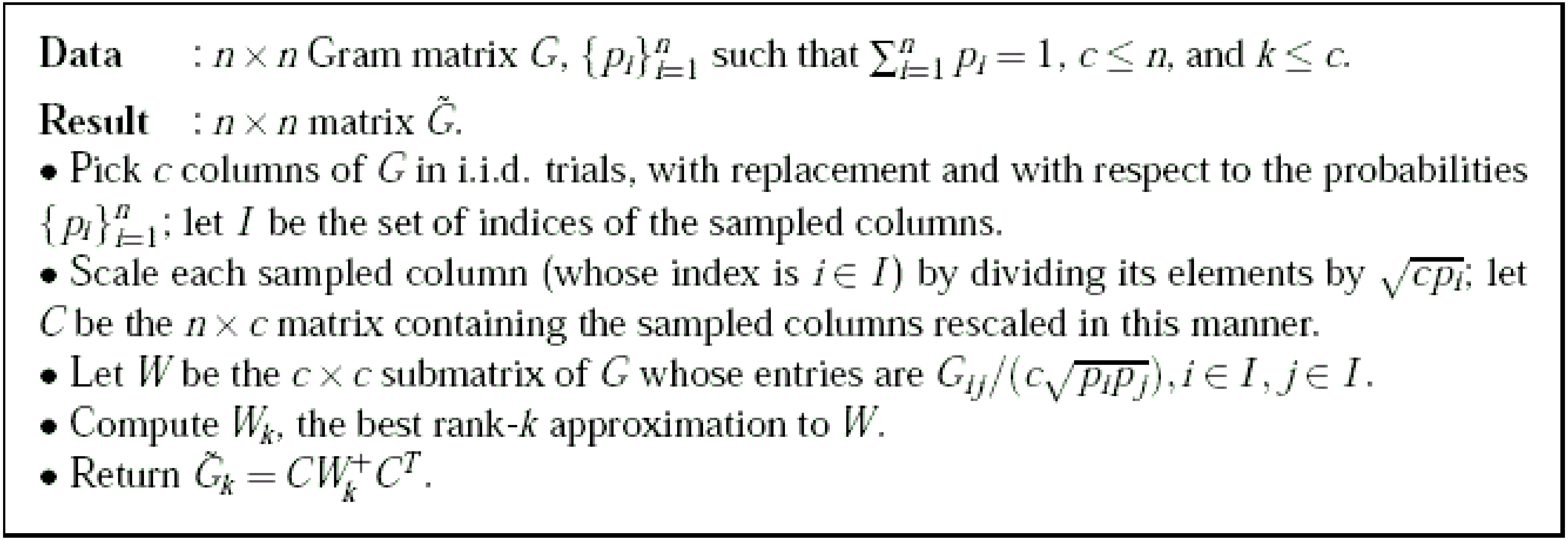
Result : $n \times n$ matrix $\tilde{G}$.

- Pick c columns of G in i.i.d. trials, with replacement and with respect to the probabilities $\{p_i\}_{i=1}^{n}$; let I be the set of indices of the sampled columns.
- Scale each sampled column (whose index is $i \in I$) by dividing its elements by $\sqrt{cp_i}$; let C be the $n \times c$ matrix containing the sampled columns rescaled in this manner.
- Let W be the $c \times c$ submatrix of G whose entries are $G_{ij}/(c\sqrt{p_i p_j}), i \in I, j \in I$.
- Compute W_k, the best rank-k approximation to W.
- Return $\tilde{G}_k = CW_k^{+}C^T$.

- Sampling scheme: $$p_i = \frac{G_{ii}^2}{\sum_{j=1}^{n} G_{jj}^2} = \frac{|X^{(i)}|^2}{\|X\|_F^2}$$
  - proofs rely on decomposing G into $X^T X$ and sampling from X
  - minimizes expected error of approx matrix mult (Frobenius)

- Scaling: makes approx matrix multiplication unbiased

Main Algorithm



Data : $n \times n$ Gram matrix G, $\{p_i\}_{i=1}^{n}$ such that $\sum_{i=1}^{n} p_i = 1$, $c \leq n$, and $k \leq c$.

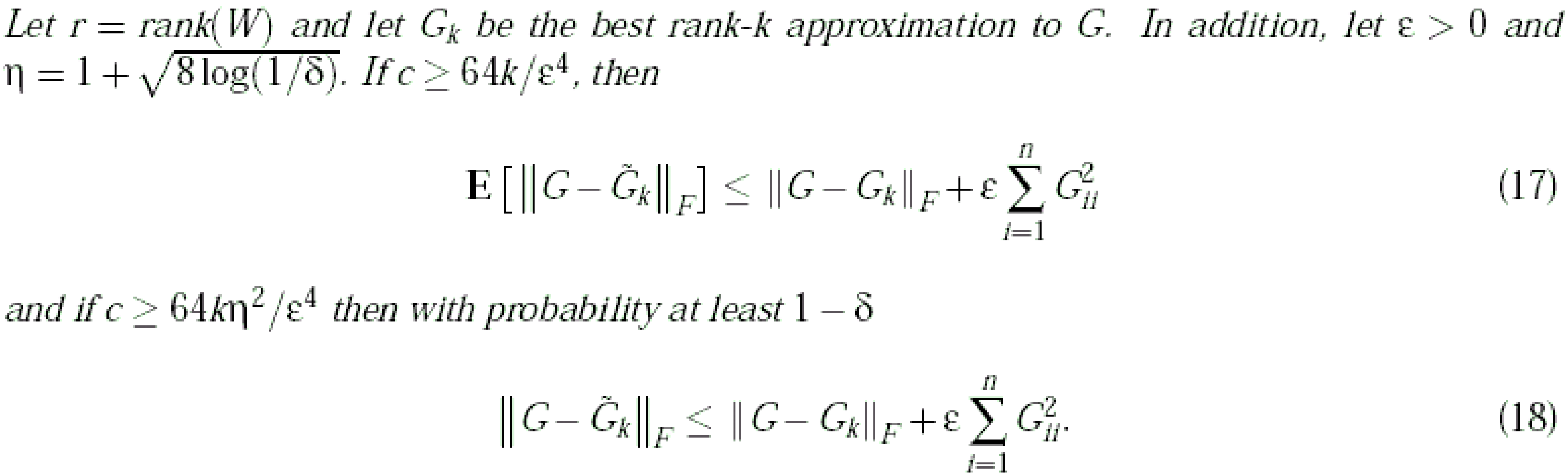
Result : $n \times n$ matrix $\tilde{G}$.

- Pick c columns of G in i.i.d. trials, with replacement and with respect to the probabilities $\{p_i\}_{i=1}^{n}$; let I be the set of indices of the sampled columns.
- Scale each sampled column (whose index is $i \in I$) by dividing its elements by $\sqrt{cp_i}$; let C be the $n \times c$ matrix containing the sampled columns rescaled in this manner.
- Let W be the $c \times c$ submatrix of G whose entries are $G_{ij}/(c\sqrt{p_i p_j}), i \in I, j \in I$.
- Compute W_k, the best rank-k approximation to W.
- Return $\tilde{G}_k = C W_k^{+} C^T$.

- Bounds (in expectation and with high probability):

$$\|G-\tilde{G}_k\|_\xi \leqslant \|G-G_k\|_\xi + \epsilon \sum_{i=1}^{n} G_{ii}^2, \ \xi = 2, F$$

Frobenius Norm Bounds



Let $r = rank(W)$ *and let* G_k *be the best rank-*k *approximation to* G. *In addition, let* $\varepsilon > 0$ *and* $\eta = 1+\sqrt{8\log(1/\delta)}$. *If* $c \geq 64k/\varepsilon^4$*, then*

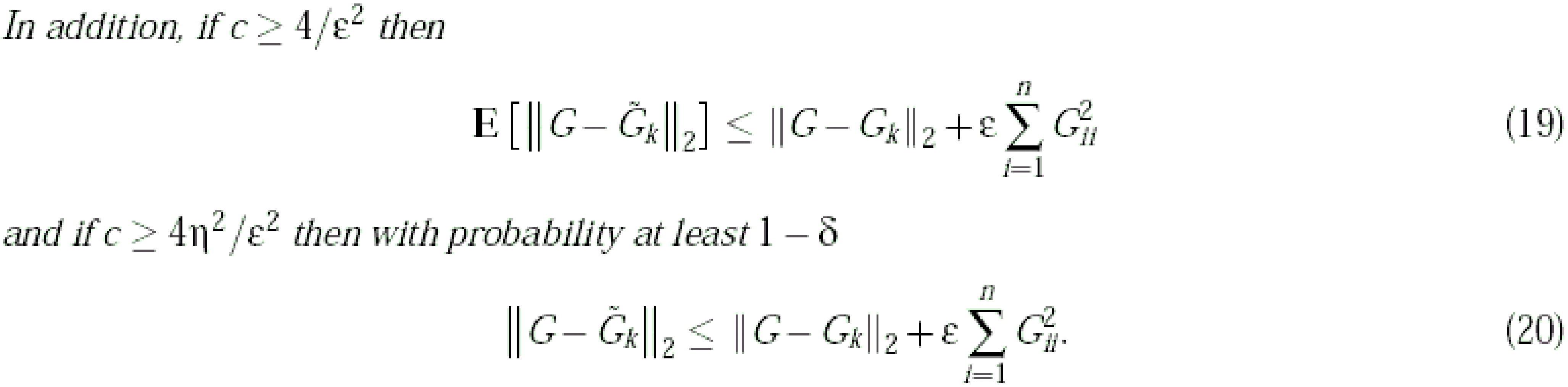
$$\mathbf{E}\left[\left\|G-\tilde{G}_k\right\|_F\right] \leq \|G-G_k\|_F + \varepsilon\sum_{i=1}^{n} G_{ii}^2 \tag{17}$$

and if $c \geq 64k\eta^2/\varepsilon^4$ *then with probability at least* $1-\delta$

$$\left\|G-\tilde{G}_k\right\|_F \leq \|G-G_k\|_F + \varepsilon\sum_{i=1}^{n} G_{ii}^2. \tag{18}$$

- Example – if $\delta = 0.1,\ \ G_{ii} = 1$:
  - $c \geq \frac{938k}{\epsilon^4}$ (implies $\epsilon \geq 1$?)
  - $\left\|G-\tilde{G}_k\right\|_F \leqslant \left\|G-G_k\right\|_F + \epsilon n$

Spectral Bounds



In addition, if $c \geq 4/\varepsilon^2$ *then*

$$\mathbf{E}\left[\left\|G-\tilde{G}_k\right\|_2\right] \leq \|G-G_k\|_2+\varepsilon\sum_{i=1}^{n} G_{ii}^2 \tag{19}$$

and if $c \geq 4\eta^2/\varepsilon^2$ *then with probability at least* $1-\delta$

$$\left\|G-\tilde{G}_k\right\|_2 \leq \|G-G_k\|_2+\varepsilon\sum_{i=1}^{n} G_{ii}^2. \tag{20}$$

- Example – if $\delta = 0.1,\ \ G_{ii} = 1$:
  - $c \geq \frac{59}{\epsilon^2}$
  - $\left\|G-\tilde{G}_k\right\|_2 \leqslant \left\|G-G_k\right\|_2 + \epsilon n$

Proof Sketch of Spectral Bound

- Column selection matrix: $S \in \mathbb{R}^{nxc}$
  - $S_{ij} = 1$ if i^th column of G chosen at trial j; $S_{ij} = 0$ otherwise
- Scaling matrix: $D \in \mathbb{R}^{cxc}$, $D_{ii} = 1 / \sqrt{cp_{i_t}}$
- Define W and C:
  - $C = GSD$
  - $W = (SD)^T GSD = DS^T GSD$
    - intersection of chosen columns/rows scaled by $\frac{1}{c\sqrt{p_{i_t} p_{j_t}}}$

Proof Sketch of Spectral Bound

- Define column-sampled and rescaled version of X
  - $C_x = XSD, \; C_x \in \mathbb{R}^{mxc}$
  - SVD: $C_x = \hat{U} \hat{\Sigma} \hat{V}^T$
- Lemma: If $\tilde{G}_k = C W_k^+ C^T$ then: $\left\| G - \tilde{G}_k \right\|_2 = \left\| X - \hat{U}_k \hat{U}_k^T X \right\|_2^2$

Proof Sketch of Spectral Bound

- Define column-sampled and rescaled version of X
  - $C_x = XSD, \quad C_x \in \mathbb{R}^{mxc}$
  - SVD: $C_x = \hat{U}\hat{\Sigma}\hat{V}^T$

- Lemma: If $\tilde{G}_k = C W_k^+ C^T$ then: $\left\|G - \tilde{G}_k\right\|_2 = \left\|X - \hat{U}_k \hat{U}_k^T X\right\|_2^2$
  - Proof: $(1)\ \ W = C_x^T C_x = \hat{V}\hat{\Sigma}^2\hat{V}\ ;\ \ W_k = \hat{V}_k \hat{\Sigma}_k^2 \hat{V}_k$

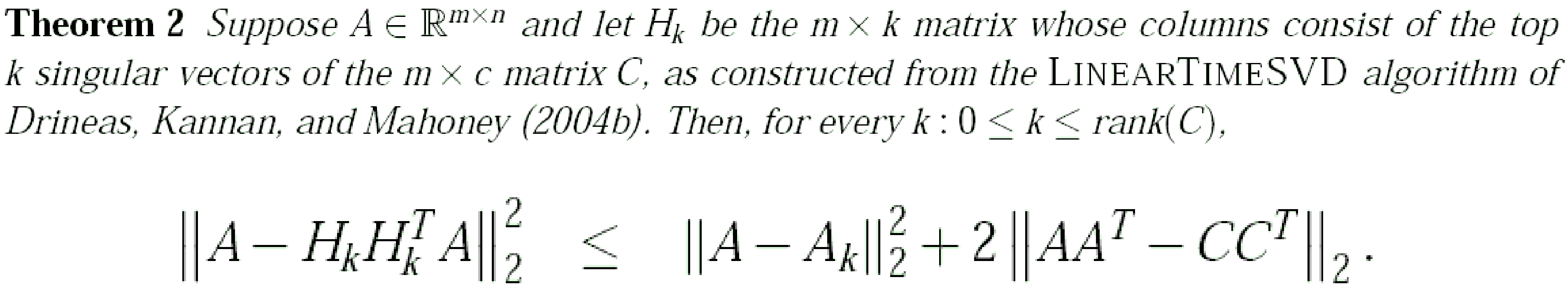
$$
\begin{aligned}
(2)\ \ \tilde{G}_k &= GSD(W_k)^+(GSD)^T \\
&= X^T C_x SD(W_k)^+(X^T C_x SD)^T \\
&= X^T \hat{U}\hat{\Sigma}\hat{V}^T(\hat{V}\hat{\Sigma}_k^2\hat{V}^T)^+\hat{V}\hat{\Sigma}\hat{U}^T X \\
&= X^T \hat{U}_k \hat{U}_k^T X
\end{aligned}
$$

$$(3)\ \ X^T X - X^T \hat{U}_k \hat{U}_k^T X = (X - \hat{U}_k \hat{U}_k^T X)^T (X - \hat{U}_k \hat{U}_k^T X)$$

$(4)\ \ \|\Omega\|_2^2 = \|\Omega^T \Omega\|_2$ for any matrix Ω

Proof Sketch of Spectral Bound

- Lemma: If $\tilde{G}_k = C W_k^+ C^T$ then: $\left\|G - \tilde{G}_k\right\|_2 = \left\|X - \hat{U}_k \hat{U}_k^T X\right\|_2^2$

- 

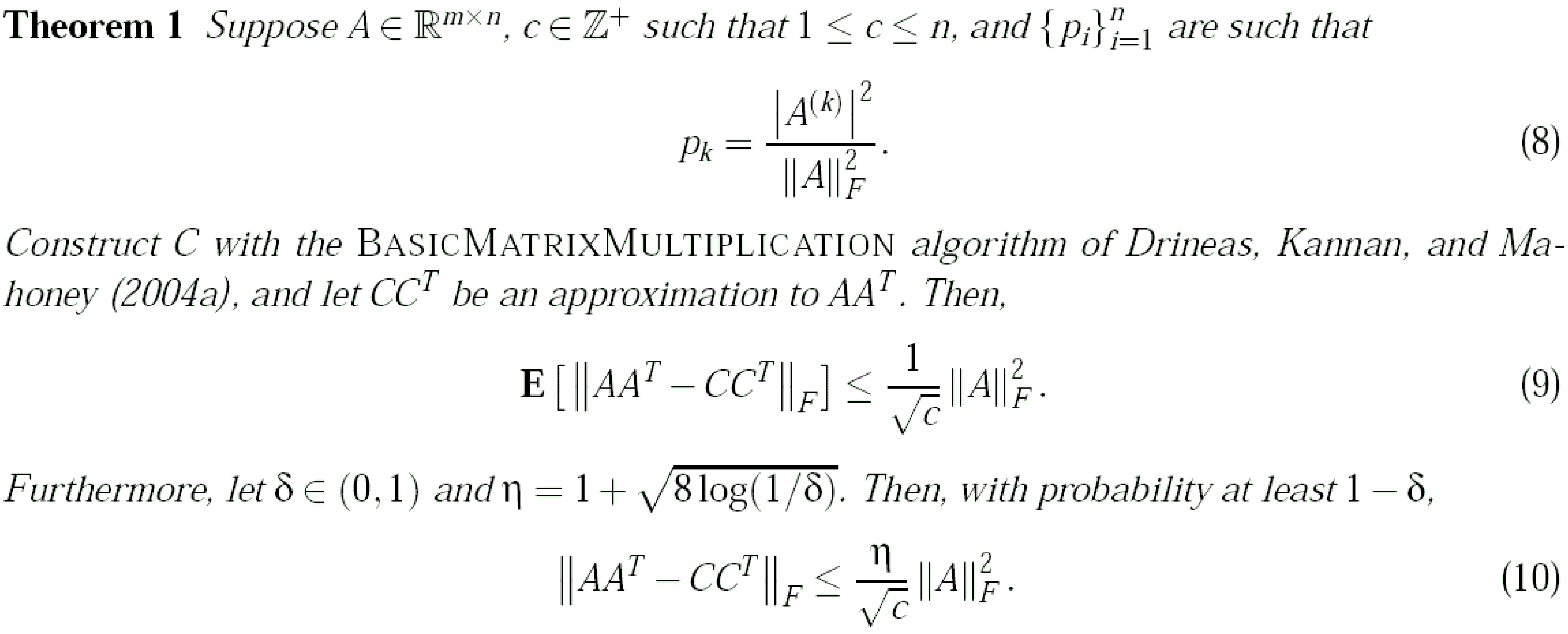
Theorem 2 *Suppose $A \in \mathbb{R}^{m \times n}$ and let H_k be the $m \times k$ matrix whose columns consist of the top k singular vectors of the $m \times c$ matrix C, as constructed from the* LinearTimeSVD *algorithm of Drineas, Kannan, and Mahoney (2004b). Then, for every $k : 0 \leq k \leq rank(C)$,*

$$\left\|A - H_k H_k^T A\right\|_2^2 \quad \leq \quad \|A - A_k\|_2^2 + 2\left\|AA^T - CC^T\right\|_2 .$$

- Combining Lemma with Theorem 2:

$$\begin{aligned}\left\|G - \tilde{G}_k\right\|_2 &\leqslant \left\|X - X_k\right\|_2^2 + 2\left\|X X^T - C_X C_X^T\right\|_2 \\ &\leqslant \left\|G - G_k\right\|_2 + 2\left\|X X^T - C_X C_X^T\right\|_2\end{aligned}$$

Proof Sketch of Spectral Bound

- Last slide: $\left\|G-\tilde{G}_k\right\|_2 \leqslant \left\|G-G_k\right\|_2 + 2\left\|X X^T - C_X C_X^T\right\|_2$
- 

Theorem 1 *Suppose* $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{Z}^+$ *such that* $1 \le c \le n$, *and* $\{p_i\}_{i=1}^n$ *are such that*

$$p_k = \frac{\left|A^{(k)}\right|^2}{\|A\|_F^2}. \tag{8}$$

Construct C *with the* BasicMatrixMultiplication *algorithm of Drineas, Kannan, and Mahoney (2004a), and let* CC^T *be an approximation to* AA^T. *Then,*

$$\mathbf{E}\left[\left\|AA^T - CC^T\right\|_F\right] \le \frac{1}{\sqrt{c}}\|A\|_F^2. \tag{9}$$

Furthermore, let $\delta \in (0,1)$ *and* $\eta = 1 + \sqrt{8\log(1/\delta)}$. *Then, with probability at least* $1-\delta$,

$$\left\|AA^T - CC^T\right\|_F \le \frac{\eta}{\sqrt{c}}\|A\|_F^2. \tag{10}$$

- Apply Theorem 1 to 2nd term on right to get final bound
  - note: $p_i = \dfrac{G_{ii}^2}{\sum_{j=1}^n G_{jj}^2} = \dfrac{|X^{(i)}|^2}{\|X\|_F^2}$; $\|X\|_F^2 = \sum_{j=1}^n G_{jj}^2$

Eigenfunction Problem

- Eigenfunction of a linear operator returns from operator as a scaled factor of itself
- Eigenfunction Problem:

$$\int_D K(x,s)\Phi(s)\,ds = \lambda\Phi(x),\ \ x \in D$$

  - relationship to discrete eigenvector problem:

$$\begin{array}{r} \\ \text{row } x \longrightarrow \\ \\ \\ \end{array}\begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}\begin{bmatrix} \vdots \\ \Phi(x) \\ \vdots \\ \vdots \end{bmatrix} = \lambda\Phi(x)$$

Nyström Method

- Quadrature-based method for numerical integration

- Quadrature rule: $\int_a^b y(s)\,ds = \sum_{j=1}^{n} w_j\, y(s_j)$
  - $\{s_j\}$ = quadrature points
  - $\{w_j\}$ = weights

- Apply to eigenfunction problem, assuming $D=[a,b]$:

$$\int_a^b K(x,s)\Phi(s)\,ds \approx \sum_{j=1}^{n} w_j k(x,s_j)\tilde{\phi}(s_j) = \tilde{\lambda}\,\tilde{\phi}$$

  - $\tilde{\lambda},\ \tilde{\phi}$ = approximate eigenvalue, eigenfunction
  - Nyström Method provides solution for $\tilde{\lambda},\ \tilde{\phi}$

Nyström Method (cont)

- last slide: $\int_a^b K(x,s)\Phi(s)\,ds \approx \sum_{j=1}^{n} w_j k(x,s_j)\tilde{\phi}(s_j) = \tilde{\lambda}\,\tilde{\phi}$

- Define set of Nyström points, $\{x_i\}$, and substitute:

$$\sum_{j=1}^{n} w_j k(x_i, s_j)\tilde{\phi}(s_j) = \tilde{\lambda}\,\tilde{\phi}(x_i)$$

  - eigendecomposition to get pairs ($\tilde{\phi}_m, \tilde{\lambda}_m$)
  - $\{x_i\}$ often set equal to $\{s_j\}$ to maintain symmetry

- Extend ($\tilde{\phi}_m, \tilde{\lambda}_m$) over entire domain:

$$\bar{\phi}_m(x) = \frac{1}{\tilde{\lambda}_m}\sum_{j=1}^{n} w_j k(x, s_j)\tilde{\phi}_m(s_j)$$

- $\bar{\phi}_m(x)$ is Nyström extension of $\tilde{\phi}_m$ and approximates $\Phi_m(x)$

Nyström Method applied to G

- Nyström extension: $\bar{\phi}_m(x) = \frac{1}{\tilde{\lambda}_m} \sum_{j=1}^{n} w_j k(x, s_j) \tilde{\phi}_m(s_j)$

- Recall: $G \approx \tilde{G} = C W^{+} C^{T}$

$$= C \hat{V} \Sigma^{-1} \hat{V}^{T} C^{T} \quad [W = \hat{V} \Sigma \hat{V}^{T}]$$

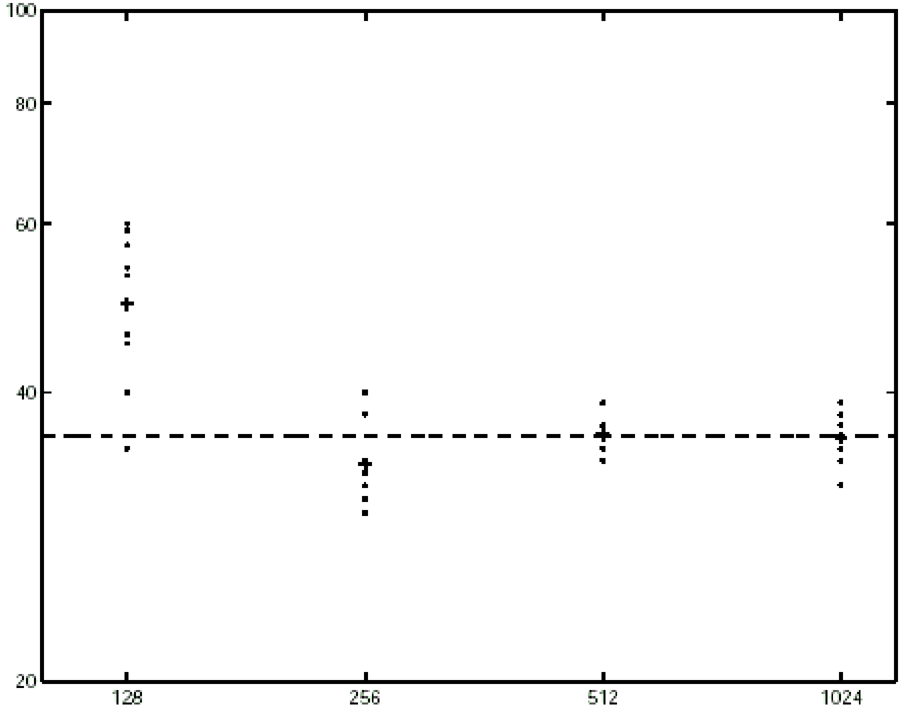
$$= C \hat{V} \Sigma^{-1} \Sigma \Sigma^{-1} \hat{V}^{T} C^{T}$$

$$= \bar{U} \Sigma \bar{U}^{T}$$

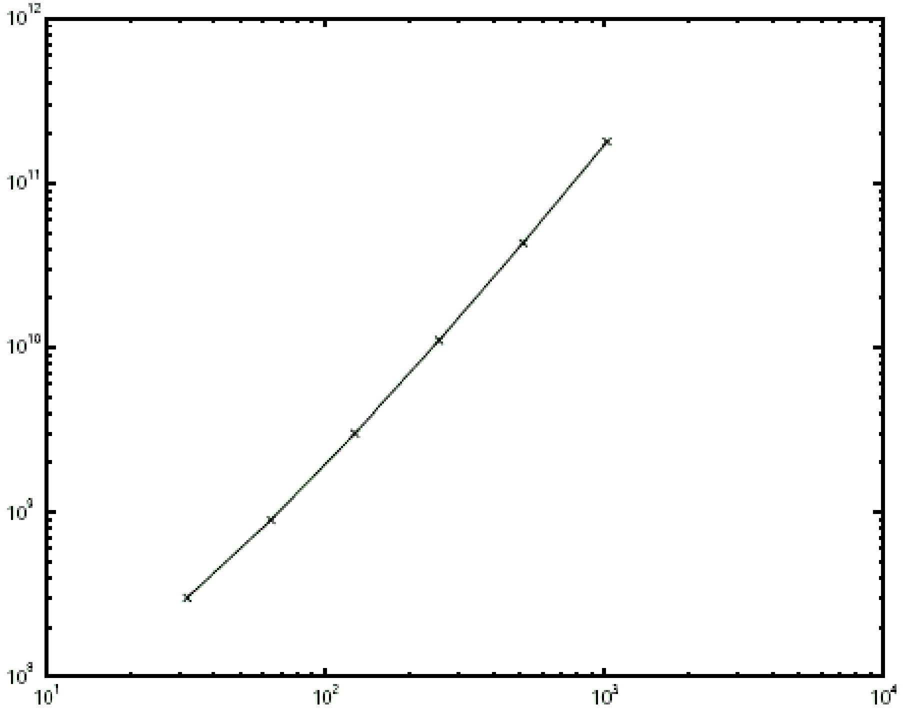
- $\bar{U} = C \hat{V} \Sigma^{-1}$: Nyström extension of solution on W to full set of data points
  - same form as Nyström extension we just derived, with $\{x_i\} = \{s_j\}$ and quadrature weights $\{w_j\}$ equal 1

Experiment 1: Full G vs Nyström Approx

- Classification using GP classifiers [Williams & Seeger, NIPS, 2001]
  - requires inverse of Gram matrix
  - USPS handwritten digits (7291 train, 2007 test)
  - Discriminate class “4” from rest



(a)

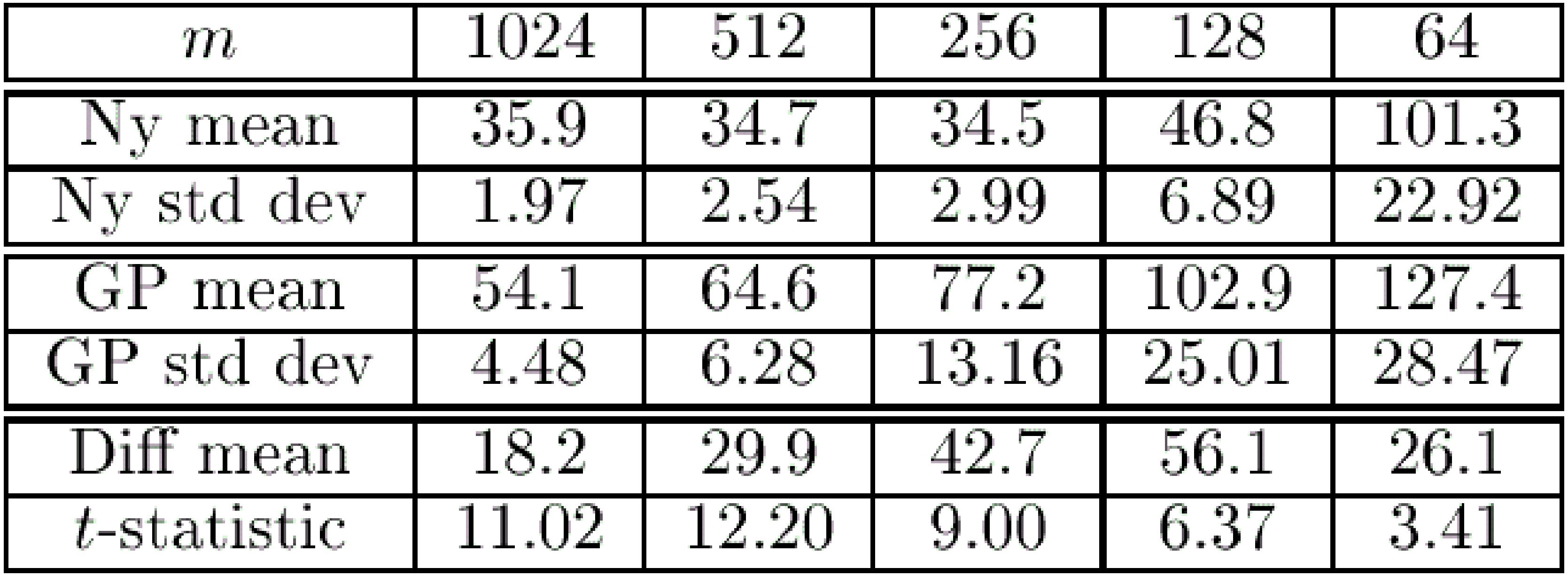


(b)

Experiment 2

- Compare: [Williams & Seeger, NIPS, 2001]
  - Exact GP classifier on *m* points
  - Nyström classifier on *m* points, extended to all points

- Nyström classifier does better!



m	1024	512	256	128	64
Ny mean	35.9	34.7	34.5	46.8	101.3
Ny std dev	1.97	2.54	2.99	6.89	22.92
GP mean	54.1	64.6	77.2	102.9	127.4
GP std dev	4.48	6.28	13.16	25.01	28.47
Diff mean	18.2	29.9	42.7	56.1	26.1
t-statistic	11.02	12.20	9.00	6.37	3.41