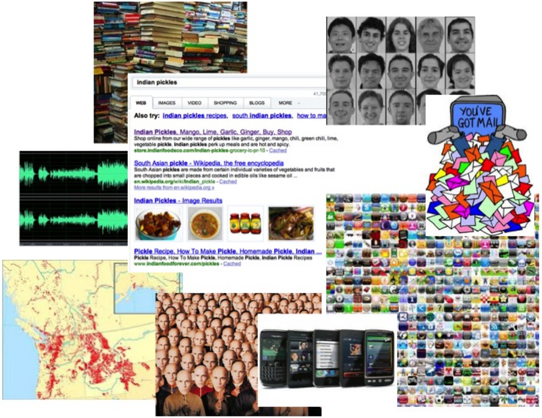
YAHOO!

Low Rank Matrix Approximation

PRESENTED BY Edo Liberty - April 24, 2015

Collaborators: Nir Ailon, Steven Zucker, Zohar Karnin, Dimitris Achlioptas, Per-Gunnar Martinsson, Vladimir Rokhlin, Mark Tygert, Christos Boutsidis, Franco Woolfe, Maxim Sviridenko, Dan Garber, Yoelle Maarek, Amit Singer, David Woodruff, Jeff Phillips, Mina Ghashami, Huan Wang, Daniel Hsu

Matrix Data



Often our data is represented by a large matrix.

Matrix Data

We will think of $A \in \mathbb{R}^{d \times n}$ as n column vectors in $\mathbb{R}^d$ and typically $n \gg d$.

Typical web scale data:

Data	**Columns**	**Rows**	d	n	**sparse**
Textual	Documents	Words	10^5 - 10^7	$> 10^{10}$	yes
Actions	Users	Types	10^1 - 10^4	$> 10^7$	yes
Visual	Images	Pixels, SIFT	10^5 - 10^6	$> 10^8$	no
Audio	Songs, tracks	Frequencies	10^5 - 10^6	$> 10^8$	no
Machine Learning	Examples	Features	10^2 - 10^4	$> 10^6$	yes/no
Financial	Prices	Items, Stocks	10^3 -10^5	$> 10^6$	no

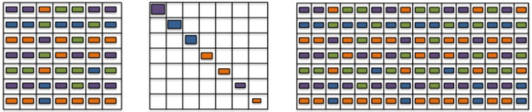
Matrix Data

Low rank matrix approximation is helpful for

- Dimension reduction
- Signal processing
- Compression
- Classification
- Regression
- Clustering
- ...

Singular Value Decomposition (SVD)

$$A = U\Sigma V^T$$



$U \in \mathbb{R}^{d\times d}$ $\Sigma \in \mathbb{R}^{d\times d}$ $V^T \in \mathbb{R}^{d\times n}$

$U^T U = I_d$ $\Sigma_{1,1} \geq \dots \geq \Sigma_{d,d} \geq 0$ $V^T V = I_d$

Best rank k Approximation

$$A_k = U_k \Sigma_k V_k^T$$

$U_k \in \mathbb{R}^{d\times k}$ $\Sigma_k \in \mathbb{R}^{k\times k}$ $V_k^T \in \mathbb{R}^{k\times n}$

$U_k^T U_k = I_k$ $\Sigma_{1,1} \geq \dots \geq \Sigma_{k,k} \geq 0$ $V_k^T V_k = I_k$

$B = A_k$ minimizes $\|A - B\|_2$ and $\|A - B\|_F$ among all rank k matrices.

Best rank k Approximation

Block power methods and Lanczos like methods:

- $\tilde{O}(k)$ passes over the matrix.
- $\tilde{O}(ndk)$ operations

$\tilde{O}(\cdot)$ hides logarithmic factors and spectral gap dependencies.

By first computing AA^T

- $\Omega(d^2)$ space
- $O(nd^2)$ operations

Assuming $d = o(n)$ and naive matrix matrix multiplication.

Matrix Approximation

Let $P_k^A = U_k U_k^T$ be the best rank k projection of the columns of A

$$\|A - P_k^A A\|_2 = \|A - A_k\|_2 = \sigma_{k+1}$$

Let P_k^B be the best rank k projection for B

$$\|A - P_k^B A\|_2 \leq \sigma_{k+1} + \sqrt{2\|AA^T - BB^T\|} \qquad \text{[FKV04]}$$

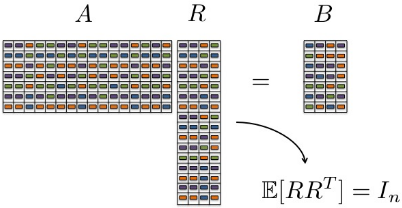
From this point on, our goal is to find B which is:

1. $\|AA^T - BB^T\| \leq \varepsilon \|AA^T\|$
2. Small, $B \in \mathbb{R}^{d \times \ell}$ and $\ell \ll d$
3. Computationally easy to obtain from A

Random projection based algorithms

Random projection

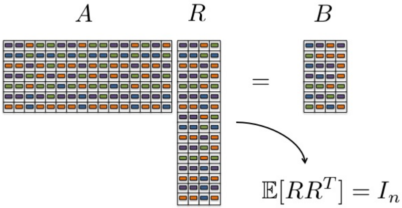
1. Output $B = AR$ [Sar06, WLRT08, CW09]
2. Where $R \in \mathbb{R}^{n \times \ell}$ such that $R_{i,j} \sim \mathcal{N}(0, 1/\ell)$.



Random projection

[Sar06, WLRT08, CW09]

1. Output $B = AR$
2. Where $R \in \mathbb{R}^{n \times \ell}$ such that $R_{i,j} \sim \mathcal{N}(0, 1/\ell)$.



Note that $\mathbb{E}[BB^T] = \mathbb{E}[ARR^T A^T] = A\ \mathbb{E}[RR^T]A^T = AA^T$

Random projection

Johnson-Lindenstrauss property [JL84, FM87, DG99]

The matrix R exhibits the Johnson-Lindenstrauss property. For any $y \in \mathbb{R}^n$

$$\Pr\left[\left|\|yR\|^2 - \|y\|^2\right| > \varepsilon\|y\|^2\right] \leq e^{-c\ell\varepsilon^2}$$

If $\ell = \tilde{O}(Rank(A)/\varepsilon^2)$ then by the union bound we have

$$\|A^TA - B^TB\| = \sup_{\|x\|=1}\left|\|xA\|^2 - \|xAR\|^2\right| \leq \varepsilon\|AA^T\|$$

This gives us exactly what we need!

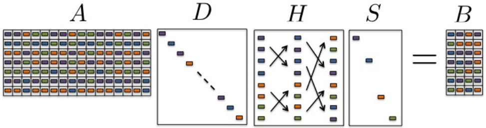
Random projection

- 1 pass
- $O(nd\ell)$ operations

Fast random projection

- This can be accelerated by making R sparser [Ach03, Mat08, DKS10, KN10].
- But in general, R cannot be "much sparser" [KN10, NNW12, NN13, NN14].

Faster Johnson-Lindenstrauss transforms require very different machinery [AC06, AL09, AC10, LAS11, AL11, KW11, AR14].



Fast random projection

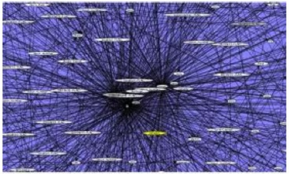
- 1 pass (by row)
- $O(nd\log(\ell))$ operations

Matrix approximation in the streaming setting

Data is dynamically aggregated

Sometimes we get one column at a time (row operations impossible...)

Data is dynamically aggregated



Sometimes, we cannot even store the entire matrix.

Streaming Matrices

Note that AA^T can be trivially computed from the stream of columns A_i

$$AA^T = \sum_{i=1}^{n} A_i A_i^T$$

In words, AA^T is the sum of outer products of the columns of A.

Streaming Matrices

Note that AA^T can be trivially computed from the stream of columns A_i

$$AA^T = \sum_{i=1}^{n} A_i A_i^T$$

In words, AA^T is the sum of outer products of the columns of A.

Naïve solution

Compute AA^T in time $O(nd^2)$ and space $O(d^2)$. Compute the best rank-k projection for AA^T in $o(nd^2)$.

Which is hopeless when d is large!

Column Sampling

Sample only ℓ columns where [FKV04, AW02, Ver, DK03, RV07, Oli10]

$$\ell \in O\left(\frac{r\log(r)}{\varepsilon^2}\right)$$

Each column of B is $A^{(j)}/p_j$ w.p. $p_j = \|A^{(j)}\|_2^2/\|A\|_F^2$

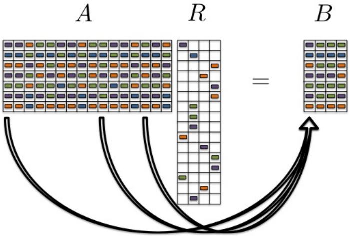
Column sampling based on ℓ_2^2 norm

1. It can be performed in a column stream, $O(nnz(A))$ operations
2. The result is sparse if the data is sparse, potentially $o(d\ell)$ space

$r = \|A\|_F^2/\|A\|_2^2$ is the numeric/stable rank of A.

Feature Hashing

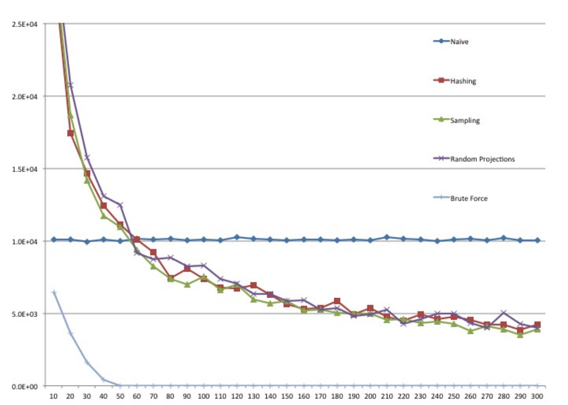
Use the "count-sketch" matrix R that contains one $\{-1, 1\}$ per row.



1. It can be applied in streaming in $O(nnz(A))$ operations
2. The result is dense, $\Omega(d\ell)$ space
3. Has some other surprising properties...

[CCFC02, WDL+09, CW12]

Experiments



The error term $\|AA^T - BB^T\|$ reduces like $1/\sqrt{\ell}$.

Frequent Directions

Frequent Directions

Lemma from [Lib13]

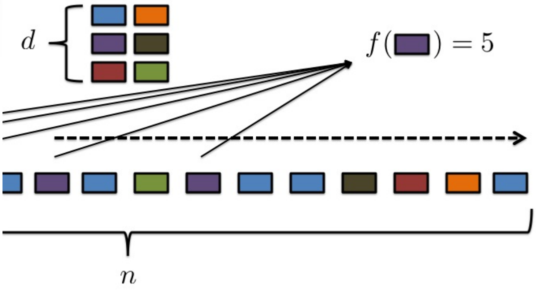
One can maintain a matrix B with only $\ell = O(1/\varepsilon)$ columns s.t.

$$\|AA^T - BB^T\|_2 \leq \varepsilon \|A\|_F^2$$

Intuition:

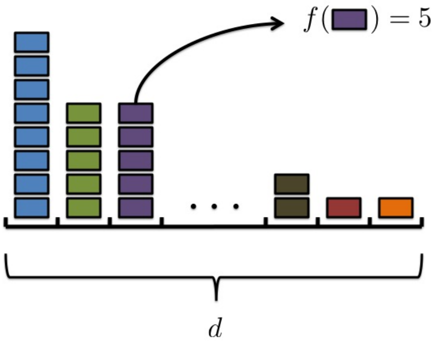
Extend Frequent-items [MG82, DLOM02, KPS03, MAEA05]

Frequent Items



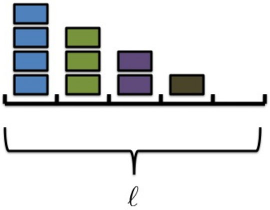
Obtain the frequency $f(i)$ of each item in the stream of items

Frequent Items



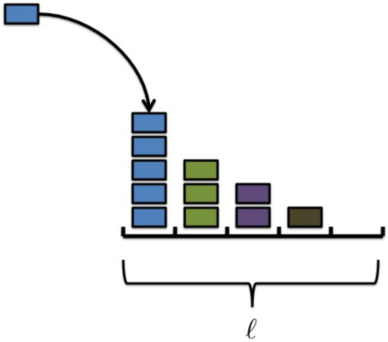
With d counters it's easy but not good enough (IP addresses, queries....)

Frequent Items



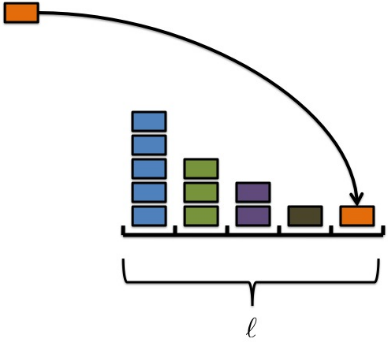
(Misra-Gries) Lets keep **less than** a fixed number of counters ℓ.

Frequent Items



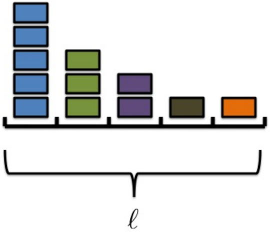
If an item has a counter we add 1 to that counter.

Frequent Items



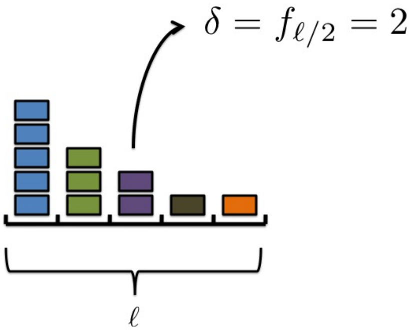
Otherwise, we create a new counter for it and set it to 1

Frequent Items



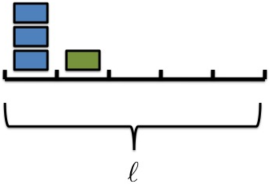
But now we do not have less than ℓ counters.

Frequent Items



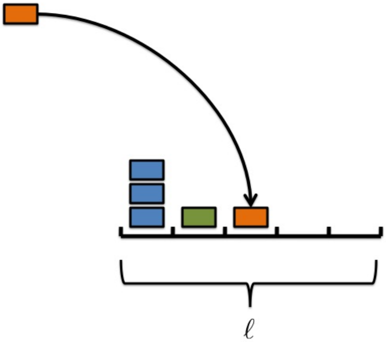
Let δ be the median counter value at time t

Frequent Items



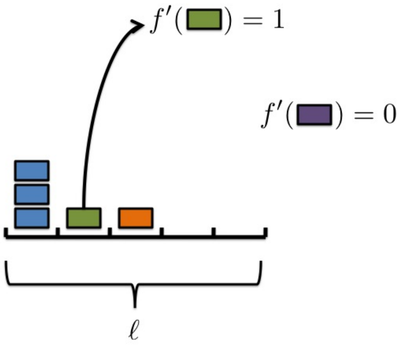
Decrease all counters by δ (or set to zero if less than δ)

Frequent Items



And continue...

Frequent Items



The approximated counts are f'

Frequent Items

- We increase the count by only 1 for each item appearance.

$$f'(i) \leq f(i)$$

- Because we decrease each counter by at most δ_t at time t

$$f'(i) \geq f(i) - \sum_t \delta_t$$

- Calculating the total approximated frequencies:

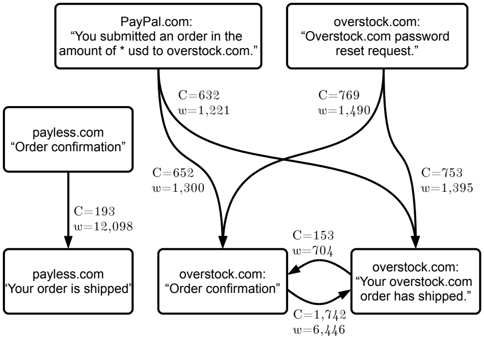
$$0 \leq \sum_i f'(i) \leq \sum_t 1 - (\ell/2) \cdot \delta_t = n - (\ell/2) \cdot \sum_t \delta_t$$

$$\sum_t \delta_t \leq 2n/\ell$$

- Setting $\ell = 2/\varepsilon$ yields

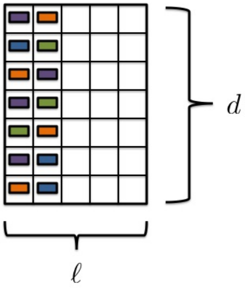
$$|f(i) - f'(i)| \leq \varepsilon n$$

Email threading



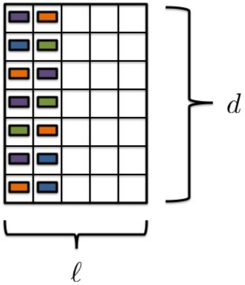
Find all email pairs such that $\Pr(e_1|e_2) \geq \theta$ [AKLM13].

Frequent Directions



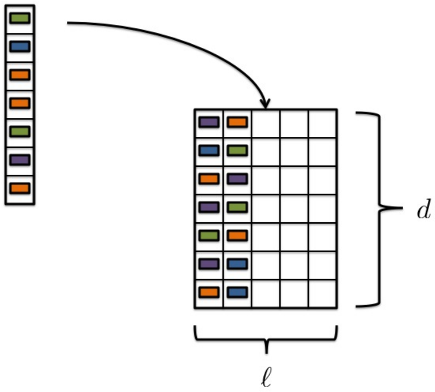
We keep a sketch of at most ℓ columns

Frequent Directions



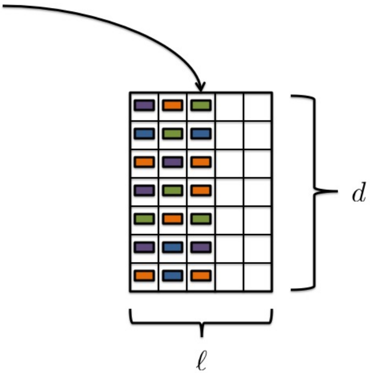
We maintain the invariant that some columns are empty (zero valued)

Frequent Directions



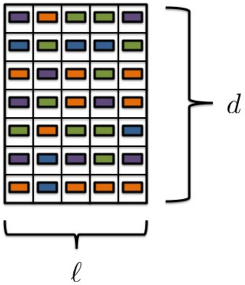
Input vectors are simply stored in empty columns

Frequent Directions



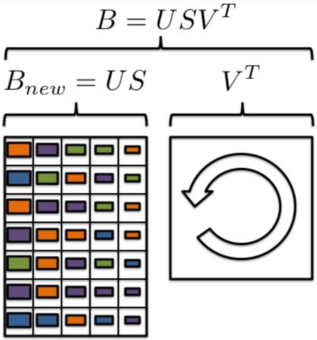
Input vectors are simply stored in empty columns

Frequent Directions



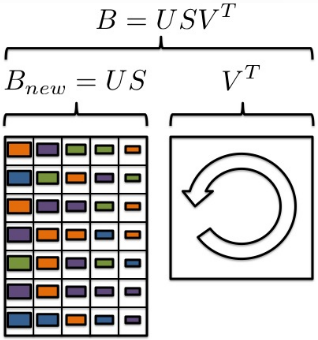
When the sketch is 'full' we need to zero out some columns...

Frequent Directions



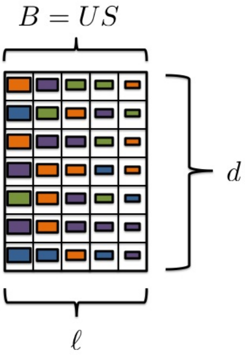
Using the SVD we compute $B = USV^T$ and set $B_{new} = US$

Frequent Directions



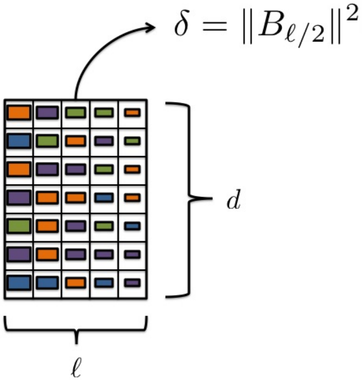
Note that $BB^T = B_{new}B_{new}^T$ so we don't "lose" anything

Frequent Directions



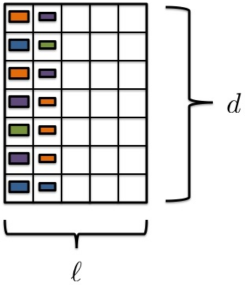
The columns of B are now orthogonal and in decreasing magnitude order

Frequent Directions



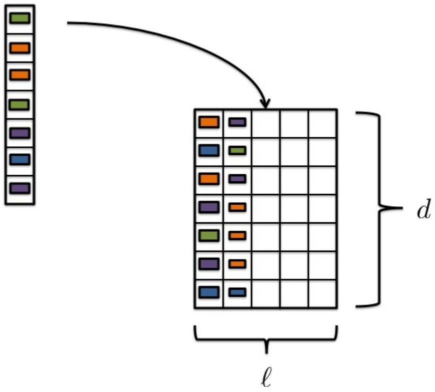
Let $\delta = \|B_{\ell/2}\|^2$

Frequent Directions



Reduce column ℓ_2^2-norms by δ (or nullify if less than δ)

Frequent Directions



Start aggregating columns again...

Frequent Directions

Input: $\ell, \quad A \in \mathbb{R}^{d \times n}$
$B \leftarrow$ all zeros matrix $\in \mathbb{R}^{d \times \ell}$
for $i \in [n]$ **do**
    Insert A_i into a zero valued column of B
    if B has no zero valued colums **then**
        $[U, \Sigma, V] \leftarrow SVD(B)$
        $\delta \leftarrow \sigma^2_{\ell/2}$
        $\check{\Sigma} \leftarrow \sqrt{\max(\Sigma^2 - I_\ell \delta, 0)}$
        $B \leftarrow U\check{\Sigma}$ # At least half the columns of B are zero.
Return: B

Bounding the error

We first bound $\|AA^T - BB^T\|$

$$\begin{aligned}\sup_{\|x\|=1} \|xA\|^2 - \|xB\|^2 &= \sup_{\|x\|=1} \sum_{t=1}^{n} [\langle x, A_t \rangle^2 + \|xB^{t-1}\|^2 - \|xB^t\|^2] \\ &= \sup_{\|x\|=1} \sum_{t=1}^{n} [\|xC^t\|^2 - \|xB^t\|^2] \\ &\leq \sum_{t=1}^{n} \|{C^t}^T C^t - {B^t}^T B^t\| \cdot \|x\|^2 \leq \sum_{t=1}^{n} \delta_t\end{aligned}$$

Which gives:

$$\|AA^T - BB^T\| \leq \sum_{t=1}^{n} \delta_t$$

Bounding the error

We compute the Frobenius norm of the final sketch.

$$
\begin{aligned}
0 \leq \|B\|_F^2 &= \sum_{t=1}^{n} [\|B^t\|_F^2 - \|B^{t-1}\|_F^2] \\
&= \sum_{t=1}^{n} [(\|C^t\|_F^2 - \|B^{t-1}\|_F^2) - (\|C^t\|_F^2 - \|B^t\|_F^2)] \\
&= \sum_{t=1}^{n} \|A_t\|^2 - tr({C^t}^T C^t - {B^t}^T B^t) \leq \|A\|_F^2 - (\ell/2) \sum_{t=1}^{n} \delta_t
\end{aligned}
$$

Which gives:

$$
\sum_{t=1}^{n} \delta_t \leq 2\|A\|_F^2 / \ell
$$

Bounding the error

We saw that:

$$\|AA^T - BB^T\| \leq \sum_{t=1}^{n} \delta_t$$

and that:

$$\sum_{t=1}^{n} \delta_t \leq 2\|A\|_F^2/\ell$$

Setting $\ell = 2/\varepsilon$ yields

$$\|AA^T - BB^T\| \leq \varepsilon\|A\|_F^2 \ .$$

The two proofs are very similar...

Stronger bounds

Lemma: covariance approximation guarantee [GP14, GLPW15]

$\|A^T A - B^T B\|_2 \leq \|A - A_k\|_F^2/(\ell - k)$ for any $k < \ell$.

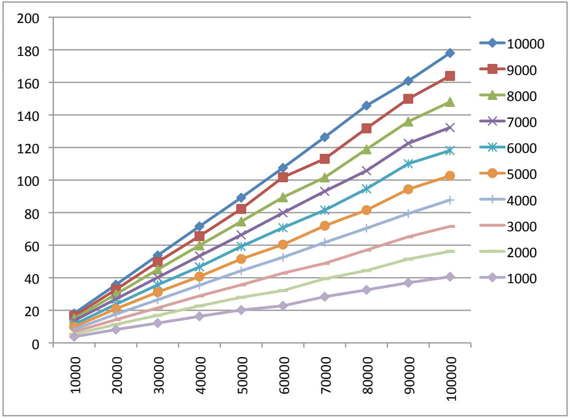
Lemma: projection approximation guarantee [GP14]

$\|A - \pi_{B_k}(A)\|_F^2 \leq \left(1 + \frac{k}{\ell - k}\right)\|A - A_k\|_F^2$ for any $k < \ell$.

Lemma: space optimality [Woo14, GLPW15]

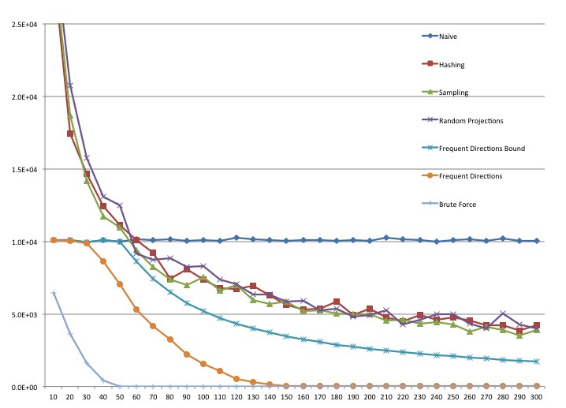
Frequent directions is space optimal. Any algorithm (randomized or not) with matching guaranties must require as much space, up to a word-size factor.

Frequent Directions



Slower than hashing or sampling but still very fast.

Frequent Directions

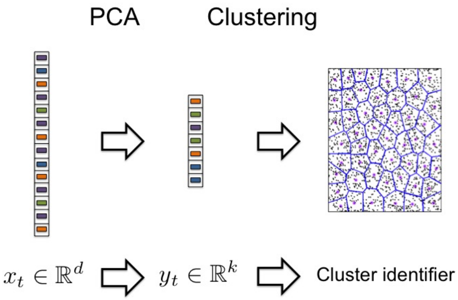


The error term $\|AA^T - BB^T\|$ reduces like $1/\ell$

Matrix approximation online

Online PCA

Consider clustering the reduced dimensional vectors online (e.g. [Mey01, LSS14])



The PCA algorithm must output y_t **before** receiving x_{t+1}.

Online PCA, prior models

Regret minimization: Minimizes $\sum_t \|x_t - P_{t-1}x_t\|^2$ where P_{t-1} is committed to before receiving x_t [WK06, NKW13]

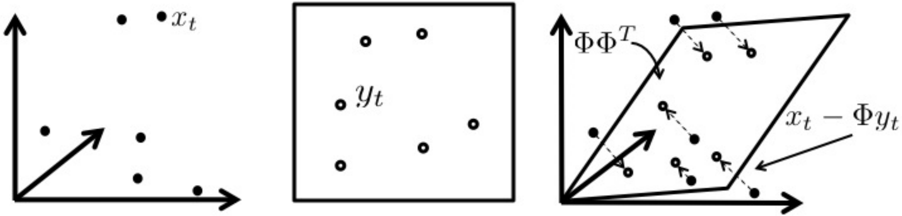
Random projection: can guarantee online that $\|(X - (XY^+)Y\|_F^2$ is small [CW09, Sar06]

Stochastic model: Assumes x_t are drown i.i.d. from an unknown distribution [OK85, ACS13, MCJ13, BDF13]

Principal Component Analysis

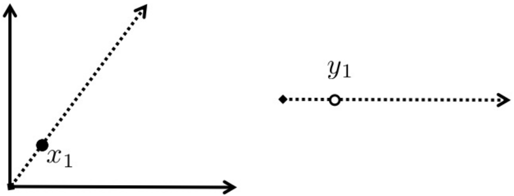
Given a set of vectors $x_t \in \mathbb{R}^d$ the goal is to map them to $y_t \in \mathbb{R}^k$ that minimize:

$$\min_{\{\Phi | \Phi^T \Phi = I_k\}} \sum_i \|x_t - \Phi y_t\|_2^2$$



Online regression

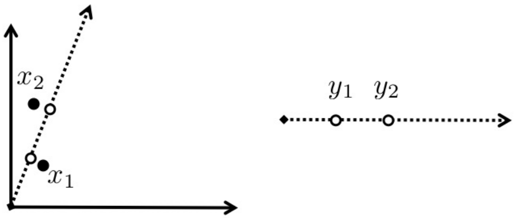
Note that this is non trivial even when $d = 2$ and $k = 1$.



For x_1 there aren't many options...

Online regression

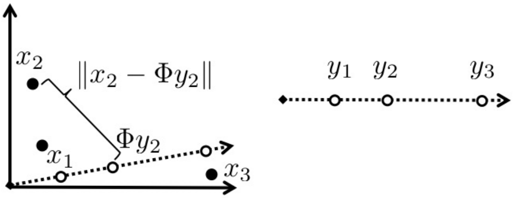
Note that this is non trivial even when $d = 2$ and $k = 1$.



For x_2 this is already a non standard optimization problem

Online regression

Note that this is non trivial even when $d = 2$ and $k = 1$.



In general, the mapping $x_i \mapsto y_i$ is not necessarily linear.

Online PCA algorithms

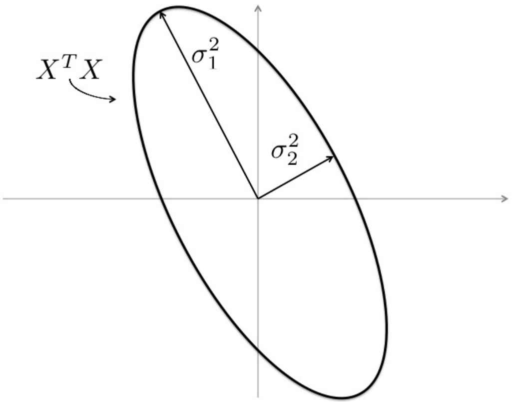
Lemma: online PCA with Frobenius bounds [BGK**L**15]

$\min_\Phi \|X - \Phi Y\|_F^2 \leq \|X_k\|_F^2 + \varepsilon \|X\|_F^2$ with target dimension $\ell \in O(k/\varepsilon^2)$

Lemma: improved online PCA with spectral bounds [K**L**15]

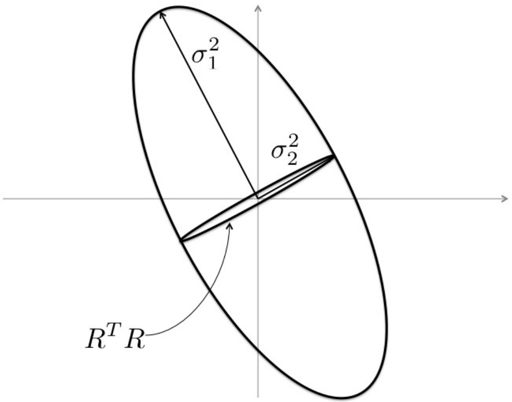
$\min_\Phi \|X - \Phi Y\|_2^2 \leq \sigma_{k+1}^2 + \varepsilon \sigma_1^2$ with target dimension $\ell = \tilde{O}(\frac{k}{\varepsilon^2})$

Online PCA algorithm intuition



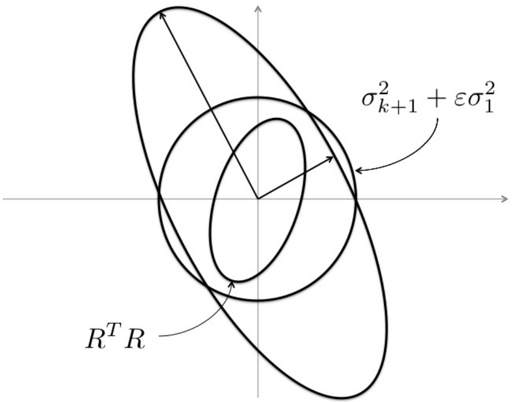
The covariance matrix $X^T X$ visualized as an ellipse.

Online PCA algorithm intuition



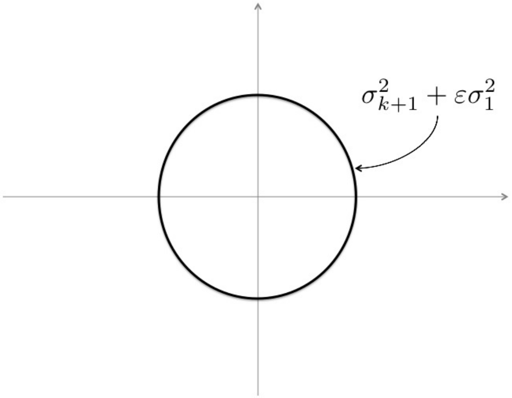
The optimal residual is $R = X - X_k$

Online PCA algorithm intuition



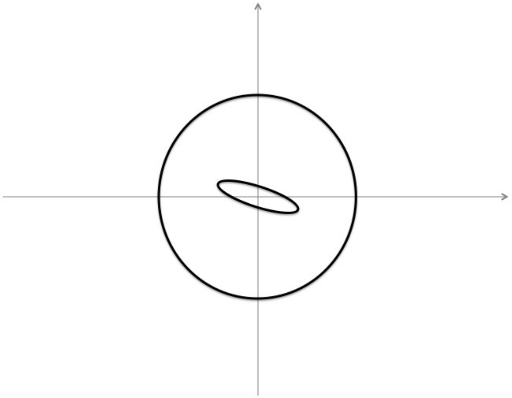
Any residual $R = X - \Phi Y$ such that $\|R^T R\| \leq \sigma_{k+1}^2 + \varepsilon\sigma_1^2$ would work

Online PCA algorithm intuition



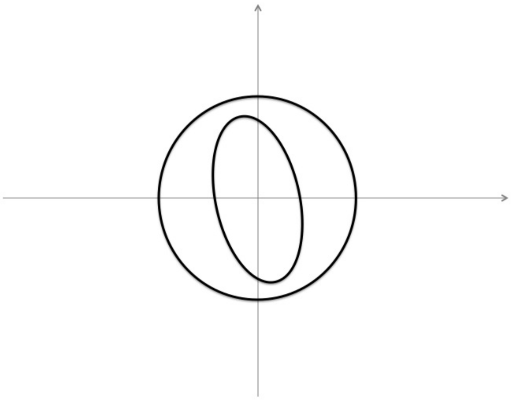
Let us assume we know $\Delta = \sigma_{k+1}^2 + \varepsilon\sigma_1^2$.

Online PCA algorithm intuition



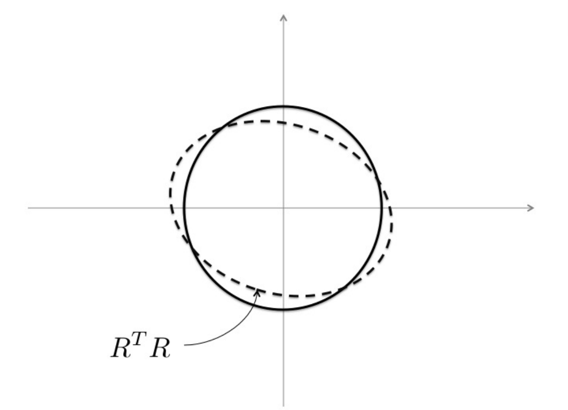
We start with mapping $x_t \mapsto 0$ and $R_{[1:t]} = X_{[1:t]}$

Online PCA algorithm intuition



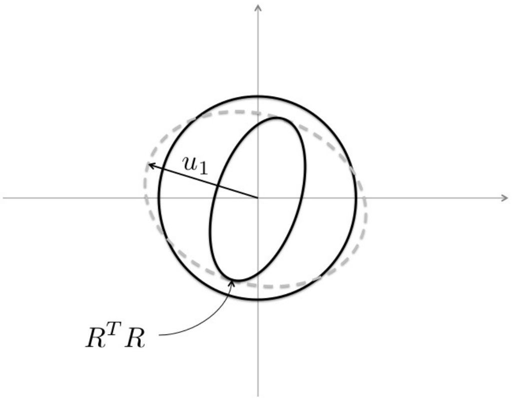
This is continued as long as $\|R^T R\| \leq \Delta$

Online PCA algorithm intuition



When $\|R^T R\| > \Delta$ we update the projection to prevents it from happening

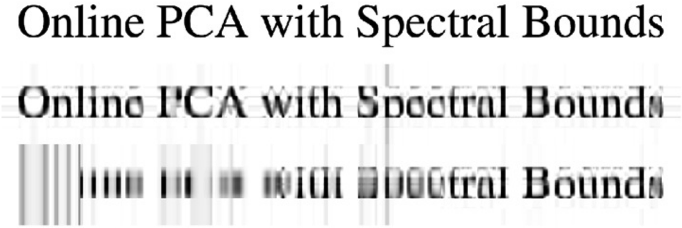
Online PCA algorithm intuition



We commit to a new online PCA direction u_i such that $\|R^T R\| \leq \Delta$ again.

Online PCA with Spectral Bounds

input: X
$U \leftarrow$ all zeros matrix
for $x_t \in X$ **do**
  if $\|(I - UU^T)X_{1:t}\|^2 \geq \sigma_{k+1}^2 + \varepsilon\sigma_1^2$
    Add the top left singular vector of $(I - UU^T)X_{1:t}$ to U
  yield $y_t = U^T x_t$



Open questions

- Reduce running time of Frequent Directions (there is some progress on that)
- Reduce running time of online PCA
- Reduce target dimension of online PCA, is it possible?
- Can we avoid the doubling trick in online PCA if we allow *scaled* isometric reconstructions?

Thank you!

Nir Ailon and Bernard Chazelle.
Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform.
In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.

Nir Ailon and Bernard Chazelle.
Faster dimension reduction.
Commun. ACM, 53(2):97–104, 2010.

Dimitris Achlioptas.
Database-friendly random projections: Johnson-Lindenstrauss with binary coins.
J. Comput. Syst. Sci., 66(4):671–687, 2003.

Raman Arora, Andy Cotter, and Nati Srebro.
Stochastic optimization of pca with capped msg.
In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1815–1823. 2013.

Nir Ailon, Zohar Shay Karnin, Edo Liberty, and Yoelle Maarek.
Threading machine generated email.
In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 405–414, 2013.

Nir Ailon and Edo Liberty.
Fast dimension reduction using rademacher series on dual bch codes.
Discrete Comput. Geom., 42(4):615–630, 2009.

Nir Ailon and Edo Liberty.
An almost optimal unrestricted fast johnson-lindenstrauss transform.
In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 185–191, 2011.

Nir Ailon and Holger Rauhut.
Fast and rip-optimal transforms.
Discrete & Computational Geometry, 52(4):780–798, 2014.

Rudolf Ahlswede and Andreas Winter.
Strong converse for identification via quantum channels.
IEEE Transactions on Information Theory, 48(3):569–579, 2002.

Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund.
The fast convergence of incremental pca.
In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3174–3182. 2013.

Christos Boutsidis, Dan Garber, Zohar Shay Karnin, and Edo Liberty.
Online principal components analysis.
In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 887–901, 2015.

Moses Charikar, Kevin Chen, and Martin Farach-Colton.
Finding frequent items in data streams.
In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, ICALP '02, pages 693–703, London, UK, UK, 2002. Springer-Verlag.

Kenneth L Clarkson and David P Woodruff.
Numerical linear algebra in the streaming model.
In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.

Kenneth L. Clarkson and David P. Woodruff.
Low rank approximation and regression in input sparsity time.
CoRR, abs/1207.6365, 2012.

S. DasGupta and A. Gupta.
An elementary proof of the Johnson-Lindenstrauss lemma.
Technical Report, UC Berkeley, 99-006, 1999.

Petros Drineas and Ravi Kannan.
Pass efficient algorithms for approximating large matrices, 2003.

Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós.
A sparse johnson: Lindenstrauss transform.
In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 341–350, 2010.

Erik D. Demaine, Alejandro López-Ortiz, and J. Ian Munro.
Frequency estimation of internet packet streams with limited space.
In *Proceedings of the 10th Annual European Symposium on Algorithms*, ESA '02, pages 348–360, London, UK, UK, 2002. Springer-Verlag.

Alan Frieze, Ravi Kannan, and Santosh Vempala.
Fast monte-carlo algorithms for finding low-rank approximations.
J. ACM, 51(6):1025–1041, November 2004.

P. Frankl and H. Maehara.
The Johnson-Lindenstrauss lemma and the sphericity of some graphs.
Journal of Combinatorial Theory Series A, 44:355–362, 1987.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff.
Frequent directions : Simple and deterministic matrix sketching.
CoRR, abs/1501.01711, 2015.

Mina Ghashami and Jeff M. Phillips.
Relative errors for deterministic low-rank matrix approximations.
In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 707–717, 2014.

W. B. Johnson and J. Lindenstrauss.
Extensions of Lipschitz mappings into a Hilbert space.
Contemporary Mathematics, 26:189–206, 1984.

Zohar Shay Karnin and Edo Liberty.
Online pca with spectral bounds.
In progress, 2015.

Daniel M. Kane and Jelani Nelson.
A derandomized sparse johnson-lindenstrauss transform.
Electronic Colloquium on Computational Complexity (ECCC), 17:98, 2010.

Richard M. Karp, Christos H. Papadimitriou, and Scott Shenker.
A simple algorithm for finding frequent elements in streams and bags.
ACM Transactions on Database Systems, 28:2003, 2003.

Felix Krahmer and Rachel Ward.
New and improved johnson-lindenstrauss embeddings via the restricted isometry property.
SIAM J. Math. Analysis, 43(3):1269–1281, 2011.

Edo Liberty, Nir Ailon, and Amit Singer.
Dense fast random projections and lean walsh transforms.
Discrete & Computational Geometry, 45(1):34–44, 2011.

Edo Liberty.
Simple and deterministic matrix sketching.
In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 581–588, New York, NY, USA, 2013. ACM.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko.
An algorithm for online k-means clustering.
CoRR, abs/1412.5721, 2014.

Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi.
Efficient computation of frequent and top-k elements in data streams.
In Thomas Eiter and Leonid Libkin, editors, *Database Theory - ICDT 2005*, volume 3363 of *Lecture Notes in Computer Science*, pages 398–412. Springer Berlin Heidelberg, 2005.

Jirí Matousek.
On variants of the johnson-lindenstrauss lemma.
Random Struct. Algorithms, 33(2):142–156, 2008.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain.
Memory limited, streaming pca.
In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2886–2894. 2013.

Adam Meyerson.
Online facility location.
In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 426–431, 2001.

Jayadev Misra and David Gries.
Finding repeated elements.
Technical report, Ithaca, NY, USA, 1982.

Jiazhong Nie, Wojciech Kotlowski, and Manfred K. Warmuth.
Online pca with optimal regrets.
In *ALT*, pages 98–112, 2013.

Jelani Nelson and Huy L. Nguyen.
Sparsity lower bounds for dimensionality reducing maps.
In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 101–110, 2013.

Jelani Nelson and Huy L. Nguyên.
Lower bounds for oblivious subspace embeddings.
In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pages 883–894, 2014.

Jelani Nelson, Huy L. Nguyên, and David P. Woodruff.
On deterministic sketching and streaming for sparse recovery and norm estimation.
In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 627–638, 2012.

Erkki Oja and Juha Karhunen.
On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix.
Journal of Mathematical Analysis and Applications, 106(1):69 – 84, 1985.

Roberto Imbuzeiro Oliveira.
Sums of random hermitian matrices and an inequality by rudelson.
arXiv:1004.3821v1, April 2010.

Mark Rudelson and Roman Vershynin.
Sampling from large matrices: An approach through geometric functional analysis.
J. ACM, 54(4), July 2007.

Tamas Sarlos.
Improved approximation algorithms for large matrices via random projections.
In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, Washington, DC, USA, 2006.

Roman Vershynin.
A note on sums of independent random matrices after ahlswede-winter.
Lecture Notes.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg.
Feature hashing for large scale multitask learning.
In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1113–1120, New York, NY, USA, 2009. ACM.

Manfred K. Warmuth and Dima Kuzmin.
Randomized PCA algorithms with regret bounds that are logarithmic in the dimension.
In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1481–1488, 2006.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert.
A fast randomized algorithm for the approximation of matrices.
Applied and Computational Harmonic Analysis, 25(3):335 – 366, 2008.

David P. Woodruff.
Low rank approximation lower bounds in row-update streams.
In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1781–1789, 2014.