Sublinear Time Low Rank Approximation of PSD Matrices

Cameron Musco
MIT

David Woodruff
CMU

Low Rank Approximation

- A is an n x d matrix
  - Think of n points in R^d
- E.g., A is a customer-product matrix
  - $A_{i,j}$ = how many times customer i purchased item j
- A is typically well-approximated by low rank matrix
  - E.g., high rank because of noise
- Goal: find a low rank matrix approximating A
  - Easy to store, quick to multiply, data more interpretable

What is a Good Low Rank Approximation?

<u>Singular Value Decomposition (SVD)</u>

Any matrix A = UΣV

- U has orthonormal columns
- Σ is diagonal with non-increasing positive entries down the diagonal
- V has orthonormal rows

- Truncated SVD rank-k approximation: $A_k = U_k \Sigma_k V_k$

$$\begin{pmatrix} \\ \mathbf{A} \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \mathbf{U}_k \\ \\ \end{pmatrix} \begin{pmatrix} \Sigma_k \end{pmatrix} \begin{pmatrix} \quad \mathbf{V}_k \quad \end{pmatrix} + \begin{pmatrix} \\ \mathbf{E} \\ \\ \end{pmatrix}$$

What is a Good Low Rank Approximation?

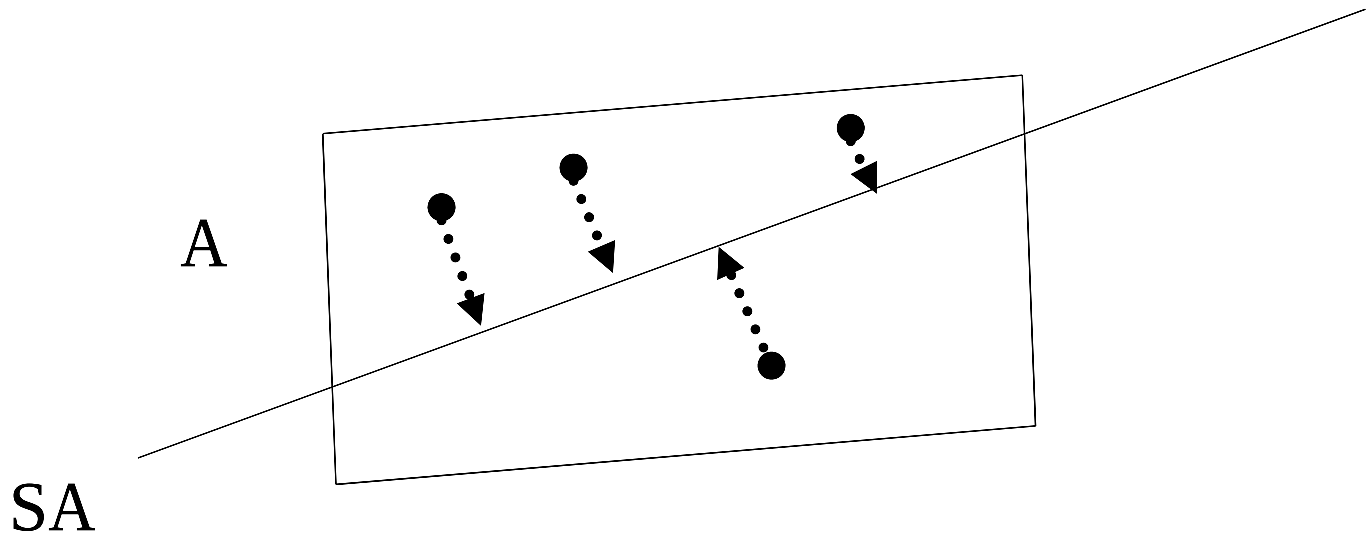
- $A_k = \text{argmin}_{\text{rank k matrices B}} |A-B|_F$
- $|C|_F = (\Sigma_{i,j} C_{i,j}^2)^{1/2}$
- Computing A_k exactly is expensive

Approximate Low Rank Approximation

- Goal: output a rank k matrix A', so that
  - $|A-A'|_F \leq (1+\varepsilon) |A-A_k|_F$
- Can do this in nnz(A) + (n+d)*poly(k/ε) time w.h.p. [CW13]

Solution to Low-Rank Approximation

- Given n x d input matrix A
- Compute SA using a sketching matrix S with k/ε << n rows. SA takes random linear combinations of rows of A



- Project rows of A onto SA, then find best rank-k approximation to points inside of SA

What is the Matrix S?

- S can be a k/ε x n matrix of i.i.d. normal random variables
- [S06] S can be an O~(k/ε) x n Fast Johnson Lindenstrauss Matrix
- [CW13] S can be a poly(k/ε) x n CountSketch matrix

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

SA can be computed in nnz(A) time

Caveat: Projecting the Points onto SA is Slow

- Current algorithm
  - Compute S*A
  - Project each of the rows onto S*A
  - Find best rank-k approximation of projected points inside of rowspace of S*A

- Bottleneck is step 2

- Approximate the projection
  - Fast algorithm for approximate regression

$$\min_X |X(SA)-A|_F^2 \Rightarrow \min_X |X(SA)R-AR|_F^2$$

  - nnz(A) + (n + d)*poly(k/ε) time

Structure-Preserving Low Rank Approximation

- Let A be an arbitrary n x n matrix
- Suppose we also require our rank-k approximation A' to be positive semidefinite (PSD)
  - A' is symmetric and all eigenvalues are non-negative
- Covariance matrices, kernel matrices, Laplacians are PSD
- Roundoff errors may make a PSD matrix non-PSD
  - We do not assume A is PSD but want A' to be PSD

Structure-Preserving Low Rank Approximation

- Goal: output a rank-k PSD matrix A' for which $|A-A'|_F$ is small
- Can assume A is symmetric
  - $A = A^{sym} + A^{asym}$, where $A^{sym}_{i,j} = \frac{A_{i,j}+A_{j,i}}{2}$ and $A^{asym}_{i,j} = \frac{A_{i,j}-A_{j,i}}{2}$
  - $|A - A'|^2{}_F = |A^{sym} - A'|_F^2 + |A^{asym}|_F^2$
  - Compute A^{sym} in nnz(A) time
- What is the best PSD rank-k approximation $A_{k,+}$ to A?
- $A_{k,+}$ is obtained by zeroing out all but its top k positive eigenvalues

PSD Low Rank Approximation

- [CW17]: In nnz(A) + n poly(k/ ε) time, can find a PSD rank-k A' so that
  - $|A-A'|_F \leq (1+\varepsilon)\ |A-A_{k,+}|_F$
- Previous work
  - [KMT09] Nystrom method based on uniform sampling requires incoherence assumptions on A
  - [GM13] Weaker $|A-A'|_F \leq |A-A_{k,+}|_F + \epsilon|A - A_{k,+}|_*$ bound, where $|.|_*$ is the nuclear norm
  - [WLZ16] Running time at least n^2k/ϵ and A' has a larger rank k/ϵ

How Good Are These Algorithms?

- For general matrices A, there is an nnz(A) time lower bound for relative error approximation

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \infty & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Lower bounds hold even to estimate $|A|_F^2$ up to relative error

- Similar nnz(A) time lower bound holds for outputting a relative error PSD low rank approximation to an arbitrary matrix A

What if Your Input Matrix is Itself PSD?

- Let A be an arbitrary n x n PSD matrix

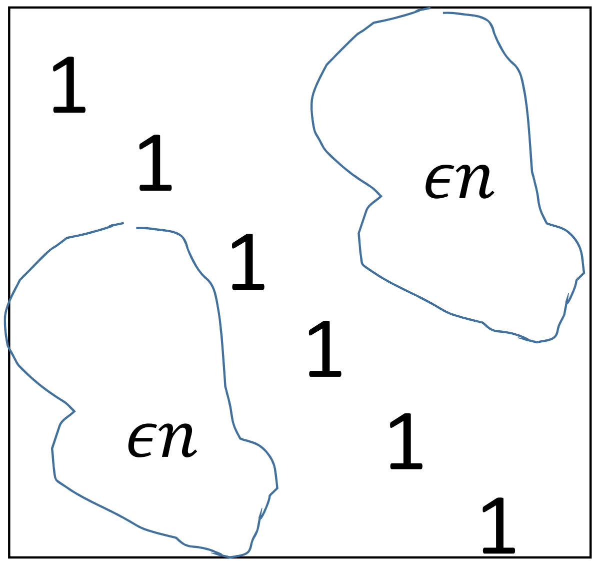
- Covariance matrices, kernel matrices, Laplacians are PSD
  - Want to approximate them for efficiency

- Is there an nnz(A) time lower bound for low rank approximation of PSD matrices?

- Is there an nnz(A) time lower bound for estimating the norm $|\mathrm{A}|_{\mathrm{F}}^2$ of a PSD matrix?

Estimating the Norm of a PSD Matrix

- $|A|_F^2 = \left|BB^T\right|_F^2 = \sum_{i,j} < B_i, B_j >^2$, where A = BB^T
- $< B_i, B_j >^2 \leq |B_i|_2^2 \cdot \left|B_j\right|_2^2 \leq \max_{i,j} < B_i, B_i >^2$
- If $|B_i|_2^2 = 1$ for all i, then
  - (1) $< B_i, B_j >^2 \leq 1$ for all i and j
  - (2) if $\sum_{i \neq j} < B_i, B_j >^2 \geq \epsilon \sum_i < B_i, B_i >^2$ then $\sum_{i \neq j} < B_i, B_j >^2 \geq \epsilon n$
- Uniformly sampling $n \cdot \text{poly}(\frac{1}{\epsilon})$ terms $< B_i, B_j >^2$ for $i \neq j$ suffices for estimating $\sum_{i \neq j} < B_i, B_j >^2$



$(1) < B_i, B_j >^2 \leq 1$ for all i,j

$$(2) \sum_{i \neq j} < B_i, B_j >^2 \geq \epsilon n$$

Conditions imply uniformly sampling $n \cdot \text{poly}\left(\frac{1}{\epsilon}\right)$ entries works

- When $|B_i|_2 \neq 1$ for all i, sample an entry with probability $p_{i,j} = |B_i|^2 \cdot |B_j|^2 / |B|_F^4$
- Let X = $< B_i, B_j >^2 / p_{i,j}$ if entry i,j is sampled
- $E[X] = \sum_{i.j} p_{i,j} < B_i, B_j >^2 / p_{i,j} = \sum_{i,j} < B_i, B_j >^2 = |B^T B|_F^2 = |A|_F^2$
- $\text{Var}[X] = \sum_{i.j} p_{i,j} < B_i, B_j >^4 / p_{i,j}^2 \leq n \cdot |A|_F^4$

Sublinear Time Low Rank Approximation of PSD Matrices

- Our Result: Given an n x n PSD matrix A, in $n \cdot k^2 \cdot poly(\frac{1}{\epsilon})$ time we can output a (factorization of a) rank-k matrix A' for which w.h.p.

$$|A - A'|_F \leq (1 + \epsilon)|A - A_k|_F$$

- The number of entries read is $n \cdot k \cdot poly(\frac{1}{\epsilon})$

- Lower Bound: Any algorithm requires reading $\Omega(n \cdot k \cdot \frac{1}{\epsilon})$ entries

Starting Point: Connection to Adaptive Sampling

Adaptively sample a column proportional to its distance to the span of columns chosen so far [DV06]:

- C ← ∅
- For i = 1, 2, …, $\frac{\mathrm{poly}(k)}{\epsilon}$
- Sample a column A_i with probability $\frac{|A_i - P_C A_i|_2^2}{|A - P_C A|_F^2}$
- C ← C ∪ $\{A_i\}$
- End

- There is a k-dimensional subspace V inside the span of C so that

$$|A - P_V A|_F^2 \leq (1 + \epsilon)|A - A_k|_F^2$$

Connection to Adaptive Sampling

- Computing the sampling probabilities and finding $P_V A$ only requires knowing inner products between columns of A and C

- Algorithm needs $n \cdot \frac{poly(k)}{\epsilon} \ll n^2$ inner products

- Since A is PSD, $A = B^T B$, and given A, all inner products between columns (or rows) of B have been precomputed!

- Run adaptive sampling algorithm using A to output $P_V B$:
$$|B - P_V B|_F^2 \leq (1 + \epsilon)|B - B_k|_F^2$$

- But $P_V A$ can be an arbitrarily bad low rank approximation to A…

Projection Cost-Preserving Sketches

- Instead, sample a set C of columns of A which not only contains a good rank-k approximation, but is a *Projection Cost-Preserving Sketch (PCP):*

[CEMMP15] There is a diagonal rescaling matrix S so that for all k-dimensional projection matrices P: $|\mathrm{SC}(\mathrm{I}-\mathrm{P})|_\mathrm{F}^2 = (1 \pm \epsilon)|\mathrm{A}(\mathrm{I}-\mathrm{P})|_\mathrm{F}^2$

- If P approximately minimizes the LHS, then P approximately minimizes the RHS

- Find a PCP C of $\mathrm{poly}\left(\frac{\mathrm{k}}{\epsilon}\right)$ columns of A and output its top k left singular vectors

- If C can be found by reading $\mathrm{n} \cdot \mathrm{k} \cdot \mathrm{poly}\left(\frac{1}{\epsilon}\right)$ entries of A, we are done

Building a PCP

- How should we sample the columns C?

- [LMP13,KLM+14,AM15] Ridge leverage scores:

$$\tau_i(A) = a_i^T \left(AA^T + \frac{|A - A_k|_F^2}{k} I \right)^{-} a_i$$

- Give a “smooth” rank-k version of standard leverage scores

- They are the standard leverage scores of [A; $(|A - A_k|_F/\sqrt{k}) \cdot I$]

Ridge Leverage Scores

[CMM16] Let $\beta \in (0,1)$. Suppose $\tilde{\tau}_i \geq \beta \tau_i$ for all i, and $p_i = \frac{\tilde{\tau}_i}{\sum_j \tilde{\tau}_j}$. Sample a set C of $t = O((k \log k)/(\beta \epsilon^2))$ columns of A with replacement, where the i-th column of C is $\frac{A_j}{(tp_j)^{\frac{1}{2}}}$ with probability p_j. With high probability,

$$(1-\epsilon)CC^T - \frac{\epsilon}{k}|A - A_k|_F^2 \cdot I \preccurlyeq AA^T \preccurlyeq (1+\epsilon)CC^T + \frac{\epsilon}{k}|A - A_k|_F^2 \cdot I$$

- Multiplicative/additive generalization of a subspace embedding
- Proof uses stable rank version of matrix Bernstein bound

Ridge Leverage Scores

[CMM16] Let $\beta \in (0,1)$. Suppose $\tilde{\tau_i} \geq \beta \tau_i$ for all i, and $p_i = \frac{\tilde{\tau_i}}{\sum_j \tilde{\tau_j}}$. Sample a set C of $t = O((k \log k)/(\beta \epsilon^2))$ columns of A, where the i-th column of C equals $\frac{A_j}{(tp_j)^{\frac{1}{2}}}$ with probability p_j. With high probability, C is a PCP:

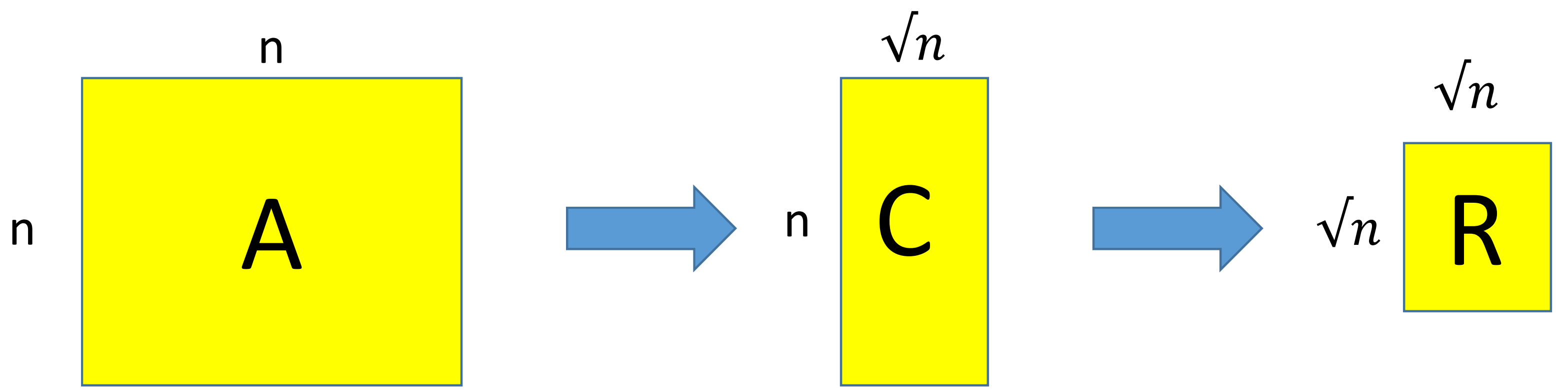
For all k-dimensional projection matrices P, $|C(I-P)|_F^2 = (1 \pm \epsilon)|A(I-P)|_F^2$

- Proof uses the multiplicative/additive generalization of a subspace embedding
- But how do we quickly get approximations $\tilde{\tau_i}$ to the ridge leverage scores?
- Recall $\tau_i(A) = a_i^T \left(AA^T + \frac{|A-A_k|_F^2}{k} I\right)^{-} a_i$

Ridge Leverage Scores

- Unclear how to obtain good approximations to the $\tau_i(A)$
- Instead, since $A = BB^T$, A is the "kernel matrix" of a linear kernel, use [MM16] which shows how to approximate the $\tau_i(B)$ up to a $\Theta(1)$ factor with $n \cdot k$ kernel evaluations. Each evaluation corresponds to a single entry of A
- We show for all i, $\tau_i(A) \leq \frac{\sqrt{n}}{\sqrt{k}} \tau_i(B)$
- Sampling $(nk)^{.5}\text{poly}\left(\frac{1}{\epsilon}\right)$ columns of A according to the $\tau_i(B)$ gives a PCP C for A!
- But we still need to sample at least $n^{.5}$ columns of A…

Reduction to a Small Square Submatrix



- C is an n x $(\mathrm{nk})^{.5}\mathrm{poly}(\frac{1}{\epsilon})$ reweighted column submatrix of A
- We show, using that C is a PCP of A, that its rank-k *standard row leverage scores* are within a factor of $\left(\frac{\mathrm{n}}{\mathrm{k}}\right)^{.5}$ of the *ridge leverage scores* of B
- Implies sampling a reweighted subset R of $(\mathrm{nk})^{.5}\mathrm{poly}\left(\frac{1}{\epsilon}\right)$ rows of C is a PCP for C!

Processing the Small Matrix R

- Since R is small, can use sketching techniques to find its approximate top k right singular vectors
- Since R is a PCP for C, can use sampling techniques based on its top k right singular vectors give approximate top k left singular vectors of C
- Since C is a PCP for A, its top k left singular vectors span a good low rank approximation to A
- Overall time is $\widetilde{\mathrm{O}}(\mathrm{nk})\mathrm{poly}\left(\frac{1}{\epsilon}\right)$

Conclusions

- First sublinear time algorithm for relative error low rank approximation of PSD matrices, bypassing an nnz(A) lower bound for general matrices
- Tight $\widetilde{\Theta}(\mathrm{nk})$ bounds for constant ϵ
- Spectral norm error impossible in sublinear time, but can find a rank-k A' with $|\mathrm{A}-\mathrm{A}'|_2^2 \leq (1+\epsilon)|\mathrm{A}-\mathrm{A_k}|_2^2 + \frac{\epsilon}{\mathrm{k}}|\mathrm{A}-\mathrm{A_k}|_\mathrm{F}^2$ in $\mathrm{n} \cdot \mathrm{poly}(\frac{\mathrm{k}}{\epsilon})$ time
- Can output a PSD rank-k matrix A' in $\mathrm{n} \cdot \mathrm{poly}(\frac{\mathrm{k}}{\epsilon})$ time
- Open questions: (1) tighter dependence on ϵ, (2) other families of matrices?