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 $A = U\Sigma V$

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

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

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

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]

- Proof based on sparse random projections

How Good Is this Algorithm?

- 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

What if Your Input Matrix is Itself PSD?

- Let A be an arbitrary n x n PSD matrix

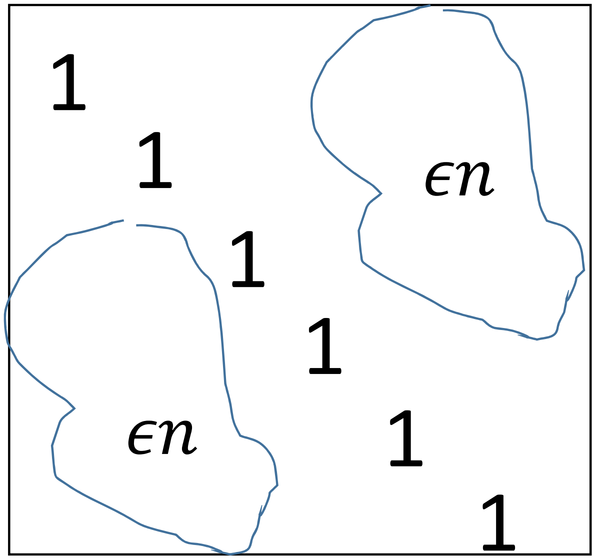
- Covariance matrices, kernel matrices, Laplacians are PSD
  - Want to approximate them for efficiency

- 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
- $< B_i, B_j >^2 \le |B_i|_2^2 \cdot \left|B_j\right|_2^2$
- If $|B_i|_2^2 = 1$ for all i, then
  - (1) $< B_i, B_j >^2 \le 1$ for all i and j
  - (2) if $\sum_{i \neq j} < B_i, B_j >^2 \ge \epsilon \sum_i < B_i, B_i >^2$ then $\sum_{i \neq j} < B_i, B_j >^2 \ge \epsilon n$
- 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 $n \cdot k^2 \cdot \text{poly}(\frac{1}{\epsilon})$ 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 $n \cdot k \cdot \text{poly}(\frac{1}{\epsilon})$

- 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]:

- $C \leftarrow \emptyset$
- For i = 1, 2, …, $\frac{k^2}{\epsilon}$
- Sample a column A_i with probability $\frac{|A_i - P_C A_i|_2^2}{|A - P_C A|_F^2}$
- $C \leftarrow C \cup \{A_i\}$
- 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

- The adaptive sampling algorithm only requires knowing inner products between columns of A and C

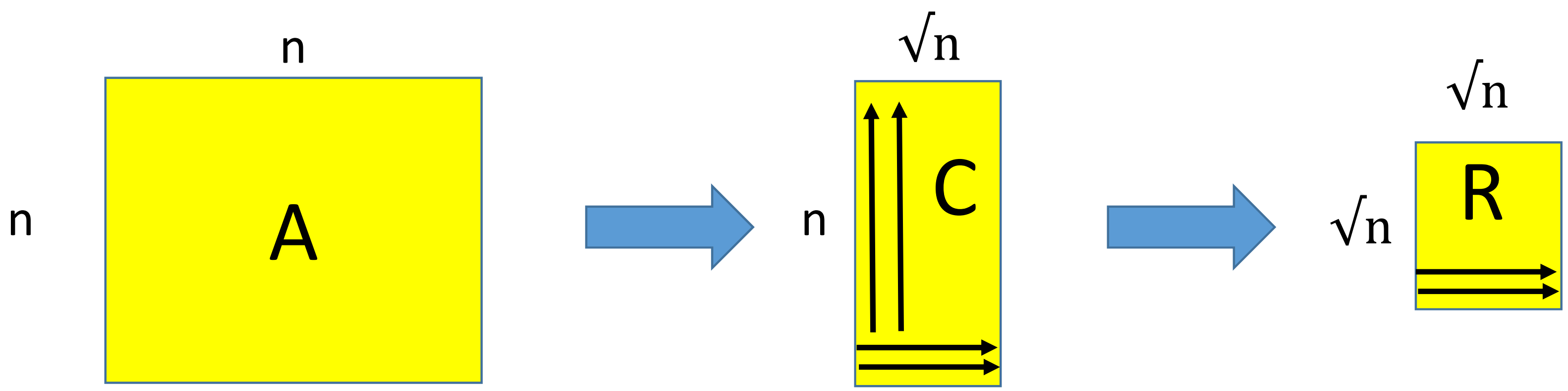
- Algorithm needs $n \cdot \frac{k^2}{\epsilon} \ll n^2$ inner products

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

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

- Setting $\epsilon = 1/\sqrt{n}$, $B^T P_V B$ can be shown to be a good approximation to A

Improving the Running Time



- Show how to compute sampling probabilities of columns and rows of A in $\widetilde{O}(nk)\mathrm{poly}\left(\frac{1}{\epsilon}\right)$ time to reduce A to a $\sqrt{n}$ x $\sqrt{n}$ matrix R

- Sampling probabilities are the "ridge leverage scores" of B, where $A = B^T B$
  - Can be computed in $\widetilde{O}(nk)$ time given A

- R is a small matrix, and can spend nnz(R) time to find its top k principal components

Conclusions

- 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 $|\mathrm{A}-\mathrm{A}'|_2^2 \le (1+\epsilon)|\mathrm{A}-\mathrm{A_k}|_2^2 + \frac{\epsilon}{\mathrm{k}}|\mathrm{A}-\mathrm{A_k}|_\mathrm{F}^2$ 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?