Input Sparsity and Hardness for Robust Subspace Approximation

Kenneth Clarkson and David Woodruff

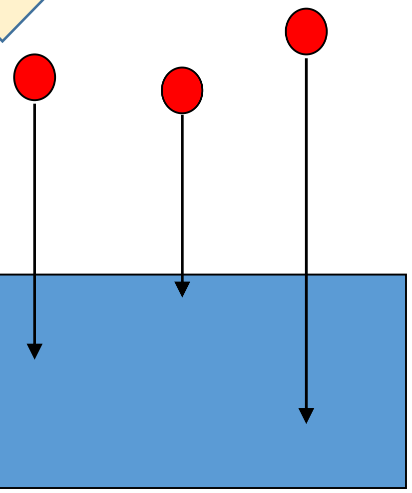
IBM Research-Almaden

Singular Value Decomposition

Abuse notation and use V to be a subspace and the d x k matrix with orthonormal columns spanning the subspace

- Given an n x d matrix A, think of the rows $a_1, a_2, \ldots, a_n$ as poin
- Find k-dimensional subspace V of R^d minimizing

$$\sum_i \left|a_i - a_i \ VV^T\right|_2^2 = \sum_i d(a_i, V)^2$$



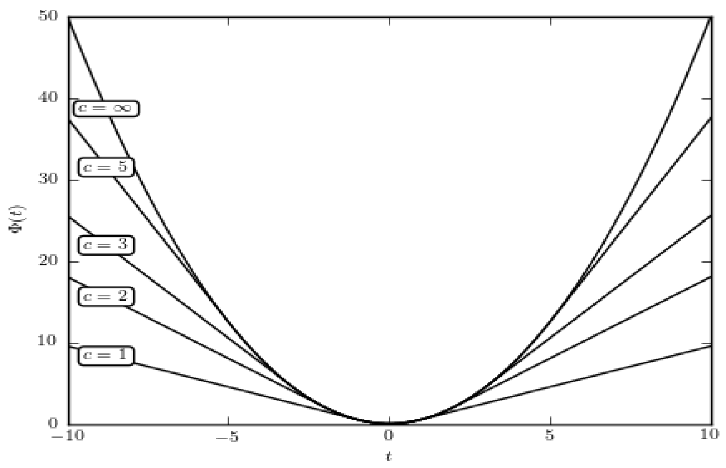
- Optimal V is given by the span of top k right singular values of A
- V can be found using $\min(n^2 d, nd^2)$ arithmetic operations
- Can find a V’ of dimension k for which

$$\sum_i d(a_i, V')^2 \leq (1+\epsilon) \min_{k-\dim V} \sum_i d(a_i, V)^2$$

in O(nnz(A)) + (n+d) poly(k/ϵ) [CW13]. See [MM13, NN13] for further optimizations

Robust Statistics

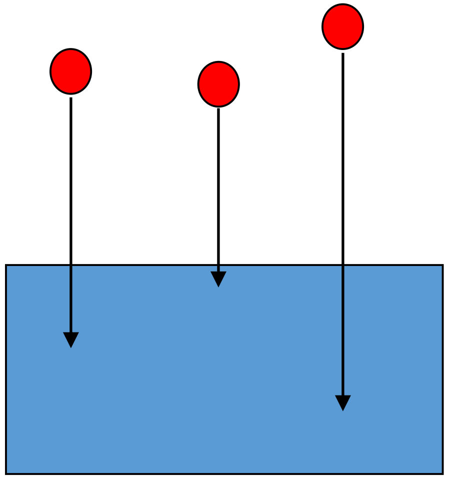
- For many problems, sum of squared distances is too sensitive to outliers
- Other problems, such as regression $\min_{x \text{ in } R^d} |Ax - b|$ often study more "robust" norms
  - E.g., $\min_{x \text{ in } R^d} |Ax - b|_1 = \sum_i |(Ax - b)_i|$
  - Sometimes, norms are not used, e.g., M-estimators: $\min_{x \in R^d} \sum_i M\big((Ax - b)_i\big)$
  - Huber estimator: $M(x) = \frac{x^2}{2\tau}$ if $x \leq \tau$, otherwise $M(x) = |x| - \tau/2$
  - Huber enjoys smoothness properties of l_2^2 and robustness properties of l_1
  - Can compute a $(1 + \epsilon)$-approximation to Huber regression in nnz(A) + poly(d/ϵ) time [CW15]
  - Similar results for regression for wide class of "nice" M-estimators [CW15]



Robust Forms of Low Rank Approximation

- (Basis Independence) if you rotate R^d by rotation matrix W, obtaining new points $a_1 W, a_2 W, \ldots, a_n W$, the cost is preserved
- This rules out approximating A by a rank-k matrix B which minimizes $\sum_i |a_i - b_i|_1$, where $b_1, \ldots, b_n$ are the rows of B
  - E.g., if B has rank 0, then $\sum_i |a_i|_1 \neq \sum_i |a_i W|_1$ for most rotations W
- Cost function studied in [DZHZ06, SV07,DV07,FL11,VX12]:

$$\min_{k-\dim V} \sum_i d(a_i, V)^p = \min_{k-\dim V} \sum_i |a_i - a_i^T V V^T|_2^p$$



- This is rotationally invariant, and

  for p in [1,2) is more robust than the SVD

Prior Work on this Cost Function

- A k-dimensional space V' is a (1+ ϵ)-approximation if
$$\sum_i d(a_i, V')^p \leq (1+\epsilon) \min_{k-\dim V} \sum_i d(a_i, V)^p$$
- For constant $1 \leq p < \infty$,
  - can output a k-dimensional space V' which is a (1+ ϵ)-approximation in $n \cdot d$ ·poly(k/ϵ) + exp(poly(k/ϵ)) time [KV07]
  - (Weak Coreset) can obtain a poly(k/ϵ)-dimensional space V' which *contains* a k-dimensional space V'' which is a (1+ ϵ)-approximation in $n \cdot d$ ·poly(k/ϵ) time [DV07, FL11]
- For p > 2,
  - the problem is NP-hard to approximate up to a fixed constant factor γ_p [DTV10, GRSW12].
  - there is a poly(nd) time algorithm achieving $\sqrt{2}\gamma_p$-approximation [DTV10]

Open Questions from Prior Work

- We are interested in $1 \leq p < 2$, since these are more robust than the SVD

1. (Exponential Term) Is the exp(poly(k/ϵ)) in the running times necessary, or is it possible to have an algorithm running in time polynomial in n,d,k,1/ϵ?
2. (Input Sparsity) Can one achieve input sparsity time, i.e., a leading order term in the time complexity of nnz(A), as in the case of p = 2?
3. (M-Estimators) What about other loss functions, e.g., M-estimators

$$\min_{k-\dim V} \sum_i M(|a_i - a_i^T V V^T|_2)$$

Can one obtain any algorithm for low rank approximation for M-estimators?

Our Contributions (Hardness)

- We show the first hardness for p in [1, 2), namely, for any p in [1,2) it is NP-hard to obtain a (1+1/d)-approximation in poly(nd) time (answers an open question of Kannan and Vempala)

- Implies there is no poly(n,d,k,1/ϵ) time algorithm unless P = NP

- Together with previous work, shows there is a "singularity" at p = 2: for every $1 \leq p < \infty$, the problem is NP-hard unless p = 2

- Open Question: we do not know if the problem is NP-hard for fixed constant ϵ

Our Contributions (Input Sparsity)

- For p in [1,2) we achieve an algorithm running in time

$$\text{nnz}(A) + (n+d)\text{poly}(k/\epsilon) + \exp(\text{poly}(k/\epsilon))$$

- nnz(A) time is required for algorithms achieving relative error, and is optimal when nnz(A) > $(n+d)\text{poly}(k/\epsilon) + \exp(\text{poly}(k/\epsilon))$

- (Weak Coreset) For p in [1,2), can find a poly(k/ϵ)-dimensional subspace V' which contains a k-dimensional subspace V'' of R^{d} which is a (1+ϵ)-approximation in nnz(A) + $(n+d)\text{poly}(k/\epsilon)$ time

Our Contributions (M-Estimators)

- We give the first results for low rank approximation with M-Estimator losses (previous empirical results in [DZHZ06])

- An M-estimator M(x) is nice if
  1. (even) $\mathrm{M(x) = M(-x)}$, with $\mathrm{M(0) = 0}$
  2. (monotonic) $\mathrm{M(a) \geq M(b)}$ for $|\mathrm{a}| \geq |\mathrm{b}|$
  3. (polynomially bounded) There is a constant $\mathrm{C_M > 0}$ so that for all $|\mathrm{a}| \geq |\mathrm{b}|$
  $$\frac{\mathrm{C_M a}}{\mathrm{b}} \leq \frac{\mathrm{M(a)}}{\mathrm{M(b)}} \leq \left(\frac{\mathrm{a}}{\mathrm{b}}\right)^2$$
  4. (square-root subadditive) $\mathrm{M(a)^{1/2} + M(b)^{1/2} \geq M(a+b)^{1/2}}$

Our Contributions for Nice M-Estimators

- For a parameter $L = (\log n)^{O(\log k)}$, we reduce the problem to

$$\min_{\mathrm{rank}(X)=k} \sum_i M(|\widehat{a_i}XB - c_i|_2),$$

where $\hat{A}, B, C$ have dimensions in $\mathrm{poly}(L, \frac{1}{\epsilon}, \log n)$, in nnz(A) log n + (n+d) poly(L/ε) time

- (Large Approximation) In O(nnz(A)) + (n+d) poly(k) time, we find a space of dimension poly(k log n) whose cost is within a factor L of the best k-dimensional space

- (Weak Coreset) In O(nnz(A)) + (n+d) poly(L/ε) time, can find a space of dimension poly(L/ε) that contains a k-dimensional space which is a $(1 + \epsilon)$-approximation

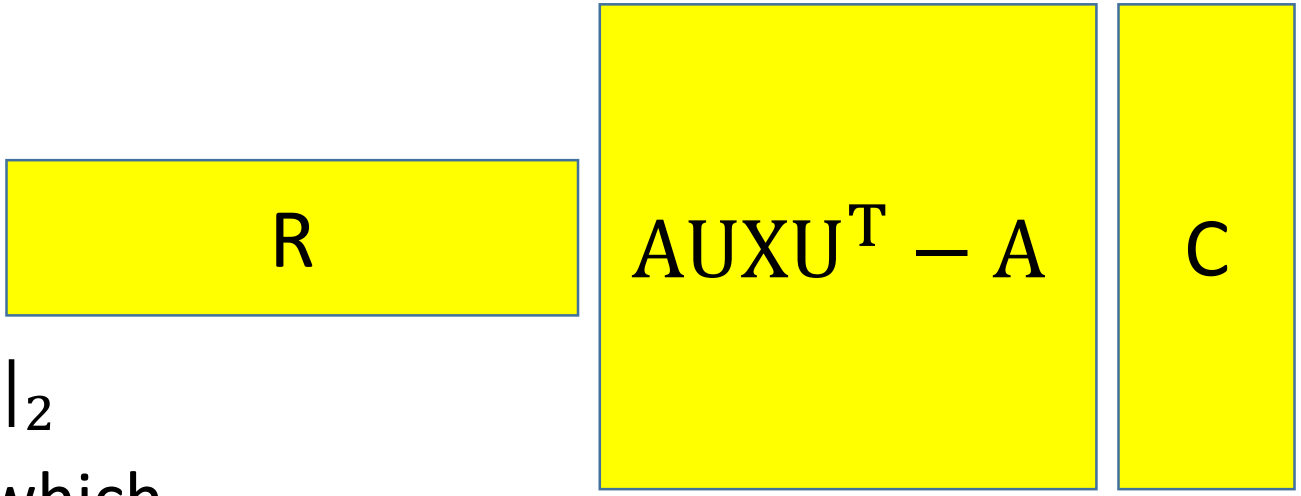
- Open Question: we do not know how to solve the small problem and avoid a factor-L approximation or a bi-criteria solution, though heuristics can be run

Talk Outline

1. Algorithm for p = 1

Due to time constraints, please see the paper for the hardness result, and adaptations of the algorithm to p in (1,2) and M-estimators

Algorithm for p = 1



- For a matrix A, let $|A|_v = \sum_i |a_i|_2$
- Would like to compute a V for which

$$|A - AVV^T|_v \leq (1+\epsilon) \min_{\mathrm{rank}(W)=k} |A - AWW^T|_v$$

- (Strategy)
  - Find poly(k/ε) x n matrix R and a d x poly(k/ε) matrix C
  - Find d x poly(k/ε) matrix U with orthonormal columns
  - If the poly(k/ε) x poly(k/ε) matrix X is the solution to

$$\min_{\mathrm{rank-k\ projectors}\ X} |RA\,UXU^T C - RAC|_v$$

then UXU^T is the desired projection matrix

Why Reduce to a Small Problem?

- Solve $\min_{\text{rank}-k \text{ projectors } X} |RA\,UXU^T C \, - RAC|_v$ using polynomial optimization

- Given c polynomial inequalities each of degree at most d in m variables: $p_1(x_1, \ldots, x_m) \geq \beta_1, \ldots, p_c(x_1, \ldots, x_m) \geq \beta_c$, can determine if there is a solution using $(cd)^{O(m)}$ arithmetic operations [BPR96]

- Since X has dimensions poly(k/ε) x poly(k/ε), one can create a small number of variables and solve the problem in exp(poly(k/ε)) time
  - Technicalities: need a lower bound on the cost given it is non-zero

Steps in Our Algorithm

- Suffices to reduce to $\min_{\text{rank}-k \text{ projectors } X} |RA\, UXU^T C - RAC|_v$
- Suppose we find a weak coreset, i.e., a subspace U of R^d of dimension poly(k/ε) which contains a k-dimensional subspace which is a (1+ε)-approximation
- Projection onto the k-dimensional subspace can be written as UXU^T where X has rank k
- Reduces the original problem to $\min_{\text{rank}(X)=k} |A\, UXU^T - A|_v$
- We are then done if we find small matrices R and C for which $\min_{\text{rank}(X)=k} |RA\, UXU^T C - RAC\Big|_v \le (1+\epsilon) \min_{\text{rank}(X)=k} |A\, UXU^T - A\Big|_v$

Sketching Matrices for the v-Norm

- Consider the problem $\min_X |XB - A|_v$ where B has rank r
- The rows x_i in the optimal X can be solved via n regression problems
$$\min_{x_i} |x_i B - a_i|_2$$
- Would like to reduce this to a smaller problem $\min_X |XBS - AS|_v$
- (Subspace Embeddings) There are d x poly(r/ε) random matrices S for which simultaneously for all x,
$$|xBS - a_i S|_2 = (1 \pm \epsilon)|xB - a_i|_2$$
with probability $\geq 1 - \mathrm{poly}\left(\frac{\epsilon}{r}\right)$
- S can be a matrix of i.i.d. Gaussians or Randomized FFT [S06]
- For faster computation, S can be the CountSketch matrix [CW13]

The CountSketch Matrix [CCFC04]

- S is d x poly(r/ε)

- S is extremely sparse!
  - Only a single non-zero per row
  - Non-zero location chosen uniformly at random
  - On that location it is 1 w.pr. ½ and -1 w.pr. ½
  - For a matrix B, $B \cdot S$ computable in nnz(B) time

- [CW13] Simultaneously for all x,

$$|xBS - a_i S|_2 = (1 \pm \epsilon)|xB - a_i|_2$$

with probability $\geq 1 - \text{poly}\left(\frac{\epsilon}{r}\right)$

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

Sketching Matrices for the v-Norm

- Want to solve $\min_X |XB - A|_v$
- The rows x_i in the optimal X can be solved via n regression problems
$$\min_{x_i} |x_i B - a_i|_2$$
- There exist d x poly(r/ε) random matrices S for which simultaneously for all x,
$$|xBS - a_i S|_2 = (1 \pm \epsilon)|xB - a_i|_2$$
with probability $\geq 1 - \text{poly}\left(\frac{\epsilon}{r}\right)$

Can we just output $X' = \text{argmin}_X |XBS - AS|_v$?

- No! To be correct on all n regression problems requires error probability 1/n, so the number of rows of S is poly(k/ε) log n, which later causes our polynomial optimization problem to have at least poly(k/ε) log n variables…

Structural Lemma

- Let X^* be the minimizer to $\min_X |XB - A|_v$
- Can show $|X^*BS - AS|_v \leq (1+\epsilon)|X^*B - A|_v$ with constant probability
- Uses a second moment argument

- For $X' = \text{argmin}_X |XBS - AS|_v$ to satisfy $|X'B - A|_v \leq (1+\epsilon)|X^*B - A|_v$, it suffices to show for all X,

$$|XBS - AS|_v \geq (1-\epsilon)|XB - A|_v$$

- (Structural Lemma) for all X, it holds that $|XBS - AS|_v \geq (1-\epsilon)|XB - A|_v$

- Intuition: S will be a subspace embedding for most $[B, A_i]$ pairs, so for most i, we will have $|X_iBS - A_iS|_v \geq (1-\epsilon)|X_iB - A_i|_v$

Structural Lemma

$|x_1 BS - a_1 S|_2$

$|x_2 BS - a_2 S|_2$

$|x_3 BS - a_3 S|_2$

…

For i = 1, …, n, say i is bad if S is not a subspace embedding for $[B, a_i]$, otherwise i is good

For a good i,
$|x_i B - a_i|_2 \geq (1-\epsilon)|x_i B - a_i|_2$

$$E[\sum_{\text{bad } i} |x_i^* B - a_i|_2] \leq \text{poly}\left(\frac{\epsilon}{r}\right) |X^* B - A|_v$$

Structural Lemma

- Previous slide shows we can condition on X^* not contracting
- What about those X for which $|x_i B - a_i|_2$ is large on those i when the subspace embedding fails?
- Suppose we additionally condition on the single event:

For all x, $|xBS|_2 = (1 \pm \epsilon)|xB|_2$

- (Triangle Inequality)
  - $|x_i BS - a_i S|_2 \geq |x_i BS - x_i^* BS|_2 - |x_i^* BS - a_i S|_2$
    $\geq (1-\epsilon)|x_i B - x_i^* B|_2 - |x_i^* BS - a_i S|_2$
    $\geq (1-\epsilon)(|x_i B - a_i|_2 - |x_i^* B - a_i|_2) - |x_i^* BS - a_i S|_2$
    $\geq (1-\epsilon)|x_i B - a_i|_2 - |x_i^* B - a_i|_2 - |x_i^* BS - a_i S|_2$
- $\sum_{\text{bad } i} |x_i^* B - a_i|_2$ is small
- $\sum_{\text{bad } i} |x_i^* BS - a_i S|_2$ is small, otherwise $|X^* BS - AS|_v > (1+\epsilon)|X^* B - A|_v$

Using the Structural Lemma

- Two steps of our algorithm:
  - Find a weak coreset to reduce the original problem to
    $$\min_{\mathrm{rank}(X)=k} |A\, UXU^T - A|_v$$
  - Find small matrices R and C on the left and right for which

$$\min_{\mathrm{rank}(X)=k} \left|RA\, UXU^T C - RAC\right|_v \le (1+\epsilon) \min_{\mathrm{rank}(X)=k} \left|A\, UXU^T - A\right|_v$$

- By structural lemma, if X' = arg $\min_{\mathrm{rank}(X)=k} |A\, UXU^T S - AS|_v$ then
  $$\left|AUX'U^T - A\right|_v \ge (1-\epsilon) \min_{\mathrm{rank}(X)=k} |A\, UXU^T - A|_v$$
- Set C = S

Finishing the Small Matrices Step

- Given a weak coreset, we've reduced the problem to $\min_{\mathrm{rank}(X)=k} |A\, UXU^T S \ - AS|_v$
- Dvoretsky's theorem: for an appropriately scaled $d \times \frac{d}{\epsilon^2}$ Gaussian matrix G, the mapping $y \rightarrow yG$ satisfies w.h.p, simultaneously for all y, $|yG|_1 = (1 \pm \epsilon)|y|_2$
- $\left|AUXU^T S \ - AS\right|_v = (1 \pm \epsilon)\left|AUXU^T G \ - ASG\right|_1$, where $|.|_1$ is entry-wise 1-norm
- Columns of $AUXU^T G \ - ASG$ are in a poly $\left(\frac{k}{\epsilon}\right)$-dimensional subspace so we can apply known sampling for the 1-norm to sample poly $\left(\frac{k}{\epsilon}\right)$ rows R so that for all X,

$$\left|RAUXU^T G \ - RASG\right|_1 = (1 \pm \epsilon)\left| AUXU^T G \ - ASG\right|_1,\text{ or}$$

$$\left|RAUXU^T \ - RAS\right|_v = (1 \pm \epsilon)\left| AUXU^T \ - AS\right|_v$$

The Weak Coreset

- Two steps of our algorithm:
  - Find a weak coreset to reduce the original problem to

$$\min_{\mathrm{rank}(X)=k} |A\,UXU^T - A|_v$$

  - Find small matrices R and C on the left and right for which

$$\min_{\mathrm{rank}(X)=k} \left|RA\,UXU^T C - RAC\right|_v \leq (1+\epsilon) \min_{\mathrm{rank}(X)=k} \left|A\,UXU^T - A\right|_v$$

- Done with finding small matrices, we just need a weak coreset

The Weak Coreset

- Structural Lemma: if $X' = \underset{X}{\text{argmin}}|XBS - AS|_v$, then with large constant probability, $|X'B - A|_v \leq (1+\epsilon)|X^*B - A|_v$, where the number of rows of S is poly(rank(B)/ε)

- Apply structural lemma with $B = A_k$, where A_k is the best rank-k approximation to A in the v-norm
  - S has poly(k/ε) rows
  - Since $X' = \underset{X}{\text{argmin}}|XA_kS - AS|_v$ satisfies $X_i' = AS\,(A_kS)^-$, there is a rank-k space in the column space of AS which is a $(1+\epsilon)$-approximation

- If X' = arg $\underset{\text{rank}-k\,X}{\min}|ASX - A|_v$, it is a $(1+\epsilon)$-approximation

The Weak Coreset

- We've reduced the original problem to $\min_{\text{rank}-k\ X} |ASX - A|_v$

- By known sampling techniques for ℓ_1 and Dvoretsky's theorem, can quickly find a matrix T for which if X'' = arg $\min_{\text{rank}-k\ X} |TASX - TA|_v$,

  then $|ASX'' - A|_v \leq 4 \min_{\text{rank}-k\ X} |ASX - A|_v$

- X'' = arg $\min_{\text{rank}-k\ X} |TASX - TA|_v$ is in the row span of TA

- Row span of TA is a 4-approximation

The Weak Coreset

- (Adaptive Sampling) [DV07] shows how to take a $\mathrm{poly}\left(\frac{k}{\epsilon}\right)$-dimensional subspace TA of R^d, which is an O(1)-approximation, and obtain a $\mathrm{poly}\left(\frac{k}{\epsilon}\right)$-dimensional subspace of R^d containing a $(1+\epsilon)$-approximation

- We show how to implement this procedure in nnz(A) time, improving the previous nnz(A)*poly(k/ε) time

- [DV07] sample a row a_i of A proportional to its distance to TA, then sample another row a_j of A proportional to its distance to span(TA, a_i), etc. We show we can sample all rows proportional to their distance to the original TA
  - Our sampling is non-adaptive

Algorithm Summary

1. Compute AS for a d x poly(k/ϵ) CountSketch matrix S
2. Compute TAS where T samples poly(k/ϵ) rows of AS using known sampling for ℓ_1
3. Feed TA into a non-adaptive sampling algorithm to obtain a weak coreset U, reducing the problem to
$$\min_{\mathrm{rank}(X)=k} \left| A\, UXU^T - A \right|_v$$
4. Find small matrices R and C to reduce the problem to
$$\min_{\mathrm{rank}(X)=k} \left| RA\, UXU^T C - RAC \right|_v$$
5. Solve the problem using polynomial optimization

Conclusions

- First input sparsity time algorithm for robust low rank approximation with cost measure

$$\min_{k-\dim V} \sum_i d(a_i, V)^p = \min_{k-\dim V} \sum_i |a_i - a_i^T V V^T|_2^p$$

- Generalize the algorithm to give the first near-input sparsity time algorithms for a wide class of M-estimators
- Show first hardness for p in [1,2), so there can be no polynomial time algorithm in n, d, k, and 1/ε unless P = NP
  - Helps explain why we need the exp(poly(k/ε)) term in our time complexity
- Improve [CW15] for regression with M-estimator losses, showing for a wide class how to obtain (1+ϵ)-approximation in nnz(A) time