CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2024

Lecture 10: Sampling and preconditioning for least squares

Outline

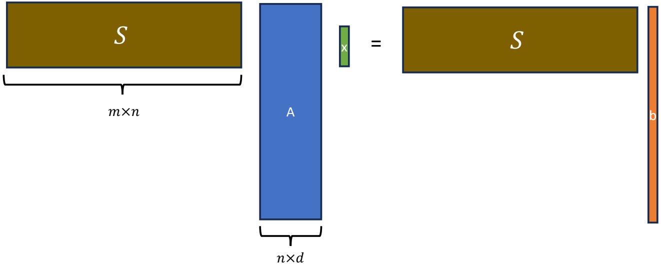
Sketch and solve

Recall:

- Generate a sketching matrix $\boldsymbol{S} \in \mathbb{R}^{m \times n}$.
- Compute sketches $\boldsymbol{SA}$ and $\boldsymbol{Sb}$.
- Solve:

$$\tilde{\boldsymbol{x}} = \min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{SAx} - \boldsymbol{Sb}\|_2^2.$$

- Typically, $m = \text{poly}(d/\epsilon)$.



Subspace embedding for sketch and solve

Sketch and solve

Suppose $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a subspace ϵ-embedding for $span([\boldsymbol{A}\ b])$.
Let,

$$\boldsymbol{x}^* = \min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$$
$$\tilde{\boldsymbol{x}} = \min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{S}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_2,$$

for $\epsilon \leq 1/3$, we have

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2 \leq (1 + 3\epsilon)\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|_2$$

Implies, we have $O(1/\epsilon^2)$ dependency on the error tolerance.

Alternate proof

Sketch and solve

If $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a Countsketch matrix with $m = O(d^2/\epsilon)$ or SRHT with $m = O(d \log d/\epsilon)$, or Gaussian sketch with $m = O(d/\epsilon)$, then

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2 \leq (1+\epsilon)\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|_2$$

Alternate proof

Sketch and solve

If $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a Countsketch matrix with $m = O(d^2/\epsilon)$ or SRHT with $m = O(d \log d/\epsilon)$, or Gaussian sketch with $m = O(d/\epsilon)$, then

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2 \leq (1+\epsilon)\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|_2$$

Proof: Let us consider an orthonormal basis $\boldsymbol{U}$ for $\boldsymbol{A}$.
Let, $\boldsymbol{U}\tilde{\boldsymbol{y}} = \boldsymbol{A}\tilde{\boldsymbol{x}}$ and $\boldsymbol{U}\boldsymbol{y}^* = \boldsymbol{A}\boldsymbol{x}^*$. Then,

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|_2^2 + \|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\|_2^2$$

and

$$\|\boldsymbol{U}\tilde{\boldsymbol{y}} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{b}\|_2^2 + \|\boldsymbol{U}\tilde{\boldsymbol{y}} - \boldsymbol{U}\boldsymbol{y}^*\|_2^2$$

Need to show that $\|\boldsymbol{U}(\tilde{\boldsymbol{y}} - \boldsymbol{y}^*)\|_2^2 = \|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*\|_2^2 = O(\epsilon)\|\boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{b}\|_2^2$.

For a subspace embedding $\boldsymbol{S}$, we have

$$\|\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U} - \boldsymbol{I}\|_2 \leq \frac{1}{2}.$$

Hence,

$\|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*\|_2 \leq$

For a subspace embedding $\boldsymbol{S}$, we have

$$\|\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U} - \boldsymbol{I}\|_2 \leq \frac{1}{2}.$$

Hence,

$\|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*\|_2 \leq$

By normal equation, we have

$$\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U}\tilde{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{b},$$

so,

$$\|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*\|_2 \leq 2\|\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{b})\|_2.$$

For a subspace embedding $\boldsymbol{S}$, we have

$$\|\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U} - \boldsymbol{I}\|_2 \leq \frac{1}{2}.$$

Hence,

$\|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*\|_2 \leq$

By normal equation, we have

$$\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U}\tilde{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{b},$$

so,

$$\|\tilde{\boldsymbol{y}} - \boldsymbol{y}^*\|_2 \leq 2\|\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{b})\|_2.$$

For $\boldsymbol{S}$ with the choice of m, we have

$$\Pr\left[\|\boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{b})\|_F \geq 3\frac{\sqrt{\epsilon}}{d}\|\boldsymbol{U}\|_F\|\boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{b}\|_F\right] \leq \delta.$$

Sampling for least squares

- We can consider sampling rows of $[\boldsymbol{A}\ \boldsymbol{b}]$.
- Recall leverage scores.

Leverage scores

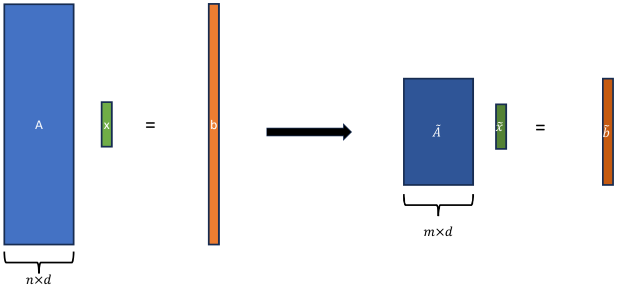
Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, and an orthonormal basis $\boldsymbol{U}$ for $span(\boldsymbol{A})$, for $i \in [n]$, the ith *leverage score*

$$\ell_i(\boldsymbol{A}) = \sup_{\boldsymbol{x}} \frac{(\boldsymbol{A}_{i*}\boldsymbol{x})^2}{\|\boldsymbol{A}\boldsymbol{x}\|^2} = \|\boldsymbol{U}_{i*}\|^2.$$

Sampling for least squares

Algorithm:

- Compute the row-leverage scores of $\boldsymbol{A}$, ℓ_i, $i = 1, \ldots, n$.
- Pick m rows of $\boldsymbol{A}$ and the corresponding elements of $\boldsymbol{b}$ with respect to the probabilities $p_i = \ell_i / d$ to $i \in [n]$.
- Rescale sampled rows of $\boldsymbol{A}$ and sampled elements of $\boldsymbol{b}$ by $1/\sqrt{mp_i}$.
- Solve the induced problem.



Leverage score sampling is subspace embedding

Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $r = \mathrm{rank}(\boldsymbol{A})$, and $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ be a sampling matrix with probabilities $p_i = \ell_i/r$, and $\boldsymbol{S}_{i*} = \boldsymbol{e}_j/\sqrt{mp_j}$ with $\Pr(j = i) = p_i$. If $m = O(r\log(r/\delta)/\epsilon^2)$, then $\boldsymbol{S}$ is ϵ-subspace embedding of $span(\boldsymbol{A})$ with probability $1-\delta$.

Leverage score sampling is subspace embedding

Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $r = \operatorname{rank}(\boldsymbol{A})$, and $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ be a sampling matrix with probabilities $p_i = \ell_i/r$, and $\boldsymbol{S}_{i*} = \boldsymbol{e}_j/\sqrt{mp_j}$ with $\Pr(j = i) = p_i$. If $m = O(r\log(r/\delta)/\epsilon^2)$, then $\boldsymbol{S}$ is ϵ-subspace embedding of $span(\boldsymbol{A})$ with probability $1-\delta$.

Proof: Let $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ be orthonormal with $span(\boldsymbol{U}) = span(\boldsymbol{A})$.

For $k \in [m]$, let $\boldsymbol{X}_k = m\boldsymbol{U}^\top[\boldsymbol{S}_{k*}]^\top\boldsymbol{S}_{k*}\boldsymbol{U} - \boldsymbol{I}$, so

$$\frac{1}{m}\sum_k \boldsymbol{X}_k = \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U} - \boldsymbol{I},$$

and for ϵ-embedding, we need to bound its spectral norm.

Matrix Chernoff

Let $\boldsymbol{X}_k$ for $k \in [m]$ be i.i.d copies of symmetric random $\boldsymbol{X} \in \mathbb{R}^{r \times r}$ with $\gamma, \sigma^2 > 0$, $\mathbb{E}[\boldsymbol{X}] = 0$, $\|\boldsymbol{X}\|_2 \leq \gamma$, and $\|\mathbb{E}[\boldsymbol{X}^2]\|_2 \leq \sigma^2$. Then for $\epsilon > 0$,

$$\Pr(\|\frac{1}{m}\sum_k \boldsymbol{X}_k\|_2 \geq \epsilon) \leq 2r \exp(-m\epsilon^2/(\sigma^2 + \gamma\epsilon/3)).$$

Apply to

$$\boldsymbol{X} = \frac{1}{p_j}[\boldsymbol{U}_{j*}]^\top \boldsymbol{U}_{j*} - \boldsymbol{I} \text{ with } Pr(j = i) = p_i = \ell_i/r = \|\boldsymbol{U}_{i*}\|_2^2/r.$$

We have

$\mathbb{E}[\boldsymbol{X}] =$

$\|\boldsymbol{X}\|_2 \leq$

$\mathbb{E}[\boldsymbol{X}^2] =$

Matrix Chernoff

Let $\boldsymbol{X}_k$ for $k \in [m]$ be i.i.d copies of symmetric random $\boldsymbol{X} \in \mathbb{R}^{r \times r}$ with $\gamma, \sigma^2 > 0$, $\mathbb{E}[\boldsymbol{X}] = 0$, $\|\boldsymbol{X}\|_2 \leq \gamma$, and $\|\mathbb{E}[\boldsymbol{X}^2]\|_2 \leq \sigma^2$. Then for $\epsilon > 0$,

$$\Pr(\|\frac{1}{m}\sum_k \boldsymbol{X}_k\|_2 \geq \epsilon) \leq 2r\exp(-m\epsilon^2/(\sigma^2 + \gamma\epsilon/3)).$$

Apply to

$$\boldsymbol{X} = \frac{1}{p_j}[\boldsymbol{U}_{j*}]^\top \boldsymbol{U}_{j*} - \boldsymbol{I} \text{ with } Pr(j = i) = p_i = \ell_i/r = \|\boldsymbol{U}_{i*}\|_2^2/r.$$

We have

$\mathbb{E}[\boldsymbol{X}] =$

$\|\boldsymbol{X}\|_2 \leq$

$\mathbb{E}[\boldsymbol{X}^2] =$

so, $\|\mathbb{E}[\boldsymbol{X}^2]\|_2 \leq r - 1$.

Computing the leverage scores

- To compute the leverage scores exactly, we need $\boldsymbol{U}$, i.e., compute the SVD of $\boldsymbol{A}$.
- Naive cost $O(nd^2)$.
- Can be approximately estimated in $O(nnz(\boldsymbol{A})\log n + d^3)$ time.

Algorithm:
Given $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, a subspace ϵ-embedding $\boldsymbol{S}_1 \in \mathbb{R}^{m\times n}$ for $\boldsymbol{A}$, and a JL matrix $\boldsymbol{S}_2 \in \mathbb{R}^{d\times m'}$. so that $\|\boldsymbol{x}^\top \boldsymbol{S}_2 = (1\pm\epsilon)\|\boldsymbol{x}\|$ for n vectors, so $m' = O(\log(n)/\epsilon^2)$, then:

1. $\boldsymbol{W} = \boldsymbol{S}_1\boldsymbol{A}$; //compute sketch
2. $[\boldsymbol{Q},\boldsymbol{R}] = qr(\boldsymbol{W})$; // change of basis
3. $\boldsymbol{Z} = \boldsymbol{A}(\boldsymbol{R}^{-1}\boldsymbol{S}_2)$; // sketch of $\boldsymbol{A}\boldsymbol{R}^{-1}$
4. return $\|\boldsymbol{Z}_{i*}\|_2^2$ for $i \in [n]$

Correctness

- $\boldsymbol{AR}^{-1}$ has singular values in $[1-\epsilon, 1+\epsilon]$.
  For all $\boldsymbol{x}$, $\|\boldsymbol{AR}^{-1}\boldsymbol{x}\| = (1 \pm \epsilon)\|\boldsymbol{S}_1\boldsymbol{AR}^{-1}\boldsymbol{x}\| = (1 \pm \epsilon)\|\boldsymbol{Qx}\| = (1 \pm \epsilon)\|\boldsymbol{x}\|$
- Let $\boldsymbol{U}$ be orthonormal with $span(\boldsymbol{U}) = span(\boldsymbol{A})$.
- $\boldsymbol{AR}^{-1}$ is like $\boldsymbol{U}$.

Correctness

- $\boldsymbol{AR}^{-1}$ has singular values in $[1-\epsilon, 1+\epsilon]$.
  For all $\boldsymbol{x}$, $\|\boldsymbol{AR}^{-1}\boldsymbol{x}\| = (1 \pm \epsilon)\|\boldsymbol{S}_1\boldsymbol{AR}^{-1}\boldsymbol{x}\| = (1 \pm \epsilon)\|\boldsymbol{Qx}\| = (1 \pm \epsilon)\|\boldsymbol{x}\|$
- Let $\boldsymbol{U}$ be orthonormal with $span(\boldsymbol{U}) = span(\boldsymbol{A})$.
- $\boldsymbol{AR}^{-1}$ is like $\boldsymbol{U}$.
- Pick $\boldsymbol{T}$ such that $\boldsymbol{AR}^{-1}\boldsymbol{T} = \boldsymbol{U}$.
- $\boldsymbol{T}$ has singular values $(1 \pm \epsilon)$.

Correctness

- $\boldsymbol{AR}^{-1}$ has singular values in $[1-\epsilon, 1+\epsilon]$.
  For all $\boldsymbol{x}$, $\|\boldsymbol{AR}^{-1}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{S}_1\boldsymbol{AR}^{-1}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{Qx}\| = (1\pm\epsilon)\|\boldsymbol{x}\|$
- Let $\boldsymbol{U}$ be orthonormal with $span(\boldsymbol{U}) = span(\boldsymbol{A})$.
- $\boldsymbol{AR}^{-1}$ is like $\boldsymbol{U}$.
- Pick $\boldsymbol{T}$ such that $\boldsymbol{AR}^{-1}\boldsymbol{T} = \boldsymbol{U}$.
- $\boldsymbol{T}$ has singular values $(1\pm\epsilon)$.
  For all $\boldsymbol{x}$,
  $\|\boldsymbol{Tx}\| = \|\boldsymbol{QTx}\| = \|\boldsymbol{S}_1\boldsymbol{AR}^{-1}\boldsymbol{Tx}\| = (1\pm\epsilon)\|\boldsymbol{AR}^{-1}\boldsymbol{Tx}\| = (1\pm\epsilon)\|\boldsymbol{Ux}\| = (1\pm\epsilon)\|\boldsymbol{x}\|$
- Then $\boldsymbol{T}^{-1}$ has singular values $(1\pm 2\epsilon)$ for $\epsilon < 1/2$.

Correctness

- $\boldsymbol{A}\boldsymbol{R}^{-1}$ has singular values in $[1-\epsilon, 1+\epsilon]$.
  For all $\boldsymbol{x}$, $\|\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{S}_1\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{Q}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{x}\|$
- Let $\boldsymbol{U}$ be orthonormal with $span(\boldsymbol{U}) = span(\boldsymbol{A})$.
- $\boldsymbol{A}\boldsymbol{R}^{-1}$ is like $\boldsymbol{U}$.
- Pick $\boldsymbol{T}$ such that $\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{T} = \boldsymbol{U}$.
- $\boldsymbol{T}$ has singular values $(1\pm\epsilon)$.
  For all $\boldsymbol{x}$,
  $\|\boldsymbol{T}\boldsymbol{x}\| = \|\boldsymbol{Q}\boldsymbol{T}\boldsymbol{x}\| = \|\boldsymbol{S}_1\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{T}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{T}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{U}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{x}\|$
- Then $\boldsymbol{T}^{-1}$ has singular values $(1\pm 2\epsilon)$ for $\epsilon < 1/2$.
- Hence, our output $\|\boldsymbol{e}_i^\top\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{S}_2\|^2 = (1\pm O(\epsilon))\|\boldsymbol{e}_i^\top\boldsymbol{U}\|^2$.

Correctness

- $\boldsymbol{A}\boldsymbol{R}^{-1}$ has singular values in $[1-\epsilon, 1+\epsilon]$.
  For all $\boldsymbol{x}$, $\|\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{S}_1\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{Q}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{x}\|$
- Let $\boldsymbol{U}$ be orthonormal with $span(\boldsymbol{U}) = span(\boldsymbol{A})$.
- $\boldsymbol{A}\boldsymbol{R}^{-1}$ is like $\boldsymbol{U}$.
- Pick $\boldsymbol{T}$ such that $\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{T} = \boldsymbol{U}$.
- $\boldsymbol{T}$ has singular values $(1\pm\epsilon)$.
  For all $\boldsymbol{x}$,
  $\|\boldsymbol{T}\boldsymbol{x}\| = \|\boldsymbol{Q}\boldsymbol{T}\boldsymbol{x}\| = \|\boldsymbol{S}_1\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{T}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{T}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{U}\boldsymbol{x}\| = (1\pm\epsilon)\|\boldsymbol{x}\|$
- Then $\boldsymbol{T}^{-1}$ has singular values $(1\pm 2\epsilon)$ for $\epsilon < 1/2$.
- Hence, our output $\|\boldsymbol{e}_i^\top\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{S}_2\|^2 = (1\pm O(\epsilon))\|\boldsymbol{e}_i^\top\boldsymbol{U}\|^2$.

$$
\begin{aligned}
\|\boldsymbol{e}_i^\top\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{S}_2\|^2 &= (1\pm\epsilon)\|\boldsymbol{e}_i^\top\boldsymbol{A}\boldsymbol{R}^{-1}\|^2 = (1\pm\epsilon)\|\boldsymbol{e}_i^\top\boldsymbol{U}\boldsymbol{T}^{-1}\|^2 \\
&= (1\pm\epsilon)(1\pm 2\epsilon)\|\boldsymbol{e}_i^\top\boldsymbol{U}\|^2
\end{aligned}
$$

Computational cost

1. $\boldsymbol{W} = \boldsymbol{S}_1 \boldsymbol{A}$; $//O(nnz(\boldsymbol{A})s$
2. $[\boldsymbol{Q}, \boldsymbol{R}] = qr(\boldsymbol{W})$; $//\ O(d^2 m)$
3. $\boldsymbol{Z} = \boldsymbol{A}(\boldsymbol{R}^{-1}\boldsymbol{S}_2)$; $//\ O(d^2 m' + nnz(\boldsymbol{A})m')$
4. return $\|\boldsymbol{Z}_{i*}\|_2^2$ for $i \in [n]$ $//\ O(nm')$

Computational cost

1. $\boldsymbol{W} = \boldsymbol{S}_1\boldsymbol{A}$; $\quad //O(nnz(\boldsymbol{A})s$
2. $[\boldsymbol{Q}, \boldsymbol{R}] = qr(\boldsymbol{W})$; $\quad //\ O(d^2m)$
3. $\boldsymbol{Z} = \boldsymbol{A}(\boldsymbol{R}^{-1}\boldsymbol{S}_2)$; $\quad //\ O(d^2m' + nnz(\boldsymbol{A})m')$
4. return $\|\boldsymbol{Z}_{i*}\|_2^2$ for $i \in [n]$ $//\ O(nm')$

If $\boldsymbol{A}$ is dense, we use SRHT and fast JL.
If $\boldsymbol{A}$ is sparse, we can use OSNAP.
Total cost is :

$$O(nnz(\boldsymbol{A})(m' + s) + d^2(m + m') = O((nnz(\boldsymbol{A})\log n + d^3 \log d)/\epsilon^2).$$

Further Reading:
Drineas, Petros, et al. "Fast approximation of matrix coherence and statistical leverage." The Journal of Machine Learning Research 13.1 (2012): 3475-3506.

Preconditioning for least squares

- Solving least squares regression exactly requires $O(nd^2 + d^3)$ cost.
- Using sketching or sampling : $O((nnz(\boldsymbol{A}) \log n + d^3 \log d)/\epsilon)$.
- However, we only get an approximate solution:

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2 \leq (1+\epsilon)\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|_2$$

- For *machine precision regression*, we need reduce the dependence on ϵ to logarithmic.
- With iterative methods, such as the general class of Krylov or conjugate-gradient type algorithms :

$$\frac{\|\boldsymbol{A}(\boldsymbol{x}^{(m)} - \boldsymbol{x}^*)\|^2}{\|\boldsymbol{A}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*)\|^2} \leq 2\left(\frac{\sqrt{\kappa(\boldsymbol{A}^\top \boldsymbol{A})} - 1}{\sqrt{\kappa(\boldsymbol{A}^\top \boldsymbol{A})} + 1}\right)^m.$$

So, need $m = O(\kappa(\boldsymbol{A}) \log(1/\epsilon))$ to get an ϵ error.

Preconditioning for least squares

- Pre-conditioning reduces the number of iterations needed for a given accuracy.
- Find a non-singular matrix $\boldsymbol{R}$, such that $\kappa((\boldsymbol{A}\boldsymbol{R}^{-1})^\top \boldsymbol{A}\boldsymbol{R}^{-1})$ is small.
- Applying CG method to $\boldsymbol{A}\boldsymbol{R}^{-1}$ would converge quickly.

Preconditioning for least squares

- Pre-conditioning reduces the number of iterations needed for a given accuracy.
- Find a non-singular matrix $\boldsymbol{R}$, such that $\kappa((\boldsymbol{A}\boldsymbol{R}^{-1})^\top \boldsymbol{A}\boldsymbol{R}^{-1})$ is small.
- Applying CG method to $\boldsymbol{A}\boldsymbol{R}^{-1}$ would converge quickly.
- Idea is similar to approximate leverage scores computation.

Apply a (sparse) subspace embedding matrix $\boldsymbol{S}$ to $\boldsymbol{A}$.
Compute $\boldsymbol{R}$ as $[\boldsymbol{Q}, \boldsymbol{R}] = qr(\boldsymbol{S}\boldsymbol{A})$.
We know that $\boldsymbol{A}\boldsymbol{R}^{-1}$ has singular values in$[1-\epsilon_0, 1+\epsilon_0]$ (almost orthonormal).

$$\kappa(\boldsymbol{A}\boldsymbol{R}^{-1}) \leq \frac{1+\epsilon_0}{1-\epsilon_0}.$$

After m iterations of CG, we have: $\|\boldsymbol{A}\boldsymbol{R}^{-1}(\boldsymbol{x}^{(m)} - \boldsymbol{x}^*)\|^2 \leq 2\epsilon_0^m \|\boldsymbol{x}^*\|^2$

Iterative Refinement

Given $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\boldsymbol{b} \in \mathbb{R}^n$, and a subspace ϵ_0-embedding $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ for $\boldsymbol{A}$,

1. $m = O(\log(1/\epsilon))$
2. $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{A}$;
3. $[\boldsymbol{Q}, \boldsymbol{R}] = qr(\boldsymbol{W})$;
4. $\boldsymbol{x}^{(0)} \leftarrow 0$;
5. for $j = 0, 1, \dots, m$:

$$\boldsymbol{x}^{(j+1)} \leftarrow \boldsymbol{x}^{(j)} + (\boldsymbol{R}^\top)^{-1}\boldsymbol{A}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}^{(j)})$$

6. return $\boldsymbol{R}^{-1}\boldsymbol{x}^{(m+1)}$

Cost:

For SRHT or OSNAP: $O(nnz(\boldsymbol{A})\log(n/\epsilon) + d^3\log^2 d + d^2\log(1/\epsilon))$

For Countsketch: $O((nnz(\boldsymbol{A}) + d^4)\log(1/\epsilon))$.

Sketch based preconditioning

Let $\boldsymbol{x}^{(j+1)} \leftarrow \boldsymbol{x}^{(j)} + (\boldsymbol{R}^\top)^{-1}\boldsymbol{A}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}^{(j)})$.
We have

$$\begin{aligned} \boldsymbol{A}\boldsymbol{R}^{-1}(\boldsymbol{x}^{(j+1)} - \boldsymbol{x}^*) &= \boldsymbol{A}\boldsymbol{R}^{-1}\left(\boldsymbol{x}^{(j)} + (\boldsymbol{R}^\top)^{-1}\boldsymbol{A}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}^{(j)}) - \boldsymbol{x}^*\right) \\ &= \\ &= \end{aligned}$$

Sketch based preconditioning

Let $\boldsymbol{x}^{(j+1)} \leftarrow \boldsymbol{x}^{(j)} + (\boldsymbol{R}^\top)^{-1}\boldsymbol{A}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}^{(j)})$.
We have

$$\boldsymbol{A}\boldsymbol{R}^{-1}(\boldsymbol{x}^{(j+1)} - \boldsymbol{x}^*) = \boldsymbol{A}\boldsymbol{R}^{-1}\left(\boldsymbol{x}^{(j)} + (\boldsymbol{R}^\top)^{-1}\boldsymbol{A}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{x}^{(j)}) - \boldsymbol{x}^*\right)$$

$$=$$

$$=$$

where $\boldsymbol{A}\boldsymbol{R}^{-1} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$. We know $\boldsymbol{A}\boldsymbol{R}^{-1}$ has singular values in$[1-\epsilon_0, 1+\epsilon_0]$.
So, diagonal entries of $\Sigma - \Sigma^3$ are at most $\sigma_i(1-(1-\epsilon_0)^2) \leq 3\sigma_i\epsilon_0$ for $\epsilon_0 \leq 1$. Hence,

$$\|\boldsymbol{A}\boldsymbol{R}^{-1}(\boldsymbol{x}^{(m+1)} - \boldsymbol{x}^*)\| \leq 3\epsilon_0\|\boldsymbol{A}\boldsymbol{R}^{-1}(\boldsymbol{x}^{(m)} - \boldsymbol{x}^*)\|$$

and by choosing $\epsilon_0 = 1/2$, say, $O(\log(1/\epsilon))$ iterations suffice to attain ϵ relative error.

Further Reading

- Avron, Haim, Petar Maymounkov, and Sivan Toledo. "Blendenpik: Supercharging LAPACK's least-squares solver." SIAM Journal on Scientific Computing 32.3 (2010): 1217-1236.
- Clarkson, Kenneth L., and David P. Woodruff. "Low-rank approximation and regression in input sparsity time." Journal of the ACM (JACM) 63.6 (2017): 1-45.

Questions?