Randomized linear algebra

Yuxin Chen

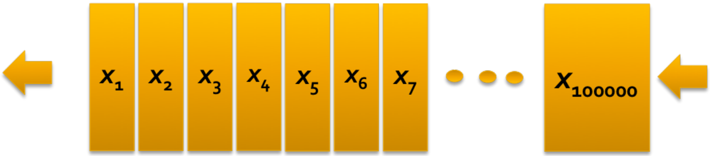
Wharton Statistics & Data Science, Spring 2022

Outline

- Approximate matrix multiplication
- Least squares approximation
- Low-rank matrix approximation

Main reference: "*Lecture notes on randomized linear algebra*,"
Michael W. Mahoney, 2016

Efficient large-scale data processing



When processing large-scale data (in particular, streaming data), we desire methods that can be performed with

- a few (e.g. one or two) passes of data
- limited memory (so impossible to store all data)
- low computational complexity

Key idea: dimension reduction via random sketching

- **random sampling:** randomly downsample data
  - often relies on the information of data
- **random projection:** rotates / projects data to lower dimensions
  - often data-agnostic

Approximate matrix multiplication

Matrix multiplication: a fundamental algebra task

Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$, compute or approximate $\boldsymbol{AB}$

Algorithm 6.1 Vanilla algorithm for matrix multiplication

for $i = 1, \cdots, m$ **do**
    for $k = 1, \cdots, n$ **do**
        $M_{i,k} = \boldsymbol{A}_{i,:}\boldsymbol{B}_{:,k}$
return $\boldsymbol{M}$

Computational complexity: $O(mnp)$, or $O(n^3)$ if $m = n = p$

For simplicity, we shall assume $m = n = p$ unless otherwise noted.

Faster matrix multiplication?

- **Strassen algorithms:** exact matrix multiplication
  - Computational complexity $\approx O(n^{2.8})$
  - For various reasons, rarely used in practice

- Approximate solution?

A simple randomized algorithm

View $\boldsymbol{AB}$ as a sum of rank-one matrices (or outer products)

$$\boldsymbol{AB} = \sum_{i=1}^{n} \boldsymbol{A}_{:,i} \boldsymbol{B}_{i,:}$$

Idea: randomly sample L rank-one components

Algorithm 6.2 Basic randomized algorithm for matrix multiplication

1: **for** $l = 1, \cdots, L$ **do**

2: Pick $i_l \in \{1, \cdots, n\}$ i.i.d. with prob. $\mathbb{P}\{i_l = k\} = p_k$

3: **return**

$$\boldsymbol{M} = \sum_{l=1}^{L} \frac{1}{L p_{i_l}} \boldsymbol{A}_{:,l} \boldsymbol{B}_{l,:}$$

- $\{p_k\}$: importance sampling probabilities

A simple randomized algorithm

Rationale: $\boldsymbol{M}$ is an *unbiased* estimate of $\boldsymbol{AB}$, i.e.

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{M}\right] &= \sum_{l=1}^{L}\sum_{k}\mathbb{P}\left\{i_l = k\right\}\frac{1}{Lp_k}\boldsymbol{A}_{:,k}\boldsymbol{B}_{k,:} \\ &= \sum_{k}\boldsymbol{A}_{:,k}\boldsymbol{B}_{k,:} = \boldsymbol{AB}\end{aligned}$$

Clearly, the approximation error (e.g. $\|\boldsymbol{AB}-\boldsymbol{M}\|$) depends on $\{p_k\}$

Importance sampling probabilities

- **Uniform sampling** ($p_k \equiv \frac{1}{n}$): one can choose the sampling set before looking at data, so it's implementable via 1 pass over data

Intuitively, one may prefer biasing towards larger rank-1 components

- **Nonuniform sampling**

$$p_k = \frac{\|\boldsymbol{A}_{:,k}\|_2 \|\boldsymbol{B}_{k,:}\|_2}{\sum_l \|\boldsymbol{A}_{:,l}\|_2 \|\boldsymbol{B}_{l,:}\|_2}$$

  - $\{p_k\}$ can be computed using one pass and $O(n)$ memory

Optimal sampling probabilities?

Let's measure the approximation error by $\mathbb{E}\left[\|\boldsymbol{M}-\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2\right]$

As it turns out, $\mathbb{E}\left[\|\boldsymbol{M}-\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2\right]$ is minimized by

$$p_k = \frac{\|\boldsymbol{A}_{:,k}\|_2\|\boldsymbol{B}_{k,:}\|_2}{\sum_l \|\boldsymbol{A}_{:,l}\|_2\|\boldsymbol{B}_{l,:}\|_2} \tag{6.1}$$

Thus, we call (6.1) the optimal sampling probabilities

Justification of the optimality of (6.1)

Since $\mathbb{E}[\boldsymbol{M}] = \boldsymbol{A}\boldsymbol{B}$, one has

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{M}-\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2\right] &= \mathbb{E}\left[\sum_{i,j}\left(M_{i,j}-\boldsymbol{A}_{i,:}\boldsymbol{B}_{:,j}\right)^2\right] = \sum_{i,j}\mathsf{Var}[M_{i,j}] \\
&= \frac{1}{L}\sum_k\sum_{i,j}\frac{A_{i,k}^2B_{k,j}^2}{p_k} - \frac{1}{L}\sum_{i,j}\left(\boldsymbol{A}_{i,:}\boldsymbol{B}_{:,j}\right)^2 \qquad \text{(check)} \\
&= \frac{1}{L}\sum_k\frac{1}{p_k}\|\boldsymbol{A}_{:,k}\|_2^2\|\boldsymbol{B}_{k,:}\|_2^2 - \frac{1}{L}\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2 \qquad (6.2)
\end{aligned}$$

In addition, Cauchy-Schwarz yields $\left(\sum_k p_k\right)\left(\sum_k \frac{\alpha_k}{p_k}\right) \geq \left(\sum_k\sqrt{\alpha_k}\right)^2$, with equality attained if $p_k \propto \sqrt{\alpha_k}$. This implies

$$\mathbb{E}\left[\|\boldsymbol{M}-\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2\right] \geq \frac{1}{L}\left(\sum_k\|\boldsymbol{A}_{:,k}\|_2\|\boldsymbol{B}_{k,:}\|_2\right)^2 - \frac{1}{L}\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2,$$

where the lower bound is achieved when $p_k \propto \|\boldsymbol{A}_{:,k}\|_2\|\boldsymbol{B}_{k,:}\|_2$

Error concentration

Practically, one often hopes that the approximation error is absolutely controlled most of the time. In other words, we desire an estimator which is sufficiently close to the truth with high probability

For approximate matrix multiplication, two error metrics are of particular interest

- Frobenius norm bound: $\|\boldsymbol{M} - \boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}$
- spectral norm bound: $\|\boldsymbol{M} - \boldsymbol{A}\boldsymbol{B}\|$

invoke matrix concentration inequalities to control these metrics

Recall: the matrix Bernstein inequality

Theorem 6.1 (Matrix Bernstein)

Let $\left\{ \boldsymbol{X}_l \in \mathbb{R}^{d_1 \times d_2} \right\}$ *be a sequence of independent zero-mean random matrices. Assume each random matrix satisfies* $\|\boldsymbol{X}_l\| \leq R$. *Define* $V := \max \left\{ \left\| \mathbb{E}\left[\sum_{l=1}^{L} \boldsymbol{X}_l \boldsymbol{X}_l^\top \right] \right\|, \left\| \mathbb{E}\left[\sum_{l=1}^{L} \boldsymbol{X}_l^\top \boldsymbol{X}_l \right] \right\| \right\}$. *Then,*

$$\mathbb{P}\left\{ \left\| \sum_{l=1}^{L} \boldsymbol{X}_l \right\| \geq \tau \right\} \leq (d_1 + d_2) \exp\left(\frac{-\tau^2/2}{V + R\tau/3} \right)$$

Frobenius norm error of matrix multiplication

Theorem 6.2

Suppose $p_k \geq \frac{\beta \|\boldsymbol{A}_{:,k}\|_2 \|\boldsymbol{B}_{k,:}\|_2}{\sum_l \|\boldsymbol{A}_{:,l}\|_2 \|\boldsymbol{B}_{l,:}\|_2}$ *for some quantity* $0 < \beta \leq 1$. *If* $L \gtrsim \frac{\log n}{\beta}$, *then with prob. exceeding* $1 - O(n^{-10})$,

$$\|\boldsymbol{M} - \boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} \lesssim \sqrt{\frac{\log n}{\beta L}} \|\boldsymbol{A}\|_{\mathrm{F}} \|\boldsymbol{B}\|_{\mathrm{F}}$$

Proof of Theorem 6.2

Clearly, $\mathrm{vec}(\boldsymbol{M}) = \sum_{l=1}^{L} \boldsymbol{X}_l$, where $\boldsymbol{X}_l = \sum_{k=1}^{n} \frac{1}{Lp_k} \boldsymbol{A}_{:,k} \otimes \boldsymbol{B}_{k,:}^{\top} \mathbb{1}\{i_l = k\}$. These matrices $\{\boldsymbol{X}_l\}$ obey

$$\|\boldsymbol{X}_l\|_2 \leq \max_k \frac{1}{Lp_k} \|\boldsymbol{A}_{:,k}\|_2 \|\boldsymbol{B}_{k,:}\|_2 \asymp \frac{1}{\beta L} \sum_{k=1}^{n} \|\boldsymbol{A}_{:,k}\|_2 \|\boldsymbol{B}_{k,:}\|_2 =: R$$

$$\mathbb{E}\left[\sum_{l=1}^{L} \|\boldsymbol{X}_l\|_2^2\right] = L \sum_{k=1}^{n} \mathbb{P}\{i_l = k\} \frac{\|\boldsymbol{A}_{:,k}\|_2^2 \|\boldsymbol{B}_{k,:}\|_2^2}{L^2 p_k^2} \leq \underbrace{\frac{\left(\sum_{k=1}^{n} \|\boldsymbol{A}_{k,:}\|_2 \|\boldsymbol{B}_{k,:}\|_2\right)^2}{\beta L}}_{=: V}$$

Invoke matrix Bernstein to arrive at

$$\|\boldsymbol{M} - \boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} = \left\|\sum_{l=1}^{L} (\boldsymbol{X}_l - \mathbb{E}[\boldsymbol{X}_l])\right\|_2 \lesssim \sqrt{V \log n} + R \log n$$

$$\asymp \sqrt{\frac{\log n}{\beta L}} \left(\sum_{k=1}^{n} \|\boldsymbol{A}_{k,:}\|_2 \|\boldsymbol{B}_{k,:}\|_2\right) \leq \sqrt{\frac{\log n}{\beta L}} \|\boldsymbol{A}\|_{\mathrm{F}} \|\boldsymbol{B}\|_{\mathrm{F}} \quad \text{(Cauchy-Schwarz)}$$

Spectral norm error of matrix multiplication

Theorem 6.3

Suppose $p_k \geq \frac{\beta \|\boldsymbol{A}_{:,k}\|_2^2}{\|\boldsymbol{A}\|_{\mathrm{F}}^2}$ *for some quantity* $0 < \beta \leq 1$, *and* $L \gtrsim \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{\beta \|\boldsymbol{A}\|^2 \log n}$. *Then the estimate* $\boldsymbol{M}$ *returned by Algorithm 6.2 obeys*

$$\|\boldsymbol{M} - \boldsymbol{A}\boldsymbol{A}^\top\| \lesssim \sqrt{\frac{\log n}{\beta L}} \|\boldsymbol{A}\|_{\mathrm{F}} \|\boldsymbol{A}\|$$

with prob. exceeding $1 - O(n^{-10})$

- If $L \gtrsim \underbrace{\frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{\|\boldsymbol{A}\|^2}}_{\text{stable rank}} \frac{\log n}{\varepsilon^2 \beta}$, then $\|\boldsymbol{M} - \boldsymbol{A}\boldsymbol{A}^\top\| \lesssim \varepsilon \|\boldsymbol{A}\|^2$
- *can be generalized to approximate* $\boldsymbol{A}\boldsymbol{B}$ *(Magen, Zouzias '11)*

Proof of Theorem 6.3

Write $\boldsymbol{M} = \sum_{l=1}^{L} \boldsymbol{Z}_l$, where $\boldsymbol{Z}_l = \sum_{k=1}^{n} \frac{1}{Lp_k} \boldsymbol{A}_{:,k} \boldsymbol{A}_{:,k}^\top \mathbb{1}\{i_l = k\}$. These matrices satisfy

$$\|\boldsymbol{Z}_l\|_2 \leq \max_k \frac{\|\boldsymbol{A}_{:,k}\|_2^2}{Lp_k} \leq \frac{1}{\beta L} \|\boldsymbol{A}\|_{\mathrm{F}}^2 =: R$$

$$\begin{aligned}
\left\| \mathbb{E}\left[\sum_{l=1}^{L} \boldsymbol{Z}_l \boldsymbol{Z}_l^\top \right] \right\| &= \left\| L \sum_{k=1}^{n} \mathbb{P}\{i_l = k\} \frac{\|\boldsymbol{A}_{:,k}\|_2^2}{L^2 p_k^2} \boldsymbol{A}_{:,k} \boldsymbol{A}_{:,k}^\top \right\| \\
&= \frac{1}{\beta L} \|\boldsymbol{A}\|_{\mathrm{F}}^2 \left\| \boldsymbol{A}\boldsymbol{A}^\top \right\| \\
&\leq \frac{1}{\beta L} \|\boldsymbol{A}\|_{\mathrm{F}}^2 \|\boldsymbol{A}\|^2 =: V
\end{aligned}$$

Invoke matrix Bernstein to conclude that with high prob.,

$$\begin{aligned}
\left\| \boldsymbol{M} - \boldsymbol{A}\boldsymbol{A}^\top \right\| = \left\| \sum_{l=1}^{L} (\boldsymbol{Z}_l - \mathbb{E}[\boldsymbol{Z}_l]) \right\| &\lesssim \sqrt{V \log n} + B \log n \\
&\asymp \sqrt{\frac{\log n}{\beta L}} \|\boldsymbol{A}\|_{\mathrm{F}} \|\boldsymbol{A}\|
\end{aligned}$$

Matrix multiplication with one-sided information

What if we can only use the information about $\boldsymbol{A}$?

For example, suppose $p_k \geq \frac{\beta \|\boldsymbol{A}_{:,k}\|_2^2}{\|\boldsymbol{A}\|_{\mathrm{F}}^2}$. In this case, matrix Bernstein does NOT yield sharp concentration. But we can still use Markov's inequality to get some useful bound

Matrix multiplication with one-sided information

More precisely, when $p_k \geq \frac{\beta \|\boldsymbol{A}_{:,k}\|_2^2}{\|\boldsymbol{A}\|_{\mathrm{F}}^2}$, it follows from (6.2) that

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{M}-\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2\right] &= \frac{1}{L}\sum_k \frac{1}{p_k}\|\boldsymbol{A}_{:,k}\|_2^2\|\boldsymbol{B}_{k,:}\|_2^2 - \frac{1}{L}\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2 \\
&\leq \frac{1}{\beta L}\left(\sum_k \|\boldsymbol{B}_{k,:}\|_2^2\right)\|\boldsymbol{A}\|_{\mathrm{F}}^2 \\
&= \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2}{\beta L}
\end{aligned}$$

Hence, Markov's inequality yields that with prob. at least $1-\frac{1}{\log n}$,

$$\|\boldsymbol{M}-\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2 \leq \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2\log n}{\beta L} \tag{6.3}$$

Least squares approximation

Least squares (LS) problems

Given $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ $(n \gg d)$ and $\boldsymbol{b} \in \mathbb{R}^n$, find the "best" vector s.t. $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{b}$, i.e.

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$$

If $\boldsymbol{A}$ has full column rank, then

$$\boldsymbol{x}_{\mathsf{ls}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{b} = \boldsymbol{V}_A \boldsymbol{\Sigma}_A^{-1} \boldsymbol{U}_A^\top \boldsymbol{b}$$

where $\boldsymbol{A} = \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^\top$ is the SVD of $\boldsymbol{A}$.

Methods for solving LS problems

Direct methods: computational complexity $O(nd^2)$

- *Cholesky decomposition:* compute upper triangular matrix $\boldsymbol{R}$ s.t. $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$, and solve $\boldsymbol{R}^\top \boldsymbol{R} \boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{b}$

- *QR decomposition:* compute QR decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ ($\boldsymbol{Q}$: orthonormal; $\boldsymbol{R}$: upper triangular), and solve $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^\top \boldsymbol{b}$

Iterative methods: computational complexity $O\big(\frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})} \log \frac{1}{\varepsilon}\big)$

- *conjugate gradient* ...

Randomized least squares approximation

Basic idea: generate a sketching / sampling matrix $\boldsymbol{\Phi}$ (e.g. via random sampling, random projection), and solve instead

$$\widetilde{\boldsymbol{x}}_{\mathsf{ls}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad \|\boldsymbol{\Phi}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\|_2$$

Goal: find $\boldsymbol{\Phi}$ s.t.

$$\begin{aligned} \widetilde{\boldsymbol{x}}_{\mathsf{ls}} \;&\approx\; \boldsymbol{x}_{\mathsf{ls}} \\ \|\boldsymbol{A}\widetilde{\boldsymbol{x}}_{\mathsf{ls}}-\boldsymbol{b}\|_2 \;&\approx\; \|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}-\boldsymbol{b}\|_2 \end{aligned}$$

Which sketching matrices enable good approximation?

We will start with two deterministic conditions that promise reasonably good approximations (Drineas et al '11)

Which sketching matrices enable good approximation?

Let $\boldsymbol{A} = \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^\top$ be the SVD of $\boldsymbol{A}$ ...

- **Condition 1 (approximate isometry)**

$$\sigma_{\min}^2(\boldsymbol{\Phi} \boldsymbol{U}_A) \geq \frac{1}{\sqrt{2}} \tag{6.4}$$

  - says that $\boldsymbol{\Phi}\boldsymbol{U}_A$ is an approximate isometry / rotation
  - $1/\sqrt{2}$ can be replaced by other positive constants

Which sketching matrices enable good approximation?

Let $\boldsymbol{A} = \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^\top$ be the SVD of $\boldsymbol{A}$ ...

- **Condition 2 (approximate orthogonality)**

$$\left\| \boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} (\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}) \right\|_2^2 \leq \frac{\varepsilon}{2} \|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}\|_2^2 \tag{6.5}$$

  - says that $\boldsymbol{\Phi}\boldsymbol{U}_A$ is roughly orthogonal to $\boldsymbol{\Phi} \underbrace{(\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b})}_{=(\boldsymbol{U}_A \boldsymbol{U}_A^\top - \boldsymbol{I})\boldsymbol{b}}$
  - even though this condition depends on $\boldsymbol{b}$, one can find $\boldsymbol{\Phi}$ satisfying this condition without using any information about $\boldsymbol{b}$

Can these conditions be satisfied?

Two extreme examples

1. $\mathbf{\Phi} = \boldsymbol{I}$, which satisfies

$$\begin{cases} \sigma_{\min}\left(\mathbf{\Phi}\boldsymbol{U}_A\right) & = \sigma_{\min}\left(\boldsymbol{U}_A\right) = 1 \\ \left\|\boldsymbol{U}_A^\top \mathbf{\Phi}^\top \mathbf{\Phi}\left(\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}\right)\right\|_2 & = \left\|\boldsymbol{U}_A^\top \left(\boldsymbol{I} - \boldsymbol{U}_A \boldsymbol{U}_A^\top\right)\boldsymbol{b}\right\|_2 = 0 \end{cases}$$

   - easy to construct; hard to solve the subsampled LS problem

Can these conditions be satisfied?

Two extreme examples

2. $\boldsymbol{\Phi} = \boldsymbol{U}_A^\top$, which satisfies

$$\begin{cases} \sigma_{\min}\left(\boldsymbol{\Phi}\boldsymbol{U}_A\right) & = \sigma_{\min}\left(\boldsymbol{I}\right) = 1 \\ \left\|\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\left(\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}\right)\right\|_2 & = \left\|\boldsymbol{U}_A^\top (\boldsymbol{I} - \boldsymbol{U}_A \boldsymbol{U}_A^\top)\boldsymbol{b}\right\|_2 = 0 \end{cases}$$

- hard to construct (i.e. compute $\boldsymbol{U}_A$); easy to solve subsampled LS problem

Quality of approximation

We'd like to assess the quality of approximation w.r.t. both fitting error and estimation error

Lemma 6.4

Under Conditions 1-2, the solution $\widetilde{\boldsymbol{x}}_{\mathsf{ls}}$ to the subsampled LS problem obeys

(i) $\|\boldsymbol{A}\widetilde{\boldsymbol{x}}_{\mathsf{ls}} - \boldsymbol{b}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}\|_2$

(ii) $\|\widetilde{\boldsymbol{x}}_{\mathsf{ls}} - \boldsymbol{x}_{\mathsf{ls}}\|_2 \leq \frac{\sqrt{\varepsilon}}{\sigma_{\min}(\boldsymbol{A})}\|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}\|_2$

Proof of Lemma 6.4(i)

The subsampled LS problem can be rewritten as

$$\begin{aligned}\min_{\boldsymbol{x}\in\mathbb{R}^d} \|\boldsymbol{\Phi b} - \boldsymbol{\Phi A x}\|_2^2 &= \min_{\boldsymbol{\Delta}\in\mathbb{R}^d} \left\|\boldsymbol{\Phi b} - \boldsymbol{\Phi A}\left(\boldsymbol{x}_{\mathsf{ls}} + \boldsymbol{\Delta}\right)\right\|_2^2 \\ &= \min_{\boldsymbol{\Delta}\in\mathbb{R}^d} \left\|\boldsymbol{\Phi}\left(\boldsymbol{b} - \boldsymbol{A x}_{\mathsf{ls}}\right) - \boldsymbol{\Phi A \Delta}\right\|_2^2 \\ &= \min_{\boldsymbol{z}\in\mathbb{R}^d} \Big\|\boldsymbol{\Phi}\left(\boldsymbol{b} - \boldsymbol{A x}_{\mathsf{ls}}\right) - \boldsymbol{\Phi} \underbrace{\boldsymbol{U}_A \boldsymbol{z}}_{=\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_{\mathsf{ls}})} \Big\|_2^2.\end{aligned}$$

Therefore, the optimal solution $\boldsymbol{z}_{\mathsf{ls}}$ obeys

$$\boldsymbol{z}_{\mathsf{ls}} = \left(\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{U}_A\right)^{-1} \left(\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top\right) \boldsymbol{\Phi}\left(\boldsymbol{b} - \boldsymbol{A x}_{\mathsf{ls}}\right).$$

Combine Conditions 1-2 to obtain

$$\|\boldsymbol{z}_{\mathsf{ls}}\|_2^2 \le \left\|\left(\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{U}_A\right)^{-1}\right\|^2 \left\|\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\left(\boldsymbol{b} - \boldsymbol{A x}_{\mathsf{ls}}\right)\right\|_2^2 \le 2\varepsilon \|\boldsymbol{b} - \boldsymbol{A x}_{\mathsf{ls}}\|_2^2$$

Proof of Lemma 6.4(i) (cont.)

Previous bounds further yield

$$\begin{aligned}
\left\|\boldsymbol{b}-\boldsymbol{A}\widetilde{\boldsymbol{x}}_{\mathsf{ls}}\right\|_2^2 &= \Big\|\underbrace{\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}}_{\perp \boldsymbol{U}_A}+\underbrace{\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}-\boldsymbol{A}\widetilde{\boldsymbol{x}}_{\mathsf{ls}}}_{\in \mathsf{range}(\boldsymbol{U}_A)}\Big\|_2^2 \\
&= \left\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\right\|_2^2+\left\|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}-\boldsymbol{A}\widetilde{\boldsymbol{x}}_{\mathsf{ls}}\right\|_2^2 \\
&= \left\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\right\|_2^2+\left\|\boldsymbol{U}_A\boldsymbol{z}_{\mathsf{ls}}\right\|_2^2 \\
&\le \left\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\right\|_2^2+\left\|\boldsymbol{z}_{\mathsf{ls}}\right\|_2^2 \\
&\le (1+2\varepsilon)\left\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\right\|_2^2
\end{aligned}$$

Finally, we conclude the proof by recognizing that $\sqrt{1+2\varepsilon} \le 1+\varepsilon$.

Proof of Lemma 6.4(ii)

From the proof of Lemma 6.4(i), we know $\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{A}\widetilde{\boldsymbol{x}}_{\mathsf{ls}} = \boldsymbol{U}_A \boldsymbol{z}_{\mathsf{ls}}$ and $\|\boldsymbol{z}_{\mathsf{ls}}\|_2^2 \leq \varepsilon \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\|_2^2$. These reveal that

$$
\begin{aligned}
\|\boldsymbol{x}_{\mathsf{ls}} - \widetilde{\boldsymbol{x}}_{\mathsf{ls}}\|_2^2 &\leq \frac{\|\boldsymbol{A}(\boldsymbol{x}_{\mathsf{ls}} - \widetilde{\boldsymbol{x}}_{\mathsf{ls}})\|_2^2}{\sigma_{\min}^2(\boldsymbol{A})} \\
&= \frac{\|\boldsymbol{U}_A \boldsymbol{z}_{\mathsf{ls}}\|_2^2}{\sigma_{\min}^2(\boldsymbol{A})} \\
&\leq \frac{\|\boldsymbol{z}_{\mathsf{ls}}\|_2^2}{\sigma_{\min}^2(\boldsymbol{A})} \\
&\leq \frac{\varepsilon \big\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\big\|_2^2}{\sigma_{\min}^2(\boldsymbol{A})}
\end{aligned}
$$

Quality of approximation (cont.)

By imposing further assumptions on $\boldsymbol{b}$, we can connect the error bound with $\|\boldsymbol{x}_{\mathsf{ls}}\|_2$

Lemma 6.5

Suppose $\|\boldsymbol{U}_A \boldsymbol{U}_A^\top \boldsymbol{b}\|_2 \geq \gamma \|\boldsymbol{b}\|_2$ *for some* $0 < \gamma \leq 1$. *Under Conditions 1-2, the solution* $\widetilde{\boldsymbol{x}}_{\mathsf{ls}}$ *to the subsampled LS problem obeys*

$$\|\boldsymbol{x}_{\mathsf{ls}} - \widetilde{\boldsymbol{x}}_{\mathsf{ls}}\|_2 \leq \sqrt{\varepsilon}\, \kappa(\boldsymbol{A}) \sqrt{\gamma^{-2} - 1} \|\boldsymbol{x}_{\mathsf{ls}}\|_2$$

where $\kappa(\boldsymbol{A})$: *condition number of* $\boldsymbol{A}$

- $\|\boldsymbol{U}_A \boldsymbol{U}_A^\top \boldsymbol{b}\|_2 \geq \gamma \|\boldsymbol{b}\|_2$ says a nontrivial fraction of the energy of $\boldsymbol{b}$ lies in range$(\boldsymbol{A})$

Proof of Lemma 6.5

Since $\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} = (\boldsymbol{I} - \boldsymbol{U}_A\boldsymbol{U}_A^\top)\boldsymbol{b}$, one has

$$\begin{aligned}
\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\|_2^2 &= \|(\boldsymbol{I} - \boldsymbol{U}_A\boldsymbol{U}_A^\top)\boldsymbol{b}\|_2^2 \\
&= \|\boldsymbol{b}\|_2^2 - \|\boldsymbol{U}_A\boldsymbol{U}_A^\top\boldsymbol{b}\|_2^2 \\
&\leq \left(\gamma^{-2} - 1\right)\|\boldsymbol{U}_A\boldsymbol{U}_A^\top\boldsymbol{b}\|_2^2 && (\text{since } \|\boldsymbol{U}_A\boldsymbol{U}_A^\top\boldsymbol{b}\|_2 \geq \gamma\|\boldsymbol{b}\|_2) \\
&= \left(\gamma^{-2} - 1\right)\|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}}\|_2^2 && (\text{since } \boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} = \boldsymbol{U}_A\boldsymbol{U}_A^\top\boldsymbol{b}) \\
&\leq \left(\gamma^{-2} - 1\right)\sigma_{\max}^2(\boldsymbol{A})\,\|\boldsymbol{x}_{\mathsf{ls}}\|_2^2
\end{aligned}$$

This combined with Lemma 6.4(ii) concludes the proof.

Connection with approximate matrix multiplication

Condition 1 can be guaranteed if

$$\left\| \boldsymbol{U}_A^\top (\boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \boldsymbol{U}_A - \underbrace{\boldsymbol{U}_A^\top \boldsymbol{U}_A}_{=\boldsymbol{I}} \right\| \leq 1 - \frac{1}{\sqrt{2}}$$

Condition 2 can be guaranteed if

$$\left\| \boldsymbol{U}_A^\top (\boldsymbol{\Phi}^\top \boldsymbol{\Phi}) (\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}) - \underbrace{\boldsymbol{U}_A^\top (\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b})}_{=\boldsymbol{U}_A^\top (\boldsymbol{I} - \boldsymbol{U}_A \boldsymbol{U}_A^\top)\boldsymbol{b} = \boldsymbol{0}} \right\|_2^2 \leq \frac{\varepsilon}{2} \underbrace{\|\boldsymbol{U}_A\|^2}_{=1} \|\boldsymbol{A}\boldsymbol{x}_{\mathsf{ls}} - \boldsymbol{b}\|_2^2$$

Both conditions can be viewed as approximate matrix multiplication (by designing $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$)

A (slow) random projection strategy

Gaussian sampling: let $\mathbf{\Phi} \in \mathbb{R}^{r \times n}$ be composed of i.i.d. Gaussian entries $\mathcal{N}(0, \frac{1}{r})$

- Conditions 1-2 are satisfied with high prob. if $r \gtrsim \frac{d \log d}{\varepsilon}$ (exercise)
- implementing Gaussian sketching is expensive (computing $\mathbf{\Phi} \boldsymbol{A}$ takes time $\Omega(nrd) = \Omega(nd^2 \log d)$)

Another random subsampling strategy

Let's begin with Condition 1 and try Algorithm 6.2 with optimal sampling probabilities ...

Another random subsampling strategy

Leverage scores of $\boldsymbol{A}$ are defined to be $\|(\boldsymbol{U}_A)_{:,i}\|_2$ $(1 \leq i \leq n)$

Nonuniform random subsampling: set $\boldsymbol{\Phi} \in \mathbb{R}^{r \times n}$ to be a (weighted) random subsampling matrix s.t.

$$\mathbb{P}\left(\boldsymbol{\Phi}_{i,:} = \frac{1}{\sqrt{r p_k}} \boldsymbol{e}_k^\top\right) = p_k, \quad 1 \leq k \leq n$$

with $p_k \propto \|(\boldsymbol{U}_A)_{i,:}\|_2^2$

- still slow: needs to compute (exactly) leverage scores

Fast and data-agnostic sampling

Can we design data-agnostic sketching matrix $\boldsymbol{\Phi}$ (i.e. independent of $\boldsymbol{A}$, $\boldsymbol{b}$) that allows fast computation while satisfying Conditions 1-2?

Subsampled randomized Hadamard transform (SRHT)

An SRHT matrix $\boldsymbol{\Phi} \in \mathbb{R}^{r \times n}$ is

$$\boldsymbol{\Phi} = \boldsymbol{R}\boldsymbol{H}\boldsymbol{D}$$

- $\boldsymbol{D} \in \mathbb{R}^{n \times n}$: diagonal matrix, whose entries are random $\{\pm 1\}$
- $\boldsymbol{H} \in \mathbb{R}^{n \times n}$: Hadamard matrix (scaled by $1/\sqrt{n}$ so it's orthonormal)
- $\boldsymbol{R} \in \mathbb{R}^{r \times n}$: uniform random subsampling

$$\mathbb{P}\left(\boldsymbol{R}_{i,:} = \sqrt{\frac{n}{r}}\boldsymbol{e}_k^\top\right) = \frac{1}{n}, \quad 1 \leq k \leq n$$

Subsampled randomized Hadamard transform

Key idea of SRHT:

- use $\boldsymbol{H}\boldsymbol{D}$ to "uniformize" leverage scores (so that $\{\|(\boldsymbol{H}\boldsymbol{D}\boldsymbol{U}_A)_{i,:}\|_2\}$ are more-or-less identical)
- subsample rank-one components uniformly at random

Uniformization of leverage scores

Lemma 6.6

For any fixed matrix $\boldsymbol{U} \in \mathbb{R}^{n\times d}$, *one has*

$$\max_{1\leq i\leq n} \|(\boldsymbol{HDU})_{i,:}\|_2 \lesssim \frac{\log n}{\sqrt{n}} \|\boldsymbol{U}\|_{\mathrm{F}}$$

with prob. exceeding $1 - O(n^{-9})$

- $\boldsymbol{HD}$ preconditions $\boldsymbol{U}$ with high prob.; more precisely,

$$\frac{\|(\boldsymbol{HDU})_{i,:}\|_2^2}{\sum_{l=1}^{n} \|(\boldsymbol{HDU})_{l,:}\|_2^2} = \frac{\|(\boldsymbol{HDU})_{i,:}\|_2^2}{\|\boldsymbol{U}\|_{\mathrm{F}}^2} \lesssim \frac{\log^2 n}{n} \qquad (6.6)$$

Proof of Lemma 6.6

For any fixed matrix $\boldsymbol{U} \in \mathbb{R}^{n\times d}$, one has

$$(\boldsymbol{HDU})_{i,:} = \sum_{j=1}^{n} \underbrace{h_{i,j} D_{j,j}}_{\text{random on } \{\pm\frac{1}{\sqrt{n}}\}} \boldsymbol{U}_{j,:},$$

which clearly satisfies $\mathbb{E}\left[(\boldsymbol{HDU})_{i,:}\right] = \mathbf{0}$. In addition,

$$V := \mathbb{E}\left[\sum_{j=1}^{n} \|h_{i,j} D_{j,j} \boldsymbol{U}_{j,:}\|_2^2\right] = \frac{1}{n}\sum_{j=1}^{n} \|\boldsymbol{U}_{j,:}\|_2^2 = \frac{1}{n}\|\boldsymbol{U}\|_{\mathrm{F}}^2$$

$$B := \max_j \|h_{i,j} D_{j,j} \boldsymbol{U}_{j,:}\|_2 = \frac{1}{\sqrt{n}} \max_j \|\boldsymbol{U}_{j,:}\|_2 \le \frac{1}{\sqrt{n}}\|\boldsymbol{U}\|_{\mathrm{F}}$$

Invoke matrix Bernstein to demonstrate that with prob. $1 - O(n^{-10})$,

$$\|(\boldsymbol{HDU})_{i,:}\|_2 \lesssim \sqrt{V \log n} + B \log n \lesssim \frac{\log n}{\sqrt{n}} \|\boldsymbol{U}\|_{\mathrm{F}}$$

Theoretical guarantees for SRHT

When uniform subsampling is adopted, one has $p_k = 1/n$. In view of Lemma 6.6,

$$p_k \geq \beta \frac{\|(\boldsymbol{HDU}_A)_{i,:}\|_2^2}{\sum_{l=1}^n \|(\boldsymbol{HDU}_A)_{l,:}\|_2^2}$$

with $\beta \asymp \log^{-2} n$. Apply Theorem 6.3 to yield

$$\begin{aligned} \left\|\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{U}_A - \boldsymbol{I}\right\| &= \left\|\boldsymbol{U}_A^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{U}_A - \boldsymbol{U}_A^\top \boldsymbol{U}_A\right\| \\ &= \left\|(\boldsymbol{U}_A^\top \boldsymbol{D}^\top \boldsymbol{H}^\top)\boldsymbol{R}^\top \boldsymbol{R}\,(\boldsymbol{HDU}_A) - (\boldsymbol{U}_A^\top \boldsymbol{D}^\top \boldsymbol{H}^\top)\,(\boldsymbol{HDU}_A)\right\| \\ &\leq 1/2 \end{aligned}$$

when $r \gtrsim \frac{\|\boldsymbol{HDU}_A\|_{\mathrm{F}}^2}{\|\boldsymbol{HDU}_A\|^2} \frac{\log n}{\beta} \asymp d \log^3 n$. This establishes Condition 1

Theoretical guarantees for SRHT

Similarly, Condition 2 is satisfied with high prob. if $r \gtrsim \frac{d \log^3 n}{\varepsilon}$ (exercise)

Back to least squares approximation

Preceding analysis suggests following algorithm

Algorithm 6.3 Randomized LS approximation (uniform sampling)

1: Pick $r \gtrsim \frac{d \log^3 n}{\varepsilon}$, and generate $\boldsymbol{R} \in \mathbb{R}^{r \times n}$, $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ (as desribed before)

2: **return** $\widetilde{\boldsymbol{x}} = (\boldsymbol{RHDA})^{\dagger} \boldsymbol{RHDb}$

- computational complexity:

$$O\Big(\underbrace{nd \log \frac{n}{\varepsilon}}_{\text{compute } \boldsymbol{HDA}} + \underbrace{\frac{d^3 \log^3 n}{\varepsilon}}_{\text{solve subsampled LS } (rd^2)} \Big)$$

An alternative approach: nonuniform sampling

Key idea of Algorithm 6.3 is to uniformize leverage scores followed by uniform sampling

Alternatively, one can also start by estimating leverage scores, and then apply nonuniform sampling accordingly

Fast approximation of leverage scores

Key idea: apply SRHT (or other fast Johnson-Lindenstrass transform) in appropriate places

$$\begin{aligned}
\|\boldsymbol{U}_{i,:}\|_2^2 &= \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2^2 = \|\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{U}^\top\|_2^2 \\
&= \|\boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{A}^\dagger\|_2^2 \\
&\approx \|\boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{\Phi}_1^\top\|_2^2
\end{aligned}$$

where $\boldsymbol{\Phi}_1 \in \mathbb{R}^{r_1 \times n}$ is SRHT matrix

Issue: $\boldsymbol{A}\boldsymbol{A}^\dagger$ is expensive to compute; can we compute $\boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{\Phi}_1^\top$ in a fast manner?

Aside: pseudo inverse

A userful observation: $\boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{\Phi}^{\top} \approx \boldsymbol{A}(\boldsymbol{\Phi}\boldsymbol{A})^{\dagger}$, where $\boldsymbol{\Phi} \in \mathbb{R}^{r \times n}$ be SRHT matrix with sufficiently large $r \gg \frac{d \text{poly} \log n}{\varepsilon^2}$

It can be shown that (check Mahoney's lecture notes)

$$\left\| (\boldsymbol{\Phi}\boldsymbol{U}_A)^{\dagger} - (\boldsymbol{\Phi}\boldsymbol{U}_A)^{\top} \right\| \leq \varepsilon$$

$$\text{and} \quad (\boldsymbol{\Phi}\boldsymbol{A})^{\dagger} = \boldsymbol{V}_A \boldsymbol{\Sigma}_A^{-1} (\boldsymbol{\Phi}\boldsymbol{U}_A)^{\dagger}$$

These mean

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{\Phi}\boldsymbol{A})^{\dagger} &= \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^{\top} \boldsymbol{V}_A \boldsymbol{\Sigma}_A^{-1} (\boldsymbol{\Phi}\boldsymbol{U}_A)^{\dagger} \approx \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^{\top} \boldsymbol{V}_A \boldsymbol{\Sigma}_A^{-1} (\boldsymbol{\Phi}\boldsymbol{U}_A)^{\top} \\ &= \boldsymbol{U}_A \boldsymbol{U}_A^{\top} \boldsymbol{\Phi}^{\top} = \boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{\Phi}^{\top} \end{aligned}$$

Fast approximation of leverage scores

Continuing our key idea: apply SRHT (or other fast Johnson-Lindenstrass transform) in appropriate places

$$
\begin{aligned}
\|\boldsymbol{U}_{i,:}\|_2^2 &\approx \|\boldsymbol{e}_i^\top \boldsymbol{A}(\boldsymbol{\Phi}_1 \boldsymbol{A})^\dagger\|_2^2 \\
&\approx \|\boldsymbol{e}_i^\top \boldsymbol{A}(\boldsymbol{\Phi}_1 \boldsymbol{A})^\dagger \boldsymbol{\Phi}_2\|_2^2
\end{aligned}
$$

where $\boldsymbol{\Phi}_1 \in \mathbb{R}^{r_1 \times n}$ and $\boldsymbol{\Phi}_2 \in \mathbb{R}^{r_1 \times r_2}$ ($r_2 \asymp \text{poly} \log n$) are both SRHT matrices

Fast approximation of leverage scores

Algorithm 6.4 Leverage scores approximation

1: Pick $r_1 \gtrsim \frac{d \log^3 n}{\varepsilon}$ and $r_2 \asymp \text{poly} \log n$

2: Compute $\boldsymbol{\Phi}_1 \boldsymbol{A} \in \mathbb{R}^{r_1 \times d}$ and its QR decompsotion, and let $\boldsymbol{R}_{\Phi_1 A}$ be the "R" matrix from QR

3: Construct $\boldsymbol{\Psi} = \boldsymbol{A} \boldsymbol{R}_{\Phi_1 A}^{-1} \boldsymbol{\Phi}_2$

4: **return** $\ell_i = \|\boldsymbol{\Psi}_{i,:}\|_2$

- computational complexity: $O\Big(\frac{nd \text{poly} \log n}{\varepsilon^2} + \frac{d^3 \text{poly} \log n}{\varepsilon^2} \Big)$

Least squares approximation (nonuniform sampling)

Algorithm 6.5 Randomized LS approximation (nonuniform sampling)

1: Run Algorithm 6.4 to compute approximate leverage scores $\{\ell_k\}$, and set $p_k \propto \ell_k^2$
2: Randomly sample $r \gtrsim \frac{d\,\mathsf{poly}\log n}{\varepsilon}$ rows of $\boldsymbol{A}$ and elements of $\boldsymbol{b}$ using $\{p_k\}$ as sampling probabilities, rescaling each by $1/\sqrt{rp_k}$. Let $\boldsymbol{\Phi A}$ and $\boldsymbol{\Phi b}$ be the subsampled matrix and vector
3: **return** $\widetilde{\boldsymbol{x}}_{\mathsf{ls}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \|\boldsymbol{\Phi A x} - \boldsymbol{\Phi b}\|_2$

informally, Algorithm 6.5 returns a reasonably good solution with prob. $1 - O(1/\log n)$

Low-rank matrix approximation

Low-rank matrix approximation

Question: given a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, how to find a rank-$k$ matrix that well approximates $\boldsymbol{A}$

- One can compute SVD of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, then return

$$\boldsymbol{A}_k = \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{A}$$

  where $\boldsymbol{U}_k$ consists of top-k singular vectors
- In general, takes time $O(n^3)$, or $O(kn^2)$ (by power methods)
- Can we find faster algorithms if we only want "good approximation"?

Randomized low-rank matrix approximation

Strategy: find a matrix C (via, e.g., subsampling columns of A), and return

$$\underbrace{CC^{\dagger}A}_{\text{project } A \text{ onto column space of } C}$$

Question: how well can $CC^{\dagger}A$ approximate A?

A simple paradigm

Algorithm 6.6

1: **input:** data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, subsampled matrix $\boldsymbol{C} \in \mathbb{R}^{n \times r}$
2: **return** $\boldsymbol{H}_k$ as top-k left singular vectors of $\boldsymbol{C}$

- As we will see, quality of approximation depends on size of

$$\underbrace{\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top}_{\text{connection with matrix multiplication}}$$

Quality of approximation (Frobenius norm)

One can also connect spectral-norm error with product of matrices

Lemma 6.7

The output of Algorithm 6.6 satisfies

$$\left\|\boldsymbol{A}-\boldsymbol{H}_k\boldsymbol{H}_k^\top\boldsymbol{A}\right\|_{\mathrm{F}}^2 \leq \left\|\boldsymbol{A}-\boldsymbol{U}_k\boldsymbol{U}_k^\top\boldsymbol{A}\right\|_{\mathrm{F}}^2 + 2\sqrt{k}\left\|\boldsymbol{A}\boldsymbol{A}^\top-\boldsymbol{C}\boldsymbol{C}^\top\right\|_{\mathrm{F}}$$

where $\boldsymbol{U}_k \in \mathbb{R}^{n\times k}$ contains top-k left singular vectors of $\boldsymbol{A}$

- This holds for any $\boldsymbol{C}$
- Approximation error depends on the error in approximating product of two matrices

Proof of Lemma 6.7

To begin with, since $\boldsymbol{H}_k$ is orthonormal, one has

$$\left\|\boldsymbol{A} - \boldsymbol{H}_k\boldsymbol{H}_k^\top\boldsymbol{A}\right\|_{\mathrm{F}}^2 = \left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2 - \left\|\boldsymbol{H}_k^\top\boldsymbol{A}\right\|_{\mathrm{F}}^2$$

Next, letting $\boldsymbol{h}_i = (\boldsymbol{H}_k)_{:,i}$ yields

$$\begin{aligned}
\left|\left\|\boldsymbol{H}_k^\top\boldsymbol{A}\right\|_{\mathrm{F}}^2 - \sum_{i=1}^k \sigma_i^2(\boldsymbol{C})\right| &= \left|\sum_{i=1}^k \left\|\boldsymbol{A}^\top\boldsymbol{h}_i\right\|_2^2 - \sum_{i=1}^k \left\|\boldsymbol{C}\boldsymbol{h}_i\right\|_2^2\right| \\
&= \left|\sum_{i=1}^k \left\langle \boldsymbol{h}_i\boldsymbol{h}_i^\top, \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top \right\rangle\right| \\
&= \left|\left\langle \boldsymbol{H}_k\boldsymbol{H}_k^\top, \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top \right\rangle\right| \\
&\le \left\|\boldsymbol{H}_k\boldsymbol{H}_k^\top\right\|_{\mathrm{F}} \left\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\right\|_{\mathrm{F}} \\
&\le \sqrt{k}\left\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\right\|_{\mathrm{F}}
\end{aligned}$$

Proof of Lemma 6.7

In addition,

$$
\begin{aligned}
\left|\sum_{i=1}^{k}\sigma_i^2(\boldsymbol{C})-\sum_{i=1}^{k}\sigma_i^2(\boldsymbol{A})\right| &= \left|\sum_{i=1}^{k}\Big\{\sigma_i(\boldsymbol{C}\boldsymbol{C}^\top)-\sigma_i(\boldsymbol{A}\boldsymbol{A}^\top)\Big\}\right| \\
&\le \sqrt{k}\sqrt{\sum_{i=1}^{n}\{\sigma_i(\boldsymbol{C}\boldsymbol{C}^\top)-\sigma_i(\boldsymbol{A}\boldsymbol{A}^\top)\}^2} \quad \text{(Cauchy-Schwarz)} \\
&\le \sqrt{k}\,\big\|\boldsymbol{C}\boldsymbol{C}^\top-\boldsymbol{A}\boldsymbol{A}^\top\big\|_{\mathrm{F}} \quad \text{(Wielandt-Hoffman inequality)}
\end{aligned}
$$

Finally, one has $\|\boldsymbol{A}-\boldsymbol{U}_k\boldsymbol{U}_k^\top\boldsymbol{A}\|_{\mathrm{F}}^2=\|\boldsymbol{A}\|_{\mathrm{F}}^2-\sum_{i=1}^{k}\sigma_i^2(\boldsymbol{A})$.

Combining above results establishes the claim

Quality of approximation (spectral norm)

Lemma 6.8

The output of Algorithm 6.6 satisfies

$$\left\|\boldsymbol{A}-\boldsymbol{H}_k\boldsymbol{H}_k^\top\boldsymbol{A}\right\|^2 \le \left\|\boldsymbol{A}-\boldsymbol{U}_k\boldsymbol{U}_k^\top\boldsymbol{A}\right\|^2 + 2\left\|\boldsymbol{A}\boldsymbol{A}^\top-\boldsymbol{C}\boldsymbol{C}^\top\right\|$$

where $\boldsymbol{U}_k \in \mathbb{R}^{n\times k}$ *contains top-*k *left singular vectors of* $\boldsymbol{A}$

Proof of Lemma 6.8

First of all,

$$\begin{aligned}
\|A - H_k H_k^\top A\| &= \max_{x:\|x\|_2=1} \left\| x^\top (I - H_k H_k^\top) A \right\|_2 \\
&= \max_{x:\|x\|_2=1, x \perp H_k} \left\| x^\top A \right\|_2
\end{aligned}$$

Additionally, for any $x \perp H_k$,

$$\begin{aligned}
\left\| x^\top A \right\|_2^2 &= \left| x^\top C C^\top x + x^\top (A A^\top - C C^\top) x \right| \\
&\leq \left| x^\top C C^\top x \right| + \left| x^\top (A A^\top - C C^\top) x \right| \\
&\leq \sigma_{k+1}(C C^\top) + \|A A^\top - C C^\top\| \\
&\leq \sigma_{k+1}(A A^\top) + 2\|A A^\top - C C^\top\| \\
&= \|A - U_k U_k^\top A\|^2 + 2\|A A^\top - C C^\top\|.
\end{aligned}$$

This concludes the proof.

Back to low-rank matrix approximation

To ensure $\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top$ is small, we can do random subsampling / projection as before. For example:

Algorithm 6.7

1: **for** $l = 1, \cdots, r$ **do**
2: Pick $i_l \in \{1, \cdots, n\}$ i.i.d. with prob. $\mathbb{P}\{i_l = k\} = p_k$
3: Set $\boldsymbol{C}_{:,l} = \frac{1}{\sqrt{r p_{i_l}}} \boldsymbol{A}_{:,l}$
4: **return** $\boldsymbol{H}_k$ as top-k left singular vectors of $\boldsymbol{C}$

Back to low-rank matrix approximation

Invoke Theorems 6.2 and 6.3 to see that with high prob.:

- If $r \gtrsim \frac{k \log n}{\beta \varepsilon^2}$, then

$$\big\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\big\|_{\mathrm{F}}^2 \le \|\boldsymbol{A} - \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{A}\|_{\mathrm{F}}^2 + \varepsilon \|\boldsymbol{A}\|_{\mathrm{F}}^2 \qquad (6.7)$$

- If $r \gtrsim \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{\|\boldsymbol{A}\|^2} \frac{\log n}{\beta \varepsilon^2}$, then

$$\big\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\big\|^2 \le \|\boldsymbol{A} - \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{A}\|^2 + \varepsilon \|\boldsymbol{A}\|^2 \qquad (6.8)$$

An improved multi-pass algorithm

Algorithm 6.8 Multi-pass randomized SVD

$\mathcal{S} = \{\}$
for $l = 1, \cdots, t$ **do**
    $\boldsymbol{E}_l = \boldsymbol{A} - \boldsymbol{A}_{\mathcal{S}} \boldsymbol{A}_{\mathcal{S}}^{\dagger} \boldsymbol{A}$
    Set $p_k \geq \frac{\beta \|(\boldsymbol{E}_l)_{:,k}\|_2^2}{\|\boldsymbol{E}_l\|_{\mathrm{F}}^2}$, $1 \leq k \leq n$
    Randomly select r column indices with sampling prob. $\{p_k\}$ and append to $\mathcal{S}$
return $\boldsymbol{C} = \boldsymbol{A}_{\mathcal{S}}$

An improved multi-pass algorithm

Theorem 6.9

Suppose $r \gtrsim \frac{k \log n}{\beta \varepsilon^2}$. *With high prob.,*

$$\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{C}^{\dagger}\boldsymbol{A}\|_{\mathrm{F}}^2 \leq \frac{1}{1-\varepsilon}\|\boldsymbol{A} - \boldsymbol{U}_k\boldsymbol{U}_k^{\top}\|_{\mathrm{F}}^2 + \varepsilon^t \|\boldsymbol{A}\|_{\mathrm{F}}^2$$

Proof of Theorem 6.9

We will prove it by induction. Clearly, the claim holds for $t = 1$ (according to (6.7)). Assume

$$\Big\|\underbrace{\boldsymbol{A} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^{\dagger}\boldsymbol{A}}_{:=\boldsymbol{E}_t}\Big\|_{\mathrm{F}}^2 \leq \frac{1}{1-\varepsilon}\|\boldsymbol{A} - \boldsymbol{U}_k\boldsymbol{U}_k^{\top}\boldsymbol{A}\|_{\mathrm{F}}^2 + \varepsilon^{t-1}\|\boldsymbol{A}\|_{\mathrm{F}}^2,$$

and let $\boldsymbol{Z}$ be the matrix of the columns of $\boldsymbol{E}_t$ included in the sample. In view of (6.7),

$$\left\|\boldsymbol{E}_t - \boldsymbol{Z}\boldsymbol{Z}^{\dagger}\boldsymbol{E}_t\right\|_{\mathrm{F}}^2 \leq \|\boldsymbol{E}_t - (\boldsymbol{E}_t)_k\|_{\mathrm{F}}^2 + \varepsilon\|\boldsymbol{E}_t\|_{\mathrm{F}}^2,$$

with $(\boldsymbol{E}_t)_k$ the best rank-k approximation of $\boldsymbol{E}_t$. Combining the above two inequalities yields

$$\begin{aligned}\left\|\boldsymbol{E}_t - \boldsymbol{Z}\boldsymbol{Z}^{\dagger}\boldsymbol{E}_t\right\|_{\mathrm{F}}^2 &\leq \|\boldsymbol{E}_t - (\boldsymbol{E}_t)_k\|_{\mathrm{F}}^2 \\ &\quad + \frac{\varepsilon}{1-\varepsilon}\|\boldsymbol{A} - \boldsymbol{U}_k\boldsymbol{U}_k^{\top}\boldsymbol{A}\|_{\mathrm{F}}^2 + \varepsilon^t\|\boldsymbol{A}\|_{\mathrm{F}}^2 \end{aligned} \tag{6.9}$$

Proof of Theorem 6.9 (cont.)

If we can show that

$$\boldsymbol{E}_t - \boldsymbol{Z}\boldsymbol{Z}^\dagger \boldsymbol{E}_t = \boldsymbol{A} - \boldsymbol{C}^t(\boldsymbol{C}^t)^\dagger \boldsymbol{A} \tag{6.10}$$

$$\|\boldsymbol{E}_t - (\boldsymbol{E}_t)_k\|_{\mathrm{F}}^2 \le \|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2 \tag{6.11}$$

then substitution into (6.9) yields

$$\begin{aligned}
\left\|\boldsymbol{A} - \boldsymbol{C}^t(\boldsymbol{C}^t)^\dagger \boldsymbol{A}\right\|_{\mathrm{F}}^2 &\le \|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2 + \frac{\varepsilon}{1-\varepsilon}\|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2 + \varepsilon^t \|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2 \\
&= \frac{1}{1-\varepsilon}\|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2 + \varepsilon^t \|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2
\end{aligned}$$

We can then use induction to finish proof

Proof of Theorem 6.9 (cont.)

It remains to justify (6.10) and (6.11).

To begin with, (6.10) follows from the definition of $\boldsymbol{E}_t$ and the fact $\boldsymbol{Z}\boldsymbol{Z}^\dagger \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger = \boldsymbol{0}$, which gives

$$\boldsymbol{C}^t(\boldsymbol{C}^t)^\dagger = \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger + \boldsymbol{Z}\boldsymbol{Z}^\dagger$$

Proof of Theorem 6.9 (cont.)

To show (6.11), note that $(\boldsymbol{E}_t)_k$ is best rank-k approximation of $\boldsymbol{E}_t$. This gives

$$\|\boldsymbol{E}_t - (\boldsymbol{E}_t)_k\|_{\mathrm{F}}^2 = \left\| \left(\boldsymbol{I} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger\right)\boldsymbol{A} - \left(\left(\boldsymbol{I} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger\right)\boldsymbol{A}\right)_k \right\|_{\mathrm{F}}^2$$

$$\leq \left\| \left(\boldsymbol{I} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger\right)\boldsymbol{A} - \left(\boldsymbol{I} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger\right)\boldsymbol{A}_k \right\|_{\mathrm{F}}^2$$

(since $\left(\boldsymbol{I} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger\right)\boldsymbol{A}_k$ is rank-k)

$$= \left\| \left(\boldsymbol{I} - \boldsymbol{C}^{t-1}(\boldsymbol{C}^{t-1})^\dagger\right)(\boldsymbol{A} - \boldsymbol{A}_k) \right\|_{\mathrm{F}}^2$$

$$\leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}^2,$$

where $\boldsymbol{A}_k$ is best rank-k approximation of $\boldsymbol{A}$. Substitution into (6.9) establishes the claim for t

Multiplicative error bounds

So far, our results read

$$\|A - CC^{\dagger}A\|_{\mathrm{F}}^2 \le \|A - A_k\|_{\mathrm{F}}^2 + \text{additive error}$$
$$\|A - CC^{\dagger}A\|^2 \le \|A - A_k\|^2 + \text{additive error}$$

In some cases, one might prefer multiplicative error guarantees, e.g.

$$\|A - CC^{\dagger}A\|_{\mathrm{F}} \le (1+\varepsilon)\|A - A_k\|_{\mathrm{F}}$$

Two types of matrix decompositions

- *CX decomposition*: let $\boldsymbol{C} \in \mathbb{R}^{n\times r}$ consist of r columns of $\boldsymbol{A}$, and return

$$\hat{\boldsymbol{A}} = \boldsymbol{C}\boldsymbol{X}$$

  for some matrix $\boldsymbol{X} \in \mathbb{R}^{r\times n}$

- *CUR decomposition*: let $\boldsymbol{C} \in \mathbb{R}^{n\times r}$ (resp. $\boldsymbol{R} \in \mathbb{R}^{r\times n}$) consist of r columns (resp. rows) of $\boldsymbol{A}$, and return

$$\hat{\boldsymbol{A}} = \boldsymbol{C}\boldsymbol{U}\boldsymbol{R}$$

  for some matrix $\boldsymbol{U} \in \mathbb{R}^{r\times r}$

Generalized least squares problem

$$\text{minimize}_{\boldsymbol{X}} \quad \|\boldsymbol{B} - \boldsymbol{A}\boldsymbol{X}\|_{\mathrm{F}}^2$$

where $\boldsymbol{X}$ is matrix (rather than vector)

- generalization of over-determined ℓ_2 regression
- optimal solution: $\boldsymbol{X}^{\mathsf{ls}} = \boldsymbol{A}^{\dagger}\boldsymbol{B}$
- if rank$(\boldsymbol{A}) \leq k$, then $\boldsymbol{X}^{\mathsf{ls}} = \boldsymbol{A}_k^{\dagger}\boldsymbol{B}$

Generalized least squares approximation

Randomized algorithm: construct a optimally weighted subsampling matrix $\boldsymbol{\Phi} \in \mathbb{R}^{r \times n}$ with $r \gtrsim \frac{k^2}{\epsilon^2}$ and compute

$$\widetilde{\boldsymbol{X}}^{\mathsf{ls}} = (\boldsymbol{\Phi}\boldsymbol{A})^{\dagger}\boldsymbol{\Phi}\boldsymbol{B}$$

Then informally, with high probability,

$$\|\boldsymbol{B} - \boldsymbol{A}\widetilde{\boldsymbol{X}}^{\mathsf{ls}}\|_{\mathrm{F}} \leq (1+\epsilon)\left\{\min_{\boldsymbol{X}} \|\boldsymbol{B} - \boldsymbol{A}\boldsymbol{X}\|_{\mathrm{F}}\right\}$$

$$\|\boldsymbol{X}^{\mathsf{ls}} - \widetilde{\boldsymbol{X}}^{\mathsf{ls}}\|_{\mathrm{F}} \leq \frac{\epsilon}{\sigma_{\min}(\boldsymbol{A}_k)}\left\{\min_{\boldsymbol{X}} \|\boldsymbol{B} - \boldsymbol{A}\boldsymbol{X}\|_{\mathrm{F}}\right\}$$

Randomized algorithm for CX decomposition

Algorithm 6.9 Randomized algorithm for constructing CX matrix decompositions

1: Compute / approximate sampling probabilities $\{p_i\}_{i=1}^n$, where $p_i = \frac{1}{k}\|(\boldsymbol{U}_{A,k})_{:,i}\|_2^2$

2: Use sampling probabilities $\{p_i\}$ to construct a rescaled random sampling marix $\boldsymbol{\Phi}$

3: Construct $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{\Phi}^\top$

Theoretical guarantees

Theorem 6.10

Suppose $r \gtrsim \frac{k \log k}{\varepsilon^2}$*, then Algorithm 6.9 yields*

$$\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{C}^{\dagger}\boldsymbol{A}\|_{\mathrm{F}} \leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}$$

Proof of Theorem 6.10

$$
\begin{aligned}
&\|\boldsymbol{A} - \boldsymbol{C}\underbrace{\boldsymbol{C}^{\dagger}\boldsymbol{A}}_{:=\boldsymbol{X}^{\mathsf{ls}}}\|_{\mathrm{F}} \\
&= \|\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{\Phi}^{\top})(\boldsymbol{A}\boldsymbol{\Phi}^{\top})^{\dagger}\boldsymbol{A}\|_{\mathrm{F}} \\
&\leq \|\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{\Phi}^{\top})(\boldsymbol{P}_{A_k}\boldsymbol{A}\boldsymbol{\Phi}^{\top})^{\dagger}\boldsymbol{P}_{A_k}\boldsymbol{A}\|_{\mathrm{F}} \quad (\boldsymbol{P}_{A_k} := \boldsymbol{U}_k\boldsymbol{U}_k^{\top}) \\
&\qquad\qquad\qquad\qquad \text{since } \boldsymbol{X}^{\mathsf{ls}} := \boldsymbol{C}^{\dagger}\boldsymbol{A} \text{ minimizes } \|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{X}\|_{\mathrm{F}} \\
&= \|\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{\Phi}^{\top})(\boldsymbol{A}_k\boldsymbol{\Phi}^{\top})^{\dagger}\boldsymbol{A}_k\|_{\mathrm{F}} \\
&\leq (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{A}_k^{\dagger}\boldsymbol{A}_k\|_{\mathrm{F}} \\
&= (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_{\mathrm{F}}
\end{aligned}
$$

Reference

- "*Lecture notes on randomized linear algebra*," M. Mahoney, 2016.
- "*Randomized algorithms for matrices and data*," M. Mahoney, Foundations and Trends in Machine Learning, 2011.
- "*Low rank matrix-valued Chernoff bounds and approximate matrix multiplication*," A. Magen, Z. Anastasios, SODA, 2011.
- "*Faster least squares approximation*," P. Drineas, M. Mahoney, S. Muthukrishnan, T. Sarlos, Numerische mathematik, 2011.
- "*The fast Johnson-Lindenstrauss transform and approximate nearest neighbors*," N. Ailon, B. Chazelle, SIAM Journal on computing, 2009.
- "*Improved analysis of the subsampled randomized Hadamard transform*," J. Tropp, Advances in Adaptive Data Analysis, 2011

Reference

- "*Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*," N. Halko, P. G. Martinsson, J. Tropp, SIAM review, 2011.