Homework 2 – Randomized matrix computations, Fall'24

Prof. D. Kressner
H. Lam

Sciper: **Name:**

Please read the following instructions carefully

- This homework consists of 3 questions, and the submission deadline is Nov 27 at 23h59.
- You can either typeset or scan your handwritten solution. Please also include this page with your signature.
- Submit your homework by email to hysan.lam@epfl.ch.
- **The submitted homework must be your own personal work and must not be copied from elsewhere.**

Your signature:

1 ► Extending the Kaczmarz method

The Kaczmarz method samples a single row of the matrix in one step to find an approximation to a linear least-squares problem. For various reasons (e.g., computational efficiency) it can be beneficial to sample an entire block of (usually consecutive) rows in one step.

Suppose we want to solve a *consistent* least-squares problem

$$\min_{x\in\mathbb{R}^m} \|Ax-b\|_2, \quad A\in\mathbb{R}^{d\times n}, b\in\mathbb{R}^d. \tag{1}$$

Let $U=\{u_1,\dots,u_m\}$ with $u_1\dot{\cup}\cdots\dot{\cup}u_m=\{1,\dots,d\}$ be a partition of the row indices of A. We aim at approximating the optimal solution x^* of (1) iteratively. At the jth iteration, we select a subset of the row indices from U, denoted by $u=u_j$, and project the current iterate x_{j-1} onto the solution space $A_ux=b_u$. Here, A_u denotes the submatrix obtained by by selecting the rows indexed by u from A. This gives

$$x_j = x_{j-1} + A_u^\dagger(b_u - A_u x_{j-1}),$$

where $\dagger$ denotes the pseudoinverse of a matrix. In this exercise, we will assume that the submatrices satisfy the following property:

$$\lambda_{\max}(A_uA_u^T)\le\beta \quad \text{for every } u\in U.$$

Suppose that at each iteration we sample u uniformly at random from U.

a) Prove that

$$\|x_j-x^*\|_2^2 = \|(I-A_u^\dagger A_u)(x_{j-1}-x^*)\|_2^2$$

b) Prove that for any fixed vector $y\in\mathbb{R}^n$,

$$\mathbb{E}\|(I-A_u^\dagger A_u)y\|_2^2 \le \left[1-\frac{\sigma_{\min}^2(A)}{\beta m}\right]\|y\|_2^2$$

and conclude that

$$\mathbb{E}\|x_j-x^*\|_2^2 \le \left[1-\frac{\sigma_{\min}^2(A)}{\beta m}\right]^j\|x_0-x^*\|_2^2.$$

2 ► Sampling graph Laplacians

Let G be a connected, undirected weighted graph on a set V of n vertices with edges E. The Graph Laplacian $L\in\mathbb{R}^{n\times n}$ is defined as

$$L=\sum_{(i,j)\in E} w_{ij}(E_{ii}+E_{jj}-E_{ji}-E_{ij}),$$

where w_{ij} denotes the weight of the edge connecting nodes i and j. The matrix E_{ij} has a 1 at entry (i,j) and 0 everywhere else. In this exercise, we construct a Laplacian S of a sparse graph to approximate the Laplacian L in the sense that S is an ϵ-spectral approximation of L, that is,

$$(1-\epsilon)L\le S\le(1+\epsilon)L.$$

First, we define a map

$$\Psi(M)=(L^\dagger)^{1/2}M(L^\dagger)^{1/2} \quad \text{for } M\in\mathbb{R}^{n\times n}$$

a) Let S be the Laplacian of another graph on the same vertex set V. Prove that $\Psi(L)$ is an orthogonal projection, and if

$$\|\Psi(S-L)\|_2\le\epsilon,$$

then S is an ϵ-spectral approximation of L.

To construct a sparse Laplacian that approximates L, we introduce a random elementary Laplacian:

$$R = \frac{w_{ij}}{p_{ij}}(E_{ii} + E_{jj} - E_{ji} - E_{ij}) \text{ with probability } p_{ij} = \frac{w_{ij}}{n-1}\|\Psi(E_{ii} + E_{jj} - E_{ji} - E_{ij})\|_2,$$

and take the average

$$S = \frac{1}{K}\sum_{i=1}^{K} R_i \quad \text{where } R_i \sim R \text{ i.i.d.}$$

b) Show that

$$\sum_{(i,j)\in E} p_{ij} = 1.$$

Hint: First show that $\|\Psi(E_{ii} + E_{jj} - E_{ji} - E_{ij})\|_2 = \operatorname{trace}(\Psi(E_{ii} + E_{jj} - E_{ji} - E_{ij}))$ and show that $\operatorname{trace}(\Psi(L)) = n - 1$.

c) Suppose $0 < \epsilon \leq 1$, using a result from Lecture 4, show that if we take

$$K \geq 4\epsilon^{-2}(n-1)\log(2n),$$

then

$$\mathbb{E}\|\Psi(S - L)\|_2 \leq \epsilon.$$

Conclude that there is a graph on the same vertex set with at most $4\epsilon^{-2}(n-1)\log(2n)$ edges and its Laplacian is an ϵ-spectral approximation of L.

3 ► Variance reduction technique for randomized trace estimation

In this question, we aim to reduce the variance of the Girard-Hutchinson estimator. Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite. Suppose we use N independent Gaussian random vectors for sampling. The Girard-Hutchinson estimator returns an approximation given by

$$\operatorname{trace}(A) \approx \operatorname{trace}_N(A) = \frac{1}{N}\sum_{i=1}^{N}(X^{(i)})^\top A X^{(i)}, \quad X^{(i)} \sim N(0, I_n) \text{ i.i.d.}$$

a) Show that

$$\mathbb{E}[\operatorname{trace}_N(A)] = \operatorname{trace}(A), \quad \text{and} \quad \operatorname{Var}[\operatorname{trace}_N(A)] = \frac{2}{N}\|A\|_F^2.$$

Conclude that, by Chebyshev's inequality, we need $N = \mathcal{O}(\frac{1}{\delta\epsilon^2})$ samples to obtain

$$\Pr\{|\operatorname{trace}_N(A) - \operatorname{trace}(A)| \geq \epsilon \cdot \operatorname{trace}(A)\} \leq \delta$$

An idea to improve the Girard-Hutchinson estimator is to apply it to a matrix with a smaller Frobenius norm to reduce variance. By the trace property:

$$\operatorname{trace}(A) = \operatorname{trace}(A - X) + \operatorname{trace}(X) \quad \forall X \in \mathbb{R}^{n\times n}.$$

If X is a low-rank matrix such that $\|A - X\|_F$ is small, we can approximate the trace of A by applying Girard-Hutchinson estimator to $A - X$ and calculate the $\operatorname{trace}(X)$ directly (why it is inexpensive to calculate $\operatorname{trace}(X)$ directly?). That is,

$$\operatorname{trace}(A) \approx \operatorname{trace}_N(A - X) + \operatorname{trace}(X) := H(A).$$

To obtain a low-rank apporixmation X we can use randomized low-rank approximation. Suppose $\Omega \in \mathbb{R}^{n\times(k+p)}$ and $\Psi \in \mathbb{R}^{n\times(k+p+\ell)}$ are independent standard Gaussian random matrices with $p, \ell \geq 2$, we obtain low-rank approximation X of A by

$$X = P_{A\Omega,\Psi}A \quad \text{where} \quad P_{A\Omega,\Psi} := A\Omega(\Psi^\top A\Omega)^\dagger\Psi^\top.$$

b) Assume $A\Omega$ and $\Psi^\top A\Omega$ have full rank and let $QR = A\Omega$ be a QR decomposition of $A\Omega$. Show that

$$P_{A\Omega,\Psi} = Q(\Psi^\top Q)^\dagger \Psi^\top, \tag{2}$$

and it is a projection (but not orthogonal projection) onto range($A\Omega$). Also,

$$\begin{aligned}
\|(I - P_{A\Omega,\Psi})A\|_F^2 &= \|(I - Q(\Psi^\top Q)^\dagger \Psi^\top)A\|_F^2 \\
&= \|(I - QQ^\top)A\|_F^2 + \|[Q(\Psi^\top Q)^\dagger \Psi^\top - QQ^\top]A\|_F^2 \\
&= \|(I - QQ^\top)A\|_F^2 + \|Q(\Psi^\top Q)^\dagger (\Psi^\top Q_\perp) Q_\perp^\top A\|_F^2
\end{aligned}$$

c) Using results from the lecture, show that

$$\mathbb{E}\|(I - P_{A\Omega,\Psi})A\|_F^2 \leq \Big(1 + \frac{k+p}{\ell - 1}\Big)\Big(1 + \frac{k}{p-1}\Big)\|A - A_k\|_F^2,$$

where A_k denotes the best rank-k approximation of A.

d) Bonus question: Take $k = N$, $p = k+1$ and $\ell = 2(k+1)$. Show that

$$\mathbb{E}[H(A)] = \operatorname{trace}(A), \quad \text{and} \quad \operatorname{Var}([H(A)]) \leq \frac{8}{N^2}\operatorname{trace}(A)^2.$$

Conclude that, by Chebyshev's inequality, we need

$$N = \mathcal{O}\Big(\sqrt{\frac{1}{\delta\epsilon^2}}\Big) \tag{3}$$

samples to obtain

$$\Pr\{|H(A) - \operatorname{trace}(A)| \geq \epsilon \cdot \operatorname{trace}(A)\} \leq \delta.$$

Comment and compare with the bound for the Girard-Hutchinson estimator. Can you Improve the dependence of δ on (3)?

Hint: You can use the fact that

$$\|A - A_k\|_F \leq \frac{1}{\sqrt{k}}\operatorname{trace}(A),$$

for a symmetric positive semidefinite matrix A.

e) Implement the new estimator with the parameters specified in part d), and the Girard-Hutchinson estimator. Test their performance using the matrix $A = Q^\top \Lambda Q \in \mathbb{R}^{1000\times 1000}$, where $\Lambda_{ii} = i^{-c}$ and Q is an orthogonal matrix obtained by orthogonalizing a random Gaussian matrix. Compare the relative error vs N for different values of $c = 0.5, 1, 1.5, 2$.

Hint: Use the representation (2) and avoid forming the $(\Psi^\top Q)^\dagger$ directly in the algorithm, use backslash instead.