Second Interim Report

Randomized Numerical Linear Algebra: Sketching Methods for Matrix Problems

Ostap Ivanochko, Maryan Petlyovanyj , Morozov Vladyslav

02 February 2026

1 Introduction and Project Aim

This report describes the intermediate progress of our final project in Linear Algebra. The project investigates *randomized numerical linear algebra* (RandNLA) methods for efficient approximate solutions to core matrix problems: least squares, low-rank approximation, and unbiased sketching. The primary goal is to understand the theory, implement the algorithms, and evaluate their performance in terms of error and runtime on both synthetic and real datasets.

The main components of the project are:

- Analysis and implementation of deterministic least squares and sketch-and-solve methods.
- Low-rank approximation using classical SVD and randomized sketching.
- Exploration and implementation of nearly-unbiased sketches and estimator averaging.

In this interim report, we summarize the progress on each task, the algorithms used, and the datasets identified for testing.

2 Project Tasks and Team Responsibilities

- **Ostap Ivanochko:** Read foundational RandNLA literature, implement deterministic least squares, and implement sketch-and-solve variants.
- **Maryan Petlyovanyj:** Focus on low-rank approximation; implement deterministic and randomized approaches for SVD and study their behavior.
- **Morozov Vladyslav:** Implement several nearly-unbiased sketching algorithms, estimator averaging, and perform empirical evaluation.

This division aligns with the major problem classes in randomized linear algebra and allows parallel progress toward the final deliverables.

3 Problem Description and Approaches

We consider three related matrix problems: least squares, low-rank approximation, and matrix sketching. These problems arise frequently in machine learning, data analysis, and scientific computing.

3.1 Least Squares

Given a matrix $A \in \mathbb{R}^{n\times d}$ and vector $b \in \mathbb{R}^n$, the least squares problem is

$$x_{\text{opt}} = \arg\min_x \|Ax - b\|_2^2.$$

Deterministic solutions include QR decomposition and normal equations. Randomized approaches use a sketch matrix S to reduce the size of the problem:

$$\min_x \|SAx - Sb\|_2^2.$$

Different sketch constructions (e.g., random projections, sampling) allow trade-offs between accuracy and efficiency.

3.2 Low-Rank Approximation

The low-rank approximation of a matrix $A \in \mathbb{R}^{m\times n}$ seeks

$$A_k = \arg\min_{\text{rank}(X)\leq k} \|A - X\|_F^2,$$

which is given by the truncated singular value decomposition (SVD):

$$A_k = U_k \Sigma_k V_k^\top.$$

RandNLA techniques use projection/sketch matrices to approximate the singular subspace with lower computational cost.

3.3 Nearly-Unbiased Sketches and Averaging

Beyond basic sketching, advanced methods aim to reduce estimator bias and variance. Averaging multiple sketch estimates or using more refined sampling distributions (e.g., leverage score sampling, determinantal point processes) are examples of these techniques.

4 Linear Algebra Concepts

This section explains the core linear algebra foundations behind our chosen approaches.

4.1 Sketching for Least Squares

The exact least squares solution can be written as

$$x_{\text{opt}} = (A^\top A)^{-1} A^\top b.$$

Sketching replaces A and b with compressed versions SA and Sb, where $S \in \mathbb{R}^{s\times n}$ is a sketching matrix with $s \ll n$. The sketched problem

$$x_s = \arg\min_x \|SAx - Sb\|_2^2$$

can be solved more cheaply when s is much smaller than n, with theoretical guarantees on the distance between x_s and x_{opt} for appropriate sketch choices.

4.2 Low-Rank Approximation

The Eckart–Young–Mirsky theorem states that the best rank-k approximation of A in the Frobenius norm is given by its truncated SVD. Randomized SVD approximates the leading singular vectors using a random projection:

$$Y = A\Omega, \quad Q = \text{orth}(Y), \quad B = Q^\top A.$$

Then the SVD of B provides approximate singular vectors for A with reduced cost.

4.3 Sketch Variance and Sampling

For sketch-and-solve or approximate multiplication, probabilities guiding the sampling (e.g., proportional to row/column norms or leverage scores) influence estimator bias and variance.

5 Pseudocode and Implementation Details

5.1 Sketch-and-Solve Least Squares

```
Input: A  R^{n×d}, b  R^n, sketch size s, sketch matrix S  R^{s×n}
Compute: A_s = SA, b_s = Sb
Solve: x_s = argmin ||A_s x - b_s||_2^2
Return x_s
```

5.2 Low-Rank Approximation (Randomized SVD)

```
Input: A  R^{m×n}, rank k, oversampling p
1. Draw random matrix Ω  R^{n×(k+p)}
2. Y = A Ω
3. Compute orthonormal Q = orth(Y)
4. B = Q^T A
5. SVD: B = Ũ  V^T
6. U_k = Q Ũ(:,1:k)
```

5.3 Nearly-Unbiased Sketch Averaging

```
Input: A, B, sketch size s, number of sketches t
For i=1..t:
    Generate sketch S_i
    Compute sketch estimate C_i
Return averaged estimate C = (1/t)  C_i
```

6 Progress and Results

We have implemented:

- Deterministic and random least squares solvers.
- Deterministic and randomized low-rank SVD routines.
- Nearly-unbiased sketch algorithms with estimator averaging.
- Empirical experiments evaluating runtime vs accuracy across sketch sizes s.

Our initial experiments on synthetic data show that:

- For unstructured Gaussian data, approximate multiplication requires large s to achieve acceptable error, often removing the runtime advantage.
- For structured or spiky data, sketches achieve better errors with smaller s and retain speed benefits over exact GEMM.

7 Datasets for Experiments

To evaluate methods on realistic data, we identified the following real datasets:

- **YearPredictionMSD** – regression dataset with 500k instances and 90 features.
- **MovieLens 100K** – user–item rating matrix useful for low-rank factorization tests.
- **MNIST** – image feature space useful for low-rank SVD experiments.

These datasets offer real structure and noise patterns, allowing meaningful comparisons between deterministic and randomized algorithms.

8 Discussion of Pros and Cons

8.1 Advantages

- RandNLA can yield substantial speedups with controlled error.
- Sketching adapts to data structure; when energy is concentrated, smaller sketches suffice.
- Nearly-unbiased sketches and averaging reduce estimator variance.

8.2 Limitations

- For unstructured data, sketch errors decay slowly, requiring large sketches.
- Overhead of sketch creation and sampling can outweigh benefits at moderate sizes.
- Implementation complexity increases for advanced sketching strategies.

9 Future Work

- Finalize evaluation on real data and incorporate comprehensive plots.
- Explore advanced sampling (e.g., determinantal processes).
- Integrate sketching into iterative solvers for least squares.
- Compare with state-of-the-art RandNLA packages.