Amirkabir University of Technology

(Tehran Polytechnic)

Project (2) Report: Topics in Mathematics and Applications "QR Decomposition for EMNIST Letters Dataset"

Zohreh Sarmeli Saeedi

March 2025

Abstract

This project focuses on the application of QR decomposition using Householder transformations and incremental Givens rotations to solve the least squares problem and classify test data from the EMNIST Letters dataset. The dataset, comprising handwritten English letters, is preprocessed, normalized, and organized into training and test sets. The Householder method achieves a classification accuracy of 68.65%, demonstrating robust performance for static datasets, while the incremental Givens method yields 57.31%, suitable for scenarios with incrementally arriving data. The report compares the two approaches, highlighting their computational efficiency, stability, and applicability based on recent literature (1; 2; 3).

1 Part 1: Householder QR Decomposition

1.1 Project Description

The objective is to apply QR decomposition using Householder transformations to the EMNIST Letters dataset, solving the least squares problem and classifying test samples. The dataset contains handwritten English letter images, preprocessed to facilitate matrix-based computations.

1.2 Project Steps

1.2.1 Data Loading

The EMNIST Letters dataset, containing images of handwritten English letters, is preprocessed as follows:

- **Image Loading**: The `load_emnist_images` function reads binary image files, converting them into a 2D array where each row represents 784 pixels, and each column corresponds to an image sample. Each image is a vector of size (1, 784).
- **Label Loading**: The `load_emnist_labels` function reads binary label files, storing class labels in a 1D array.
- **Normalization**: Pixel values are normalized from the range [0, 255] to [0, 1].

1.2.2 Data Organization and Classification

The dataset consists of 26 classes (English letters). For each class:

- 200 samples are selected for training, and 20 for testing.
- Training and test data are organized into dictionaries `train_data` and `test_data`.
- Data matrices are constructed with dimensions (784, 200) for training and (784, 20) for testing, using the transpose of each image vector.
- Labels are adjusted from 1–26 to 0–25 for consistency.

1.2.3 QR Decomposition with Householder Transformations

Householder transformations zero out sub-diagonal elements of a matrix iteratively (1). The process is:

1. Initialize matrices Q (orthogonal) and R (upper triangular).
2. Apply Householder transformations iteratively to each matrix column.
3. Update Q using the transpose of Householder matrices.
4. Store Q and R in a dictionary `qr_cache` for classification.

1.2.4 Least Squares and Classification

The least squares problem is solved to classify test samples:

1. Retrieve Q and R for each class from `qr_cache`.
2. Project test samples onto Q to compute $b = Q^T x$.
3. Solve the least squares system $Rc = b$ using `np.linalg.lstsq` to obtain coefficients c.
4. Compute reconstruction error as $||x - Qc||_2$.
5. Assign the class with the smallest error as the predicted label.

1.2.5 Model Evaluation

For each class, 20 test samples are evaluated by comparing predicted and true labels. The model accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \times 100$$

The resulting accuracy is 68.65%, indicating relatively good performance in recognizing handwritten letters, consistent with findings in **(author?)** (2).

1.3 Results

The Householder-based QR decomposition model achieves an accuracy of 68.65%, demonstrating effective classification for static datasets.

2 Part 2: Incremental Givens QR Decomposition

2.1 Project Description

This section implements incremental QR decomposition using Givens rotations for the EMNIST Letters dataset. Instead of recomputing QR decomposition for each class, Givens rotations update the existing decomposition, improving computational efficiency for incrementally arriving data (3).

2.2 Project Steps

2.2.1 Data Loading and Preprocessing

Similar to Part 1, data is loaded, normalized to [0, 1], and organized into training (200 samples per class) and test sets (20 samples per class).

2.2.2 Initial QR Decomposition

An initial QR decomposition is computed for the first 200 samples of each class using Householder transformations.

2.2.3 Givens Rotations

Givens rotations zero out specific matrix elements using 2D rotation matrices (1). The process includes:

- **Givens Coefficients**: The function `givens_rotation(a, b)` computes coefficients $c = \frac{a}{\sqrt{a^2+b^2}}$ and $s = \frac{-b}{\sqrt{a^2+b^2}}$ to eliminate b.
- **Updating QR**: The function `updated_qr(Q, R, a)` projects vector a onto the orthogonal complement of Q. If the complement is near zero, R is extended with a new column; otherwise, both Q and R are updated using Givens rotations.
- **Multiple Columns**: The function `update_qr_multiple_columns(Q, R, X_new)` incrementally adds multiple new columns to the QR decomposition.

2.2.4 Data Selection

For each class, 220 samples are selected (200 for initial training, 20 additional for incremental updates).

2.2.5 Least Squares and Classification

Test samples are projected onto Q, and the least squares problem is solved to compute distances. The class with the smallest distance is selected as the predicted label.

2.3 Results

The incremental Givens-based model achieves an accuracy of 57.31%, indicating reasonable performance for dynamic data scenarios.

3 Comparison of Methods

3.1 General Approach

- **Householder Method**: Computes QR decomposition in a single pass, suitable for static datasets (1).
- **Incremental Givens Method**: Updates QR decomposition incrementally, ideal for streaming data (3).

3.2 Applications

- **Householder**: Best for scenarios where all data is available upfront, offering higher accuracy (68.65%) and numerical stability.
- **Givens Incremental**: Suited for applications with continuous data arrival, such as online learning, but with lower accuracy (57.31%) due to potential numerical errors.

3.3 Analysis

The Householder method's higher accuracy is attributed to its direct and stable decomposition, avoiding cumulative numerical errors. In contrast, the Givens method may accumulate errors during incremental updates, impacting performance (2). However, Givens offers advantages:

- **Incremental Updates**: Efficiently incorporates new data without recomputing the entire decomposition.
- **Computational Efficiency**: Reduces computational cost for streaming data scenarios.

4 Conclusion

The Householder QR decomposition method outperforms the incremental Givens approach in accuracy (68.65% vs. 57.31%) for the EMNIST Letters dataset, making it preferable for static data. The Givens method, while less accurate, is valuable for dynamic, streaming data applications due to its efficiency in incremental updates. Future work could explore hybrid approaches to balance accuracy and adaptability (3).

References

[1] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Baltimore, MD: Johns Hopkins University Press, 2013.

[2] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 1997.

[3] Å. Björck, *Numerical Methods for Least Squares Problems.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 1996.

Appendix

The complete project implementation, along with related documentation, is available in the GitHub repository below:

`https://github.com/Zohreh004/Implementation-of-Project-2-`