AMSC660 Final Report

Department of Chemistry and Biochemistry, University of Maryland

Kecai Xuan

kxuan@umd.edu

December 11, 2024

1 Linear Discriminant Analysis

1.1 Rank of $\mathrm{S_b}$

Consider the vector set $\{m_1 - m, m_2 - m, \cdots m_c - m\}$, we have the following relation:

$$\sum_{i=1}^{c} n_i \left(m_i - m\right) = \left[\sum_{i=1}^{c} n_i m_i\right] - nm = nm - nm = 0$$

which means that there exist the nonzero linear combination parameters such that the sum of the vectors is zero. Therefore, the vector set is linearly dependent.

Since S_b is defined as

$$S_b = \sum_{i=1}^{c} n_i \left(m_i - m\right) \left(m_i - m\right)^\top$$

a weighted sum of these rank-1 matrices, it follows that the image (or output) of the linear transformation defined by S_b lies within the space spanned by the vector set $\{m_1 - m, m_2 - m, \cdots m_c - m\}$. That is because when applying S_b to any vector x, it gives the linear combination of the vectors in the set:

$$S_b x = \sum_{i=1}^{c} n_i \left(m_i - m\right) \left(m_i - m\right)^\top x$$

where $\left(m_i - m\right)^\top x$ is a scalar.

The rank of matrix S_b is equal to the dimension of its image, which is the space spanned by the vector set $\{m_1 - m, m_2 - m, \cdots m_c - m\}$ that is linearly dependent. Therefore, the rank of S_b is at most $c - 1$.

1.2 Gradient of $\mathbf{J}(\mathbf{w})$

The gradient of $J(w)$ is given by

$$\begin{aligned} \nabla J(w) &= \nabla \left(\frac{w^\top S_b w}{w^\top S_w w}\right) \\ &= \frac{2}{(w^\top S_w w)^2} \Big[(w^\top S_w w) S_b w - (w^\top S_b w) S_w w\Big] \end{aligned}$$

Let it be zero, we have $(w^\top S_w w)S_b w - (w^\top S_b w)S_w w = 0$, where $w^\top S_w w, w^\top S_b w$ are scalars. So the only way to get the equation satisfied is to have

$$S_b w = \lambda S_w w, \quad \lambda = \frac{w^\top S_b w}{w^\top S_w w}$$

1.3 Cholesky Decomposition

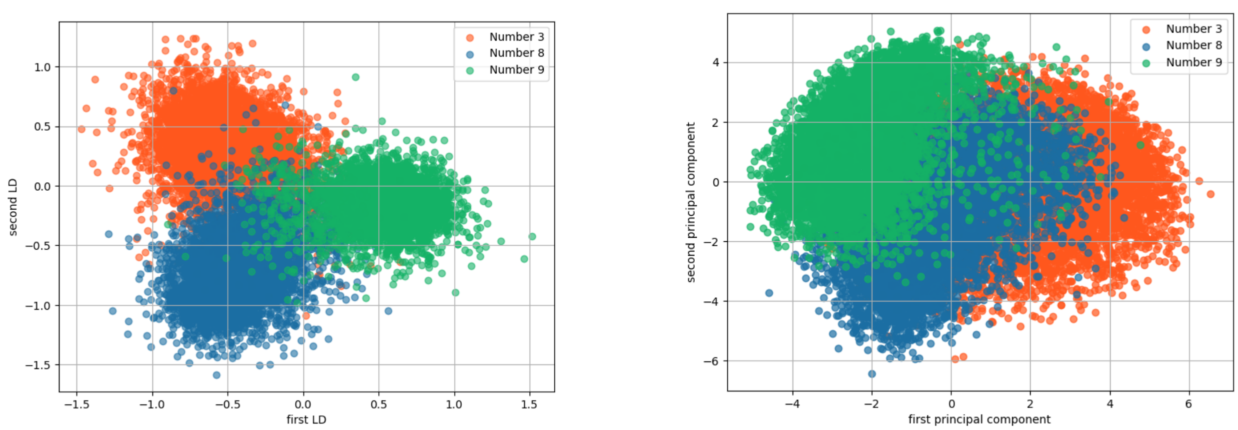
It is obvious that S_w is a SPD matrix, which can be decomposed as $S_w = LL^\top$. Define $y = L^\top w, w = L^{-\top} y$, then the equation $S_b w = \lambda S_w w$ can be rewritten as

$$S_b L^{-\top} y = \lambda L y \quad \Rightarrow \quad L^{-1} S_b L^{-\top} y = \lambda y$$

which takes the form of $Ay = \lambda y$ where $A = L^{-1} S_b L^{-\top}$ is SPD.

1.4 LDA Performance and Comparison with PCA on MINST

The projection of the data onto the first two linear discriminants and the first two principal components are shown in the following figure. For this problem, tt is clear that the linear discriminants are able to separate the data better than the principal components.



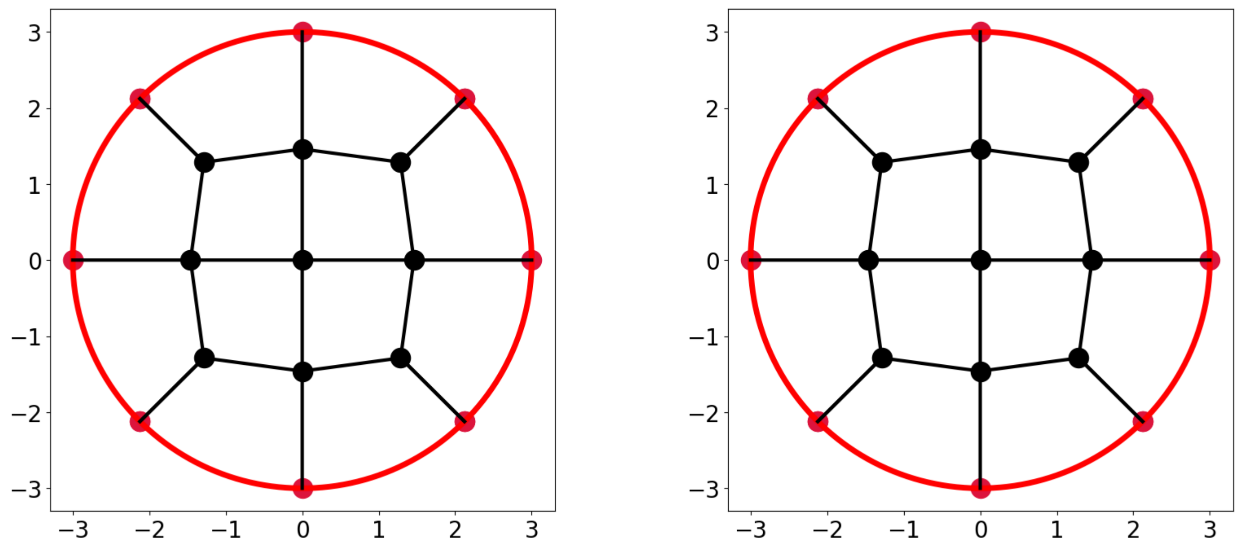
(a) LDA Performance (b) PCA Performance

Figure 1: LDA and PCA on MINST dataset

2 Adam and BFGS on Spring Optimization

2.1 Convergence Near Optimal Solution

I implemented the Adam method and BFGS method to optimize the spring system. The final configuration of the spring system is shown in fig.2a and fig.2b.



(a) Adam final configuration

(b) BFGS final configuration

Figure 2: Final Configuration of Spring System Using Adam and BFGS

The Adam method converged at:

$$\text{iteration} = 187, \quad \text{energy} = 1.4924514079509037, \quad \text{gradient norm} = 9.54947425912235 \times 10^{-7}$$

with the final vector:

$$\text{theta} = \begin{bmatrix} 2.35620 & 2.35620 & 3.14160 & 3.92699 \\ 3.92699 & 4.71239 & 5.49779 & 5.49779 \\ 6.28319 & 7.06859 & 7.06859 & 7.85399 \end{bmatrix}$$

$$\text{position} = \begin{bmatrix} -1.28645 & -1.33638 \times 10^{-5} & 1.28642 & -1.46011 \\ 2.46565 \times 10^{-8} & 1.46011 & -1.28642 & 1.33410 \times 10^{-5} \\ 1.28645 & 1.28642 & 1.46011 & 1.28645 \\ -1.33433 \times 10^{-5} & -2.12707 \times 10^{-8} & 1.33654 \times 10^{-5} & -1.28645 \\ -1.46011 & -1.28642 & & \end{bmatrix}$$

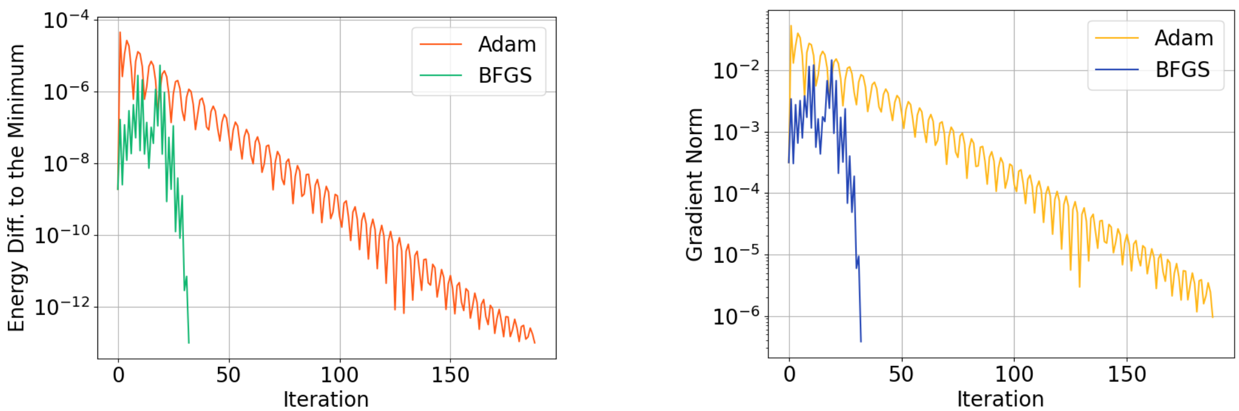
while the BFGS method converged at

$$\text{iteration} = 31, \quad \text{energy} = 1.4924514079508941, \quad \text{gradient norm} = 3.8268520084642785 \times 10^{-7}.$$

whose final vector is almost the same as the Adam method.

2.2 Performance Near Optimal Solution

The results are shown in fig.3a and fig.3b. We show that the Adam method decreases linearly when using logarithmic scale, while the BFGS method decreases even faster when the initial position is near the optimal solution.



(a) Energy vs. Iteration

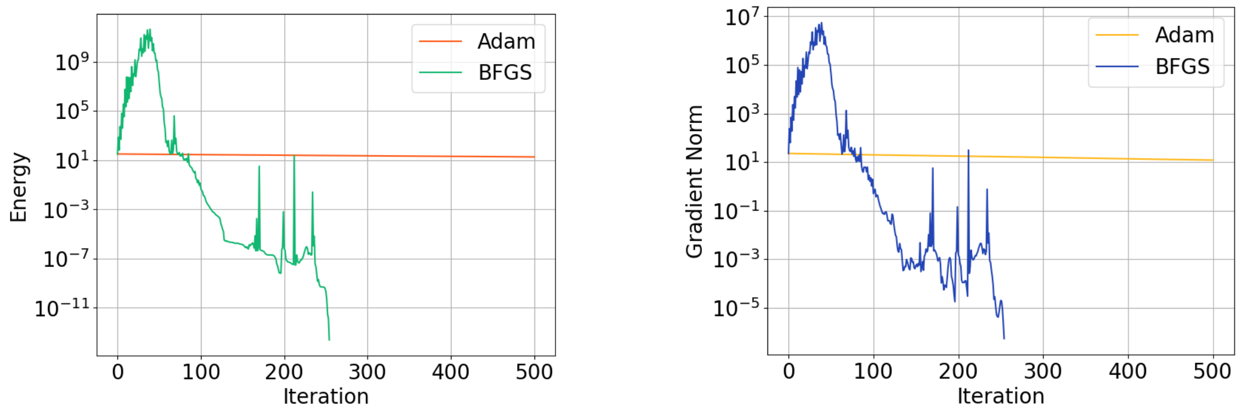
(b) Gradient Norm vs. Iteration

Figure 3: Adam and BFGS on Spring Optimization

3 Far from Optimal Solution

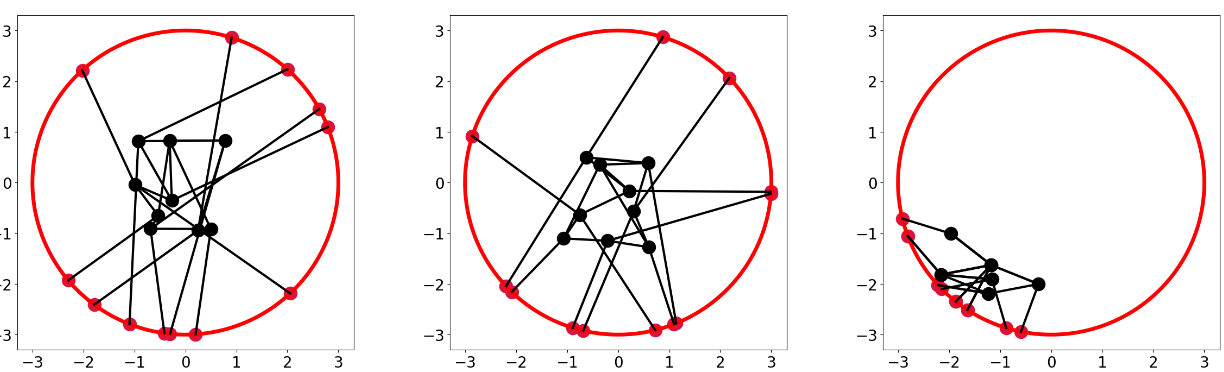
For random generated initial positions, the Adam method fails significantly. The energy and gradient norm are shown in fig.4a and fig.4b in 500 iterations, which is enough for BFGS convergence.

The initial configuration, Adam final configuration and BFGS configuration are shown in fig.**??**, fig.5b and fig.5c respectively.



(a) Energy vs. Iteration (b) Gradient Norm vs. Iteration

Figure 4: Adam and BFGS on Spring Optimization with Random Initial Positions



(a) Initial Configuration (b) Adam final configuration (c) BFGS final configuration

Figure 5: Spring System with Random Initial Positions Using Adam and BFGS