Student Name: Xinxi Zhang
NetID: XZ657
RUID: 219004759

MATH FOUND DS (16:198:501)
Homework 1

---

1. Let $\underline{\omega}$ be a column vector of unknowns, $\underline{\omega} = (\underline{\omega}_1, \underline{\omega}_2, \ldots, \underline{\omega}_d)$. **For a given vector $\underline{c}$, show that the gradient of $\underline{\omega}^T \underline{c}$ is given by $\underline{c}$. For a given symmetric matrix A, show that the gradient of $\underline{\omega}^T A \underline{\omega}$ is given by $2A\underline{\omega}$**

   Answer:

   (a) To prove the gradient of $\underline{\omega}^T \underline{c}$ is given by $\underline{c}$:
   we have:

$$\underline{\omega}^T \underline{c} = \sum_{i=1}^{d} \omega_i c_i, \quad \frac{\mathrm{d}\omega_i c_i}{\mathrm{d}\omega_i} = c_i$$

$$\Rightarrow \quad \frac{\partial(\underline{\omega}^T \underline{c})}{\partial \underline{\omega}} = c$$

$$\Rightarrow \quad \underline{\omega}^T \underline{c} \text{ is given by } \underline{c}$$

   (b) To prove the gradient of $\underline{\omega}^T A \underline{\omega}$ is given by $2A\underline{\omega}$:
   we have:

$$A = \begin{bmatrix} | & | & & | \\ \underline{a}_1 & \underline{a}_2 & \ldots & \underline{a}_d \\ | & | & & | \end{bmatrix}, \quad A^T = A, \quad a_{ij} = a_{ji}$$

$$\Rightarrow \quad A\underline{\omega} = A^T \underline{\omega} = \begin{bmatrix} \underline{a}_1^T \underline{\omega} \\ \underline{a}_2^T \underline{\omega} \\ \ldots \\ \underline{a}_d^T \underline{\omega} \end{bmatrix}$$

$$\Rightarrow \quad \underline{\omega}^T A \underline{\omega} = \sum_{i=1}^{d} \underline{\omega}_i \underline{a}^T \underline{\omega}$$

$$\begin{aligned} \Rightarrow \quad \frac{\partial \underline{\omega}^T A \underline{\omega}}{\partial \omega_i} &= \sum_{j=1, j \neq i}^{d} w_j \underline{a}_{ji} + \sum_{j=1, j \neq i}^{d} w_j \underline{a}_{ij} + 2\underline{a}_{ij}\omega_i \\ &= 2\sum_{j=1}^{d} w_j \underline{a}_{ij} \\ &= 2\underline{a}_i^T \underline{\omega} \end{aligned}$$

$$\Rightarrow \quad \frac{\partial \underline{\omega}^T A \underline{\omega}}{\partial \omega} = \begin{bmatrix} 2\underline{a}_1^T \underline{\omega} \\ 2\underline{a}_2^T \underline{\omega} \\ \dots \\ 2\underline{a}_d^T \underline{\omega} \end{bmatrix} = 2A\underline{\omega}$$

$$\Rightarrow \quad \underline{\omega}^T A \underline{\omega} \text{ is given by } 2A\underline{\omega}$$

2. (a) For a given matrix A, let $G(A)$ be the matrix of derivatives so that $G_i, j = \partial/\partial A_i, j[L]$. **Show that $G = 4A(A^T A - B)$.**

Firstly we can note that B is a symmetric matrix because B is generated by $\tilde{A}^T \tilde{A}$. So we have $B_{i,j} = B_{j,i}$.

Let:

$$A = \begin{bmatrix} | & | & | & | \\ \underline{a}_1 & \underline{a}_2 & \dots & \underline{a}_d \\ | & | & | & | \end{bmatrix}, A^T = \begin{bmatrix} - & \underline{a}_1 & - \\ - & \underline{a}_2 & - \\ - & \dots & - \\ - & \underline{a}_d & - \end{bmatrix}, B = \begin{bmatrix} | & | & | & | \\ \underline{b}_1 & \underline{b}_2 & \dots & \underline{b}_d \\ | & | & | & | \end{bmatrix},$$

$$G = 4A(A^T A - B) = 4(AA^T A - AB)$$

$$\begin{aligned} \Rightarrow \quad G_{x,y} &= 4(\sum_{i=1}^{D} A_{x,i} * \underline{a}_i \cdot \underline{a}_y - \sum_{i=1}^{D} A_{x,i} * B_{i,y}) \\ &= 4\sum_{i=1}^{D} A_{x,i} * (\underline{a}_i \cdot \underline{a}_y - B_{i,y}) \end{aligned}$$

And we have:

$$\begin{aligned} \frac{\mathrm{d}L}{\mathrm{d}A_{x,y}} &= \frac{\mathrm{d}\sum_{i=1}^{D}\sum_{j=1}^{D}(B_{i,j} - \underline{a}_i \cdot \underline{a}_j)^2}{\mathrm{d}A_{x,y}} \\ &= \frac{\mathrm{d}\sum_{i=1}^{D}\sum_{j=1}^{D}(B_{i,j}^2 + (\underline{a}_i \cdot \underline{a}_j)^2 - 2B_{i,j} * (\underline{a}_i \cdot \underline{a}_j))}{\mathrm{d}A_{x,y}} \\ &= \frac{\mathrm{d}\sum_{i=1}^{D}\sum_{j=1}^{D}((\underline{a}_i \cdot \underline{a}_j)^2 - 2B_{i,j} * (\underline{a}_i \cdot \underline{a}_j))}{\mathrm{d}A_{x,y}} \end{aligned}$$

And the $A_{x,y}$ is only contained in $\{\underline{a}_i : i = y\}$, so:

$$\begin{aligned} \frac{\mathrm{d}L}{\mathrm{d}A_{x,y}} &= \frac{\mathrm{d}\sum_{j=1}^{D}((\underline{a}_y \cdot \underline{a}_j)^2 - 2B_{x,j} * (\underline{a}_y \cdot \underline{a}_j)) + \sum_{i=1}^{D}((\underline{a}_i \cdot \underline{a}_y)^2 - 2B_{i,x} * (\underline{a}_i \cdot \underline{a}_y))}{\mathrm{d}A_{x,y}} \\ &= 2\frac{\mathrm{d}\sum_{i=1}^{D}(\underline{a}_i \cdot \underline{a}_y)^2}{\mathrm{d}A_{x,y}} - 4\frac{\mathrm{d}\sum_{i=1}^{D} B_{i,y} * (\underline{a}_i \cdot \underline{a}_y)}{\mathrm{d}A_{x,y}} \\ &= 4\sum_{i=1}^{D}(A_{i,y} * (\underline{a}_i \cdot \underline{a}_y)) - 4\sum_{i=1}^{D}(B_{i,x} * A_{i,y}) \\ &= 4\sum_{i=1}^{D}(A_{i,y} * (\underline{a}_i \cdot \underline{a}_y) - B_{i,y}) \\ &= G_{x,y} \end{aligned}$$

(b) **Generate a random B as above, taking $D = 10$,$k = 10$, and taking A as a random initial matrix, implement this process to show that L decreases over time to 0 - graph your results.**

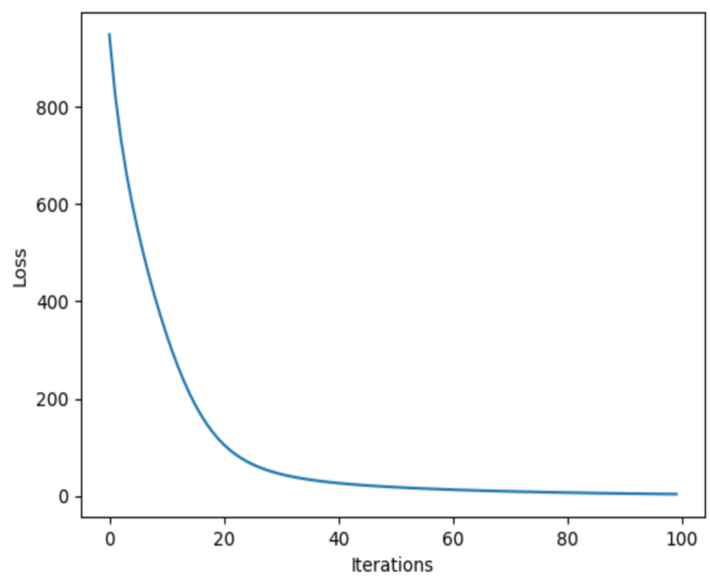
Algorithm 1 Gradient Descent

$\tilde{A} \leftarrow$ Random Matrix(k,D)
$B \leftarrow \tilde{A}^T \tilde{A}$
$A \leftarrow$ Random Matrix(k,D)
$L \leftarrow \sum_{i=1}^{D} \sum_{j=1}^{D} [B - A^T A]_{i,j}^2$
$\alpha \leftarrow$ a sufficiently small number
while $L \neq 0$ **do**
    $G = 4A(A^T A - B)$
    $A = A - \alpha G$
    $L \leftarrow \sum_{i=1}^{D} \sum_{j=1}^{D} [B - A^T A]_{i,j}^2$
end while

The Python code for the algorithm above is filed as "Gradient_Descent.py"

About α, different scales of α $(0.1, 0.01, 0.001)$ have been tried to observe if they are sufficiently small enough that the Loss can converge to 0. When the $\alpha = \{0.1, 0.01\}$, the Loss cannot converge and keep growing to infinity. When $\alpha = 0.001$, the Loss can converge to 0. However, the Loss will not converge to exactly 0 in real-time programming, so the program will be terminated when Loss is very close to 0.

And the history of Loss during the iteration is shown below:



(c) **For a given $B(D = 10, k = 10)$, do you recover the same A every time, for different initial starting points? Why or why not.**

The answer is no. The easy explanation for this is we can easily generate different A that yield the same B:

$$\text{Let:}\quad A_{i,j} = 1, \quad i,j = 1,2,\ldots,D$$
$$\bar{A}_{i,j} = -1, \quad i,j = 1,2,\ldots,D$$

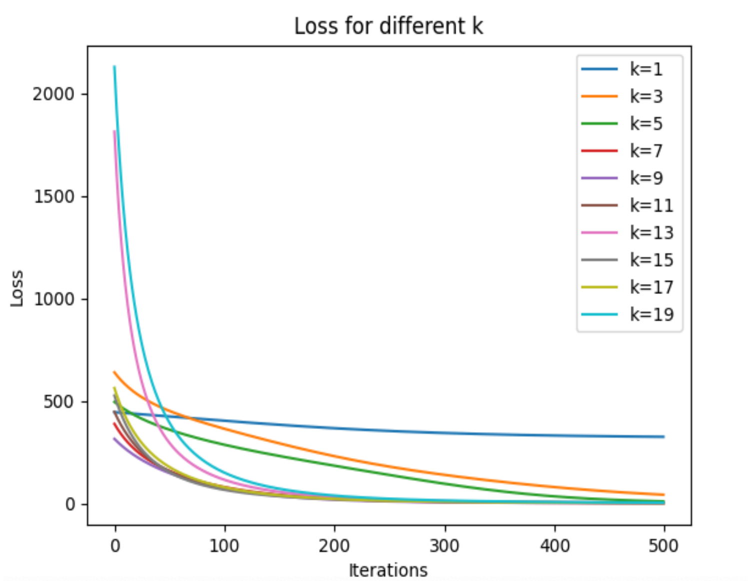
$$\Rightarrow \quad B = A^T A = \bar{A}^T \bar{A}, \quad B_{i,j} = 1, \quad i,j = 1,2,\ldots,D$$

And interestingly, if we rotate A with a rotation matrix M to generate $\bar{A}$, we will find that $\bar{A}^T \bar{A} = A^T A = B$. This is Because:

$$\bar{A}^T \bar{A} = (MA)^T (MA) = A^T M^T M A = A^T (M^T M) A = A^T I A = A^T A$$

(d) Generating B randomly as above, with $D = 5, k = 10$, suppose that the 'true' value of k is forgotten. **Try to find A for different values of k. What do you notice about the loss for different k? Can you recover the 'true' dimension?**

By altering the code of "Gradient_Descent.py", we can try to recover B with different A with different k. (The code is filed as 2_d.py) And the graph of their lost during the gradient descent is shown below:



We can see that with different k, we can still recover B. So we cannot recover the 'true' dimension. And by observing the Loss of different k, we can see that Losses for bigger k converge more quickly to 0.

(e) Think about the relationship between the columns of A and the matrix $A^T A$, and **use this to explain the results of the previous two bullet points.**

we have:

$$[A^T A]_{i,j} = \underline{a}_i \cdot \underline{a}_j$$

So the entries of $A^T A$ represent the dot products between columns of A, which means that the matrix $A^T A$ represent the relationship between columns of A.
In this case, we can try to use this information to explain bullet points (c) and (d):

For (c): We can recover B with different A because any A has the columns relationships represented by B can generate B by $A^T A$.

For (d): We cannot find the true k because B only represent the columns relationships between each A's columns. It has nothing to do with the dimension of the A's columns.

(f) **What happens if you try to take B as the identity matrix? What does the solution A represent (for any k that works)?**

$$B = I$$

$$\Rightarrow \quad \underline{a}_i \cdot \underline{a}_j = B_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\Rightarrow \quad \text{the columns in } A \text{ are } \textbf{Orthonormal} \text{ to each other}$$

(g) **What happens if you try to take B as the diagonal matrix of all 1s, except for bottom right corner which is -1? What does the solution A represent (for any k that works)?**

$$\underline{a}_i \cdot \underline{a}_j = B_{i,j} = \begin{cases} 1 & i = j \neq D \\ -1 & i = j = D \\ 0 & i \neq j \end{cases}$$

I've tried to recover the B with different k but failed. The Loss stop descending after it reach 1. And the B only can be recovered as the diagonal matrix of all 1s, except for bottom right corner which is 0.

So I don't think there is an A can recover B. Because what B tells us is that the dot product of last column of A and itself is -1, which is impossible because this represents the norm of $\underline{a}_D$ and the norm of a vector cannot be negative.

And actually, I don't think there is an A to recover B if there is a negative entire on B's diagonal.

3. Show that if M preserves norms under multiplication ($||M\underline{v}||_2 = ||\underline{v}||_2$ for all $\underline{v}$), then the columns of M must be orthonormal with respect to each other.

Firstly, we have:

$$\begin{aligned} ||M\underline{v}||_2^2 &= (M\underline{v})^T(M\underline{v}) \\ &= \underline{v}^T M^T M \underline{v} \end{aligned}$$

$$\text{Let:} \quad M' = M^T M, M = \begin{bmatrix} | & | & | & | \\ \underline{m}_1 & \underline{m}_2 & \dots & \underline{m}_d \\ | & | & | & | \end{bmatrix}$$
$$\Rightarrow \quad M'_{i,j} = \underline{m}_i \cdot \underline{m}_j$$

Because $||M\underline{v}||_2 = ||\underline{v}||_2$ for all $\underline{v}$, we can:

$$\text{Let:} \quad \underline{v}_i = \begin{cases} 1; & i = k \\ 0; & \text{elsewhere} \end{cases} \quad , k \text{ is a constant and } k \in \{1, 2, \dots, D\}$$

$$\begin{aligned} \Rightarrow \quad & \underline{v}^T M^T M \underline{v} = \underline{v}^T M' \underline{v} = \underline{m}_k \cdot \underline{m}_k = ||\underline{m}_k||_2 = ||\underline{v}||_2 = 1 \\ \Rightarrow \quad & \text{The columns of } M \text{ are normal vectors} \end{aligned}$$

Then we can:

$$\text{Let:} \quad \underline{v}_i = \begin{cases} 1; & i = k \\ 1; & i = k' \\ 0; & \text{elsewhere} \end{cases} \quad ; \ k, k' \text{ is constants that } k, k' \in \{1, 2, \dots, D\}, \text{ and } k \neq k'$$

$$\begin{aligned} \Rightarrow \quad \underline{v}^T M^T M \underline{v} &= \underline{v}^T M' \underline{v} \\ &= \underline{m}_k \cdot \underline{m}_k + \underline{m}_{k'} \cdot \underline{m}_{k'} + \underline{m}_{k'} \cdot \underline{m}_k + \underline{m}_k \cdot \underline{m}_{k'} \\ &= 2 + 2(\underline{m}_k \cdot \underline{m}_{k'}) \end{aligned}$$

And we Have:

$$\begin{aligned} & v^T M^T M \underline{v} = ||\underline{v}|| = 2 \\ & \Rightarrow \quad \underline{m}_k \cdot \underline{m}_{k'} = 0 \\ & \Rightarrow \text{The columns of } M \text{ are } \textbf{Orthogonal} \text{ to each other} \end{aligned}$$

So, if M preserves norms under multiplication ($||M\underline{v}||_2 = ||\underline{v}||_2$ for all $\underline{v}$), then the columns of M must be **orthonormal** with respect to each other.

4. (a) **Show that if $\underline{w}^T$ denotes the solution at time T of this problem, then:**

$$\underline{\omega} = R_T^{-1} U_T, \text{ where } R_T = \sum_{t=1}^{T} \underline{x}_t \underline{x}_t^T, U_t = \sum_{t=1}^{T} y_t \underline{x}_t$$

if $\underline{\omega}^T$ denotes the solution, we can know that:

$$\frac{\partial \sum_{t=1}^{T} (\underline{x}_t^T \omega - y_t)^2}{\partial \omega} = 0$$

$$\begin{aligned} \frac{\partial \sum_{t=1}^{T} (\underline{x}_t^T \omega - y_t)^2}{\partial \omega} &= \frac{\partial \sum_{t=1}^{T} ((\underline{x}_t^T \omega)^2 - 2 y_t \underline{x}_t^T \omega)}{\partial \omega} \\ &= \frac{\partial \sum_{t=1}^{T} (\underline{x}_t^T \omega)^2}{\partial \omega} - 2 \frac{\partial \sum_{t=1}^{T} y_t \underline{x}_t^T \omega}{\partial \omega} \\ &= 2 \sum_{t=1}^{T} \underline{x}_t \underline{x}_t^T \underline{\omega}_T - 2 \sum_{t=1}^{T} y_t \underline{x}_t \\ &= 2 R_T \underline{\omega}_T - 2 U_T \\ &= 0 \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad & 2 R_T \underline{\omega}_T = 2 U_T \\ \Rightarrow \quad & \underline{\omega} = R_T^{-1} U_T \end{aligned}$$

(b) **Assume that R_T^{-1} has been computed. Express R_{T+1}^{-1} in terms of R_T^{-1} and $\underline{x}_{t+1}$.**

$$\begin{aligned} R_T &= \sum_{t=1}^{T} \underline{x}_t \underline{x}_t^T \\ \Rightarrow \quad R_{T+1} &= R_T + \underline{x}_{t+1} \underline{x}_{t+1}^T \end{aligned}$$

By using the Sherman–Morrison formula:

$$\begin{aligned} R_{T+1}^{-1} &= (R_T + \underline{x}_{t+1} \underline{x}_{t+1}^T)^{-1} \\ &= R_T^{-1} - \frac{R_T^{-1} \underline{x}_{t+1} \underline{x}_{t+1}^T R_T^{-1}}{1 + \underline{x}_{t+1}^T R_T^{-1} \underline{x}_{t+1}} \end{aligned}$$

(c) Rather than re-computing a new model $\underline{\omega}_t$ at each timestep t, it would be better if we could *update* our previous model to reflect the new data **Show that:**

$$\underline{\omega}_{T+1} = \underline{\omega}_T + (y_{t+1} - \underline{x}_{t+1}^T \underline{\omega}_T) K_{T+1}$$

Let:

$$K'_{T+1} = \frac{R_T^{-1} \underline{x}_{t+1} \underline{x}_{t+1}^T R_T^{-1}}{1 + \underline{x}_{t+1}^T R_T^{-1} \underline{x}_{t+1}}$$

$$
\begin{aligned}
\Rightarrow \underline{\omega}_{T+1} &= (R_T^{-1} - K'_{T+1})(U_T + y_{t+1}\underline{x}_{t+1}) \\
&= R_T^{-1}U_T + R_T^{-1}y_{t+1}\underline{x}_{t+1} - K'_{T+1}(U_T + y_{t+1}\underline{x}_{t+1}) \\
&= \underline{\omega}_t + y_{t+1}R_T^{-1}\underline{x}_{t+1} - \frac{R_T^{-1}\underline{x}_{t+1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}\underline{x}_{t+1}^T\underline{\omega}_T - \frac{R_T^{-1}\underline{x}_{t+1}\underline{x}_{t+1}^T R_T^{-1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}y_{t+1}\underline{x}_{t+1} \\
&= \underline{\omega}_t + \frac{y_{t+1}R_T^{-1}\underline{x}_{t+1} + y_{t+1}R_T^{-1}\underline{x}_{t+1}\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}} - \frac{R_T^{-1}\underline{x}_{t+1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}\underline{x}_{t+1}^T\underline{\omega}_T \\
&- \frac{R_T^{-1}\underline{x}_{t+1}\underline{x}_{t+1}^T R_T^{-1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}y_{t+1}\underline{x}_{t+1} \\
&= \underline{\omega}_t + \frac{y_{t+1}R_T^{-1}\underline{x}_{t+1} - R_T^{-1}\underline{x}_{t+1}\underline{x}_{t+1}^T\underline{\omega}_T}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}} \\
&= \underline{\omega}_t + (y_{t+1} - x_{t+1}^T\underline{\omega}_T)\frac{R_T^{-1}\underline{x}_{t+1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}
\end{aligned}
$$

HELL YEAH I DID IT!

$$
\Rightarrow \quad K_{T+1} = \frac{R_T^{-1}\underline{x}_{t+1}}{1+\underline{x}_{t+1}^T R_T^{-1}\underline{x}_{t+1}}
$$

5. Consider the matrix

$$A = \begin{bmatrix} 1 & 0.5 \\ 0 & 1+\epsilon \end{bmatrix}$$

(a) Find the eigenvalues and eigenvectors of A, for $\epsilon \neq 0$, taking the eigenvectors to be of unit norm.

In order to find eigenvalues λ and eigenvectors $\underline{v}$ of A, we have:

$$det(A - \lambda I) = 0 \quad \Leftrightarrow \quad (1-\lambda)(1+\epsilon-\lambda) = 0$$

so the eigenvalues of A are $\lambda_1 = 1$ and $\lambda_2 = 1+\epsilon$, and then we can compute the unit eigenvectors:

$$\begin{bmatrix} 0 & 0.5 \\ 0 & \epsilon \end{bmatrix} \cdot \underline{v}_1 = 0, \quad \begin{bmatrix} -\epsilon & 0.5 \\ 0 & 0 \end{bmatrix} \cdot \underline{v}_2 = 0,$$

$$\Rightarrow \quad \underline{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \underline{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{4\epsilon^2+1}} \\ \frac{2\epsilon}{\sqrt{4\epsilon^2+1}} \end{bmatrix}$$

(b) Diagonalize A in terms of these eigenvalues and eigenvectors.

Let:

$$V = \begin{bmatrix} 1 & \frac{1}{\sqrt{4\epsilon^2+1}} \\ 0 & \frac{2\epsilon}{\sqrt{4\epsilon^2+1}} \end{bmatrix}, \quad \triangle = \begin{bmatrix} 1 & 0 \\ 0 & 1+\epsilon \end{bmatrix}$$

$$\Rightarrow \quad A = V \triangle V^{-1}$$

(c) Taking the limit as ϵ decreases from above to 0, what happens to the diagonalization matrices? What can you conclude when $\epsilon = 0$?

When Taking the limit as ϵ decreases from above to 0, we have:

$$\lim_{\epsilon \to 0} V = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad \text{which is not invertible because: } \det(\lim_{\epsilon \to 0} V) = 0$$

So we can conclude that when $\epsilon = 0$, A is not diagonalizable.