**SVD Proofs:**

Let's find this \( \mathbf{V}, \mathbf{\Sigma}, \mathbf{U} \) assuming they always exist.

Assuming \( \mathbf{AV} = \mathbf{U\Sigma} \), let's calculate what \( \mathbf{U}, \mathbf{\Sigma}, \mathbf{V} \) are.

**Prove that \( \mathbf{V} \) is the eigenbasis of \( \mathbf{A}^T \) (row space of \( \mathbf{A} \)).**

\[
\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \\
\mathbf{A}^T = \mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T = \mathbf{V} \mathbf{\Sigma} \mathbf{U}^T \\
\therefore \mathbf{A} \mathbf{A}^T = (\mathbf{V} \mathbf{\Sigma} \mathbf{U}^T)(\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T) \\
= \mathbf{V} \mathbf{\Sigma} \mathbf{U}^T \mathbf{\Sigma} \mathbf{V}^T \\
= \mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^{-1} \\
\text{or} \quad \mathbf{A} \mathbf{A}^T \mathbf{V} = \mathbf{V} \mathbf{\Sigma} \mathbf{V} \\
\therefore \mathbf{V} \text{ is the eigen vector matrix of } \mathbf{A} \mathbf{A}^T \\
\text{and } [\sigma_1, \sigma_2 \ldots]^T \text{ are the } \sqrt{\lambda_1, \sqrt{\lambda_2} \ldots} \text{ of matrix } \mathbf{A} \mathbf{A}^T. \]

\( \mathbf{A}_{m \times n} \Rightarrow \mathbf{A} \mathbf{A}^T \in n \times n \therefore \mathbf{V} \in n \times n \)

**Prove that \( \mathbf{U} \) is the eigenbasis of \( \mathbf{A} \) (col. space of \( \mathbf{A} \)).**

Now, how to find \( \mathbf{U} \)?

\[
\mathbf{A} \mathbf{A}^T = (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)(\mathbf{V} \mathbf{\Sigma} \mathbf{U}^T) = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \mathbf{\Sigma} \mathbf{U}^T \\
= \mathbf{U} \mathbf{\Sigma}^2 \mathbf{U}^T = \mathbf{U} \mathbf{\Sigma} \mathbf{U}^{-1} \\
\Rightarrow \quad \mathbf{A} \mathbf{A}^T \mathbf{U} = \mathbf{U} \mathbf{\Sigma} \mathbf{V} \\
\text{(Eigen vector of } \mathbf{A} \mathbf{A}^T). \]

\( \mathbf{U} = m \times m \)

**Prove that \( \mathbf{U} \) and \( \mathbf{V} \) are both orthogonal.**
Prove that matrix $A$ always has the SVD decomposition

\[ A^T A \cdot V = \lambda V \quad \rightarrow \quad \text{always true, } \lambda > 0 \text{ and } V \text{ is } \mathbb{I} \text{ since } A^T A \text{ is PSD.} \]

\[ A^T \left( \frac{AV}{\sqrt{\lambda}} \right) = \frac{\lambda}{\sqrt{\lambda}} V \]

Now \[ A A^T \left( \frac{AV}{\sqrt{\lambda}} \right) = A \left( \frac{\lambda}{\sqrt{\lambda}} V \right) = \lambda \left( \frac{AV}{\sqrt{\lambda}} \right) \quad \rightarrow \quad \text{This is the eigenvector eq. for } A A^T. \]

\[ \therefore \quad \text{The matrix } \left( \frac{AV}{\sqrt{\lambda}} \right) \text{ must be orthonormal, since } A A^T \text{ is PSD.} \]

Let \[ U = \frac{AV}{\sqrt{\lambda}} \text{ where } U \text{ is orthonormal.} \quad \therefore \quad AV = U \sqrt{\lambda} \]

\[ \therefore \quad A = U \sqrt{\lambda} V^{-1} = U \Sigma V^T \]
PRINCIPAL COMPONENT ANALYSIS (PCA)
**PCA: Principal Component Analysis**

**Prerequisite:** Matrix $A = \begin{bmatrix} d_1 & d_2 & \cdots & d_n \\ x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix}$

Covariance $(A) = AA^T = \begin{bmatrix} \sum x_1^2 & \sum x_1 y_1 & \cdots & \sum x_1 z_1 \\ \sum y_1 x_1 & \sum y_1^2 & \cdots & \sum y_1 z_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sum z_1 x_1 & \sum z_1 y_1 & \cdots & \sum z_1 z_1 \end{bmatrix} = \begin{bmatrix} \text{Var}(x) & \text{Cov}(x,y) & \text{Cov}(x,z) \\ \text{Cov}(y,x) & \text{Var}(y) & \text{Cov}(y,z) \\ \text{Cov}(z,x) & \text{Cov}(z,y) & \text{Var}(z) \end{bmatrix}$

**PCA's Goal:** Which basis $B$ will make the data uncorrelated?

**Ans:** Let's represent data in another orthogonal basis $B$.

When data $d_i$ is represented in this new basis $B$, it becomes, say, $z_i$.

Note: If $B$ is a Fourier basis, then $z_i$ is the Fourier transform.

$D = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix} \equiv \begin{bmatrix} d_1 & d_2 & \cdots & d_n \end{bmatrix}$

$B \cdot Z = D$

Now, to be uncorrelated, covariance of data (in new basis) should be a diagonal matrix (because uncorrelated means $\text{cov}(x,y) = 0$)

Now, data covariance (in new basis) $= ZZ^T$

$ZZ^T = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \\ v_1 & v_2 & \cdots & v_n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix} \begin{bmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ \vdots & \vdots & \vdots \\ u_n & v_n & w_n \end{bmatrix} = \Lambda$

$(B^T D)(B^T D)^T = \Lambda$

$B^T D \cdot D^T (B^T D)^T = \Lambda$

$D^T (B^T D)^T = B \Lambda$

$D^T B = B \Lambda$

Thus, the eigenvectors of the data covariance matrix gives us the desired basis vectors to decorrelate the data.
Now, to compress data $D$, basically remove the last $K$ columns of $B$ and last $K$ rows of $Z$, then take the product of the matrices $B'Z' = D'$. This $D'$ is the compressed matrix.
③ How do you know maxima or minima?

\[
\frac{\partial^2 f(x)}{\partial x^2} > 0 \quad \text{i.e.} \quad \frac{\partial^2 f(x^*)}{\partial x^2} > 0 \quad \text{at} \quad x = x^*
\]

④ Functions in higher dimensions (i.e., when \( \mathbf{x} \) is vector) \( \Rightarrow \quad f : \mathbb{R}^n \rightarrow \mathbb{R} \)

\[
\nabla f(x) \equiv \nabla f_x = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}
\]

\( \text{“nabla” or “del”} \)

\[
\nabla^2 f(x) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}
\]

Called the “Hessian” matrix

⑤ How do we find maxima/minima of such functions of vectors?

\[
\nabla f_x = 0 \quad \Rightarrow \quad \text{gives extremums}
\]

\[
\nabla^2 f_{x^*} > 0 \quad \Rightarrow \quad \text{indicates minima}
\]

Hessian is a positive definite matrix.

Matrix \( A \) is P.D. when all \( \lambda_i(A) > 0 \) or \( x^T A x > 0, \forall x \)

Positive semi definite (PSD) when \( \lambda_i(A) \geq 0, x^T A x \geq 0, \forall x \)
Note: $\nabla^2 f_{x^*} \geq 0$ is a necessary but not sufficient condition

Example: $f(x) = x^3$
$\nabla f_x = 3x^2 = 0 \Rightarrow x^* = 0$

But is $x^*$ a minima or maxima or neither?
$\nabla^2 f(x^*) = 6x \bigg|_{x=0} = 0$
But observe that $x^* = 0$ is neither a minima or maxima.

$\nabla f_x > 0$ is sufficient condition

$\nabla f_x = 0$ and $\nabla^2 f_x > 0$ gives us local minima.
But how can I get global minima?

Well, if $f_x$ is a convex $f_{x'}$, then local minima is global minima.

What's a convex $f_{x'}$?
Functions that have an upward curvature everywhere.

Intuitively: The straight line joining any two points $f(x_1)$ and $f(x_2)$ always lies above $f(y)$, where $y \in [x_1, x_2]$. 
Mathematically: \( \alpha f(x_1) + (1-\alpha)f(x_2) \geq f(\alpha x_1 + (1-\alpha)x_2), \ \alpha \in [0,1] \)

How to test for convexity? \( \nabla^2 f_x \geq 0 \iff \text{convex fns.} \)

Summary: Given \( f(x) \),
if \( \nabla^2 f(x) \geq 0 \) (i.e., Positive semi-def Hessian)
then \( f(x) \) is convex fn.
Thus \( f(x) = 0 \) gives \text{GLOBAL MINIMA}.

But here is the bad news:

- Even if \( f(x) \) is convex, in many cases, it's difficult to solve for \( \nabla f(x) = 0 \).

Example: \( f(x) = e^x + x^2 \)

Closed form solution difficult.

We need to solve such functions iteratively.
Motivates gradient descent.

Main idea: We want to start at some \( x = x_0 \)
Move \( x_0 \rightarrow x_1 \rightarrow x_2 \ldots \rightarrow x^* \)
s.t. \( x^* \) is local/global minima of \( f(x) \)

This implies: \( f(x_{k+1}) < f(x_k) \)
so from \( x_k \), we should go along a direction that decreases the value of \( f(x_k) \)
3. What direction will take us most downward?

Answer: The direction of \(-\nabla f(x_k)\).

**Proof:**

Taylor's 1st order expansion says

\[
f(y) = f(x) + \nabla f(x)^T (y-x) + O(|y-x|)
\]

\[
\therefore f(x_k + \epsilon \hat{v}) = f(x_k) + \epsilon \nabla f(x_k)^T \hat{v} + O(\epsilon)
\]

\[
\lim_{\epsilon \to 0} \frac{f(x_k + \epsilon \hat{v}) - f(x_k)}{\epsilon} = \nabla f(x_k)^T \hat{v}
\]

Rate of change of \(f(x)\) along direction \(\hat{v}\)

So what is the max and min value of \(\nabla f(x_k)^T \hat{v}\)?

By Cauchy–Schwarz inequality

\[
-\|\nabla f(x_k)\| \|\hat{v}\| \leq \nabla f(x_k)^T \hat{v} \leq \|\nabla f(x_k)\| \|\hat{v}\|
\]

\[
\therefore \text{Maximal downward direction} = -\nabla f(x_k)
\]

Thus:

\[
\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_k = \mathbf{x}_k - \alpha \nabla f(x_k)
\]

This is called "steepest gradient descent (SGD)"
Steepest Grad. Descent Algorithm :

1. \( k = 0 \); \( \alpha = \text{small positive value} \);
   \( \varepsilon = \text{very small value} \)
2. \( x[k] = \text{random vector} \)
3. Calculate \( \nabla f(x[k]) \)
4. \( x[k+1] = x[k] - \alpha \nabla f(x[k]) \)
5. if \( f(x[k+1]) - f(x[k]) < \varepsilon \) then terminate
6. \( k++ \)
7. Goto 3

Questions:

(a) Why does step size \( \alpha \) need to be small?
(b) Can you draw a case where SGD may not converge if \( \alpha \) is not small enough?
(c) Does SGD take the shortest path from \( x_0 \) to \( x^* \)?