Lecture 10 - Unit 9 CA Notes

10.1 Classifying Critical Points - 1D example

Note that in previous calculus classes, when we computed the second derivative \( f''(x) \), we are in fact computing the Hessian, which is a \( 1 \times 1 \) diagonal matrix \( H = [f''(x)] \). Recall that the eigenvalues of a diagonal matrix are on the diagonal, so in this case \( \lambda = f''(x) \). If the eigenvalues (which in a 1D case is just \( \lambda = f''(x) \)) are all positive, then it is a local minimum, and if they are all negative, then it is a local maximum, and otherwise mixed eigenvalues (in 1D this means \( \lambda = 0 \)) indicate a saddle point.

10.2 Quadratic Form

We have \( f(x) = x^T \hat{A}x - \hat{b}^T x + \hat{c} \) as a quadratic form, and for symmetric \( \hat{A} \) the critical points satisfy \( \hat{A}x = \hat{b} \), and the Hessian is \( H = \hat{A} \). For our least squares minimization problem, our objective function is \( c^T A^T A c - 2b^T A c \), which is equivalent to \((1/2)c^T A^T A c - b^T A c \). This is a quadratic form with \( \hat{A} = A^T A \) (symmetric) and \( \hat{b} = A^T b \) (and \( \hat{c} = 0 \)).

10.3 Normal Equations

A mistake that people often make is to write down \( Ax = b \), and we have stressed that \( Ax \neq b \)! They then take the incorrect starting point and subsequently write down \( A^T Ax = A^T b \), which turns out to be correct. One can think of a simple analogy, where one start with an inconsistent set of equations in \( Ax \neq b \), similar to \( 5 \neq 6 \), then apply \( A^T \) to get \( A^T Ax = A^T b \), similar to \( 0 \cdot 5 = 0 \cdot 6 \), which turns out to be consistent. What happens here is that multiplying by \( A^T \) (or 0) throws away the inconsistent part of the matrix.

The better way to think about this is to follow what we have done in class: form and solve the minimization problem on the residual (in L2 norm), and showing that the critical point \( c^* \) that satisfies \( A^T A c^* = A^T b \) is a minimum if \( A^T A \) is SPD.

10.4 Least Squares with SVD

10.4.1 A note on \( \Sigma \) in SVD for Non-square \( A \)

We often loosely refer to \( \Sigma \) as a digaonal matrix, for \( \Sigma \) in the SVD of a matrix \( A \in \mathbb{R}^{m \times n} \). In fact, the non-reduced form of \( \Sigma \) is \( m \times n \) just like \( A \) - it is roughly diagonal, but not actually diagonal if it is non-square.

Now, \( \Sigma \Sigma^T \) is \( m \times m \), and \( \Sigma^T \Sigma \) is \( n \times n \), and both are square and diagonal. If \( A \) is full column rank and \( m > n \) (tall), then \( \Sigma \Sigma^T \) has \( n \) nonzero singular values, and the rest \( m - n \) diagonal entries
are 0, and $\Sigma^T\Sigma$ has $n$ positive entries and is full rank. In the lecture notes, by $\Sigma^2$ we are referring to $\Sigma^T\Sigma$.

For example, in class we showed the case of $m = 5$ and $n = 3$, where

$$
\Sigma = \begin{bmatrix}
\sigma_1 & 0 & 0 \\
0 & \sigma_2 & 0 \\
0 & 0 & \sigma_3 \\
0 & 0 & 0
\end{bmatrix} \in \mathbb{R}^{5 \times 3}, \quad \Sigma^T = \begin{bmatrix}
\sigma_1 & 0 & 0 & 0 \\
0 & \sigma_2 & 0 & 0 \\
0 & 0 & \sigma_3 & 0
\end{bmatrix} \in \mathbb{R}^{3 \times 5},
$$

$$
\Sigma\Sigma^T = \begin{bmatrix}
\sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\
\sigma_2^2 & \sigma_3^2 & 0 \\
\sigma_3^2 & 0 & 0
\end{bmatrix} \in \mathbb{R}^{5 \times 5}, \quad \Sigma^T\Sigma = \begin{bmatrix}
\sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\
\sigma_2^2 & \sigma_3^2 & 0 \\
\sigma_3^2 & 0 & 0
\end{bmatrix} \in \mathbb{R}^{3 \times 3}.
$$

10.4.2 Showing that $A^T A$ is SPD for Full Rank $A$

Note that here we use the loose notation as explained above where $\Sigma^2 = \Sigma^T\Sigma$ has full rank (no zeros on the diagonal) for square/tall $A$.

Since $A = U\Sigma V^T$, $A^T A = V\Sigma^2 V^T$. As the lecture note shows, we can transform the eigenproblem for $A^T A$ to be $\Sigma^2 V^T v = \lambda V^T v$.

First, let’s call $w = V^T v$, then we have $\Sigma^2 w = \lambda w$. This is very easy since $\Sigma^2$ is diagonal! The eigenvalues are on the diagonal $\sigma_k^2$'s, and the eigenvectors are the standard basis vectors $e_k$'s! For example,

$$
\begin{bmatrix}
\sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\
\sigma_2^2 & \sigma_3^2 & 0 \\
\sigma_3^2 & 0 & 0
\end{bmatrix} \begin{bmatrix}1 \\ 0 \\ 0 \end{bmatrix} = \sigma_1^2 \begin{bmatrix}1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix}
\sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\
\sigma_2^2 & \sigma_3^2 & 0 \\
\sigma_3^2 & 0 & 0
\end{bmatrix} \begin{bmatrix}0 \\ 1 \\ 0 \end{bmatrix} = \sigma_2^2 \begin{bmatrix}0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix}
\sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\
\sigma_2^2 & \sigma_3^2 & 0 \\
\sigma_3^2 & 0 & 0
\end{bmatrix} \begin{bmatrix}0 \\ 0 \\ 1 \end{bmatrix} = \sigma_3^2 \begin{bmatrix}0 \\ 0 \\ 1 \end{bmatrix}.
$$

Then, when we transform it back to $v = V w$, we have $v_k = V e_k$ is the $k$-th column of $V$! Thus, the eigenvectors of $A^T A$ are the columns of $V$. Since the eigenvalues are all positive, we know that $A^T A$ is SPD.

10.4.3 Conditioning

$A^T A = V \Sigma^2 V^T$, so basically $U$ becomes $V$, and $\Sigma$ becomes $\Sigma^2$. As a side note, for symmetric matrices, the SVD has $U = V$. It is not a good idea to directly try to solve for the normal equations, since the condition number is squared as the singular values are squared. It provides cool theoretical tools for reasoning, but in practice one should not write algorithms based on this!

10.5 Understanding Least Squares

We will further clarify the dimension of the notation in the lecture notes. Suppose $A$ is $m \times n$, with $m \geq n$ (tall/square). When we write $\Sigma = \begin{bmatrix} D \\ 0 \end{bmatrix}$, we assume $D$ is diagonal $n \times n$ matrix.

Similarly, when we write $U^T b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, we assume $b_1$ is an $n$ dimensional vector, and $b_2$ is an $m-n$ dimensional vector.
Note that solving the regular linear equations $Ax = b$ via SVD automatically gets us the least squares solution $VD^{-1}b_1$:

\[
Ax = b
\]

\[
U \Sigma V^T x = b
\]

\[
\Sigma \hat{x} = \hat{b}, \quad \text{where } \hat{x} = V^T x, \hat{b} = U^T b
\]

\[
\begin{bmatrix} D \\ 0 \end{bmatrix} \hat{x} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}
\]

\[
V^T x = \hat{x} = D^{-1}b_1
\]

\[
x = VD^{-1}b_1
\]