Basic Optimization
The Jacobian of \( F(c) = \begin{pmatrix} F_1(c) \\ F_2(c) \\ \vdots \\ F_m(c) \end{pmatrix} \) has entries \( J_{ik} = \frac{\partial F_i}{\partial c_k}(c) \).

Thus, the Jacobian \( J(c) = F'(c) = \begin{pmatrix} \frac{\partial F_1}{\partial c_1}(c) & \frac{\partial F_1}{\partial c_2}(c) & \cdots & \frac{\partial F_1}{\partial c_n}(c) \\ \frac{\partial F_2}{\partial c_1}(c) & \frac{\partial F_2}{\partial c_2}(c) & \cdots & \frac{\partial F_2}{\partial c_n}(c) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial c_1}(c) & \frac{\partial F_m}{\partial c_2}(c) & \cdots & \frac{\partial F_m}{\partial c_n}(c) \end{pmatrix} \)
Gradient

• Consider the scalar (output) function $f(c)$ with multi-dimensional input $c$

• The Jacobian of $f(c)$ is $J(c) = \left( \frac{\partial f}{\partial c_1} (c) \quad \frac{\partial f}{\partial c_2} (c) \quad \cdots \quad \frac{\partial f}{\partial c_n} (c) \right)$

• The gradient of $f(c)$ is $\nabla f(c) = J^T(c) = \left( \begin{array}{c} \frac{\partial f}{\partial c_1} (c) \\ \frac{\partial f}{\partial c_2} (c) \\ \vdots \\ \frac{\partial f}{\partial c_n} (c) \end{array} \right)$

• In 1D, both $J(c)$ and $\nabla f(c) = J^T(c)$ are the usual $f'(c)$
Critical Points

• To identify critical points of $f(c)$, set the gradient to zero: $\nabla f(c) = 0$

$$
\begin{bmatrix}
\frac{\partial f}{\partial c_1}(c) \\
\frac{\partial f}{\partial c_2}(c) \\
\vdots \\
\frac{\partial f}{\partial c_n}(c)
\end{bmatrix} = 0 \quad \text{or} \quad
\begin{bmatrix}
\frac{\partial f}{\partial c_1}(c) = 0 \\
\frac{\partial f}{\partial c_2}(c) = 0 \\
\vdots \\
\frac{\partial f}{\partial c_n}(c) = 0
\end{bmatrix}
$$

• This is a system of equations:

• Any $c$ that simultaneously solves all the equations is a critical point.

• In 1D, this is the usual $f''(c) = 0$
Hessian

• The Hessian of $f(c)$ is $H(c) = J\left(\nabla f(c)\right)^T$ and has entries $H_{ik} = \frac{\partial^2 f}{\partial c_i \partial c_k}(c)$

$$
\begin{pmatrix}
\frac{\partial^2 f}{\partial c_1^2}(c) & \frac{\partial^2 f}{\partial c_1 \partial c_2}(c) & \ldots & \frac{\partial^2 f}{\partial c_1 \partial c_n}(c) \\
\frac{\partial^2 f}{\partial c_2 \partial c_1}(c) & \frac{\partial^2 f}{\partial c_2^2}(c) & \ldots & \frac{\partial^2 f}{\partial c_2 \partial c_n}(c) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial c_n \partial c_1}(c) & \frac{\partial^2 f}{\partial c_n \partial c_2}(c) & \ldots & \frac{\partial^2 f}{\partial c_n^2}(c)
\end{pmatrix}
$$

• The Hessian is $H(c) = \begin{pmatrix}
\frac{\partial^2 f}{\partial c_1^2}(c) & \frac{\partial^2 f}{\partial c_1 \partial c_2}(c) & \ldots & \frac{\partial^2 f}{\partial c_1 \partial c_n}(c) \\
\frac{\partial^2 f}{\partial c_2 \partial c_1}(c) & \frac{\partial^2 f}{\partial c_2^2}(c) & \ldots & \frac{\partial^2 f}{\partial c_2 \partial c_n}(c) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial c_n \partial c_1}(c) & \frac{\partial^2 f}{\partial c_n \partial c_2}(c) & \ldots & \frac{\partial^2 f}{\partial c_n^2}(c)
\end{pmatrix}$

• $H(c)$ is symmetric, when the order of differentiation doesn’t matter

• In 1D, this is the usual $f''(c)$
Differential Forms

• Vector valued function: \( dF(c) = J(F(c)) dc \)

• Scalar valued function (no different): \( df(c) = J(f(c)) dc \)

• Transpose second equation (above): \( df(c) = dc^T \nabla f(c) \)

• Substitute \( \nabla f \) for \( F \) (above): \( d\nabla f(c) = J(\nabla f(c)) dc \) or \( d\nabla f(c) = H^T(c) dc \)

• Take differential of \( df(c) \) transposed: \( d^2 f(c) = J(dc^T \nabla f(c)) dc \)

• Some hand waving: \( d^2 f(c) = dc^T H^T(c) dc \)
Classifying Critical Points

• Given critical point $c^*$, i.e. with $\nabla f(c^*) = 0$, the Hessian is used to classify it.
• If $H(c^*)$ is **positive definite**, then $c^*$ is a **local minimum**.
• If $H(c^*)$ is **negative definite**, then $c^*$ is a **local maximum**.
• Otherwise, $H(c^*)$ is indefinite, and $c^*$ is a saddle point.
Classifying Critical Points (in 1D)

- In 1D, given critical point \( c^* \), i.e. with \( \nabla f(c^*) = f'(c^*) = 0 \), the Hessian is used to classify it.
- In 1D, \( H(c^*) = (f''(c^*)) \) is a size 1x1 diagonal matrix with eigenvalue \( f''(c^*) \).

  - If \( H(c^*) \) is positive definite with eigenvalue \( f''(c^*) > 0 \), then \( c^* \) is a local minimum.
    - As usual, \( f''(c^*) > 0 \) implies concave up and a local min.
  - If \( H(c^*) \) is negative definite with eigenvalue \( f''(c^*) < 0 \), then \( c^* \) is a local maximum.
    - As usual, \( f''(c^*) < 0 \) implies concave down and a local max.
  - Otherwise, \( H(c^*) \) is indefinite with eigenvalue \( f''(c^*) = 0 \), and \( c^* \) is a saddle point.
    - As usual, \( f''(c^*) = 0 \) implies an inflection point (not a local extrema).
Quadratic Form

• The quadratic form of a square matrix $\tilde{A}$ is $f(c) = \frac{1}{2} c^T \tilde{A} c - \tilde{b}^T c + \tilde{c}$
  • In 1D, $f(c) = \frac{1}{2} \tilde{a} c^2 - \tilde{b} c + \tilde{c}$

• Minimize $f(c)$ by (first) finding critical points where $\nabla f(c) = 0$

• Note $\nabla f(c) = \frac{1}{2} \tilde{A} c + \frac{1}{2} \tilde{A}^T c - \tilde{b}$, since $J(c^T v) = J(v^T c) = v^T$ (the gradient is $v$)
  • Solve the symmetric system $\frac{1}{2}(\tilde{A} + \tilde{A}^T)c = \tilde{b}$ to find critical points

• When $\tilde{A}$ is symmetric, $\nabla f(c) = \tilde{A} c - \tilde{b} = 0$ is satisfied when $\tilde{A} c = \tilde{b}$
  • In 1D, the critical point is on the line of symmetry $\tilde{c} = \frac{\tilde{b}}{\tilde{a}}$

• That is, solve $\tilde{A} c = \tilde{b}$ to find the critical point
Quadratic Form

• The Hessian of \( f(c) \) is \( H = \frac{1}{2} (\tilde{A}^T + \tilde{A}) \) or just \( \tilde{A} \) when \( \tilde{A} \) is symmetric

• When \( \tilde{A} \) is SPD, the solution to \( \tilde{A}x = \tilde{b} \) is a minimum

• When \( \tilde{A} \) is symmetric negative definite, the solution to \( \tilde{A}x = \tilde{b} \) is a maximum

• When \( \tilde{A} \) is indefinite, the solution to \( \tilde{A}x = \tilde{b} \) is a saddle point

• In 1D, \( H = (\tilde{a}) \) is a size 1x1 diagonal matrix with eigenvalue \( \tilde{a} \)

• As usual, \( \tilde{a} > 0 \) implies concave up and a local min

• As usual, \( \tilde{a} < 0 \) implies concave down and a local max

• As usual, \( \tilde{a} = 0 \) implies an inflection point (not a local extrema)
Recall: Least Squares (Unit 8)

- Minimizing $\|r\|_2$ is referred to as least squares, and the resulting solution is referred to as the least squares solution
  - The least squares solution is the unique solution when $\|r\|_2 = 0$
- Minimizing $\|Dr\|_2$ is referred to as weighted least squares

- $\|r\|_2$ is minimized when $\|r\|_2^2$ is minimized
- And $\|r\|_2^2 = r \cdot r = (b - Ac) \cdot (b - Ac) = c^T A^T Ac - 2b^T Ac + b^T b$ is minimized when $c^T A^T Ac - 2b^T Ac$ is minimized
- Thus, minimize $c^T A^T Ac - 2b^T Ac$
- Similarly, for weighted least squares, minimize $c^T A^T D^2 Ac - 2b^T D^2 Ac$
Normal Equations

• $c^T A^T D^2 Ac - 2b^T D^2 Ac$ has the same minimum as $\frac{1}{2} c^T A^T D^2 Ac - b^T D^2 Ac$

• This is a quadratic form with symmetric $\tilde{A} = A^T D^2 A$ and $\tilde{b} = A^T D^2 b$

• The critical point is found from solving $\tilde{A}c = \tilde{b}$ or $A^T D^2 Ac = A^T D^2 b$

• Weighted least squares defaults to ordinary least squares when $D = I$

• So, for (unweighted) least squares, solve $A^T Ac = A^T b$

• These are called the normal equations
Hessian

• Recall: $A$ is a tall (or square) full rank matrix with size $mxn$ where $m \geq n$

• The Hessian $H = \tilde{A} = A^T A = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T = V \Lambda V^T$
  • where $\Lambda = \Sigma^T \Sigma$ is a size $nxn$ matrix of (nonzero) singular values squared
  • $HV = V\Lambda$ illustrates that $H$ has all positive eigenvalues (and so is SPD)

• That is, the critical point is indeed a minimum (as desired)

• For weighted least squares, note that if $D$ is a diagonal matrix with no zeroes on the diagonal, then $DAc = 0$ if and only if $Ac = 0$

• That is, a full column rank $A$ implies a full column rank $DA$

• Then, the SVD of $DA$ can be used to prove that $H = (DA)^T (DA)$ is SPD