Basic Optimization

Jacobian

- 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 <u>scalar</u> (output) function $f(c)$ with multi-dimensional input c
- The Jacobian of $f(c)$ is $J(c) = \begin{pmatrix} \frac{\partial f}{\partial c_1}(c) & \frac{\partial f}{\partial c_2}(c) & \cdots & \frac{\partial f}{\partial c_n}(c) \end{pmatrix}$
- The <u>gradient</u> of $f(c)$ is $\color{red}{\nabla f(c) = J^T(c)} = \begin{pmatrix} \frac{\partial f}{\partial c_1}(c) \\ \frac{\partial f}{\partial c_2}(c) \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) \end{pmatrix}$
- 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: $\nabla f(c) = 0$
- This is a system of equations: $\begin{pmatrix} \frac{\partial f}{\partial c_1}(c) \\ \frac{\partial f}{\partial c_2}(c) \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) \end{pmatrix} = 0$ or $\begin{pmatrix} \frac{\partial f}{\partial c_1}(c) = 0 \\ \frac{\partial f}{\partial c_2}(c) = 0 \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) = 0 \end{pmatrix}$
- Any c that simultaneously solves all the equations is a critical point
- In 1D, this is the usual $f'(c) = 0$

Jacobian of the Gradient

- Taking the Jacobian of the column vector gradient gives:

- The $J\big(\nabla f(c)\big) = \begin{pmatrix} \frac{\partial^2 f}{\partial c_1 \partial c_1}(c) & \frac{\partial^2 f}{\partial c_2 \partial c_1}(c) & \cdots & \frac{\partial^2 f}{\partial c_n \partial c_1}(c) \\ \frac{\partial^2 f}{\partial c_1 \partial c_2}(c) & \frac{\partial^2 f}{\partial c_2 \partial c_2}(c) & \cdots & \frac{\partial^2 f}{\partial c_n \partial c_2}(c) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial c_1 \partial c_n}(c) & \frac{\partial^2 f}{\partial c_2 \partial c_n}(c) & \cdots & \frac{\partial^2 f}{\partial c_n \partial c_n}(c) \end{pmatrix}$

- Note: $\frac{\partial^2 f}{\partial c_2 \partial c_1} = \frac{\partial}{\partial c_2}\left(\frac{\partial f}{\partial c_1}\right) = f_{c_1 c_2}$

Hessian

- The Hessian of $f(c)$ is $H(c) = J\big(\nabla f(c)\big)^T$ and has entries $H_{ik} = \dfrac{\partial f}{\partial c_i \partial c_k}(c)$
- The Hessian is $H(c) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial c_1^2}(c) & \dfrac{\partial^2 f}{\partial c_1 \partial c_2}(c) & \cdots & \dfrac{\partial^2 f}{\partial c_1 \partial c_n}(c) \\ \dfrac{\partial^2 f}{\partial c_2 \partial c_1}(c) & \dfrac{\partial^2 f}{\partial c_2^2}(c) & \cdots & \dfrac{\partial^2 f}{\partial c_2 \partial c_n}(c) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial c_n \partial c_1}(c) & \dfrac{\partial^2 f}{\partial c_n \partial c_2}(c) & \cdots & \dfrac{\partial^2 f}{\partial c_n^2}(c) \end{pmatrix}$
- $H(c)$ is symmetric, when the order of differentiation doesn't matter
- In 1D, this is the usual $f''(c)$

Differential Forms

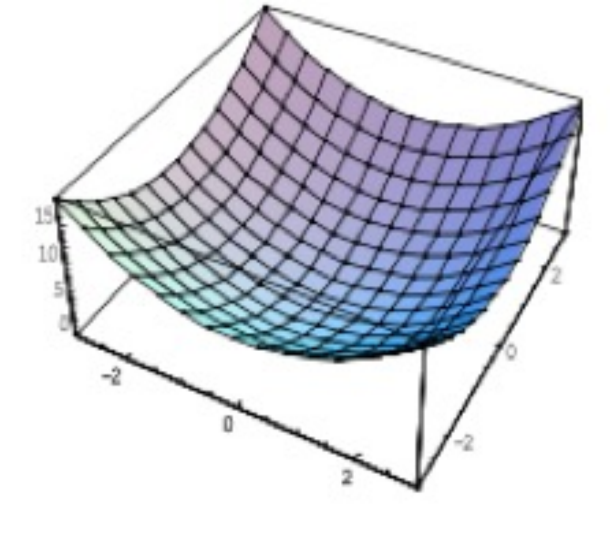
- Vector valued function: $dF(c) = J\big(F(c)\big)dc$
- Substitute ∇f for F to get: $d\nabla f(c) = J\big(\nabla f(c)\big)dc = H^T(c)dc$

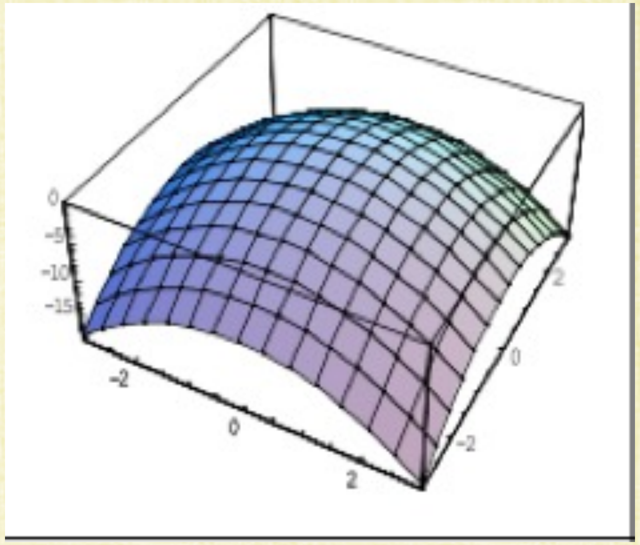
- Scalar valued function: $df(c) = J\big(f(c)\big)dc$
- Take the transpose: $df(c) = dc^T \nabla f(c)$

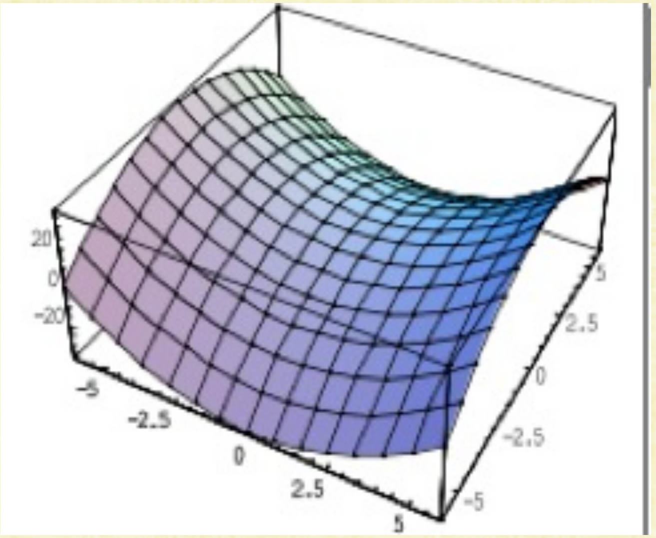
- Take (another) differential: $d^2 f(c) = J\big(dc^T \nabla f(c)\big)dc$
- Some hand waving: $d^2 f(c) = dc^T H^T(c)dc = dc \cdot H^T(c)dc$

Classifying Critical Points

- Given a critical point c^*, i.e. with $\nabla f(c^*) = 0$, the Hessian is used to classify it
- If $H(c^*)$ is <u>positive definite</u>, then c^* is a <u>local minimum</u>
- If $H(c^*)$ is <u>negative definite</u>, then c^* is a <u>local maximum</u>
- 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, $H(c^*) = \left(f''(c^*)\right)$ is a size $1x1$ diagonal matrix with eigenvalue $f''(c^*)$

- If $H(c^*)$ is <u>positive definite</u> with eigenvalue $f''(c^*) > 0$, then c^* is a <u>local minimum</u>
  - As usual, $f''(c^*) > 0$ implies concave up and a local min
- If $H(c^*)$ is <u>negative definite</u> with eigenvalue $f''(c^*) < 0$, then c^* is a <u>local maximum</u>
  - 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 <u>quadratic form</u> of a square matrix $\tilde{A}$ is $f(c) = \frac{1}{2}c^T\tilde{A}c - \tilde{b}^Tc + \tilde{c}$
  - 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 $\nabla f(c) = \frac{1}{2}\tilde{A}c + \frac{1}{2}\tilde{A}^Tc - \tilde{b}$, since $J(c^Tv) = J(v^Tc) = v^T$ (the gradient is v)
  - Solve the <u>symmetric</u> 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 $\tilde{A}c = \tilde{b}$ is a minimum
- When $\tilde{A}$ is symmetric negative definite, the solution to $\tilde{A}c = \tilde{b}$ is a maximum
- When $\tilde{A}$ is indefinite, the solution to $\tilde{A}c = \tilde{b}$ 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 (it's really a least squares solution)
  - A 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 $\|r\|_2^2 = r \cdot r = (b - Ac) \cdot (b - Ac) = c^T A^T A c - 2b^T A c + b^T b$ is minimized when $c^T A^T A c - 2b^T A c$ is minimized
- Thus, minimize $c^T A^T A c - 2b^T A c$
- For weighted least squares, minimize $c^T A^T D^2 A c - 2b^T D^2 A c$

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 $D = I$
- 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:

- Nonzero diagonal elements in D implies that $DAc = 0$ 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