1D Optimization

Part II Roadmap

- Part I – Linear Algebra (units 1-12) $Ac = b$

linearize

line search

- Part II – Optimization (units 13-20)
  - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima ← Theory
  - (units 17-18) Computing/Avoiding Derivatives
  - (unit 19) Hack 1.0: "I give up" $H = I$ and J is mostly 0 (descent methods)
  - (unit 20) Hack 2.0: "It's an ODE!?" (adaptive learning rate and momentum)

Methods (units 17-20)

Leveraging Root Finding (from unit 15)

- Relative extrema of $g(t)$ occur at critical points where $g'(t) = 0$; thus, can use root finding on g' to identify relative extrema

- Newton: $t^{q+1} = t^q - \frac{g'(t^q)}{g''(t^q)}$ (dividing by g'' is even worse than dividing by g')

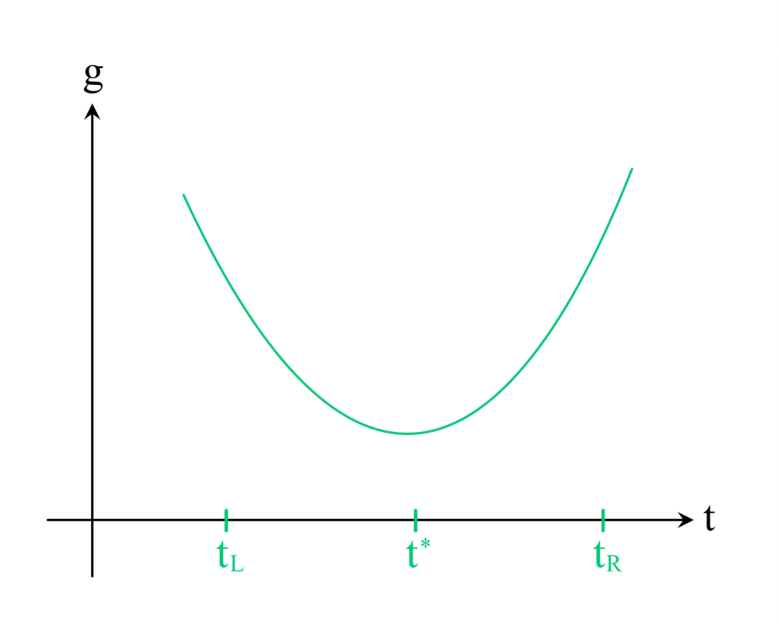
- Secant: $t^{q+1} = t^q - g'(t^q)\frac{t^q - t^{q-1}}{g'(t^q) - g'(t^{q-1})}$ (can replace g' with approximations too)

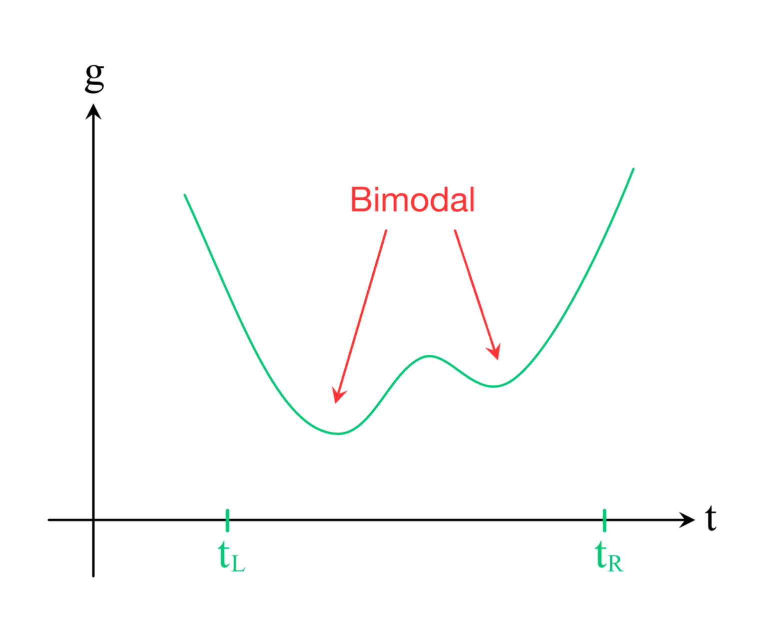
- Bisection: $g'(t_L)g'(t_R) < 0$ is the new condition

- Mixed Methods: mixing the above (as in unit 15)

Unimodal

- Unimodal means one mode (bimodal means two modes)
- In 1D optimization, this means that the function has one relative minimum
- $g(t)$ is unimodal in $[t_L, t_R]$ if and only if g is monotonically decreasing in $[t_L, t^*]$ and monotonically increasing in $[t^*, t_R]$



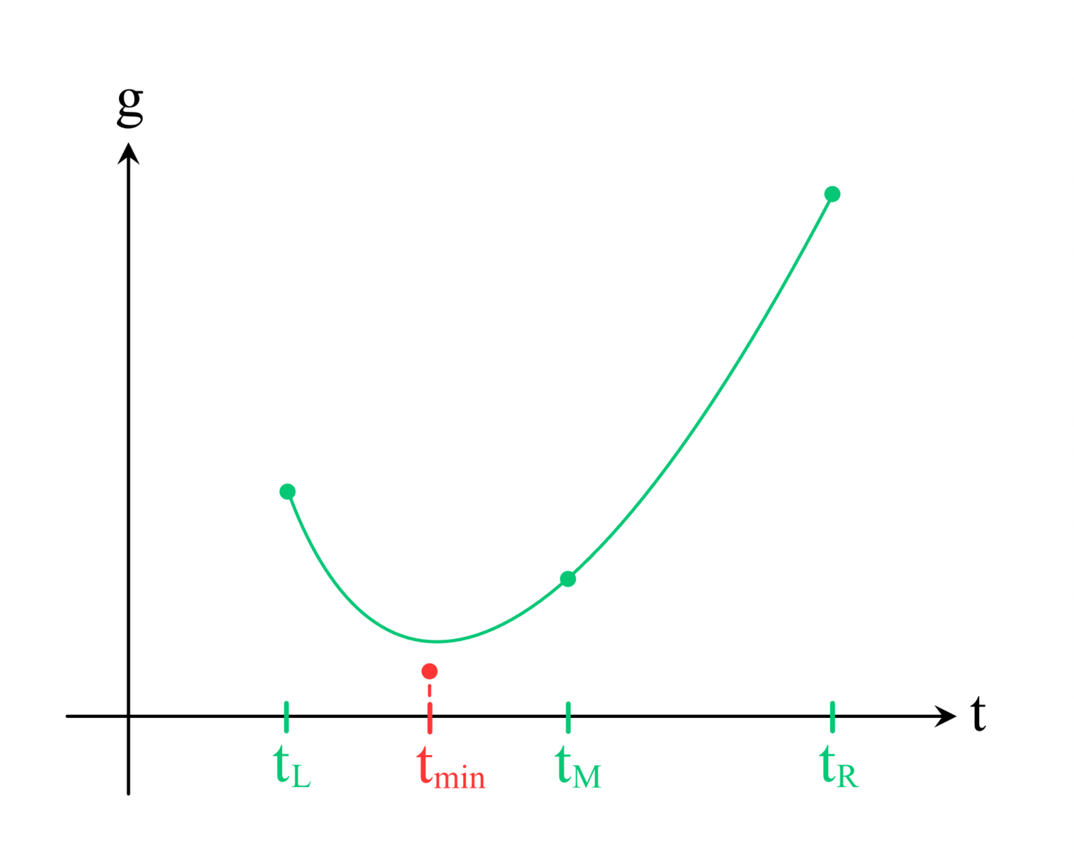


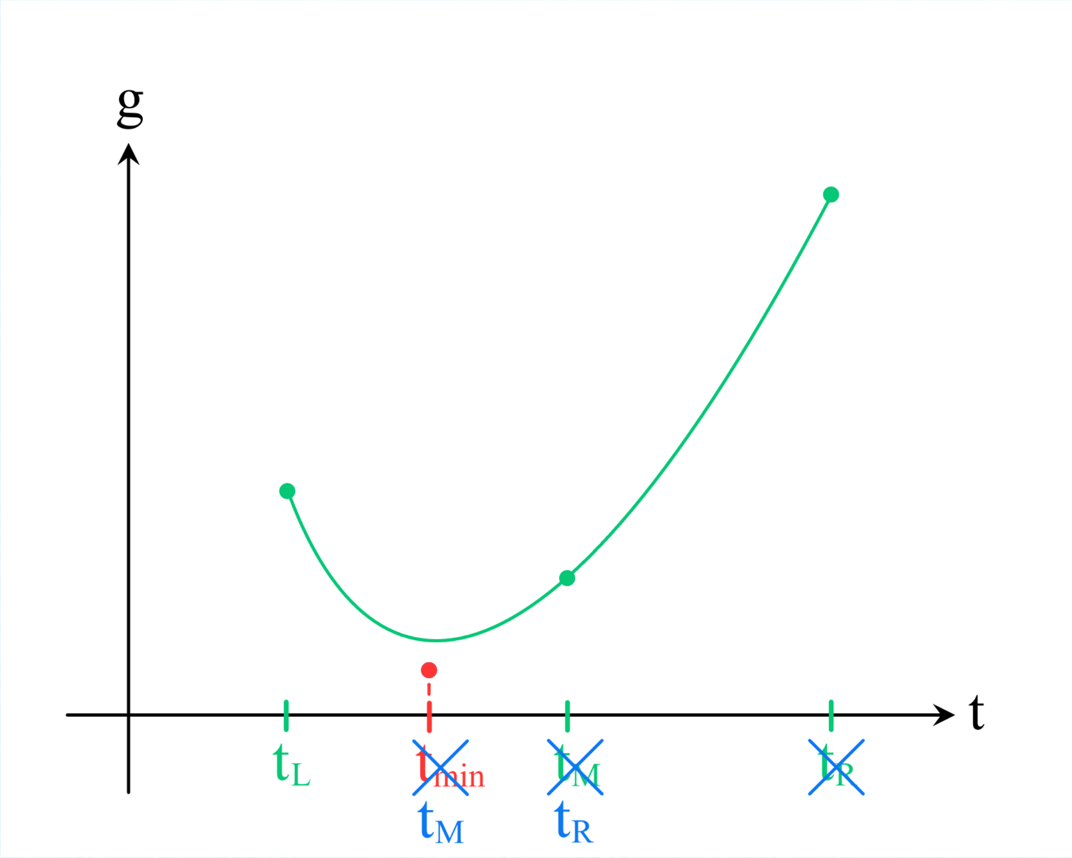
Successive Parabolic Interpolation

- Motivated by Newton/Secant (which use lines to find candidates for roots), use parabolas to find candidates for minima
- Given interval $[t_L, t_R]$ with midpoint $t_M = \frac{t_L+t_R}{2}$, create the unique parabola through t_L, t_R, and t_M
  - A unimodal g in $[t_L, t_R]$ makes this parabola concave up
  - Let t_{min} be the point where the parabola takes on its minimum value
- Assume $t_{min} < t_M$ (otherwise, simply swap their names)
- If $g(t_{min}) \leq g(t_M)$, discard $[t_M, t_R]$ which cannot contain the minimum
  - Then, set $t_R = t_M$ and $t_M = t_{min}$
- If $g(t_{min}) \geq g(t_M)$, discard $[t_L, t_{min}]$ which cannot contain the minimum
  - Then, set $t_L = t_{min}$ and $t_M = t_M$ (no change)
- Superlinear convergence rate with $p \approx 1.325$

Successive Parabolic Interpolation

- When $g(t_{min}) \leq g(t_M)$, discard $[t_M, t_R]$ and set $t_R = t_M$ and $t_M = t_{min}$





Discarding Intervals

- Bisection required only 3 points to be able to discard an interval during root finding
- Successive Parabolic Interpolation demonstrated that 4 points is enough during minimization

- Let $[t_L, t_R]$ have two intermediate points with $t_L < t_{M1} < t_{M2} < t_R$
  - If g is unimodal in $[t_L, t_R]$, one can safely discard either $[t_L, t_{M1}]$ or $[t_{M2}, t_R]$
- If $g(t_{M1}) \leq g(t_{M2})$, discard $[t_{M2}, t_R]$ which cannot contain the minimum
- If $g(t_{M1}) \geq g(t_{M2})$, discard $[t_L, t_{M1}]$ which cannot contain the minimum

Golden Section Search

- After discarding an interval, either t_{M1} or t_{M2} becomes an endpoint, and keeping the other as an interior point (efficiently) reduces evaluations of g
- Let $\delta = t_R - t_L$ be the interval size and $\lambda \in (0,.5)$ be the fraction inward of t_{M1}
- Then $t_{M1} = t_L + \lambda\delta$, and symmetric placement gives $t_{M2} = (t_L + \delta) - \lambda\delta$
- Discard the left interval (discarding the right gives the same math) to obtain $t_L^{new} = t_{M1}$ and $\delta^{new} = (1-\lambda)\delta$
- Then $t_{M2} = (t_L^{new} - \lambda\delta + \delta) - \lambda\delta = t_L^{new} + \frac{(1-2\lambda)}{1-\lambda}\delta^{new}$ can be designated as either t_{M1}^{new} or t_{M2}^{new} if $\frac{1-2\lambda}{1-\lambda}$ is equal to either λ or $1-\lambda$ (those are both quadratic equations)
- Of the four solutions, only one has $\lambda \in (0,.5)$: $\lambda = \frac{3-\sqrt{5}}{2}$ with t_{M2} becoming t_{M1}^{new}

Golden Section Search

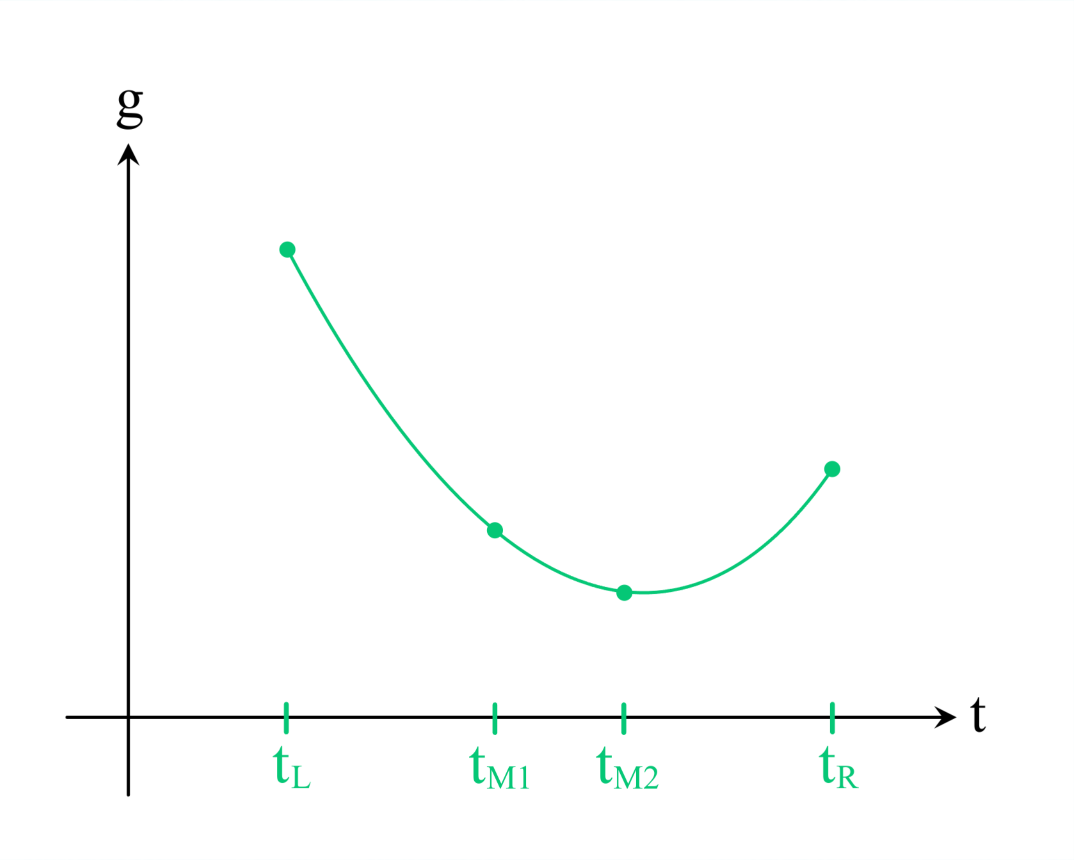
- Rewrite: $t_{M1} = (1-\lambda)t_L + \lambda t_R$ and $t_{M2} = \lambda t_L + (1-\lambda)t_R$
- Switch the parameter to the more typical $\tau = 1 - \lambda = \frac{\sqrt{5}-1}{2}$
- Then, $t_{M1} = \tau t_L + (1-\tau)t_R$ and $t_{M2} = (1-\tau)t_L + \tau t_R$

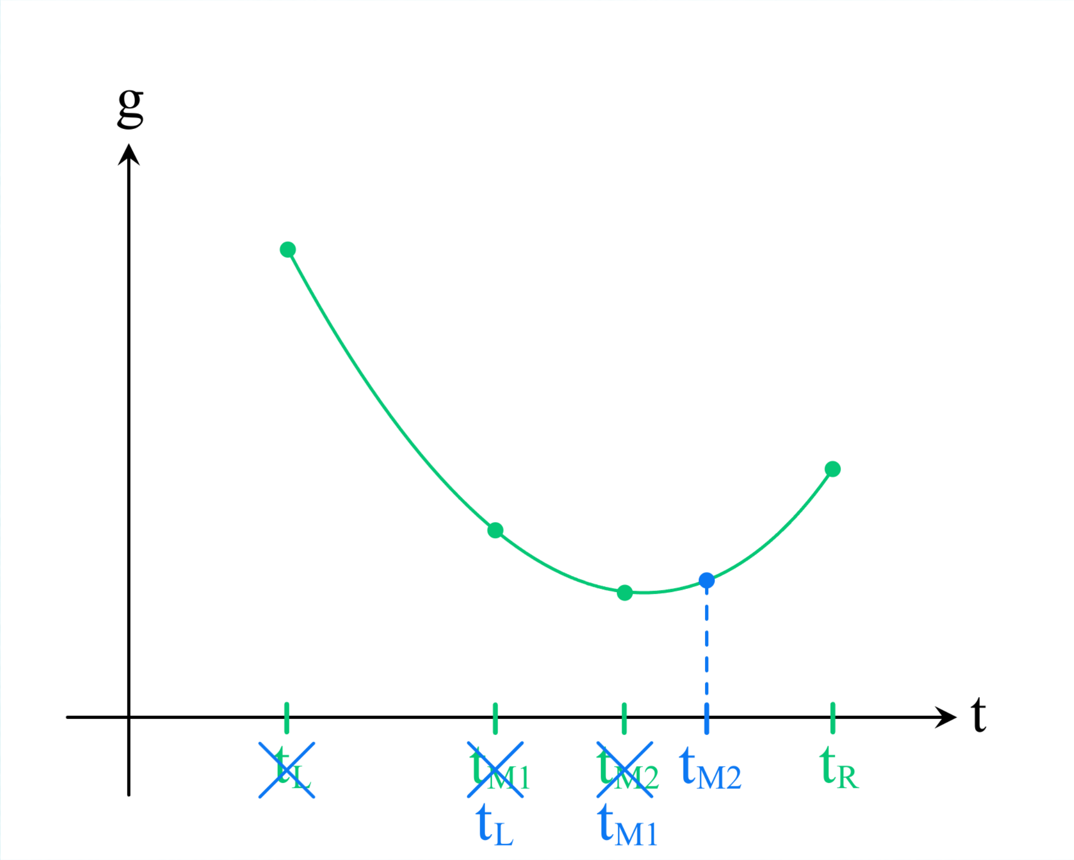
- If $g(t_{M1}) \leq g(t_{M2})$, discard $[t_{M2}, t_R]$, set $t_R = t_{M2}$, $t_{M2} = t_{M1}$, and recompute t_{M1}
- If $g(t_{M1}) \geq g(t_{M2})$, discard $[t_L, t_{M1}]$, set $t_L = t_{M1}$, $t_{M1} = t_{M2}$, and recompute t_{M2}

- Stop when the interval size is small (as usual)
- Linear convergence rate ($p = 1$) with $C = \frac{(1-\lambda)\delta}{\delta} = \tau \approx .618$

Golden Section Search

- If $g(t_{M1}) \geq g(t_{M2})$, discard $[t_L, t_{M1}]$, set $t_L = t_{M1}$, $t_{M1} = t_{M2}$, recompute t_{M2}





Mixed Methods

- Given a unimodal $[t_L, t_R]$
- Iterate with Succsessive Parabolic Interpolation as long as the iterates stay inside the interval
  - When iteration attempts to leave the interval, use prior iterates to shrink the interval as much as possible (while still containing the minima)
- If Successive Parabolic Interpolation attempts to leave the current interval, instead use Golden Section Search to continue shrinking the interval

- Leverages the speed of Successive Parabolic Interpolation, while still guaranteeing convergence via Golden Section Search

- Many/various strategies exist

Function/Derivative Requirements

- All methods require evaluation of the function g
- Root finding approaches differentiate g and solve $g'(t) = 0$ to identify critical points
  - All root finding methods require evaluation of the function, which is g' here
  - Newton (and mixed methods using Newton) requires the derivative of the function, which is g'' here

Recall: Useful Derivatives (unit 15)

- $\frac{\partial}{\partial t} c^{q+1}(t) = \Delta c^q$, since $c^{q+1}(t) = c^q + t\Delta c^q$

- $\frac{\partial}{\partial t} F\left(c^{q+1}(t)\right) = J_F\left(c^{q+1}(t)\right)\Delta c^q$ and $\frac{\partial}{\partial t} F^T\left(c^{q+1}(t)\right) = (\Delta c^q)^T J_F^T\left(c^{q+1}(t)\right)$
  - $\frac{\partial}{\partial t} F_i\left(c^{q+1}(t)\right) = (J_F)_i\left(c^{q+1}(t)\right) \Delta c^q$ where the $F_i\left(c^{q+1}(t)\right)$ are the scalar row entries of $F\left(c^{q+1}(t)\right)$

- Scalar $\hat{f}(c^{q+1}(t))$ has system $J_{\hat{f}}^T\left(c^{q+1}(t)\right) = 0$ for critical points
- $\frac{\partial}{\partial t} J_{\hat{f}}^T\left(c^{q+1}(t)\right) = H_{\hat{f}}^T\left(c^{q+1}(t)\right)\Delta c^q$ and $\frac{\partial}{\partial t} J_{\hat{f}}\left(c^{q+1}(t)\right) = (\Delta c^q)^T H_{\hat{f}}\left(c^{q+1}(t)\right)$
  - $\frac{\partial}{\partial t}\left(J_{\hat{f}}^T\right)_i (c^{q+1}(t)) = \left(H_{\hat{f}}^T\right)_i (c^{q+1}(t))\Delta c^q$

Additional Useful Derivatives

- $\frac{\partial}{\partial t} J_F\left(c^{q+1}(t)\right) = (\Delta c^q)^T H_F\left(c^{q+1}(t)\right)$
  - H_F is a rank 3 tensor of all 2nd derivatives of F
  - $\frac{\partial}{\partial t} (J_F)_i\left(c^{q+1}(t)\right) = (\Delta c^q)^T (H_F)_i\left(c^{q+1}(t)\right)$

- $\frac{\partial}{\partial t} H_{\hat{f}}^T\left(c^{q+1}(t)\right) = (\Delta c^q)^T OMG_{\hat{f}}^T\left(c^{q+1}(t)\right)$
  - $OMG_{\hat{f}}^T$ is a rank 3 tensor of all 3rd derivatives of $\hat{f}$
  - $\frac{\partial}{\partial t} \left(H_{\hat{f}}^T\right)_i \left(c^{q+1}(t)\right) = (\Delta c^q)^T \left(OMG_{\hat{f}}^T\right)_i \left(c^{q+1}(t)\right)$

Recall: Nonlinear Systems Problems (unit 15)

- Solve $J_F(c^q)\Delta c^q = (\beta - 1)F(c^q)$ for Δc^q and use $c^{q+1}(t) = c^q + t\Delta c^q$ in $F\left(c^{q+1}(t)\right) = 0$
- Option 1: find simultaneous (for all i) roots for all the $g_i(t) = F_i\left(c^{q+1}(t)\right) = 0$
  - Here, $g_i'(t) = (J_F)_i\left(c^{q+1}(t)\right)\Delta c^q$
- Option 2: find roots of $g(t) = \frac{1}{2}F^T\left(c^{q+1}(t)\right)F\left(c^{q+1}(t)\right) = 0$
  - Here, $g'(t) = \frac{1}{2}F^T\left(c^{q+1}(t)\right)J_F\left(c^{q+1}(t)\right)\Delta c^q + \frac{1}{2}(\Delta c^q)^T J_F^T\left(c^{q+1}(t)\right)F\left(c^{q+1}(t)\right)$
  - Since both terms are scalars, $g'(t) = F^T\left(c^{q+1}(t)\right)J_F\left(c^{q+1}(t)\right)\Delta c^q$

Nonlinear Systems Problems

- Solve $J_F(c^q)\Delta c^q = (\beta - 1)F(c^q)$ for Δc^q and use $c^{q+1}(t) = c^q + t\Delta c^q$ in $F\left(c^{q+1}(t)\right) = 0$
- Option 1: find simultaneous (for all i) minima for all the $g_i(t) = F_i\left(c^{q+1}(t)\right)$ aiming for roots where all $F_i\left(c^{q+1}(t)\right) = 0$
  - Here, $g_i'(t) = (J_F)_i\left(c^{q+1}(t)\right)\Delta c^q$ and $g_i''(t) = (\Delta c^q)^T (H_F)_i\left(c^{q+1}(t)\right) \Delta c^q$
- Option 2: minimize $g(t) = \frac{1}{2} F^T\left(c^{q+1}(t)\right)F\left(c^{q+1}(t)\right)$ aiming for its roots
  - Here, $g'(t) = F^T\left(c^{q+1}(t)\right)J_F\left(c^{q+1}(t)\right)\Delta c^q$
  - $g''(t) = F^T\left(c^{q+1}(t)\right)(\Delta c^q)^T H_F\left(c^{q+1}(t)\right)\Delta c^q + (\Delta c^q)^T J_F^T\left(c^{q+1}(t)\right)J_F\left(c^{q+1}(t)\right)\Delta c^q$

Recall: Optimization Problems (unit 15)

- Solve $H^T_{\hat{f}}(c^q)\Delta c^q = (\beta - 1)J^T_{\hat{f}}(c^q)$ for Δc^q and use $c^{q+1}(t) = c^q + t\Delta c^q$ in $J^T_{\hat{f}}\left(c^{q+1}(t)\right) = 0$
- <u>Option 1</u>: find simultaneous (for all i) roots for all the $g_i(t) = (J^T_{\hat{f}})_i\left(c^{q+1}(t)\right) = 0$ to find the critical points of $\hat{f}(c)$
  - Here, $g_i'(t) = (H^T_{\hat{f}})_i\left(c^{q+1}(t)\right)\Delta c^q$
- <u>Option 2</u>: find roots of $g(t) = \frac{1}{2}J_{\hat{f}}\left(c^{q+1}(t)\right)J^T_{\hat{f}}\left(c^{q+1}(t)\right) = 0$ to find or make progress towards critical points of $\hat{f}(c)$
  - Here, $g'(t) = \frac{1}{2}J_{\hat{f}}\left(c^{q+1}(t)\right)H^T_{\hat{f}}\left(c^{q+1}(t)\right)\Delta c^q + \frac{1}{2}(\Delta c^q)^T H_{\hat{f}}\left(c^{q+1}(t)\right)J^T_{\hat{f}}\left(c^{q+1}(t)\right)$
  - Since both terms are scalars, $g'(t) = J_{\hat{f}}\left(c^{q+1}(t)\right)H^T_{\hat{f}}\left(c^{q+1}(t)\right)\Delta c^q$
- <u>Option 3</u>: minimize $\hat{f}(c^{q+1}(t))$ directly (see unit 16)

Optimization Problems

- Solve $H_{\hat{f}}^T(c^q)\Delta c^q = (\beta - 1)J_{\hat{f}}^T(c^q)$ for Δc^q and use $c^{q+1}(t) = c^q + t\Delta c^q$ in $J_{\hat{f}}^T\left(c^{q+1}(t)\right) = 0$
- Option 1: find simultaneous (for all i) minima for all the $g_i(t) = (J_{\hat{f}}^T)_i\left(c^{q+1}(t)\right)$ aiming for the roots which are critical points of $\hat{f}(c)$
  - Here, $g_i'(t) = (H_{\hat{f}}^T)_i\left(c^{q+1}(t)\right)\Delta c^q$ and $g_i''(t) = (\Delta c^q)^T \left(OMG_{\hat{f}}^T\right)_i \left(c^{q+1}(t)\right) \Delta c^q$
- Option 2: minimize $g(t) = \frac{1}{2} J_{\hat{f}}\left(c^{q+1}(t)\right) J_{\hat{f}}^T\left(c^{q+1}(t)\right)$ aiming for the roots which are critical points of $\hat{f}(c)$
  - Here, $g'(t) = J_{\hat{f}}\left(c^{q+1}(t)\right) H_{\hat{f}}^T\left(c^{q+1}(t)\right)\Delta c^q$
  - $g''(t) = J_{\hat{f}}\left(c^{q+1}(t)\right)(\Delta c^q)^T OMG_{\hat{f}}^T\left(c^{q+1}(t)\right)\Delta c^q + (\Delta c^q)^T H_{\hat{f}}\left(c^{q+1}(t)\right) H_{\hat{f}}^T\left(c^{q+1}(t)\right)\Delta c^q$
- Option 3: minimize $g(t) = \hat{f}(c^{q+1}(t))$ directly
  - $g'(t) = J_{\hat{f}}\left(c^{q+1}(t)\right)\Delta c^q$ and $g''(t) = (\Delta c^q)^T H_{\hat{f}}\left(c^{q+1}(t)\right)\Delta c^q$