Unconstrained minimization: Newton's method

15.093: Optimization

Dimitris Bertsimas
Alexandre Jacquillat

Sloan School of Management
Massachusetts Institute of Technology

Reminder: descent methods for unconstrained minimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x})$$

Algorithm

1. *Initialization: starting point* $\boldsymbol{x}^0 \in \mathbb{R}^n$*, and iteration counter* $k = 0$
2. *Repeat, until termination criterion is reached*
   - 2.1 *Update iteration counter:* $k \leftarrow k + 1$
   - 2.2 *Determine a descent direction* $\boldsymbol{d}^k$*, such that* $\nabla f(\boldsymbol{x}^k)^\top \boldsymbol{d}^k < 0$
   - 2.3 *Determine a step size* $\alpha^k > 0$
   - 2.4 *Update* $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$

- Main design questions
  1. Initialization: how to determine the starting point $\boldsymbol{x}^0$?
  2. <u>Descent: how to determine the descent direction $\boldsymbol{d}^k$?</u>
  3. Line search: how to choose the step size α^k?
  4. Termination criterion; typically, $\|\nabla f(\boldsymbol{x}^k)\| \leq \eta$ for small $\eta > 0$

Newton's method

Motivation and intuition: a second-order view

- Taylor series expansion around $\boldsymbol{x}$

$$f(\boldsymbol{y}) \approx \widehat{f}(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

Motivation and intuition: a second-order view

- Taylor series expansion around $\boldsymbol{x}$

$$f(\boldsymbol{y}) \approx \widehat{f}(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$$

- Approximation of the minimization problem: $\min f(\boldsymbol{y}) \to \min \widehat{f}(\boldsymbol{y})$
- From first-order conditions, move in the Newton direction

$$\begin{aligned} \nabla \widehat{f}(\boldsymbol{y}) = \mathbf{0} &\implies \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) = \mathbf{0} \\ &\implies \boldsymbol{y} = \boldsymbol{x} - (\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x}) \end{aligned}$$

- The Newton decrement approximates the algorithm's progress:

$$f(\boldsymbol{x}) - \min_{\boldsymbol{y}} \widehat{f}(\boldsymbol{y}) = f(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x}+\boldsymbol{d}) = \frac{1}{2}\lambda(\boldsymbol{x})^2$$

Definition

- Newton direction: $\boldsymbol{d} = -(\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x})$
- Newton decrement: $\lambda(\boldsymbol{x}) = \left(\nabla f(\boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x})\right)^{1/2}$

Motivation and intuition: a first-order view

- We seek a stationary point: $\nabla f(\boldsymbol{y}) = \boldsymbol{0}$
- Taylor series expansion of $\nabla f(\boldsymbol{y})$ around $\boldsymbol{x}$

$$\nabla f(\boldsymbol{y}) \approx \widehat{g}(\boldsymbol{y}) = \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

Motivation and intuition: a first-order view

- We seek a stationary point: $\nabla f(\boldsymbol{y}) = \mathbf{0}$
- Taylor series expansion of $\nabla f(\boldsymbol{y})$ around $\boldsymbol{x}$

$$\nabla f(\boldsymbol{y}) \approx \widehat{g}(\boldsymbol{y}) = \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

- Approximation of the problem: $\nabla f(\boldsymbol{y}) = \mathbf{0} \rightarrow \widehat{g}(\boldsymbol{y}) = \mathbf{0}$
- From the equation, move in the Newton direction

$$\begin{aligned} \widehat{g}(\boldsymbol{y}) = \mathbf{0} &\implies \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) = \mathbf{0} \\ &\implies \boldsymbol{y} = \boldsymbol{x} - (\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x}) \end{aligned}$$

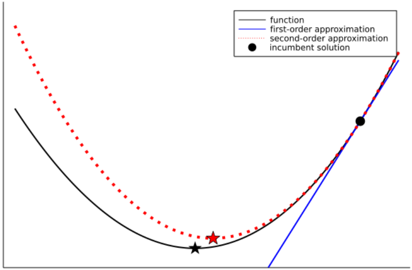
- History:
  - The Newton method was developed by Newton and Raphson in the 1600's for solving systems of equations
  - Extension to optimization by Simpson in the 1700's: $g(\boldsymbol{y}) = \nabla f(\boldsymbol{y})$

Visualization

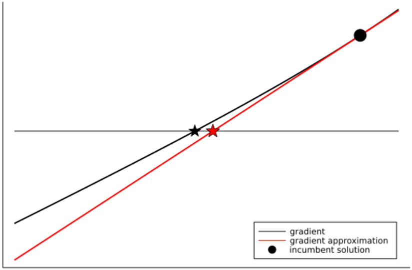
- Two equivalent ways of interpreting Newton's method
  1. Second-order view: minimization of second-order Taylor approximation
  2. First-order view: root of first-order Taylor approximation of gradient

→ By leveraging second-order (Hessian) information, we obtain a stronger approximation of the function, hence of the minimization problem

Second-order view



First-order view



Newton's method

Algorithm (Newton's method)

1. *Initialization: starting point* $\boldsymbol{x}^0 \in \mathbb{R}^n$, *and iteration counter* $k = 0$
2. *Repeat, until termination criterion is reached*
   2.1 *Update iteration counter:* $k \leftarrow k + 1$
   2.2 *Determine Newton's direction* $\boldsymbol{d}^k = -(\nabla^2 f(\boldsymbol{x}^k))^{-1} \nabla f(\boldsymbol{x}^k)$
   2.3 *Update* $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \boldsymbol{d}^k$

Newton's method

Algorithm (Newton's method)

1. *Initialization: starting point* $\boldsymbol{x}^0 \in \mathbb{R}^n$, *and iteration counter* $k = 0$
2. *Repeat, until termination criterion is reached*
   - 2.1 *Update iteration counter:* $k \leftarrow k + 1$
   - 2.2 *Determine Newton's direction* $\boldsymbol{d}^k = -(\nabla^2 f(\boldsymbol{x}^k))^{-1} \nabla f(\boldsymbol{x}^k)$
   - 2.3 *Update* $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \boldsymbol{d}^k$

→ A second-order method: use of first-order gradient and second-order Hessian information to proceed from iteration to iteration
  - More progress at each iteration than gradient descent
  - More work per iteration: $\mathcal{O}(n^3)$ operations

- "Pure" Newton method relies on a step size of 1
- Convergence criterion based on estimated improvement: $\lambda(\boldsymbol{x})^2/2 \leq \varepsilon$
- Affine invariance: Newton's method independent of problem scaling
  - Newton's method for $f(\boldsymbol{x})$ and $\widetilde{f}(\boldsymbol{y}) = f(\boldsymbol{T}\boldsymbol{y})$ yields $\boldsymbol{x}^{k+1} = \boldsymbol{T}\boldsymbol{y}^{k+1}$
  - Recall that this is not the case with gradient descent

Example: fitting a logistic regression model

$$\max \ f(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right) \right\}$$

$$\nabla f(\boldsymbol{\beta}) = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{\beta})), \quad \text{with: } p_i(\boldsymbol{X}, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

$$\nabla^2 f(\boldsymbol{\beta}) = -\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}, \text{ where: } \boldsymbol{W} = \text{diag}\left(\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \cdot \frac{1}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right)$$

- A convex optimization problem: $\max f(\boldsymbol{\beta})$ with $\nabla^2 f(\boldsymbol{\beta}) \preceq \mathbf{0}$
- Minimum at a stationary point $\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{\beta}^*)) = \mathbf{0}$
- Applying Newton's algorithm to the logistic regression model:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= \boldsymbol{\beta}^k + (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{\beta}^k)) \\ &= (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z}^k, \text{ with } \boldsymbol{z}^k = \boldsymbol{X}\boldsymbol{\beta}^k + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{\beta}^k)) \end{aligned}$$

$\rightarrow$ Interpretation as iterative re-weighted least squares:

$$\boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}} \ (\boldsymbol{z}^k - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{W} (\boldsymbol{z}^k - \boldsymbol{X}\boldsymbol{\beta})$$

Local convergence

Local convergence of Newton's method

Definition (operator norm of a matrix)

$$\|\boldsymbol{M}\| = \max\{\|\boldsymbol{M}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\}$$

Theorem

$f(\cdot)$ twice continuously differentiable, and $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. Assume that:

- $\|(\nabla^2 f(\boldsymbol{x}^*))^{-1}\| \leq \frac{1}{m}$ *for some* $m > 0$
- $\nabla^2 f(\boldsymbol{x})$ *is L-Lipschitz in the β-ball around $\boldsymbol{x}^*$ for some $\beta > 0$, $L > 0$:*

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}^*, \beta)$$

Define $\delta = \min\left\{\beta, \frac{2m}{3L}\right\}$. The following holds:

1. *If $\|\boldsymbol{x}^k - \boldsymbol{x}\| \leq \delta$, then $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}\| \leq \delta$ for all $k = 0, 1, 2, \cdots$*
2. $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| \leq \frac{3L}{2m}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2, \ \forall k = 0, 1, 2, \cdots$

Local convergence: interpretation and implications

Corollary

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \leq \frac{1}{C}\left(CX\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|\right)^{2^k}, \text{ where } C = \frac{3L}{2m}$$

- Interpretation of the result:
    - Once in a δ-neighborhood of $\boldsymbol{x}^*$, the algorithm stays there
    - Quadratic convergence within the δ-neighborhood: $\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2} \leq \frac{3L}{2m}$

  $\rightarrow$ Solution within ε of the optimum after at most

$$\left\lceil \frac{\log\left(\frac{\log(C\varepsilon)}{\log(C\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|)}\right)}{\log 2} \right\rceil \text{ iterations}$$

- Newton's method is attracted to local minima & maxima: $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$
- β, m and L are hard to estimate, but not used in the algorithm
- The algorithm and local convergence do not require the convexity of f, only that $H(\boldsymbol{x}^*)$ is nonsingular and not badly behaved near $\boldsymbol{x}^*$.

Proof of the theorem (1/2)

- Notation: $g(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ and $H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$

Lemma

$$g(\boldsymbol{x}^k) - g(\boldsymbol{x}^*) = \int_0^1 H(\boldsymbol{x}^* + t(\boldsymbol{x}^k - \boldsymbol{x}^*))(\boldsymbol{x}^k - \boldsymbol{x}^*)dt$$

- By definition of $\boldsymbol{x}^{k+1}$, we derive:

$$\begin{aligned}
\boldsymbol{x}^{k+1} - \boldsymbol{x}^* &= \boldsymbol{x}^k - \boldsymbol{x}^* - (H(\boldsymbol{x}^k))^{-1} g(\boldsymbol{x}^k) \\
&= \boldsymbol{x}^k - \boldsymbol{x}^* - (H(\boldsymbol{x}^k))^{-1} (g(\boldsymbol{x}^k) - \underbrace{g(\boldsymbol{x}^*)}_{=\boldsymbol{0}}) \\
&= (\boldsymbol{x}^k - \boldsymbol{x}^*) - (H(\boldsymbol{x}^k))^{-1} \int_0^1 H(\boldsymbol{x}^* + t(\boldsymbol{x}^k - \boldsymbol{x}^*))(\boldsymbol{x}^k - \boldsymbol{x}^*)dt \\
&= (H(\boldsymbol{x}^k))^{-1} \int_0^1 \left[H(\boldsymbol{x}^k) - H(\boldsymbol{x}^* + t(\boldsymbol{x}^k - \boldsymbol{x}^*)) \right] (\boldsymbol{x}^k - \boldsymbol{x}^*)dt
\end{aligned}$$

Proof of the theorem (2/2)

Lemma

Under the conditions of the theorem, we have

$$\|H(\boldsymbol{x})^{-1}\| \leq \frac{1}{m - L\|\boldsymbol{x} - \boldsymbol{x}^*\|}, \ \forall \boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, \beta)$$

$$\begin{aligned}
\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| &\leq \left\|H(\boldsymbol{x}^k))^{-1}\right\| \int_0^1 \left\|H(\boldsymbol{x}^k) - H(\boldsymbol{x}^* + t(\boldsymbol{x}^k - \boldsymbol{x}^*))\right\| \|\boldsymbol{x}^k - \boldsymbol{x}^*\| dt \\
&\leq \frac{1}{m - L\|\boldsymbol{x} - \boldsymbol{x}^*\|} \cdot \int_0^1 L(1-t)\|\boldsymbol{x}^k - \boldsymbol{x}^*\| dt \cdot \|\boldsymbol{x}^k - \boldsymbol{x}^*\| \\
&= \frac{L}{2(m - L\|\boldsymbol{x} - \boldsymbol{x}^*\|)} \|(\boldsymbol{x}^k - \boldsymbol{x}^*)\|^2
\end{aligned}$$

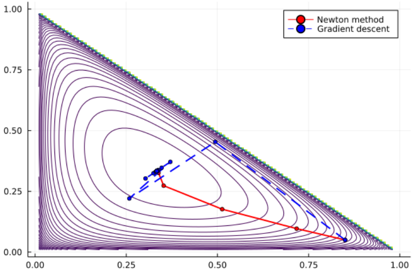
- Since $L\|(\boldsymbol{x}^k - \boldsymbol{x}^*)\| \leq 2m/3$, we obtain:
  1. $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| \leq \dfrac{2m/3}{2(m - 2m/3)} \|(\boldsymbol{x}^k - \boldsymbol{x}^*)\| = \|(\boldsymbol{x}^k - \boldsymbol{x}^*)\| \leq \delta$
  2. $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| \leq \dfrac{L}{2(m - 2m/3)} \|(\boldsymbol{x}^k - \boldsymbol{x}^*)\|^2 = \dfrac{3L}{2m} \|(\boldsymbol{x}^k - \boldsymbol{x}^*)\|^2$

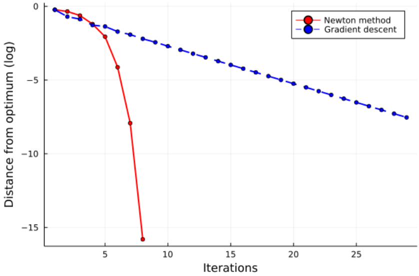
Example and illustration

$$f(x_1, x_2) = -\log(1 - x_1 - x_2) - \log x_1 - \log x_2$$

$$\nabla f(x_1, x_2) = \left(\frac{1}{1 - x_1 - x_2} - \frac{1}{x_1}; \frac{1}{1 - x_1 - x_2} - \frac{1}{x_2} \right)^\top$$

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} \left(\frac{1}{1 - x_1 - x_2}\right)^2 + \left(\frac{1}{x_1}\right)^2 & \left(\frac{1}{1 - x_1 - x_2}\right)^2 \\ \left(\frac{1}{1 - x_1 - x_2}\right)^2 & \left(\frac{1}{1 - x_1 - x_2}\right)^2 + \left(\frac{1}{x_1}\right)^2 \end{bmatrix}$$





Global convergence

Global convergence: issues and solutions

- Quadratic convergence in Newton's method is "local"
  - There is no guaranteee that $f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k)$ at each iteration
  - Newton's method is attracted to local minima & maxima: $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$
  - What happens if we start "far" away from $\boldsymbol{x}^*$?

→ Augment algorithm with line search: $\alpha^k = \arg\min_\alpha f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k)$

Proposition

If $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$, then $\boldsymbol{d} = -\nabla^2 f(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x}) \neq \mathbf{0}$ is a descent direction:

$$f(\boldsymbol{x} + \alpha \boldsymbol{d}) < f(\boldsymbol{x}) \text{ for } \alpha \text{ sufficiently small.}$$

Algorithm (Newton's method with exact line search)

1. *Initialization: starting point $\boldsymbol{x}^0 \in \mathbb{R}^n$, and iteration counter $k = 0$*
2. *Repeat, until termination criterion is reached*
   - 2.1 *Update iteration counter: $k \leftarrow k + 1$*
   - 2.2 *Determine Newton's direction $\boldsymbol{d}^k = -(\nabla^2 f(\boldsymbol{x}^k))^{-1} \nabla f(\boldsymbol{x}^k)$*
   - 2.3 *Determine α^k via exact line search, and update $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$*

Global convergence: (strong) convexity to the rescue

Theorem

$f(\cdot)$ twice continuously differentiable. Assume that:

- *f is M-smooth*
- *f is m-strongly convex*
- *$\nabla^2 f(\boldsymbol{x})$ is L-Lipschitz: $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y}$*

Newton's method with line search satisfies, for $0 < \eta \leq m^2/L$, $\gamma > 0$:

1. *If $\|\nabla f(\boldsymbol{x}^k)\| \geq \eta$: $f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) \leq -\gamma$*
2. *If $\|\nabla f(\boldsymbol{x}^k)\| < \eta$: $\|\nabla f(\boldsymbol{x}^{k+1})\| < \eta$ & $\|\nabla f(\boldsymbol{x}^{k+1})\| \leq \frac{L}{2m^2}\|\nabla f(\boldsymbol{x}^k)\|^2$*

$\rightarrow$ Two phases in Newton's method:

1. Damped phase: progress of at least η per iteration
2. Local phase: quadratic convergence within a local neighborhood

$\rightarrow$ Main objective: getting quickly into a good neighborhood

$$\frac{f(\boldsymbol{x}^0 - z^*)}{\gamma} + \log_2 \log_2 \left(\frac{\varepsilon_0}{\varepsilon}\right) \text{ iterations, for some constant } \varepsilon_0$$

Damped phase: proof

- $\nabla f(\boldsymbol{x}^k)^\top \boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k)^\top (\nabla^2 f(\boldsymbol{x}^k))^{-1} \nabla f(\boldsymbol{x}^k) = -\lambda(\boldsymbol{x}^k)^2$
- Due to the M-smoothness of the function f:

$$f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) \leq f(\boldsymbol{x}^k) + \alpha \nabla f(\boldsymbol{x}^k)^\top \boldsymbol{d}^k + \frac{M\alpha^2}{2} \|\boldsymbol{d}^k\|^2$$

$$\lambda(\boldsymbol{x}^k)^2 = \nabla f(\boldsymbol{x}^k)^\top (\nabla^2 f(\boldsymbol{x}^k))^{-1} \nabla f(\boldsymbol{x}^k) \geq \frac{1}{M} \|\nabla f(\boldsymbol{x}^k)\|^2$$

- Due to the strong convexity of the function f:

$$\lambda(\boldsymbol{x}^k)^2 = (\boldsymbol{d}^k)^\top \nabla^2 f(\boldsymbol{x}^k) \boldsymbol{d}^k \geq m \|\boldsymbol{d}^k\|^2$$

- By combining the two properties, we obtain:

$$f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) \leq f(\boldsymbol{x}^k) - \alpha \lambda(\boldsymbol{x}^k)^2 + \frac{M\alpha^2}{2m} \lambda(\boldsymbol{x}^k)^2$$

- Exact line search: $\alpha^k \in \arg\min_\alpha \; f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k)$
- By minimizing over α on both sides, we obtain:

$$f\left(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k\right) \leq f(\boldsymbol{x}^k) - \frac{m}{2M} \lambda(\boldsymbol{x}^k)^2 = f(\boldsymbol{x}^k) - \frac{m}{2M^2} \|\nabla f(\boldsymbol{x}^k)\|^2$$

$$\Longrightarrow f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) = f\left(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}\right) - f(\boldsymbol{x}^k) \leq -\frac{m}{2M^2} \eta^2 := -\gamma$$

Conclusion

Summary

Takeaway

Newton's method was originally developed to find roots of equations, and then extended to solve optimization problems.

Takeaway

Newton's method augments descent methods by leveraging second-order (Hessian) information.

Takeaway

Convergence of Newton's method is very fast (quadratic) locally.
$\Longrightarrow$ *two steps: (i) finding a solution near a stationary point; and (ii) finding the stationary point, exploiting quadratic convergence.*

Takeaway

Numerous enhancements exist to address global convergence issues.